HAL open science

Estimation par ondelettes dans les modèles partiellement linéaires

Irène Gannaz

▶ To cite this version:

Irène Gannaz. Estimation par ondelettes dans les modèles partiellement linéaires. Mathématiques [math]. Université Joseph-Fourier - Grenoble I, 2007. Français. NNT : . tel-00197146

HAL Id: tel-00197146

https://theses.hal.science/tel-00197146v1

Submitted on 14 Dec 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Joseph Fourier - Grenoble I

École Doctorale « Mathématiques, Sciences et
Technologie de l'Information, Informatique »

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Spécialité : **Mathématiques Appliquées**

préparée au Laboratoire Jean Kuntzmann

présentée et soutenue publiquement par

Irène Gannaz

le 07 décembre 2007

Estimation par ondelettes dans les modèles partiellement linéaires

---

Composition du Jury :

M.	Anestis ANTONIADIS	Université Grenoble I	Directeur de thèse
M.	Jalal FADILI	ENSICAEN	Examinateur
Mme	Irène GIJBELS	Kathioleke Universiteit Leuven	Rapportrice
M.	Anatoli IOUDITSKI	Université Grenoble I	Président

Au vu des rapports de :

Mme	Irène GIJBELS	Kathioleke Universiteit Leuven
Mme	Dominique PICARD	Université Paris VII

Remerciements

Mes remerciements s'adressent en premier lieu à Anestis Antoniadis, pour sa bonne humeur constante et ses « salut la belle » souriants. Sa vision générale de la recherche et sa confiance en mes capacités ont contribué à me faire aimer ces trois ans de recherche. Je le remercie chaleureusement pour son contact constructif si généreux sur le plan humain comme scientifique.

Je remercie également Irène Gijbels et Dominique Picard d'avoir rapporter cette thèse et d'avoir lu si attentivement le manuscrit. Leurs remarques constructives ont permis d'améliorer ce travail et donnent de nouvelles perspectives de prolongation.

Je suis aussi reconnaissante envers Anatoli Iouditski qui m'a fait l'honneur de présider mon jury. Merci enfin à Jalal Fadili pour avoir accepté d'être membre de mon jury et pour avoir suivi mon travail au long de ces trois ans.

Je ne peux passer outre dans ces remerciements tous les thésards du labo ou "assimilés" qu'il m'a été donné de rencontrer au cours de ces trois ans. Je déplore le manque d'exhaustivité des noms cités : que les oubliés me pardonnent... Je pense en premier lieu à Carine, pour les allers-retours entre les bureaux et les discussions qui s'éternisent. Cette thèse a aussi été rythmée par la profusion des jeux de mots de Yann et les soirées jeux des Oliviers : Olivier "Grumphh" qui a m'a fait découvrir le Carom et Olivier "l'espagnol" qui n'en est pas un et mériterait plutôt "le sportif". Je n'oublie pas Olivier "le parisien" qui montait parfois à Grenoble faire un peu d'ondelettes et de ski de randonnée. *Kop Khun kha* Pramote, pour ces réflexions politiques et ces intermèdes musicaux dans le bureau. Merci à tous ceux qui ont égaillé de si nombreuses pauses et soirées au cours de cette thèse : Aude la toulousaine et Vincent avec ses boutades, Claire et Cyrille les pros de l'aïkido, Basile avec sa guitare... Merci aux thésards et stagiaires dont les discussions ont animé les pauses : toute la salle 3 qui a si bien accueilli la stateuse du rez-de-chaussée, Julie et ses grands sourires, Robin, Laurent T., Adrien, Mélanie, Emilie, Damien... Merci enfin à tous ceux du labo qui m'ont offert ce beau vélo !

Je remercie aussi ma famille d'avoir été présente durant ces années, jamais bien loin (même à l'autre bout du monde) et toujours prête à faire du baby-sitting ;o). Et merci à tous les "ouvriers" de la famille et de la belle famille, nombreux, d'avoir contribué à ce que je fasse plus de statistique que de plâtre ou de plomberie durant ces deux dernières années.

Merci à Etienne d'être là tout simplement. Et d'avoir été un petit garçon si souriant et si calme. Merci à Christian, pour tout.

Table des matières

Introduction **1**

1 Le modèle partiellement linéaire **7**
1.1 Un modèle semi-paramétrique 7
1.1.1 Pourquoi ce modèle 8
1.1.2 Exemples d'application 9
1.2 Inférence statistique 10
1.2.1 Estimation par les moindres carrés pénalisés 10
1.2.2 La régression partielle de Speckman (1988) 14

2 Rappels sur les ondelettes et leur usage en régression **19**
2.1 Analyse multirésolution et ondelettes 20
2.1.1 Approximations de fonctions 21
2.1.2 Ondelettes périodiques 22
2.1.3 Espaces de Besov 23
2.2 Transformation en ondelettes 25
2.2.1 Ecriture matricielle 25
2.2.2 Algorithme pyramidal 26
2.3 Régression avec un plan d'observation équidistant 28
2.3.1 Estimations linéaires 28
2.3.2 Estimations non linéaires : seuillage des coefficients 29
2.4 Régression avec un plan d'observation non-équidistant 33
2.4.1 Aperçu de différentes approches 33
2.4.2 Hypothèses sur le plan d'observation 36
2.4.3 Approche théorique 39
2.4.4 Approche pratique 41

3 Du seuillage à l'estimation robuste **45**
3.1 Identifiabilité et transformée en ondelettes 46

3.2 Critère des moindres carrés penalisés 48
3.2.1 Conditions du premier ordre 48
3.2.2 Seuillage doux et estimateur de Huber 49
3.2.3 Vers d'autres M-estimateurs 52

4 Propriétés asymptotiques des estimateurs 57
4.1 Hypothèses 57
4.2 Propriétés asymptotiques 61
4.2.1 Cas du seuillage doux et de l'estimateur de Huber 61
4.2.2 Exemple issu de Fadili et Bullmore (2005) 63
4.3 Estimation de la variance 65
4.4 Preuves 67
4.4.1 Consistance de $\hat{\beta}_n$ avec le seuillage doux 68
4.4.2 Consistance de $\hat{\beta}_n$ avec la pénalité quadratique 82
4.4.3 Estimation de la partie fonctionnelle 85

5 Quelques algorithmes 87
5.1 Estimation conjointe des paramètres 87
5.1.1 Backfitting 88
5.1.2 Deux algorithmes pour le seuillage doux 89
5.2 Approche semi-quadratique 91
5.2.1 ARTUR, ou les poids modifiés 93
5.2.2 LEGEND, ou les résidus modifiés 94

6 Simulations et un exemple sur données réelles 97
6.1 Comparaison des algorithmes pour le seuillage doux 98
6.1.1 Exemple 1 : Fonction sinusoïdale 99
6.1.2 Exemple 2 : Fonction linéaire par morceaux 104
6.1.3 Exemple 3 : Dimension 5 106
6.2 Comparaison des différents seuillages 109
6.2.1 Validation croisée pour la pénalité quadratique 109
6.2.2 Exemple 1 110
6.2.3 Exemple 2 112
6.2.4 Exemple 3 115
6.3 Application à l'étude de données réelles issues d'une IRM fonctionnelle . 118
6.3.1 Description des données 118
6.3.2 Modélisation par un modèle partiellement linéaire 119

7 Cas de plans d'observations aléatoires 125

7.1 Principe d'estimation 125
7.2 Simulations avec un plan d'observations aléatoire 128
7.2.1 Modèle non paramétrique 128
7.2.2 Modèle partiellement linéaire 140

Conclusion et perspectives **155**

Bibliographie **157**

Notations

$\mathbb{R}^k, \mathbb{N}^k, \mathbb{Z}^k$	Ensemble des k-uplets de réels, d'entiers et d'entiers relatifs.
$\mathcal{M}_{n,p}(\mathbb{R})$	Ensembles des matrices de taille $n \times p$ à coefficients réels.
$\mathbb{P}$, $\mathbb{E}$, Var	Probabilité, Espérance, Variance.
v.a., v.a. i.i.d.	variable aléatoire, v.a. indépendantes et identiquement distribuées.
$[x]$	Partie entière du nombre x.
$\mathbb{1}_E$	Fonction indicatrice de l'ensemble E.
$\|e\|$, $e \in \mathbb{R}^k$	Norme l^2 d'un vecteur e. Si $e = (e_1 \dots e_k)^T$, alors $\|e\| = \left(\sum_{i=1}^k e_i^2\right)^{1/2}$.
$\|A\|$, $A \in \mathcal{M}_{n,p}(\mathbb{R})$	Norme de Frobenius d'une matrice A. Si A est de terme général $a_{i,j}$, alors $\|A\| = \left(\sum_{i,j} a_{i,j}^2\right)^{1/2}$.
A^T, A^{-1}	Matrice transposée et matrice inverse de la matrice A.
$\xrightarrow{\mathbb{P}}$	Convergence en probabilité
$u_n = \bigcirc(v_n)$, $u_n \in \mathbb{R}^k$, $v_n \in \mathbb{R}$	$\forall i = 1, \dots, k$, $\exists c_i > 0$, $\lvert u_{in}\rvert < c_i \lvert v_n\rvert$.
$u_n = \circ(v_n)$, $u_n \in \mathbb{R}^k$, $v_n \in \mathbb{R}$	$\forall i = 1, \dots, k$, $\lim_{n\to\infty} u_{in}/v_n = 0$.
$U_n = \bigcirc_{\mathbb{P}}(v_n)$, U_n v.a. à valeurs dans $\mathbb{R}^k$, $v_n \in \mathbb{R}$	$\forall i = 1, \dots, k$, $\forall \delta > 0$, $\exists t_{i,\delta} > 0$, $\mathbb{P}(v_n^{-1}\lvert U_{in}\rvert > t_{i,\delta}) \leqslant \delta$.
$U_n = \circ_{\mathbb{P}}(v_n)$, U_n v.a. à valeurs dans $\mathbb{R}^k$, $v_n \in \mathbb{R}$	$\forall i = 1, \dots, k$, $v_n^{-1} U_n \xrightarrow{\mathbb{P}} 0$.
$L^p(I)$	$= \{f : I \to \mathbb{R}, \int \lvert f\rvert^p < \infty\}$
W^s, $s \in \mathbb{N}$	$= \{f : I \to \mathbb{R}, I \subset \mathbb{R}$, s-fois continûment dérivable, $\int \lvert f^{(s)}\rvert^p < \infty\}$.
$\Lambda^\alpha(M)$, $\alpha \leqslant 1$	$= \{f : I \to \mathbb{R}, I \subset \mathbb{R}\, \forall (x,y) \in I^2, \lvert f(x) - f(y)\rvert \leqslant M\lvert x-y\rvert^\alpha\}$.
$\Lambda^\alpha(M)$, $\alpha > 1$	$= \{f : I \to \mathbb{R}, I \subset \mathbb{R}$ $[\alpha]$-fois continûment dérivable, $f^{([\alpha])} \in \Lambda^{\alpha-[\alpha]}(M)\}$.
SNR	Rapport Signal sur Bruit : Si $Y \in \mathbb{R}^n$ est un vecteur observé avec un bruit d'écart-type σ, $SNR = \frac{\|Y\|}{\sqrt{n}\sigma}$.

Introduction

Le problème de la régression statistique consiste à analyser l'influence de variables dites explicatives X, $X \in \mathbb{R}^p$, sur une variable y. Lorsqu'aucun a priori quant à la forme de la relation explicative n'est possible, la modélisation la plus adaptée est le modèle non paramétrique. Celui-ci conste à supposer que les variables sont régies par l'équation

$$y_i = g(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

le terme ε_i représentant un bruit. Néanmoins, ce modèle souffre d'un manque d'interprétabilité et les techniques d'estimation non paramétriques ont des performances moindres que dans un cadre paramétrique. Ce dernier point, connu en général sous l'appélation *fléau de la dimension*, a été mis en évidence par Ibragimov et Has'Minskii (1980) à l'aide de la notion de vitesse asymptotique minimax. Leurs travaux ont incité à développer, dans les dernières décennies, des modèles dits semi-paramétriques.

L'idée d'un modèle semi-paramétrique est d'émettre des hypothèses quant à la forme de la fonction g que l'on cherche à estimer. A l'aide de ces hypothèses, une reformulation du modèle est réalisée permettant de *réduire la dimension* des variables explicatives intervenant dans l'estimation fonctionnelle. Parmi la classe de modèles résultant de ce procédé se trouvent, de manière non exhaustive, le modèle additif, le modèle à direction révélatrice ou encore le modèle partiellement linéaire. Le but du présent travail est d'apporter une contribution à l'étude de ce dernier modèle, en s'intéressant à l'application de techniques non linéaires d'estimation par ondelettes dans celui-ci.

Le modèle partiellement linéaire est présenté dans le **premier chapitre**. Ce modèle s'écrit sous la forme

$$y_i = X_i^T \beta + f(T_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

avec X_i covariables de $\mathbb{R}^p$ et, dans le cadre de cette thèse, $T_i \in \mathbb{R}$. Il est ainsi constitué de deux parties distinctes : une partie linéaire et une partie fonctionnelle. La dimension des

variables explicatives de la partie fonctionnelle étant ici l'unité, le problème de fléau de la dimension cité plus haut est amoindri comparativement à la version non paramétrique. Ce modèle a connu un essor important suite aux travaux de Engle *et al.* (1986), qui mettaient notamment en exergue la capacité de modélisation offerte.

Nous nous intéressons plus particulièrement aux procédures d'estimation par moindres carrés pénalisés développées dans les années 80. La majorité des approches effectuées dans ce modèle consiste à introduire des techniques d'interpolations par des splines ou des estimateurs à noyaux afin d'estimer la partie fonctionnelle du modèle.

Les deux articles fondamentaux sur ce modèle sont ceux de Rice (1986) et Speckman (1988). Rice a montré que lorsque les variables explicatives respectivement de la partie linéaire du modèle et de la partie fonctionnelle, X et T, étaient corrélées, alors une estimation optimale et du vecteur de régression β et de la fonction f n'était pas réalisable. (La notion de corrélation entre les variables utilisée ici est définie précisément dans le chapitre.) L'auteur aboutit à la conclusion que pour atteindre la vitesse optimale dans l'estimation du vecteur de régression, un sous-lissage de l'estimation fonctionnelle est nécessaire. Les travaux de Rice ont engendré de nombreuses réflexions sur ce modèle. Parmi celles-ci, nous pouvons citer Speckman, qui retrouve des conclusions similaires. Ce dernier propose une autre approche, dite de *régression partielle*, permettant de réduire les biais dans l'estimation, et résolvant le problème du choix du paramètre de lissage. Les deux auteurs s'accordent aussi pour dénoncer les mauvais résultats que peut engendrer une procédure de validation croisée dans ce contexte.

Dans un modèle non paramétrique, la projection sur une base d'ondelettes associée à des techniques de seuillage a permis le développement d'estimateurs adaptatifs quant à la régularité de la fonction estimée. Ces procédés d'estimation présentent l'avantage d'offrir des résultats quasi-minimax sur une large gamme de fonctions mais aussi de passer outre une éventuelle validation croisée pour déterminer le paramètre de lissage. C'est pourquoi ce travail est motivé par l'introduction des schémas d'estimation par ondelettes dans le modèle partiellement linéaire.

En conséquence, le **deuxième chapitre** propose un bref rappel sur la notion d'analyse multirésolution et d'ondelette. Nous redonnons aussi la définition d'espace de Besov. Dans un second temps, ce chapitre présente les procédures d'estimation dans un modèle de régression non paramétrique lorsque les observations sont réalisées en des points équidistants de l'intervalle unité. Le seuillage des coefficients d'ondelette permet de construire des estimateurs quasi-minimax sur des boules de Besov, tout en étant adaptatifs relativement à la régularité de la fonction estimée. La généralisation de ces schémas

d'estimation au cas d'observations distribuées selon un plan aléatoire fait actuellement encore l'objet de nombreux travaux. Après avoir succintement résumé quelques articles relatifs à ce domaine, nous nous attarderons principalement sur la procédure d'estimation proposée par Kerkyacharian et Picard (2004), étant donné que celle-ci semble la plus adaptée à l'application dans un modèle partiellement linéaire.

Le **Chapitre 3** concerne l'introduction des ondelettes et du seuillage des coefficients dans un modèle partiellement linéaire. Nous nous restreignons au cadre d'un plan d'observation équidistant de la partie fonctionnelle. En s'appuyant sur les travaux de Antoniadis et Fan (2001), nous élaborons une procédure des moindres carrés pénalisés dans laquelle la pénalisation, portant sur les coefficients d'ondelettes, permet d'aboutir au seuillage souhaité. Nous établissons l'équivalence entre la minimisation du critère des moindres carrés pénalisés et des estimations successives du vecteur des paramètres de régression β par un M-estimateur et de la partie fonctionnelle par ondelettes avec seuillage des coefficients. Grâce à ces résultats, nous pouvons interpréter les estimateurs comme suit :

1. Le vecteur de paramètres β est évalué par une procédure de M-estimation dans le modèle de régression linéaire de y par X après projection sur une base d'ondelette. Il est donc estimé sans tenir compte de la présence d'une partie fonctionnelle dans le modèle. Cela revient à intégrer dans le bruit les coefficients d'ondelettes de la partie fonctionnelle.
2. La fonction f est l'estimation par ondelettes avec le seuillage désiré dans la régression non paramétrique de $y - X\hat{\beta}_n$, l'estimateur $\hat{\beta}_n$ ayant été calculé à l'étape précédente.

L'originalité de notre travail réside aussi dans le parallèle montré entre les seuillages réalisés à l'étape 2 et les M-estimateurs associés dans l'étape 1 ci-dessus. Par exemple le seuillage doux est associé au M-estimateur de Huber, le seuillage SCAD à celui de Hampel ou encore le seuillage fort à la moyenne tronquée. Ainsi, les M-estimateurs obtenus sont des estimateurs usuels introduits dans les années 80 par la théorie de l'estimation robuste.

De plus, le fait d'avoir une équivalence avec une procédure d'estimation en deux étapes ouvre des perspectives de développement intéressantes. Outre la construction d'algorithmes estimant successivement et non conjointement les paramètres (cf. Chapitre 5), elle offre la possibilité d'approfondir le comportement asymptotique des estimateurs. Ceci fait l'objet du Chapitre 4.

Dans le **quatrième chapitre**, nous nous intéressons plus précisément à deux pénalisations dans le critère des moindres carrés pénalisés : la pénalisation l^1, menant au seuillage doux

et au M-estimateur de Huber, et une pénalisation par une norme de Sobolev, de forme quadratique, menant à un lissage usuel et à un estimateur des moindres carrés pondérés.

Pour chacune de ces procédures, nous étudions les propriétés asymptotiques des estimateurs. Dans les deux cas considérés, l'estimateur de la partie fonctionnelle a un comportement quasi-minimax en probabilité pour la norme L^2 sur des espaces de Besov. Cependant, en l'absence d'hypothèse quant à la structure des variables, la vitesse de convergence de l'estimateur de la partie linéaire est d'ordre non paramétrique, ce qui n'est pas satisfaisant. Lorsque nous supposons que les covariables ont une structure similaire à celle considérée par Rice (1986) dans ce contexte, nous obtenons que les estimateurs du vecteur de régression β convergent à une vitesse quasi-minimax. Ces résultats sont donc cohérents avec ceux de Rice (1986). Le principal intérêt est que la procédure menant au seuillage doux offre un schéma adaptatif, permettant d'éviter une procédure de validation croisée. Ce choix de pénalisation est donc particulièrement intéressant.

Par ailleurs, ces estimateurs nécessitent une estimation de la variance du bruit. Une procédure basée sur une décomposition QR de la matrice de régression et sur le caractère creux des coefficients d'ondelettes sur un espace de Besov est proposée dans ce chapitre.

Les **Chapitres 5 et 6** traitent de la mise en œuvre des estimateurs dans les cas du seuillage doux, du seuillage dur et du seuillage associé à la pénalité par une norme de Sobolev. Dans le Chapitre 5, nous présentons les divers algorithmes qui seront appliqués dans le Chapitre 6. Parmi ces algorithmes, nous pouvons distinguer ceux qui estiment conjointement le vecteur de régression β et la fonction f de ceux reposant sur l'estimation en deux étapes établie au Chapitre 3 et évoquée plus haut. Le principal algorithme estimant simultanément les paramètres est le *Backfitting*, déjà proposé par Fadili et Bullmore (2005) dans le même contexte. Quant aux méthodes de calcul succesif des estimateurs, nous proposons deux algorithmes semi-quadratiques –ARTUR et LEGEND– issus de la théorie de la reconnaissance robuste. Ces deux algorithmes peuvent par exemple être rencontrés dans Dahyot et Kokaram (2004).

Lors de la comparaison sur des simulations, il apparaît que le *Backfitting* est plus performant en terme de temps de calcul que les autres procédures. L'algorithme LEGEND obtient des résultats similaires avec un temps de calcul légèrement supérieur mais du même ordre de grandeur. Concernant la qualité des résultats obtenus, l'estimation du vecteur des paramètres de régression linéaire est satisfaisante pour les seuillages fort et doux. Pour ces seuillages, l'estimation de la partie fonctionnelle donne de bons résultats pour le critère de l'erreur quadratique moyenne. Concernant le seuillage issu d'une pénalisation par une norme de Sobolev, une procédure de validation croisée est nécessaire afin de

déterminer le paramètre de lissage. L'inconvénient est le sous-lissage probable engendré par un tel schéma, souligné par Rice (1986) et Speckman (1988). Les estimations obtenues sont de moins bonne qualité et nous observons, entre autres, un sous-lissage de la partie fonctionnelle, ce qui confirme à notre avis la préférence pour des schémas adaptatifs dans un modèle partiellement linéaire.

Nous avons par ailleurs étudié en fin de Chapitre 6 un exemple sur des données réelles, gracieusement fournies par Dr. Fadili. Ces données consistent en trois signaux obtenus par Imagerie à Résonance Magnétique fonctionnelle (IRMf). L'un des signaux est issu d'une expérience avec des stimuli tandis que les deux autres sont des enregistrements de référence, où aucun stimulus n'a été réalisé. Dans la modélisation par un modèle partiellement linéaire, la partie linéaire intègre des facteurs de nuisance. La partie fonctionnelle résume donc l'activité neuronale sans le biais causé par ces nuisances. Cette modélisation s'avère peu convaincante en raison principalement du bruit important, comparativement aux modalités de la partie linéaire du modèle obtenu.

Le **Chapitre 7** généralise l'étude au cas où la partie fonctionnelle est observée sur un plan aléatoire. Nous appliquons la procédure d'estimation de Kerkyacharian et Picard (2004) exposée au Chapitre 2. Dans ce chapitre, nous mettons en application le schéma d'estimation proposé sur des simulations. Pour cela, nous utilisons l'algorithme élaboré par Vidakovic (2002) permettant de calculer les valeurs des ondelettes en des points donnés. Il apparaît que l'estimateur choisi, bien qu'ayant de bonnes propriétés asymptotiques théoriques, n'aboutit pas à des résultats satisfaisants en pratique. Le problème semble être lié au caractère aléatoire du plan d'observations qui à distance finie entraîne une inhomogénéité dans la répartition des observations entre les supports des ondelettes. L'alternative proposée pour la mise en œuvre est celle de Chesneau et Willer (2005). Les résultats obtenus lors de l'application sur des simulations sont satisfaisants pour une large gamme de densités du plan des observations.

CHAPITRE 1

LE MODÈLE PARTIELLEMENT LINÉAIRE

Sommaire

1.1 Un modèle semi-paramétrique **7**
1.1.1 Pourquoi ce modèle 8
1.1.2 Exemples d'application 9
1.2 Inférence statistique **10**
1.2.1 Estimation par les moindres carrés pénalisés 10
1.2.2 La régression partielle de Speckman (1988) 14

Ce premier chapitre présente les modèles partiellement linéaires, qui font l'objet du présent travail. Outre leur formalisme, l'intérêt de leur mise en œuvre est souligné. Dans un second temps, nous évoquons quelques résultats théoriques quant à l'inférence statistique dans ces modèles.

1.1 Un modèle semi-paramétrique

Soient $(X_1 \dots, X_n)$ et $(T_1 \dots, T_n)$ deux échantillons de covariables à valeurs respectivement dans $\mathbb{R}^p$ et $\mathbb{R}^q$. Etant donnée l'observation sur $\mathbb{R}$ de $(Y_1, \dots, Y_n)$, nous considérons le modèle de régression :

$$y_i = X_i^T \beta + f(T_i) + U_i, \qquad i = 1, \dots, n, \tag{1.1}$$

où $(U_1, \dots, U_n)$ est un échantillon aléatoire issu de la loi normale centrée de variance σ^2. Sous forme matricielle, nous noterons :

$$Y = X\beta + F(T) + U.$$

Le vecteur de bruit U est supposé indépendant des plans d'observation X et T. Une dépendance entre les variables X et T peut être envisagée. Nous désignerons par la suite β_0 la valeur du vrai paramètre.

1.1.1 Pourquoi ce modèle

Le modèle (1.1) fait partie de la classe des modèles semi-paramétriques. Le but de cette gamme de modèle est d'offrir un compromis entre le modèle linéaire et le modèle non paramétrique. En effet, le modèle paramétrique et le modèle non paramétrique présentent tous deux des inconvénients. Le problème majeur du modèle paramétrique doit être considéré d'un point de vue de modélisation ; comme le soulignent Engle *et al.* (1986), le modèle de regression est trop réducteur et ne permet pas de prendre en compte toutes les intéractions. Ainsi, ils citent l'exemple de l'influence de la température sur la consommation d'électricité : la relation entre les variables ne peut s'exprimer sous une forme linéaire, même après transformation des données. Les outils non paramétriques, en ne posant pas d'hypothèses quant à la forme du lien explicatif entre les variables, deviennent indispensables pour traiter un tel cas.

Cependant, la liberté de modélisation qu'offre le domaine non paramétrique a un coût : schématiquement, pour estimer une fonction f k-fois dérivable, la vitesse optimale est de l'ordre de $n^{-\frac{k}{2k+p}}$, le paramètre p désignant la dimension des variables explicatives. Tandis que dans le modèle de régression paramétrique, elles sont de l'ordre de $n^{-1/2}$ quel que soit le nombre de variables explicatives. Il apparaît par conséquent utile de prendre en compte la linéarité dans les modèles, afin d'augmenter la vitesse d'estimation. Ce constat, souvent rencontré sous l'expression de *fléau de la dimension*, a incité le développement depuis les années 80 d'une nouvelle gamme de modèles que sont les modèles semi-paramétriques. L'idée de ces modèles est de combiner les modèles non paramétriques et les modèles linéaires.

Le modèle partiellement linéaire que nous allons aborder ici est un des nombreux modèles semi-paramétriques existants. Ce modèle a été introduit à notre connaissance par Green et Yandell (1985) mais a particulièrement connu son essor après l'article de Engle *et al.* (1986). La capacité de modélisation de phénomènes divers ainsi que les résultats satisfaisants d'estimation statistique dans ce modèle ont contribué à son développement.

L'intérêt principal du modèle partiellement linéaire est qu'il permet de distinguer les relations linéaires et les relations non linéaires au sein d'un même modèle. L'idée est de

prendre en compte l'a priori que l'on a quant à la linéarité de certaines relations afin de réduire le coût de l'estimation qu'aurait un modèle non paramétrique, tout en gardant la complexité de modélisation sous-jacente au modèle non paramétrique pour expliquer les autres relations.

1.1.2 Exemples d'application

Un premier exemple d'application de ce modèle au milieu industriel est l'étude de la consommation électrique en fonction de la température réalisée par Engle *et al.* (1986). Les auteurs veulent étudier l'influence de la température sur la consommation électrique des ménages. Ils expliquent cette consommation par un modèle partiellement linéaire dans lequel la partie linéaire intègre les effets du prix de l'électricité, du revenu des ménages, etc... Cette modélisation permet donc d'étudier la relation entre la température et la consommation d'électricité sans que cette étude soit biaisée par la présence des covariables de la partie linéaire.

En biologie, on peut citer notamment les travaux de Green et Yandell (1985) qui analysent l'influence de la dose d'un produit donné sur l'hyperplasie de la voie biliaire des rats. La présence d'hyperplasie à la mort du rat est expliquée de manière linéaire par le logarithme de la dose reçue, le poids du rat et sa position dans la cage, et de manière non linéaire par l'âge de la mort. Il est intéressant de remarquer que la relation sur laquelle se focalise l'analyse est ici linéaire, contrairement à l'étude de Engle *et al.* (1986).

Plus récemment F. Meyer (2003) puis Fadili et Bullmore (2005) s'intéressent à la modélisation d'un signal issu d'Imagerie à Résonance Magnétique fonctionnelle par un modèle partiellement linéaire. L'idée est d'intégrer dans la partie linéaire du modèle des paramètres de nuisance, afin de conserver dans la partie fonctionnelle un signal épuré. De tels signaux seront étudiés au Chapitre 6.

L'éventail des domaines où ce modèle peut être appliqué est vaste. Hardle *et al.* (2000) illustrent son aptitude à modéliser des phénomènes dans des domaines aussi variés que l'industrie, la finance, la biologie ... Nous renvoyons à leur ouvrage pour un aperçu plus large des applications possibles.

Si ce modèle présente souvent un intérêt, on est amenés à se poser la question de savoir dans quel cas il est adapté et quelles sont les variables correspondant respectivement aux parties linéaire et fonctionnelle du modèle. A notre connaissance, seuls Chen et Chen

(1991) et Roger-Vial (2003) ont abordé ce point. Nous renvoyons à leurs travaux pour plus de détails.

1.2 Inférence statistique

L'objet du problème est d'estimer le vecteur des paramètres de regression β_0 ainsi que la fonction f au vu de l'observation des triplets $(Y_i, X_i, T_i)_{i=1,\dots,n}$.

Parmi les techniques développées dans l'estimation de la partie fonctionnelle dans un modèle partiellement linéaire, outre les splines et l'estimation à noyau, on peut citer aussi l'estimation linéaire locale (Hamilton et Truong (1997)), l'estimation par la méthode des polynômes locaux (Roger-Vial (2003)), l'estimation par séries orthogonales (Donald et Newey (1994))... A titre de remarque, l'application des techniques de bootstrap a aussi été mise en œuvre dans ce modèle par Hardle *et al.* (1999). Mais nous n'aborderons pas ces travaux ici dans un souci de concision. Ne pouvant présenter de manière exhaustive l'ensemble des approches considérées, nous nous intéressons plus particulièrement à l'estimation par minimisation d'un critère des moindres carrés pénalisés ainsi que l'approche dite de *régression partielle* de Speckman (1988), étant donné que nous avons opté pour une procédure similaire dans ce travail.

1.2.1 Estimation par les moindres carrés pénalisés

La première approche, proposée notamment par Green et Yandell (1985), Engle *et al.* (1986) ou encore Chen et Shiau (1991), est l'estimation reposant sur un critère des Moindres Carrés Pénalisés. Les estimations de β_0 et f sont obtenues en minimisant un critère de la forme :

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta - f(T_i))^2 + \lambda Pen(f), \tag{1.2}$$

où $Pen(f)$ est un indicateur de la régularité de f. L'idée d'un tel critère est d'introduire la régularité de la partie fonctionnelle de sorte que la fonction obtenue ne "colle" pas aux données mais de la contraindre à être suffisament lisse. Par ailleurs, l'introduction d'une pénalité est justifiée par la théorie bayésienne, comme le font remarquer notamment Green et Yandell (1985) et Heckman (1986).

Typiquement, la pénalité choisie est de la forme $Pen(f) = \int (f^{(m)})^2$; c'est le choix par

exemple de Heckman (1986), Rice (1986) ou Chen (1988). Cette pénalité correspond en fait à supposer que f appartient à un espace de Sobolev

$$W_m^2 := \left\{ g \; m - \text{fois différentiable, telle que} \int (g^{(m)})^2 < \infty \right\}.$$

Dans le cadre d'un modèle non paramétrique, l'optimum des moindres carrés avec une telle pénalité est atteint sur l'espace des splines de degré $2m-1$. Ainsi, il semble naturel d'approximer f par une projection sur l'espace des splines de degré $2m-1$. Considérons W matrice de projection sur un espace de splines. Alors, si $f = W^T\theta$, les conditions du premier ordre s'écrivent

$$\hat{\beta} = (X^TX)^{-1}X^T(y - W^T\hat{\theta}) \tag{1.3}$$

$$\hat{\theta} = W(y - X\hat{\beta}). \tag{1.4}$$

En remplaçant l'expression de $\hat{\theta}$ dans $\hat{\beta}$, nous obtenons :

$$\hat{\beta} = (X^T(I-S)X)^{-1}X^T(I-S)y$$

avec $S = W^TW$.

Cette approche a tout d'abord été appliquée par de nombreux auteurs parmi lesquels Green et Yandell (1985), Engle *et al.* (1986) et Heckman (1986). Notons qu'en particulier Rice (1986) et Chen et Shiau (1991) optent pour la base de Reinsch. Chen (1987) , Chen et Chen (1991) ou encore Schick (1996) proposent quant à eux d'estimer f dans un espace de splines engendré par des subdivisions de dimension supérieure à $2m-1$, adaptées au plan des observation $(T_i)_{i=1,\dots,n}$.

Premiers résultats asymptotiques

Les premiers résultats relatifs au comportement asymptotique des estimateurs construits par moindres carrés pénalisés sont à notre connaissance dus à Heckman (1986) et Rice (1986). Leurs travaux ont été élaborés dans le contexte d'une pénalité de Sobolev, associée à une projection dans un espace de splines comme évoqué ci-dessus. Nous présentons brièvement ces résultats ainsi que leur limite avant d'évoquer des méthodes d'estimation proposées pour améliorer leur performance.

Normalité asymptotique

Heckman (1986) établit la normalité asymptotique de l'estimateur $\hat{\beta}_n$:

Théorème 1.1. *(Heckman (1986))*
Supposons que f appartienne à l'espace de Sobolev W^m.
Supposons qu'existe une fonction w continue et strictement positive telle que le vecteur des covariables t_i vérifie la relation $\int_0^{t_i} w(u)du = \frac{i}{n}$.
Supposons de plus que $(X_1, \ldots, X_n)$ sont des variables aléatoires i.i.d. centrées, de matrice de covariance Σ et admettant un moment d'ordre 4.

Si $\sqrt{n}\lambda$ tend vers l'infini, alors $\hat{\beta}_n$ est un estimateur consistant de β_0.
Si de plus λ tend vers 0, ou $f^{(m)} \equiv 0$, alors $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converge en loi vers une variable aléatoire B de loi gaussienne centrée et de matrice de covariance $\sigma^2\Sigma^{-1}$.

Ce qu'il est intéressant de remarquer dans ce résultat est que l'estimateur atteint la vitesse minimax d'un modèle linéaire. Autrement dit, la présence d'une partie fonctionnelle dans le modèle n'affecte pas la vitesse de convergence de l'estimation, et nous obtenons de même un estimateur asymptotiquement sans biais.

Notons que la normalité asymptotique de $\hat{\beta}_n$ fut ensuite étenduee au cas des modèles partiellement linéaires généralisés avec une procédure des quasi-moindres carrés pénalisés par Chen (1987).

Vitesses minimax : problème de variables corrélées

Rice (1986) dans un contexte similaire s'est intéressé à ce que devenait la consistance de $\hat{\beta}_n$ établie par Heckman (1986) lorsque l'hypothèse de non corrélation entre les variables X et T était amoindrie. Plus précisément, Rice (1986), sous l'hypothèse $p = 1$, considère que les covariables X peuvent s'écrire sous la forme

$$x_i = g(T_i) + \eta_i, \tag{1.5}$$

avec $(\eta_i)_{j=1,\ldots,p}$ n-échantillon issu d'une variable centrée de variance finie (ou covariables se comportant comme tel).

Lorsque la fonction g s'annule, le cadre de l'étude est celui de Heckman (1986). Rice (1986) montre qu'alors non seulement $\hat{\beta}_n$ admet une vitesse optimale de convergence, mais que de plus l'estimation de la partie fonctionnelle atteint une vitesse minimax. L'auteur montre en fait ce résultat sous des hypothèses moins restrictives :

Théorème 1.2. *(Rice (1986))*
Supposons que f appartient à l'espace de Sobolev W^m et que les covariables x_i, $i = 1, \ldots, n$, vérifient (1.5) avec g fonction polynomiale de degré inférieur ou égal à $m - 1$. Si $\lambda = n^{-1/(1+2m)}$,

alors

$$\begin{aligned} \mathbb{E}\|\hat{\beta}_n - \beta_0\| &= \bigcirc(n^{-1/2}) \\ \mathbb{E}\|\hat{f}_n - f\| &= \bigcirc(n^{-m/(1+2m)}). \end{aligned}$$

Autrement dit, la présence de la partie non paramétrique dans le modèle n'affecte pas la qualité de l'estimation de la partie paramétrique et réciproquement.

Néanmoins, Rice (1986) considère aussi des fonctions g plus complexes et conclut que si les hypothèses sur la fonction g sont amoindries (c'est-à-dire que la corrélation entre les variables explicatives est plus forte) le biais de l'estimateur $\hat{\beta}_n$ résultant est trop important. Plus précisément, si g est m-fois continûment dérivable, il va falloir pour pouvoir atteindre la vitesse optimale dans la régression linéaire choisir un paramètre λ qui va imposer un sous-lissage dans l'estimation de f. En conséquence, l'estimation de f ne pourra être optimale. Ainsi, un compromis doit alors être réalisé entre l'optimalité dans l'estimation de la partie linéaire et celle dans l'estimation de la partie non paramétrique du modèle.

La conclusion de ces résultats est qu'on peut espérer atteindre la normalité asymptotique dans l'estimation du vecteur de régression de la partie linéaire et des vitesses minimax respectivement pour la partie linéaire et la partie fonctionnelle du modèle, sans que le fait d'être dans un modèle partiellement linéaire affecte les vitesses respectives des deux parties. Cependant, ces bons comportements asymptotiques sont valables sous la condition que la corrélation entre les variables de la partie linéaire et celles de la partie non paramétrique du modèle ne soit pas forte.

La conclusion de Rice (1986) sur le comportement des estimateurs dans le cadre de variables corrélées a engendré de nombreuses reflexions. Attardons nous notamment sur les articles de Chen et Shiau (1991) et de Speckman (1988).

Une estimation en deux étapes : Chen et Schiau (1991)

Chen et Shiau (1991) construisent une estimation en deux étapes afin de "décorréler" les variables explicatives. Cette méthode sera ensuite reprise par Chen et Chen (1991) et par Hamilton et Truong (1997) et peut aussi être trouvée dans Antoniadis (1992). Cette approche consiste à extraire des variables explicatives de la partie linéaire de nouvelles variables non corrélées avec la partie fonctionnelle. L'idée est de régresser les variables

explicatives X du modèle (1.1) par rapport aux variables T selon un modèle non paramétrique. Cela signifie que nous supposons que les variables X s'écrivent :

$$X_i = g(T_i) + R_i,\ i = 1, \ldots, n.$$

Les auteurs proposent d'estimer g selon les techniques usuelles d'estimations non paramétriques afin d'en déduire les covariables R non corrélées avec T. L'estimation par moindres carrés pénalisés (1.2) sera ensuite effectuée en considérant que la matrice de régression est non pas la matrice X initiale mais la matrice des résidus R obtenus dans la première étape.

Chen et Shiau (1991) introduisent les matrices de régression par splines S_λ. Les auteurs utilisent des paramètres de lissages λ_j distincts pour chacune des p regressions $x_{i,j} = g_j(t_i) + r_{i,j}$, $i = 1, \ldots, n$ et un paramètre de lissage λ pour le critère des moindres carrés pénalisés sur les résidus. Ils obtiennent des estimateurs de la forme :

$$\begin{aligned} \hat{\beta}_n &= (\tilde{X}^T(I_{pn} - \tilde{S})^T(I_n - S_\lambda)(I_{pn} - \tilde{S})\tilde{X})^{-1}\tilde{X}^T(I_{pn} - \tilde{S})^T(I_n - S_\lambda)y \\ \hat{f}_n &= S_\lambda(y - X\hat{\beta}_n) - (I - S_\lambda)S\tilde{X}\hat{\beta}_n, \end{aligned}$$

avec $\tilde{X}$ et $\tilde{S}$ matrices par blocs définies par : $\tilde{X} = diag(X_1 \ldots X_p)$ et $\tilde{S} = diag(S_{\lambda_1} \ldots S_{\lambda_p})$.

Chen et Shiau (1991)) aboutissent par ce moyen à des estimateurs minimax et à un $\hat{\beta}_n$ asymptotiquement normal.

1.2.2 La régression partielle de Speckman (1988)

L'article de Speckman (1988), outre le fait de proposer d'autres estimateurs, généralise aussi l'estimation par splines à une estimation par noyaux dans les modèles partiellement linéaires.

Dans le cadre d'un modèle non paramétrique, $Y = f(T) + U$, considérons que l'estimation de f s'écrit $\hat{f} = WY$. W est une matrice de lissage par noyaux ou de projection sur une base de splines, supposée symétrique. Par ailleurs, l'estimation du paramètre de régression choisie dans le modèle $Y = X\beta + U$ est celle des moindres carrés $\hat{\beta} = (X^TX)^{-1}X^TY$. Les deux options d'estimation possibles dans le modèle partiellement linéaire, outre l'estimation conjointe des deux parties, sont :

1. Si f était connue, le vecteur β serait estimé par $\hat{\beta} = (X^TX)^{-1}X^T(Y - f(T))$. Il en découle que f peut être estimée à partir du modèle $Y - X\hat{\beta} = f(T) + U$ par $\hat{f} =$

$W(Y - X\hat{\beta})$. En remplaçant f par $\hat{f}$ dans l'expression de $\hat{\beta}$, les estimations obtenues s'écrivent alors :

$$\begin{cases} \hat{\beta}_n^{(1)} & = & (X^T(I-W)X)^{-1}X^T(I-W)Y \\ \hat{f}_n^{(1)} & = & W(Y - X\hat{\beta}) \end{cases}.$$

Nous obtenons ici la forme des estimateurs étudiés par Heckman (1986), Rice (1986) et Chen (1988).

2. Si β était connu, alors f serait estimée par $\hat{f} = W(Y - X\beta)$. De même que précédemment, β est obtenu par estimation des moindres carrés dans le modèle $Y - \hat{f}(T) = X\beta + U$ puis la valeur obtenue est remplacée dans $\hat{f}$. Cette technique aboutit aux estimations suivantes :

$$\begin{cases} \hat{\beta}_n^{(2)} & = & (X^T(I-W)^2X)^{-1}X^T(I-W)^2Y \\ \hat{f}_n^{(2)} & = & W(Y - X\hat{\beta}) \end{cases}.$$

Telle que présentée par Speckman (1988), cette estimation consiste à projeter le modèle dans l'espace orthogonal à W de sorte à faire disparaître la partie non paramétrique puis à appliquer une estimation des moindres carrés pour obtenir une estimation du vecteur de regression de la partie linéaire.

La conclusion de Speckman (1988) est que la deuxième méthode est préférable. En effet, il montre que si les variances des estimateurs sont inchangées, les biais de la méthode 2 sont réduits par rapport à la méthode 1 lorsque les variables sont corrélées, c'est-à-dire lorsque les variables X vérifient (1.5) avec g n'appartenant pas à la classe des polynômes de degré inférieur ou égal à $m - 1$. Speckman montre par ailleurs la normalité asymptotique de $\hat{\beta}_n$ tel que défini en 2. Quant aux estimations des parties fonctionnelles, les deux procédures mènent à des vitesses minimax, pour le choix usuel de lissage.

Résumons les propriétés asymptotiques établies par l'auteur :

Théorème 1.3. *(Speckman (1988))*
Supposons qu'existe une fonction w continue et strictement positive telle que le vecteur des covariables t_i vérifie la relation $\int_0^{t_i} w(u)du = \frac{i}{n}$.
Supposons que f appartient à l'espace de Sobolev W^m et que les covariables x_i, $i = 1, \ldots, n$, vérifient (1.5) avec g fonction m-fois continûment dérivable.

Alors

$$\begin{aligned} \mathbb{E}\hat{\beta}_n^{(1)} - \beta_0 &\sim \lambda^m, \\ \mathbb{E}\hat{\beta}_n^{(2)} - \beta_0 &= \bigcirc(\lambda^{2m} + \lambda^m n^{-1/2}) \end{aligned}$$

et $Var(\hat{\beta}_n^{(i)}) = \bigcirc(n^{-1})$ *pour* $i = 1, 2$. *De plus,*

$$\begin{aligned} \mathbb{E}\hat{f}_n^{(i)} - f &= \bigcirc(\lambda^m) \\ Var(\hat{f}_n^{(i)}) &= \bigcirc(n^{-1}\lambda^{-1}) \end{aligned}$$

pour $i = 1, 2$.

Si de plus les variables η *admettent un moment d'ordre strictement supérieur à 2 fini, alors* $\sqrt{n}(\hat{\beta}_n^{(2)} - \mathbb{E}\hat{\beta}_n^{(2)})$ *converge en loi vers une variable aléatoire B de loi gaussienne centrée et de matrice de covariance* $\sigma^2\Sigma^{-1}$.

Lorsque le paramètre de lisssage λ est pris de telle sorte que l'estimation de la partie fonctionnelle soit minimax, c'est-à-dire $\lambda = n^{-1/(1+2m)}$, ce théorème nous confirme bien que la procédure 2 permet de construire des estimateurs des parties linéaires et fonctionnelles qui sont optimaux ; tandis que le biais de l'estimation du vecteur de paramètres β dans la procédure 1 a une vitesse non paramétrique. L'auteur souligne par ailleurs le danger de la première méthode car la nécessité de mettre en place une procédure de validation croisée pour choisir le paramètre λ risque d'entraîner un sous-lissage dans l'estimation de la partie fonctionnelle.

Conclusion : Des splines aux ondelettes...

Depuis les années 90, l'usage des ondelettes en statistique s'est répandu (cf Antoniadis (1997)). Elles offrent en effet l'avantage d'élargir le champ de régularité de la partie fonctionnelle, comparativement aux techniques de splines, en permettant d'appréhender des fonctions moins lisses et spatialement plus hétérogènes que des fonctions de Sobolev. De plus, des procédures d'estimation non linéaires, via le seuillage des coefficients, permettent d'élaborer des estimateurs quasi-minimax qui sont adpatatifs quant à la régularité de la fonction estimée. Ceci est particulièrement intéressant dans le modèle partiellement linéaire en raison des biais possibles engendrés par la validation croisée soulignés par Rice (1986) et Speckman (1988).

Récemment, l'utilisation des ondelettes dans le modèle partiellement linéaire a été proposée par divers auteurs. Une estimation linéaire a été mise en place par F. Meyer (2003) ou Xue (2003), une estimation non linéaire par Chang et Qu (2004) et Fadili et Bullmore (2005), ou encore une estimation bayésienne par simulations Monte-Carlo par Chaînes de Markov avec des techniques d'ondeletttes par Qu (2005).

L'objet du présent travail est donc d'introduire à notre tour une estimation non linéaire par ondelettes dans ce modèle. Après avoir brièvement présenté les ondelettes et leur application en régression fonctionnelle, nous décrirons ainsi les procédures d'estimation construites en introduisant des ondelettes dans le modèle partiellement linéaire. Nous donnerons ensuite les résultats quant à leurs propriétés asymptotiques et étudierons leur mise en œuvre algorithmique.

CHAPITRE 2

RAPPELS SUR LES ONDELETTES ET LEUR USAGE EN RÉGRESSION

Sommaire

2.1 Analyse multirésolution et ondelettes 20
2.1.1 Approximations de fonctions 21
2.1.2 Ondelettes périodiques 22
2.1.3 Espaces de Besov 23
2.2 Transformation en ondelettes 25
2.2.1 Ecriture matricielle 25
2.2.2 Algorithme pyramidal 26
2.3 Régression avec un plan d'observation équidistant 28
2.3.1 Estimations linéaires 28
2.3.2 Estimations non linéaires : seuillage des coefficients 29
2.4 Régression avec un plan d'observation non-équidistant 33
2.4.1 Aperçu de différentes approches 33
2.4.2 Hypothèses sur le plan d'observation 36
2.4.3 Approche théorique 39
2.4.4 Approche pratique 41

Ce chapitre propose un bref rappel sur les ondelettes ainsi que sur leur usage dans le modèle de regression non paramétrique. Dans un premier temps, nous redonnons la définition des bases d'ondelettes orthonormées ainsi que des espaces de Besov, puis nous nous intéressons à l'estimation par transformée en ondelette dans le modèle de régression lorsque le plan d'observation est équidistribué et lorsque ce n'est pas le cas. Dans cette dernière section, nous insistons particulièrement sur la procédure d'estimation développée par Kerkyacharian et Picard (2004), qui sera appliquée au Chapitre 7.

2.1 Analyse multirésolution et ondelettes

Parmi les différentes approches possibles, nous avons choisi d'introduire les ondelettes à travers la notion d'analyse multirésolution. L'idée d'une analyse multirésolution (AMR) est de définir une suite d'espaces emboîtés $(V_j)_{j\in\mathbb{Z}}$ telle que à chaque incrémentation, l'approximation résultante d'une fonction f de $L^2(\mathbb{R})$ sur cette espace est affinée. De manière rigoureuse, une analyse multirésolution se définit de la façon suivante :

Définition 2.1. Analyse multirésolution Orthonormée (AMR)
On appelle Analyse multirésolution orthonormée de $\mathbf{L}^2(\mathbb{R})$ toute suite $(V_j)_{j\in\mathbb{Z}}$ croissante d'espaces fermés de $\mathbf{L}^2(\mathbb{R})$ telle que

(i) $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ et $\overline{\cup_{j\in\mathbb{Z}} V_j} = \mathbf{L}^2(\mathbb{R})$,

(ii) $\forall\, f \in \mathbf{L}^2(\mathbb{R}), \forall\, j \in \mathbb{Z}, f \in V_j$ si et seulement si $x \mapsto f(2^{-j}x)$ appartient à V_{j+1},

(iii) Il existe φ, appelée *fonction d'échelle*, telle que la famille $\{x \mapsto \varphi(x-k)\}_{k\in\mathbb{Z}}$ soit une base orthonormée de V_0.

Nous ne traiterons pas des bases biorthonormée d'ondelettes, où les analyses multirésolution sont définies à partir de bases de Riesz et non plus de bases orthonormées. Nous renvoyons aux livres de Daubechies (1992), Y. Meyer (1992) ou Mallat (1999) pour ces dernières.

Cette définition a pour conséquence qu'à chaque degré de résolution j, la famille de fonctions $\{\varphi_{j,k} : x \to 2^{j/2}\varphi(2^j x - k)\}_{k\in\mathbb{Z}}$ forme une base orthonormée de l'espace V_j pour la norme L^2. Comme φ appartient à V_0, qui est inclus dans V_1, il en résulte que φ peut s'exprimer comme combinaison linéaire de $\{\varphi_{1,k}\}_{k\in\mathbb{Z}}$. Autrement dit, il existe une suite de réels $(h_k)_{k\in\mathbb{Z}}$ telle que

$$\forall x \in \mathbb{R},\ \varphi(x) = \sum_{k\in\mathbb{Z}} h_k \varphi(2x-k).$$

Le principe de cette relation, appelée *relation à deux échelles*, permet comme nous le verrons ultérieurement d'élaborer des algorithmes rapides de décomposition ou de reconstruction dans le contexte d'une analyse multirésolution.

Il est intéressant dès lors de pouvoir affiner la connaissance sur une fonction en augmentant le niveau de résolution sans recalculer l'intégralité des coefficients associés. On introduit alors les espaces de *détails* : passer d'une résolution j à une résolution $j+1$ revient à ajouter des détails, compris dans l'espace W_j complémentaire de V_j dans V_{j+1} ;

pour tout $j \in \mathbb{Z}$, W_j est défini par :

$$V_{j+1} = V_j \oplus W_j.$$

Une conséquence immédiate de cette définition ainsi que de la Définition 2.1 est que pour tout j_0 appartenant à $\mathbb{Z}$, l'espace $L^2(\mathbb{R})$ vérifie :

$$L^2(\mathbb{R}) = \overline{V_{j_0} \oplus \oplus_{j=j_0}^{+\infty} W_j}. \tag{2.1}$$

Il existe une fonction ψ telle que $\{x \mapsto \psi(x-k)\}_{k\in\mathbb{Z}}$ soit une base orthonormée de W_0. La fonction ψ est alors appelée *ondelette*. De même que précédemment, à tout niveau de résolution $j \in \mathbb{Z}$, la famille $\{\psi_{j,k} : x \mapsto 2^{j/2}\psi(2^j x - k)\}_{k\in\mathbb{Z}}$ forme une base orthonormée de l'espace W_j. De plus, de manière similaire, une relation à deux échelles peut être établie : comme W_0 est inclus dans V_1, il existe une suite de réels $(g_k)_{k\in\mathbb{Z}}$ telle que

$$\forall x \in \mathbb{R},\ \psi(x) = \sum_{k\in\mathbb{Z}} g_k \varphi(2x-k).$$

Un exemple simple d'analyse multirésolution est celle de Haar, générée par la fonction d'échelle $\varphi = \mathbb{1}_{[0,1)}$ et l'ondelette $\psi = \mathbb{1}_{[1/2,1)} - \mathbb{1}_{[0,1/2)}$.

2.1.1 Approximations de fonctions

En utilisant la décomposition de l'espace $L^2(\mathbb{R})$ donnée en (2.1), nous déduisons de ce qui précède que pour tout $j_0 \in \mathbb{Z}$, toute fonction f appartenant à $L^2(\mathbb{R})$ s'écrit :

$$f = \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{+\infty}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}, \tag{2.2}$$

avec $\alpha_{j_0,k} = \int f\varphi_{j_0,k}$ et $\beta_{j,k} = \int f\psi_{j,k}$.

Définition 2.2. Ordre d'une AMR
Une analyse multirésolution orthogonale est dite d'ordre N si, pour tout degré de résolution $j \in \mathbb{Z}$, tout polynôme P de degré inférieur à $N-1$ appartient à V_j, i.e. P peut s'écrire sous la forme :

$$P = \sum_{k\in\mathbb{Z}} c_{j,k}\varphi_{j,k}.$$

Un calcul simple montre que l'ordre d'une Analyse Multi-résolution orthogonale est équivalent au nombre de moments nuls de l'ondelette associée. Rappelons la propriété des moment nuls.

Définition 2.3. Nombre de moments nuls
Une fonction f admet N moments nuls si pour tout $p = 0, \ldots, N-1$,

$$\begin{aligned} &\int f(x)x^p dx = 0, \\ \text{et} \quad &\int \left| f(x)x^N \right| dx < \infty. \end{aligned}$$

L'intérêt d'une décomposition multi-échelle est que, contrairement à une décomposition dans une base de Fourier, celle-ci est localisée en temps et en fréquence. Les valeurs des coefficients de détails sont faibles lorsque la fonction est régulière, mais deviennent élevés dans le voisinage des points de discontinuité. En effet, lorsque la fonction est localement polynômiale, la propriété de représentation des polynômes associée à l'odre de l'AMR assure que les coefficients de détails vont être faibles. Cette localisation dans les voisinages des points critiques permet une meilleure détection de ces points. De plus, les valeurs étant peu élevées dans les zones de régularité et décroissant avec le degré de résolution j, le fait de ne pas conserver ces coefficients a une influence moindre sur la reconstruction de la fonction. Ce principe peut permettre de compresser l'information en ne gardant que les coefficients significatifs. On peut aussi espérer qu'une telle procédure réduira les fréquences parasites du signal lorsque celui-ci est bruité.

L'ondelette de Haar, citée plus haut, présente l'inconvénient majeur de n'avoir qu'un seul moment nul. On préférera donc en général prendre des ondelettes ayant un nombre de moments nuls plus élevés. Daubechies (1992), entre autres, propose des ondelettes ayant des nombres de moments nuls élevés, et à support compact, qui seront utilisées dans les applications. D'autres exemples de fonctions d'échelle et d'ondelettes peuvent être trouvés par exemple dans Mallat (1999) ou Hardle *et al.* (1998).

2.1.2 Ondelettes périodiques

La fonction que l'on souhaiterait estimer étant définie sur $[0,1]$, la construction d'une base d'ondelettes sur $\mathrm{L}^2([0,1])$ est possible en périodisant les fonctions φ et ψ. Soient $\tilde{\varphi}_\kappa$

et $\tilde{\psi}_\kappa$ définies par :

$$\begin{aligned}\tilde{\varphi}_\kappa &= \sum_{l\in\mathbb{Z}} \tilde{\varphi}(x+l),\\ \tilde{\psi}_\kappa &= \sum_{l\in\mathbb{Z}} \tilde{\psi}(x+l).\end{aligned}$$

Alors, le couple $(\tilde{\varphi}, \tilde{\psi})$ engendre une analyse multirésolution orthonormée sur $[0,1]$. Cette périodisation est une manière simple et intuitive de restreindre une anlayse multirésolution à un intervalle, mais son inconvénient majeur est que des problèmes de discontinuités aux bords de l'intervalle en découlent. Nous renvoyons à la thèse de Maxim (2003) pour d'autres constructions d'ondelettes sur l'intervalle et notamment d'ondelettes adaptées aux bords.

Toute fonction $f \in \mathrm{L}^2([0,1])$ peut alors se décomposer sous la forme :

$$f = \sum_{k\in\mathbf{I_{j_0}}} \alpha_{j_0,k}\tilde{\varphi}_{j_0,k} + \sum_{j\geqslant j_0}\sum_{k\in\mathbf{I}_j} \beta_{j,k}\tilde{\psi}_{j,k}$$

avec $\alpha_\kappa = \int f\tilde{\varphi}$ et $\beta_\kappa = \int f\tilde{\psi}$, et $\mathbf{I}_j = \{k = 0, 1, \ldots, 2^j - 1\}$, la restriction des indices k aux ensembles $\mathbf{I}_j$ étant due à la périodicité des fonctions considérées.

2.1.3 Espaces de Besov

Nous étudierons par la suite le cas de fonctions appartenant à des espaces de Besov. Nous rappelons par conséquent leur définition. Pour les preuves des résultats énoncés dans ce chapitre, on pourra se référer notamment à Hardle *et al.* (1998).

Les espaces de Besov sont définis à partir du module de continuité, qui est une notion plus faible que la dérivabilité, donc permet aux espaces de Besov d'être peu contraignants quant à la régularité des fonctions.

Pour $1 \leqslant \pi \leqslant \infty$, le module de continuité d'ordre π d'une fonction f, $\omega_\pi^N(f,.)$, se définit comme :

$$\omega_\pi^N(f,t) := \sup_{|h|\leqslant t} \|\Delta_h^N f\|_\pi \qquad où \qquad \Delta_h f(x) = f(x+h) - f(x) \tag{2.3}$$

$$et \qquad \Delta_h^N f = \underbrace{\Delta_h o \Delta_h o \ldots o \Delta_h}_{N\ fois}\ f. \tag{2.4}$$

Les espaces de Besov sont ensuite construits de la façon suivante :

Définition 2.4. Espaces de Besov
Soient $0 < s < \infty$, $0 < \pi \leqslant \infty$ et $0 < q \leqslant \infty$. On définit l'espace de Besov $\mathcal{B}^s_{\pi,r}$ comme l'ensemble des fonctions f appartenant à $\mathbf{L}^\pi(\mathbb{R})$ telles que

$$\left(\int_0^\infty \left(\frac{\omega_\pi^N(f,t)}{t^s}\right)^q \frac{dt}{t}\right)^{1/q} < \infty,$$

avec N entier tel que $s < N$ et la modification usuelle pour $q = \infty$.

Remarque : La définition ci-dessus semble dépendre de l'entier N considéré, mais la caractérisation d'un espace de Besov en terme d'ondelettes permet en réalité de montrer qu'il n'en est rien.

Les espaces de Besov sont particulièrement adaptés à l'estimation par projection sur une base d'ondelettes en raison de leur caractérisation par rapport à une telle base.

Proposition 2.5. *Caractérisation des espaces de Besov par ondelettes*
Soient φ une fonction d'échelle issue d'une Analyse Multi-Résolution et ψ l'ondelette associée. Supposons que φ vérifie la propriété des N moments nuls et qu'il existe une constante C telle que ψ vérifie

$$\begin{aligned} \sum_{k\in\mathbb{Z}} |\psi(x-k)| &\leqslant C \quad p.s., \\ \sum_{k\in\mathbb{Z}} |D^N\psi(x-k)| &\leqslant C \quad p.s. \end{aligned}$$

Soit $0 < s < N$, $1 \leqslant \pi \leqslant \infty$ et $0 < r \leqslant \infty$ et $f = \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\varphi_{j_0,k} + \sum_{j\geqslant j_0}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}$ appartenant à $\mathbf{L}^\pi(\mathbb{R})$.

Alors, $f \in \mathcal{B}^s_{\pi,r}$ si et seulement si $\sum_{j\in\mathbb{N}} 2^{j(s+1/2-1/\pi)r}\left(\sum_{k\in\mathbb{Z}}|\beta_{j,k}|^\pi\right)^{r/\pi} < \infty$. Avec les modifications usuelles si π ou r est infini.

Ainsi dès que l'ondelette considérée est suffisamment régulière et vérifie une condition de moments nuls, les espaces de Besov sont caractérisés par le comportement des coefficients de la décomposition sur une base d'ondelettes. Cette propriété fondamentale incite à estimer les fonctions appartenant à des espaces de Besov. C'est pourquoi dans toute la suite, nous considérerons que la fonction f que l'on cherche à estimer appartient à un ensemble $\mathcal{B}^s_{\pi,r}$. On remarquera en particulier que Y. Meyer (1992) et Daubechies (1992) proposent des bases d'ondelettes vérifiant les propriétés nécessaires à la caractérisation par ondelettes.

Les espaces de Besov peuvent alors être munis de la norme

$$\|f\|_{s,\pi,r} = \|f\|_{\mathbf{L}^\pi} + \sum_{j\in\mathbb{N}} 2^{j(s+1/2-1/\pi)r} \left(\sum_{k\in\mathbb{Z}} |\beta_{j,k}|^\pi \right)^{r/\pi}.$$

La boule $\{\|f\|_{s,\pi,r} \leqslant M\}$ sera notée par la suite $\mathcal{B}^s_{\pi,r}(M)$. Par abus de notation, les espaces de Besov définis sur $\mathbb{R}$ seront confondus par la suite avec les espaces définis sur un intervalle afin de ne pas prêter à confusion.

2.2 Transformation en ondelettes

Cette section donne dans un premier temps l'écriture matricielle de la transformée en ondelette et de la transformée inverse associée. Ensuite, elle décrit comment ces transformées peuvent être mises en œuvre algorithmiquement.

2.2.1 Ecriture matricielle

Soit $(V_j)_{j\in\mathbb{N}}$ une Analyse Multi Résolution orthogonale sur $[0;1]$ engendrée par la fonction d'échelle φ. On note $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$. Alors $V_j = Vect(\varphi_{j,k}, k = 0 \dots 2^j - 1)$. Soit ψ l'ondelette associée. On utilise des notations similaires pour $\psi_{j,k}$ et les espaces engendrés par $\{\psi_{j,k}\}_k$ sont notés W_j.

La matrice W de projection sur l'espace $V_J = V_{j_0} \oplus_{j=j_0}^{J-1} W_j$ est une matrice orthogonale de taille $n \times n$. Nous distinguerons par la suite les deux blocs engendrés respectivement par les fonctions d'échelle et les ondelettes.

Pour un vecteur quelconque $e = (e_1 \dots e_n)^T$ la transformée en ondelette $d = We$ est de la forme :

$$We = \begin{pmatrix} s_{j_0,0} & \dots & s_{j_0,2^{j_0}-1} & w_{j_0,0} & \dots & w_{j_0,2^{j_0}-1} & \dots & w_{J-1,2^{J-1}-1} \end{pmatrix}^T$$

où $s_{j,k}$ et $w_{j,k}$sont les coefficients empiriques respectivement d'échelle et d'ondelette de e

de degré (j,k). L'approximation suivante permet de mieux appréhender ces coefficients :

$$\begin{aligned} s_{j_0,k} &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i \varphi_{j_0,k}(t_i) \quad \text{pour } k = 0\ldots 2^{j_0}-1 \\ w_{j,k} &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i \psi_{j,k}(t_i) \quad \text{pour } \left\{ \begin{array}{lcl} j &=& j_0 \ldots J-1, \\ k &=& 0\ldots 2^j-1. \end{array} \right. \end{aligned}$$

Cette approximation servira notamment dans l'établissement des propriétés asymptotiques.

La transformée étant orthogonale, le vecteur e peut ensuite être obtenu à partir de sa transformée en ondelette d par $e = W^T d$.

2.2.2 Algorithme pyramidal

Nous présentons ici la mise en œuvre algorithmique du calcul des coefficients d'ondelettes et de la reconstruction d'un signal à partir de la donnée des coefficients.

Les relations à deux échelles données ci-dessus permettent d'établir une relation entre les coefficients à différentes échelles. Mallat (1989) propose deux algorithmes issus de ce constat permettant l'un de calculer les coefficients de la décomposition de f à des niveaux de plus en plus fin et l'autre de calculer les coefficients à des niveaux de plus en plus grossiers. Ces algorithmes sont appelés respectivement *algorithme de décomposition* et *algorithme de reconstruction*.

D'après ce qui précède, il existe deux suites de réels $(h_k)_{k\in\mathbb{Z}}$ et $(g_k)_{k\in\mathbb{Z}}$ telles que

$$\begin{aligned} \forall x \in \mathbb{R}, \quad \varphi(x) &= \textstyle\sum_{k\in\mathbb{Z}} h_k \varphi(2x-k) \\ \psi(x) &= \textstyle\sum_{k\in\mathbb{Z}} g_k \varphi(2x-k). \end{aligned}$$

Sous forme matricielle, ces équations donnent :

$$\forall j \in \mathbb{Z},\ \Phi_j = H_j \Phi_{j+1} \ \text{ et } \ \Psi_j = G_j \Phi_{j+1}.$$

Ces relations établissent un lien entre les coefficients d'échelle s_j et d'ondelettes w_j de différents niveaux de résolutions. Il en résulte en effet que

$$\begin{aligned} \forall j \in \mathbb{Z}, \quad s_j = H_j s_{j+1} \ \text{ et } \ w_j &= G_j w_{j+1} \\ s_{j+1} &= H_j s_j + G_j w_j. \end{aligned}$$

La dernière de ces équations permet de reconstruire le signal à partir des coefficients d'échelles et d'ondelettes de diverses résolutions tandis que les premières servent au calcul des coefficients d'échelles et d'ondelettes d'un signal. Ces algorithmes permettent de décomposer et de reconstruire de manière efficace un signal en coefficients d'ondelettes. Ce sont eux qui seront appliqués lors de la mise en œuvre pratique.

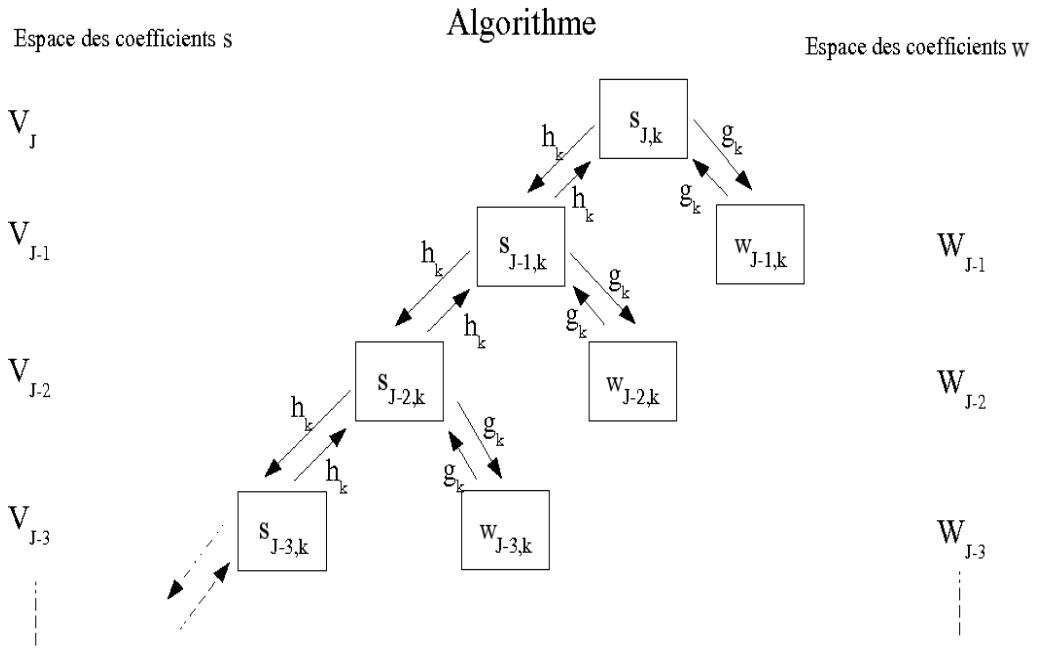

FIG. 2.1: Algorithmes de décomposition et de reconstruction.

Nous n'avons pas construit les algorithmes présentés ici mais nous appliquons ceux du module WaveLab de Matlab, élaborés par Donoho et ses collaborateurs au Département de Statistiques de l'Université de Stanford.

2.3 Régression non paramétrique avec ondelettes : cas d'un plan d'observation équidistant

Dans cette section, nous nous intéressons au problème de l'estimation par ondelettes d'une fonction f définie sur $[0,1]$ et à valeurs réelles dans le modèle suivant :

$$y_i = f(t_i) + v_i,\ i = 1, \dots, n,\ n = 2^J,\ J \in \mathbb{N}, \tag{2.5}$$

où le plan $(t_i)_{i=1,\dots,n}$ désigne le plan équidistant sur $[0,1]$: $t_i = \frac{i}{n}$, et les erreurs $(v_i)_{i=1,\dots,n}$ sont i.i.d. de loi normale $\mathcal{N}(0, \sigma_v^2)$. Nous noterons par ailleurs $Y = (Y_1 \dots Y_n)^T$ le vecteur des observations.

A partir de la représentation en ondelettes donnée en (2.2), nous pouvons élaborer une procédure d'estimation dans ce modèle. Nous distinguerons les estimateurs linéaires des estimateurs non linéaires : les premiers consistent à estimer f par une estimation à noyau, le noyau étant élaboré à l'aide d'ondelettes, tandis que les seconds essayent de ne pas considérer dans l'estimation les coefficients de moindre importance relativement à l'information sur f en introduisant des techniques de seuillage.

2.3.1 Estimations linéaires

Considérons que nous avons à notre disposition les observation $(t_i, y_i)_{i=1,\dots,n}$ ainsi qu'une base d'ondelettes orthogonale générerée par le couple (φ, ψ). L'estimateur linéaire de f construit à partir de la décomposition (2.2) consiste à estimer la projection de f sur l'espace V_{j_1} de la façon suivante :

$$\hat{f} = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k} \varphi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j_0,k} \psi_{j_0,k} \tag{2.6}$$

où les coefficients $\hat{\alpha}_{j_0,k}$ et $\hat{\beta}_{j,k}$ sont les coefficients empiriques respectivement d'échelle et d'ondelette, donnés par

$$\hat{\alpha}_{j_0,k} = \frac{1}{n} \sum_{i=1}^{n} y_i \varphi_{j_0,k}(t_i), \quad \text{et} \quad \hat{\beta}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} y_i \psi_{j,k}(t_i)\ ; \tag{2.7}$$

$\hat{\alpha}_{j_0,k}$ et $\hat{\beta}_{j,k}$ sont les estimateurs naturels des coefficients $\alpha_{j_0,k} = \int_0^1 f \varphi_{j_0,k}$ et $\beta_{j,k} = \int_0^1 f \psi_{j,k}$.

Le paramètre j_1 joue le rôle d'un paramètre de lissage : plus j_1 est grand, plus le nombre de coefficients intervenant dans l'estimation, donc de détails, est grand. Mais trop de coefficients peut avoir pour conséquence un sous-lissage dans l'estimation de la fonction.

Le choix optimal de j_1, menant à une vitesse minimax sur un espace de Besov $\mathcal{B}^s_{\pi,r}$, est $2^{j_1} = n^{\frac{s}{1+2s}}$. Ce choix dépend de la régularité de la fonction f que l'on cherche à estimer. L'estimateur n'est donc pas adaptatif. En outre, les limites de l'estimation linéaire ont été mises en évidence, comme cela peut être vu dans Donoho et Johnstone (1998). Ceci incite à élaborer des techniques non linéaires d'estimation. Les procédures de seuillage, présentées ci-après, répondent à cette attente.

Remarque : Nous avons dit en introduction que cet estimateur pouvait s'écrire comme un estimateur à noyau. Ceci résulte du fait que l'on peut écrire l'estimateur en (2.6) sous la forme

$$\forall x \in [0,1],\ \hat{f}_n(x) = \sum_{i=1}^{n} K_{j_0,j_1}(x,t_i)Y_i, \tag{2.8}$$

où K_{j_0,j_1} est défini par

$$\forall (u,v) \in [0,1]^2, K_{j_0,j_1}(u,v) = \frac{1}{n}\sum_{k=0}^{2^{j_0}-1} \varphi_{j_0,k}(u)\varphi_{j_0,k}(v) + \frac{1}{n}\sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} \psi_{j,k}(u)\psi_{j,k}(v).$$

Grâce à cette écriture, nous pouvons notamment appliquer les résultats de Speckman (1988) dans le cadre d'un modèle partiellement linéaire lorsqu'une estimation par ondelettes linéaire est réalisée.

2.3.2 Estimations non linéaires : seuillage des coefficients

L'estimation avec seuillage a été introduite en 1994 par Donoho et Johnstone. L'idée est de ne pas tenir compte des coefficients d'ondelettes de faible valeur, en considérant qu'ils transcrivent du bruit plus qu'une information sur le signal. En effet, si la représentation en ondelettes de la fonction f que l'on cherche à estimer est creuse (comme cela est le cas par exemple lorsque f appartient à un espace de Besov, défini en Section 2.1.3) cela signifie que seul un nombre réduit de coefficients véhiculent l'information nécessaire à l'estimation de f.

Principe et différents types de seuillage

Lorsque nous écrivons les coefficients d'ondelettes empiriques intervenant dans l'estimation de la fonction f, donnée en (2.6), nous pouvons remarquer qu'ils se décomposent en deux termes :

$$\forall (j,k),\ \hat{\beta}_{j,k} = \beta_{j,k} + \frac{1}{n}\sum_{i=1}^{n} v_i \psi_{j,k}(t_i).$$

Cette décomposition est immédiate en remplaçant $f(t_i)$ par son expression dans le modèle (2.5).

Si le niveau de bruit est élevé comparativement au coefficient de détail $\beta_{j,k}$, alors, il devient raisonnable de penser que le deuxième terme dans l'estimation $\hat{\beta}_{j,k}$ est prédominant. La représentation creuse d'un signal en coefficients d'ondelettes incite alors à ne prendre que les coefficients relativement élevés. Le principe du seuillage est ainsi de se donner un seuil à partir duquel on considérera que l'on peut conserver les coefficients dans l'estimation.

De façon plus formelle, nous introduisons l'estimateur seuillé de f comme suit :

$$\hat{f} = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} \gamma(\hat{\beta}_{j_0,k}, \lambda_j)\psi_{j_0,k}, \tag{2.9}$$

la fonction $\gamma(.,\lambda)$ désignant la fonction de seuillage de seuil λ.

Les seuillages les plus connus sont le seuillage fort (*hard thresholding*) :

$$\gamma_H(u,\lambda) = u\mathbb{1}_{|u|>\lambda}$$

et le seuillage doux (*soft thresholding*) :

$$\gamma_S(u,\lambda) = signe(u)(|u|-\lambda)_+.$$

De nombreux types de seuillages peuvent être rencontrés, comme nous verrons par la suite. Citons notamment la procédure de seuillage SCAD, qui réalise en quelque sorte un compromis entre les seuillages fort et doux. Rencontrée par exemple dans Fan et Li (2001) ou Antoniadis et Fan (2001), sa fonction de seuillage est :

$$\gamma_{SCAD}(u,\lambda) = \begin{cases} signe(u)(|u|-\lambda)_+ & \text{si } |u| \leqslant 2\lambda \\ \frac{1}{a-2}((a-1)u - a\lambda signe(u)) & \text{si } 2\lambda < |u| \leqslant a\lambda \\ signe(u)(|u|-\lambda)_+ & \text{si } |u| > a\lambda, \end{cases}$$

avec $a = 3.7$ si l'on considère les arguments bayésiens de Fan et Li (2001). La figure ci-contre illustre ces trois fonctions de seuillages. Remarquons que la particularité du seuillage fort est que la fonction associée n'est pas continue, contrairement aux seuillages doux et SCAD.

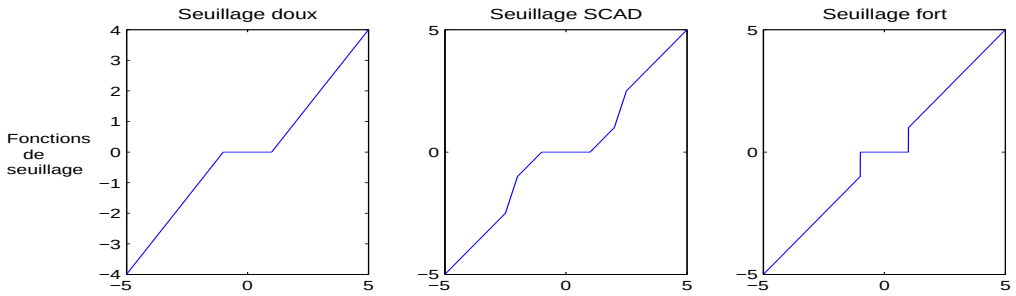

FIG. 2.2: Fonctions de seuillages des seuillages doux, SCAD et fort.

Choix du seuil et estimation de la variance

Dans la forme de l'estimation donnée en (2.9), nous avons considéré qu'à chaque niveau de résolution j correspondait un seuil λ_j. Comme nous nous intéressons par la suite à des bruits blancs i.i.d., un unique seuil peut être utilisé uniformément sur les niveaux de résolution. Cependant, lorsque l'hypothèse de non corrélation des résidus n'est plus vérifiée (ou lorsque la transformée en ondelettes n'est pas orthogonale, mais ce cas n'a pas été évoqué ici) il est nécessaire d'adapter le seuil au niveau de résolution. Mais nous verrons des procédures de seuillages avec des données i.i.d. faisant appel à des seuils dépendant du degré de résolution.

Nous utiliserons principalement par la suite le seuillage universel, qui est le plus fréquemment utillisé. Ce seuillage vaut

$$\lambda = \sigma_v \sqrt{2\log(n)}.$$

Notons qu'il entraîne généralement un surlissage de la fonction, ce qui explique qu'on trouve souvent le seuil corrigé $\lambda = \sigma_v \sqrt{2\log(n) - \log(\log(n))}$.

Avec le seuil universel, Donoho *et al.* (1995) mettent en évidence le caractère quasi-minimax de l'estimateur ainsi construit pour les seuillages doux et fort :

Théorème 2.6.
Soit $1 < p < \infty$. Supposons que la fonction d'échelle φ et l'ondelette ψ appartiennent à $\mathcal{C}^R$ et que ψ admette N moments nuls. Considérons l'estimateur à seuillage fort ou doux avec un seuil $\lambda = \sigma\sqrt{2\log(n)}$. Si f appartient à une boule de Besov $\mathcal{B}^s_{\pi,r}$ avec $0 < s < min(R, N)$ et $s + 1/2 - 1/\pi > 0$, alors il existe C constante indépendante de n telle que

$$\mathbb{E}\|\hat{f}_n - f\|_p^p \leqslant C\left(\frac{\log(n)}{n}\right)^{s/(1+2s)}.$$

Cet estimateur est quasi-minimax dans la mesure ou sa vitesse ne diffère que d'un facteur logarithmique avec la vitesse optimale au sens minimax, et il est adaptatif quant à la régularité de f car les paramètres s, π, r n'interviennent pas dans la construction de $\hat{f}_n$.

Le seuil nécessitant une estimation de l'écart-type du bruit σ_v, Donoho *et al.* (1995) proposent l'estimateur *Maximum Absolute Deviation* (MAD) appliqué aux coefficients d'ondelettes de résolution maximale. Si on note θ_{J-1} les 2^{J-1} dernières composantes du vecteur $\theta = WY$, avec W matrice de transformée orthogonale en ondelettes définie en Section 2.2.1, l'estimateur MAD s'écrit :

$$\hat{\sigma}_v = \frac{mediane\left(|\theta_{J-1} - mediane(\theta_{J-1})|\right)}{0.6745}. \tag{2.10}$$

Les bonnes propriétés de cet estimateur résultent de la robustesse de la procédure ainsi que du caractère creux de la représentation en coefficients d'ondelettes à une résolution élevée.

Moindres carrés pénalisés

Le seuillage peut aussi être vu comme une sélection des coefficients par une procédure de moindres carrés pénalisés. C'est ainsi que Antoniadis et Fan (2001) le considèrent comme solution d'un problème de minimisation de la forme :

$$\|WY - \theta\|_{l^2}^2 + \sum_{i=1}^{n} Pen_{i,\lambda}(|\theta_i|), \tag{2.11}$$

Antoniadis et Fan (2001) offrent ainsi une vision générale du seuillage. Les auteurs donnent par ailleurs les fonctions de pénalité menant aux seuillages usuels. Nous reviendrons sur ces dernières au Chapitre 3 lorsque nous introduirons le seuillage dans l'estimation de la partie fonctionnelle d'un modèle partiellement linéaire.

2.4 Régression avec un plan d'observation non-équidistant : l'approche de Kerkyacharian et Picard (2006)

Considérons maintenant le modèle suivant :

$$y_i = f(T_i) + v_i,\ i = 1, \dots, n, \tag{2.12}$$

où le plan d'observation $(T_i)_{i=1,\dots,n}$ est irrégulier sur $[0,1]$. Nous supposerons par la suite que les observations $(T_1, \dots, T_n)$ sont indépendantes et issues d'une loi de probabilité de fonction de répartition G et admettant une densité g. Les erreurs $(v_i)_{i=1,\dots,n}$ sont de nouveau supposées i.i.d. de lois normale $\mathcal{N}(0, \sigma_v^2)$.

De nombreux articles ont essayé d'étendre les techniques de régression par ondelettes du cas équidistant au cas où le plan d'observation n'est pas équidistant. Nous résumons brièvement diverses méthodes proposées dans ce but. Nous détaillerons ensuite plus précisément le schéma d'estimation élaboré par Kerkyacharian et Picard (2004) qui sera appliqué au Chapitre 7.

2.4.1 Aperçu de différentes approches

La première idée pour estimer f dans le modèle (2.12) est d'essayer d'appliquer la procédure du modèle (2.5), c'est-à-dire de considérer que le plan d'observation $(T_i)_{i=1,\dots,n}$ est équidistant. Cai et Brown (1999) montrent que dans le cas où l'échantillon $(T_i)_{i=1,\dots,n}$ est i.i.d. de loi uniforme sur $[0,1]$, l'erreur quadratique moyenne de l'estimateur obtenu est quasi-minimax sur l'ensemble des fonctions 1/2-Hölderiennes. Cependant, cette méthode semble assez restrictive, étant données l'hypothèse de loi uniforme sur le plan d'observation et la régularité élevée de la fonction f nécessaire.

Antoniadis et Fan (2001) étudient quant à eux l'estimation obtenue lorsqu'on considère une grille régulière plus fine que celle des observations. Supposons que chaque T_i s'écrit sous la forme $T_i = n_i/2^{J^*}$. Les auteurs introduisent la matrice A dont chaque ligne i correspond à la ligne n_i de la matrice de transformée en ondelettes associée au plan régulier d'ordre 2^{J^*} citée en Section 2.2.1. Ils sont alors ramenés à étudié le modèle

$$y = A\theta + v,\ v \sim \mathcal{N}(0, \sigma_v^2 I_{2^{J^*}}).$$

Le seuillage est ensuite défini par minimisation d'un critère des moindres carrés pénalisés, comme en (2.11). Une procédure en deux étapes -ROSE- est par ailleurs proposée

afin de ne pas avoir à résoudre le problème de minimisation correspondant (en général non convexe, non lisse et en grande dimension).

Toujours dans l'optique de se ramener à une grille d'observation régulière afin de pouvoir, de nouveau, appliquer les schémas décrits en Section 2.3, Hall et Turlach (1997) et Kovak et Silverman (2000) proposent d'interpoler les données. Les auteurs utilisent une interpolation par des polynômes, dépendants du plan d'observation, et considèrent un seuillage global pour Hall et Turlach (1997) et adapté à chaque coefficient pour Kovak et Silverman (2000). Sous certaines conditions quant à l'interpolation, Hall et Turlach (1997) montrent que ce shéma mène à un risque quadratique optimal pour des fonctions r-fois continûment déroivables et sous l'hypothèse d'une densité g continue par morceaux et bornée inférieurement par une constante strictement positive.

Cai et Brown (1998) supposent que pour tout i, $T_i = H^{-1}(i/n)$ et considèrent l'estimation $\hat{h}_n = n^{-1/2}\sum y_i \varphi_{J,i}$ de $h = f \circ H^{-1}$. Les auteurs définissent ensuite $\hat{f}_n$ comme la projection sur l'espace $V_J = V_{j_0} \oplus \oplus_{j=j_0}^{J} W_j$ de $\hat{h}_n \circ H$. Cette méthode permet notamment d'introduire du seuillage. Cai et Brown (1998) montrent que sur l'ensemble des fonctions s-Hölderiennes, l'estimateur obtenu atteint une vitesse quasi-minimax, en étant adaptatif quant à la régularité s, lorsque la fonction H est connue et lipschitziennne. L'inconvénient dans la mise en œuvre pratique est que cette procédure nécessite de connaître les valeurs des fonctions φ_J en des points quelconqes. Willer (2006) fait remarquer que ce schéma est équivalent à interpoler les observations à l'aide de fonctions d'échelle (composées avec la fonction de répartition des observations) puis à appliquer au modèle obtenu une procédure usuelle d'estimation par ondelettes.

Antoniadis *et al.* (1994) proposent une estimation à noyaux. Ils obtiennent un estimateur de Naradaya-Watson de noyau

$$E(T_i, .) = \sum \int_{A_i} K(., s) ds \mathbb{1}_{T_i \in A_i}$$

où $K(t,s) = N^{-1}\sum_k \varphi_{j1,K}(t)\varphi_{j1,k}(s)$ et A_i est une partition de l'intervalle adaptée au plan d'observation, le paramètre j_1 étant choisi par validation croisée. Les auteurs montrent la consistance de cette procédure sous des conditions que nous ne détaillerons pas ici. Amato *et al.* (2006) développent quant à eux une procédure des moindres carrés pénalisés menant un estimateur à noyau avec un seuillage par blocs des coefficients. Outre l'algorithme d'implémentation, les auteurs montrent l'optimalité du risque quadratique sur des espaces de Besov.

Antoniadis et Pham (1998) s'inspirent de l'estimateur de Naradaya-Watson. Si h désigne

la fonction fg, alors l'estimateur défini par les auteurs s'écrit $\hat{f}_n = \frac{\hat{h}_n}{\hat{g}_n}$ avec $\hat{h}_n$ et $\hat{g}_n$ estimateurs par ondelettes respectifs de h et g. L'estimation de g est donnée par la théorie usuelle d'estimation de densité par ondelettes. Quant à l'estimateur de h, le coefficient d'ondelette d'ordre (j,k) est estimé par $\frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{j,k}(T_i)$. Les auteurs établissent les propriétés asymptotiques dans le cadre d'estimateurs linéaires. Notons que cette méthode suppose de connaître la valeur d'une fonction d'échelle ou d'une ondelette en un point quelconque de l'intervalle unité. Antoniadis et Pham (1998) utilisent pour cela un résultat de Chui (1992), et on peut aussi utiliser Vidakovic (2002).

Pensky et Vidakovic (2001) proposent quant à eux une nouvelle estimation des coefficients d'ondelettes dans un cadre non équidistribué en constatant que le coefficient d'ondelette d'ordre (j,k) de la fonction f s'écrit $\beta_{j,k} = \mathbb{E}[Y\frac{\psi_{j,k}(T)}{g(T)}]$. Chesneau (2007) reprend leur principe d'estimation en ajoutant un seuillage sur les coefficients. Il montre l'optimalité en norme L^p de ce shéma pour un seuil non adaptatif lorsque la densité g vérifie $\frac{1}{g} \in L^{\max(p,2)}$.

Une autre idée développée notamment par Sardy *et al.* (1998), Delouille et Sachs (2001) ou encore Delouille et Sachs (2001) est d'adapter l'analyse multirésolution au plan d'observation. Ces articles modifient la base de Haar afin qu'elle soit adaptée non plus aux intervalles dyadiques de $[0,1]$ mais à la partition engendrée par le plan d'observation. La construction engendre une analyse multirésolution non orthonormée, ce qui explique l'introduction d'une correction des coefficients par Delouille et Sachs (2001). Delouille et Sachs (2001) et Delouille et Sachs (2001) proposent par ailleurs des estimateurs seuillés. Sur l'ensemble des fonctions s-Hölderiennes à variations bornées, ces derniers obtiennent une estimation adaptative ayant une vitesse en $n^{-s/(1+2s)}$ pour le risque L^2 associé à la mesure engendrée par le plan d'observation, sous des hypothèses relativement fortes quant à la régularité de la distribution de ce plan.

Kerkyacharian et Picard (2004) envisagent aussi de modifier la base de l'analyse multirésolution afin de l'adapter aux données. Cependant leur procédure s'applique à tout type d'ondelettes sans restriction à la base de Haar comme les travaux cités ci-dessus. Le grand atout de leur approche est qu'ils montrent comment en pratique l'algorithme usuel du cas équidistant peut être appliqué. De plus, par rapport au problème de l'estimation dans un modèle partiellement linéaire, ils aboutissent à des transformées orthogonales, ce qui permet de conserver les propriétés de la partie linéaire d'un tel modèle. Nous ne détaillerons pas plus ici cet estimateur, celui-ci faisant l'objet du reste de la section.

En résumé, nous pouvons classer ces méthodes en trois catégories : la première est d'es-

sayer de se ramener à un modèle équidistant soit par interpolation soit en considérant une grille considérablement plus fine que la grille initiale. L'approximation résultante peut diminuer la qualité globale de l'estimateur, ce qui explique que nous ne mettions pas en œuvre ces schémas d'estimation par la suite. Nous pouvons remarquer cependant qu'ils présentent l'avantage d'être facilement adaptables au modèle partiellement linéaire dans la mesure où le caractère linéaire est conservé. La seconde option est d'appliquer un estimateur à noyaux. Mais l'introduction d'un critère des moindres carrés pénalisés (Amato *et al.* (2006)) ou la forme d'estimation de Antoniadis et Pham (1998) excluent l'approche des moindres carrés pénalisés pour laquelle nous avons opté dans ce travail et donc le lien avec la M-estimation qui sera vu au Chapitre 3. (Cette remarque est aussi valable pour Pensky et Vidakovic (2001).) Enfin, la dernière catégorie de procédures proposées revient à adapter la base de décomposition au plan d'observation. La modification de la base de Haar a pour principal inconvénient de perdre l'orthogonalité de la base. Dans un modèle partiellement linéaire, nous devrions alors justifier de l'identifiabilité du modèle projeté et obtiendrions de plus un modèle hétéroscédastique. De plus, Kerkyacharian et Picard (2004) propose une méthodologie qui n'a besoin d'estimer les ondelettes qu'en des points dyadiques. Ceci semble particulièrement intéressant, notamment par comparaison aux autres travaux qui avaient aussi recours à la composition des ondelettes avec la fonction de répartition (Cai et Brown (1998), Antoniadis et Pham (1998)).

En conclusion, la déformation des bases d'ondelettes explorée par Kerkyacharian et Picard (2004) semble la plus intéressante à appliquer dans un modèle partiellement linéaire dans la mesure où cette procédure mène à une estimation adaptative, basée sur une transformée orthogonale (ou du moins asymptotiquement) et évalue les ondelettes en des points dyadiques, ce qui réduit le coût de calcul. C'est pourquoi nous détaillons cette estimation dans la suite de la section.

2.4.2 Hypothèses sur le plan d'observation

La théorie développée par Kerkyacharian et Picard (2004) repose sur la notion de poids de Muckenhoupt. Soit $1 < p < \infty$. L'hypothèse fondamentale faite sur la distribution du plan d'observations est la suivante :

(H_p) La densité g commune aux observations $(T_i)_{i=1,\dots,n}$ est à support compact $[a, b]$ et la fonction $y \mapsto w_g(y) = \frac{1}{g(G^{-1}(y))}$ appartient à la classe $\mathcal{A}_p([a, b])$ des poids de Muckenhoupt de degré p sur [a,b], c'est-à-dire qu'il existe une constante $C > 0$

telle que pour tout intervalle I inclus dans [a,b],

$$\left(\frac{1}{|I|}\int_I w_g\right)^{1/p}\left(\frac{1}{|I|}\int_I w_g^{-p/q}\right)^{1/q} \leqslant C,$$

avec $1/p + 1/q = 1$.

Remarque : Notons que la fonction w_g intervient naturellement dans le calcul d'intégrale : pour toute fonction h définie sur $[0,1]$, $\int h \circ G(t)dt = \int_0^1 h(s)w_g(s)ds$. Cette fonction est aussi la dérivée de l'inverse généralisée de G.

Les auteurs montrent que supposer (H_p) équivaut à dire qu'il existe une constante C telle que pour tout intervalle I inclus dans $[a,b]$,

$$\left(\frac{1}{|I|}\int_I g(x)^q dx\right)^{1/q} \leqslant C\frac{1}{|I|}\int_I g(x)dx, \tag{2.13}$$

où $1/q + 1/p = 1$. L'idée est que la densité doit être répartie de manière homogène, c'est-à-dire que les réalisations doivent donner suffisamment de poids à chaque intervalle. Kerkyacharian et Picard (2004) font remarquer que la loi uniforme vérifie (H_p). Si g appartient à L^q et est bornée inférieurement par $c > 0$, alors (H_p) est vérifiée.

Nous pouvons par ailleurs établir un lien avec l'hypothèse de Chesneau (2007), $\frac{1}{g} \in L^{\max(p,2)-1}$. Soit r, $1 < r < 1 + (\max(p,2)-1)/q^2$ et introduisons $p' \geqslant 1$ tel que $1/p' = 1/q - q(r-1)/(\max(p,2)-1)$. L'inégalité de Hölder donne, pour tout intervalle I,

$$\begin{aligned}\left(\int_I g^q\right)^{1/q} &= \left(\int_I g^{qr}\,(1/g)^{q(r-1)}\right)^{1/q}\\ &\leqslant \left(\int_I g^{qrp'}\right)^{1/p'}\left(\int_I\left(\frac{1}{g}\right)^{\max(p,2)-1}\right)^{q(r-1)/(\max(p,2)-1)}.\end{aligned}$$

Comme $r > 1$, nous avons $qrp' > 1$ et par conséquent l'inégalité de Jensen implique que $\left(\int_I g^{qrp'}\right)^{1/p'} \leqslant \left(\int_I g\right)^{qr}$. En utilisant que g est une densité donc à valeur positive et d'intégrale 1, et que $qr > 1$, nous obtenons

$$\left(\int_I g^q\right)^{1/q} \leqslant \left(\int_I g\right)\left\|\frac{1}{g}\right\|_{\max(p,2)-1}^{q(r-1)}.$$

Ainsi la propriété de Muckenhoupt est vérifiée avec une constante C égale à

$$C = \max\left(1, \left\|\frac{1}{g}\right\|_{\max(p,2)-1}^{(\max(p,2)-1)/q}\right)|b-a|^{1/p}.$$

Nous obtenons donc la propriété suivante :

Proposition 2.7. *Soit $p > 1$. Si g est à support compact et $1/g \in L^{\max(p,2)-1}$, alors la propriété (H_p) est vérifiée.*

Néanmoins, l'hypothèse (H_p) est moins restrictive que l'hypothèse de Chesneau (2007) : montrons qu'il existe des fonctions g à support compact vérifiant (H_p) telles que $1/g$ n'appartienne pas à $L^{\max(p,2)-1}$. Considérons pour cela des fonctions avec une singularité. Nous avons le résultat suivant :

Proposition 2.8. *Soient $[a,b]$ un intervalle de $\mathbb{R}$ et x_0 un point de $[a,b]$. Pour tout $\alpha > 0$, la fonction g définie sur $[a,b]$ qui à x associe $|x - x_0|^\alpha$ vérifie l'hypothèse (H_p).*

Ce résultat étaye nos propos dans la mesure où, lorsque l'exposant α est inférieur à la quantité $1/\max(p,2)$, la fonction g correspondante n'appartient pas à l'espace $L^{\max(p,2)-1}$. Ceci montre que l'hypothèse (H_p) semble relativement peu restrictive et permet notamment à la densité du plan d'observation d'avoir des discontinuités.

Démonstration. Si $\alpha \geqslant 1$, alors pour tout $p > 1$, la fonction g appartient à $L^{\max(p,2)}$, donc la propriété (H_p) est vérifiée. Nous étudions par conséquent le cas $\alpha < 1$.

Commençons par étudier la fonction $g : x \mapsto |x|^\alpha$. Soit $I = [c,d]$ un intervalle de $\mathbb{R}$ inclus dans $[a,b]$. Nous allons montrer que (2.13) est vérifiée. Distinguons selon que 0 est ou non dans l'intervalle I.

- Si $I = [c,d]$, avec $0 \leqslant c < d$. Le membre de gauche de l'inégalité (2.13) s'écrit :

$$\begin{aligned}
\left(\frac{1}{|I|}\int_I g^q\right)^{1/q} &= \frac{1}{(\alpha q+1)^{1/q}}\frac{(d^{\alpha q+1}-c^{\alpha q+1})^{1/q}}{(d-c)^{1/q}} \\
&= \frac{\alpha+1}{(\alpha q+1)^{1/q}}A^{1/q}\left(\frac{1}{|I|}\int_I g\right) \\
\text{avec} \qquad & A = \frac{(d-c)^q}{d-c}\frac{d^{\alpha q+1}-c^{\alpha q+1}}{(d^{\alpha+1}-c^{\alpha+1})^q}.
\end{aligned}$$

  Il suffit pour pouvoir conclure de montrer que la quantité A peut être bornée uniformément indépendamment de I. Tout d'abord, nous avons les inégalités

$$\begin{aligned}
d^{\alpha+1}-c^{\alpha+1} &\geqslant dd^\alpha - cd^\alpha = (d-c)d^\alpha \\
d^{\alpha q+1}-c^{\alpha q+1} &\leqslant dc^{\alpha q} - cc^{\alpha q} = (d-c)c^{\alpha q}.
\end{aligned}$$

  D'où

$$A \leqslant (\frac{c}{d})^{\alpha q}.$$

  Nous pouvons ainsi majorer A par 1.

- Si $I = [-c, d]$, avec $0 \leqslant c, d$. Sans perte de généralité, nous pouvons supposer que $c \leqslant d$ par parité de la fontion g. De même que précédemment, nous pouvons écrire :

$$\begin{aligned} \left(\frac{1}{|I|}\int_I g^q\right)^{1/q} &= \frac{\alpha+1}{(\alpha q+1)^{1/q}} A^{1/q}\left(\frac{1}{|I|}\int_I g\right) \\ \text{avec} \qquad A &= \frac{(d+c)^q}{d+c}\frac{d^{\alpha q+1}+c^{\alpha q+1}}{(d^{\alpha+1}+c^{\alpha+1})^q}. \end{aligned}$$

  Nous avons

$$\begin{aligned} c^{\alpha+1}+d^{\alpha+1} &\geqslant d^{\alpha+1} \\ c^{\alpha q+1}+d^{\alpha q+1} &\leqslant 2d^{\alpha q+1}. \end{aligned}$$

  D'où $A \leqslant \frac{2(c+d)^q d^{\alpha q+1}}{(c+d)d^{\alpha q+q}}$. On en déduit que $A \leqslant 2(\frac{d+c}{d})^q \frac{d}{c+d} \leqslant 2^{1+q}$.

Ainsi, l'inégalité (2.13) est vérifiée en prenant $C = 2^{1+1/q}\frac{\alpha+1}{(\alpha q+1)^{1/q}}$.

La généralisation au cas où la singularité est en un point x_0 quelconque est immédiate. Il suffit en effet de remarquer que pour tout $\beta > 0$, pour tout $c \leqslant d$,

$$\frac{1}{d-c}\int_c^d |x-x_0|^\beta dx = \frac{1}{(d-x_0)-(c-x_0)}\int_{c-x_0}^{d-x_0} |u|^\beta du.$$

□

2.4.3 Approche théorique

Considérons le modèle (2.12) défini en début de section et supposons que la densité g commune aux T_i vérifie l'hypothèse (H_p).

L'idée de Kerkyacharian et Picard (2004) est de se ramener à une fonction définie sur [0,1] à l'aide de G en considérant la fonction $f \circ G^{-1}$, où G^{-1} est l'inverse généralisée de G. (Notons que cette approche avait déjà été envisagée par Cai et Brown (1998).) Considérons $\psi_{j,k}$ une base d'ondelettes. Le coefficient d'ordre (j,k) de $f \circ G^{-1}$ vaut $\beta_{j,k} = \int f \circ G^{-1}(s)\psi_{j,k}(s)ds$. Par changement de variable il est aussi égal à $\beta_{j,k} = \int f(t)\psi_{j,k} \circ G(t)g(t)dt$. Alors, lorsque G est connu, les coefficients peuvent être estimés par

$$\beta^*_{j,k,n} = \frac{1}{n}\sum_{i=1}^n Y_i \psi_{j,k} \circ G(T_i) \; ; \tag{2.14}$$

en effet, les v_i étant centrés, nous avons :

$$\mathbb{E}[\beta^*_{j,k,n}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f(T_i)\psi_{j,k} \circ G(T_i)\right] \approx \int f(t)\psi_{j,k} \circ G(t)g(t)dt = \beta_{j,k}.$$

Sous l'hypothèse (H_p) ci-dessus, la famille $\{\psi_{j,k} \circ G,\ j \geqslant -1,\ k \in \mathbb{Z}\}$ vérifie la propriété de Temlyakov ainsi que la propriété de seuillage. En considérant les fonctions $\psi_{j,k} \circ G$, nous sommes donc dans une trame très proche du cadre usuel des ondelettes.

Afin de se placer dans un cadre similaire et d'établir les propriétés d'estimation, Kerkyacharian et Picard (2004) introduisent des espaces de Besov pondérés $\mathcal{B}^G_{s,\pi,r}$. Le principe de ces espaces de Besov est de calculer le module de continuité par rapport à la mesure induite par la fonction de répartition G. Plus précisément, dans la Définition 2.4, le module de continuité ω^N_π défini en (2.3) est remplacé par $\omega^{N,G}_\pi$ défini par :

$$\omega^{N,G}_p(f,t) := \sup_{|h|\leqslant t} \left(\int |\Delta^N_h f(u)|^\pi w_g(u) du \right)^{1/\pi},$$

avec w_g donné par l'hypothèse (H_p) et la modification usuelle si π est infini. L'hypothèse (H_p) ci-dessus sert à ce que la définition des espaces pondérés résultante ait un sens et à établir le caractère creux des coefficients d'ondelette associés à cet espace. De même qu'en Section 2.1.3, les espaces $\mathcal{B}^G_{s,\pi,r}$ obtenus peuvent être caractérisés par le comportement des coefficients d'ondelettes : les coefficients d'ondelettes $\beta_{j,k}$ d'une fonction h appartenant à un tel espace vérifient

$$\sum_j 2^{j(s+1/2)} \left(\sum_k |\beta_{j,k}|^p w(I_{j,k})^{1/p} \right)^q \ < \ \infty,$$

avec $w(I_{j,k}) = \int_{I_{j,k}} w_g$, l'intervalle $I_{j,k}$ étant le support de $\psi_{j,k}$ et w_g la fonction introduite dans (H_p). Lorsque la loi des T_i est la loi uniforme, nous retrouvons les espaces de Besov non pondérés. Nous pouvons aussi remarquer que quelque soit G, supposer que $f \circ G^{-1}$ appartient à l'espace de Besov $\mathcal{B}^s_{\infty,\infty}$ est équivalent à supposer que f appartient à l'espace pondéré $\mathcal{B}^G_{s,\infty,\infty}$.

Définissons

$$f^* = \alpha^*_{0,0,n} \varphi \circ G + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta^*_{j,k,n} \mathbb{1}_{\{|\beta^*_{j,k,n}| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ G,$$

avec κ constante, $2^{j_1} \sim \sqrt{\frac{n}{\log(n)}}$, les coefficients $\beta^*_{j,k,n}$ donnés en (2.14) et $\alpha^*_{0,0,n}$ construit sur le même principe. Les auteurs montrent que si f est bornée et $f \in \mathcal{B}^G_{s,\pi,r}$ avec $\pi \geqslant p > 1$ et $s \geqslant 1/2$, alors

$$\mathbb{E}\|f^* - f\|_p \ \leqslant \ C \left(\frac{\log(n)}{n} \right)^{-\frac{s}{1+2s}}.$$

Lorsque g est telle que $\frac{1}{g} \in L^{\max(p,2)-1}$, Chesneau (2007) montre que ce résultat est quasi-minimax. Kerkyacharian et Picard (2004) donnent aussi une caractérisation maxiset de l'estimateur f^* que nous ne présentons pas ici.

2.4.4 Approche pratique

L'estimateur f^* ne peut être calculé au vu des données car G n'est pas connu. Kerkyacharian et Picard (2004) proposent alors une estimation en 2 étapes. Les auteurs séparent l'échantillon en deux parties égales : une pour estimer la fonction de répartition G et l'autre pour en déduire l'estimation de f.

1. Nous estimons G à partir de $T_1 \dots T_{[n/2]}$:

$$\hat{G}_{[n/2]}(t) = \frac{1}{[n/2]} \sum_{i=1}^{[n/2]} \mathbb{1}_{\{T_i \leqslant t\}}.$$

2. L'estimation de f ne repose ensuite que sur les observations $T_{[n/2]+1} \dots T_n$ afin d'avoir l'indépendance entre les données utilisées et l'estimation de G. Nous estimons les coefficients d'ondelette $\beta_{j,k}$ en remplaçant G par $\hat{G}_{[n/2]}$ dans (2.14) :

$$\hat{\beta}_{j,k,n} = \frac{1}{n - [n/2]} \sum_{i=[n/2]+1}^{n} Y_i \psi_{j,k} \circ \hat{G}_{[n/2]}(T_i).$$

Ce qui mène à l'estimateur $\hat{f}_n$:

$$\hat{f}_n = \hat{\alpha}^{\varphi}_{0,0,n} \circ \hat{G}_{[n/q]} + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k,n} \mathbb{1}_{\{|\hat{\beta}_{j,k,n}| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ \hat{G}_{[n/q]}, \tag{2.15}$$

avec $2^{j_1} \sim \sqrt{\frac{n}{\log(n)}}$.

Kerkyacharian et Picard établissent le comportement asymptotique de l'estimateur $\hat{f}_n$ ainsi construit :

Théorème 2.9.
Supposons que nous avons le modèle (2.12) avec un plan d'observation $(T_i)_{i=1,\dots,n}$ i.i.d. de densité g vérifiant l'hypothèse (H_p), $p > 1$.
Soit $s \geqslant 1/2$. Si f est bornée et appartient à l'espace $\mathcal{B}^G_{s,\infty,\infty}$, alors l'estimateur $\hat{f}_n$ défini dans l'équation (2.15) ci-dessus vérifie :

$$\mathbb{E}\|\hat{f}_n - f\|_p \leqslant C \left(\frac{\log(n)}{n}\right)^{-\frac{s}{1+2s}}.$$

La vitesse obtenue est quasi-minimax lorsque le plan d'observation est équidistant, ou vérifie l'hypothèse de Chesneau (2007) citée plus haut. Mais nous avons vu que l'hypothèse faite sur le plan est moindre dans le contexte de ce théorème. Les auteurs ont ainsi développé une estimation menant à une vitesse de convergence quasi-minimax, dès lors que la fonction à estimer est suffisamment régulière, et que la loi régissant le plan (T_i) est suffisamment homogène, sans que cette dernière hypothèse soit très restrictive. Remarquons toutefois que les hypothèses quant à la régularité de f sont plus restrictives que pour l'estimateur f^*, mais l'estimateur $\hat{f}_n$ est, lui, utilisable sur données réelles.

En pratique, pour réaliser l'étape 2 ci-dessus, nous considérons le plan $T_{[n/2]+1}, \dots, T_n$ ordonné, c'est-à-dire qu'à une permutation près des indices, nous prenons $T_{[n/2]+1} \leqslant \dots \leqslant T_n$.
Désignons par $\hat{G}^{-1}_{[n/2]}$ la fonction suivante :

$$\hat{G}^{-1}_{[n/2]} \ : \ \begin{array}{lcl} \{1/[n/2],\ 2/[n/2],\ \dots 1\} & \to & \mathcal{P}(\{T_{[n/2]+1} \dots T_n\}) \\ i/[n/2] & \mapsto & \left\{T_j,\ \hat{G}_{[n/2]}(T_j) = i/[n/2]\right\}. \end{array}$$

Si $\hat{G}^{-1}_{[n/2]}(i/[n/2]) = \emptyset$ nous posons $\tilde{Y}_i = 0$. Ensuite, si $\hat{G}^{-1}_{[n/2]}(i/[n/2]) = \{T_{j_1}, \dots, T_{j_i}\}$, nous considérons $\tilde{Y}_i = Y_{j_1} + Y_{j_2} + \dots + Y_{j_i}$. Alors nous avons

$$\hat{\beta}_{j,k,n} = \frac{1}{[n/2]} \sum_{l=[n/2]+1}^{n} Y_l \psi_{j,k} o \hat{G}_{[n/2]}(T_l) = \frac{1}{[n/2]} \sum_{i=1}^{[n/2]} \tilde{Y}_i \psi_{j,k}(i/[n/2]).$$

Ainsi, nous pouvons estimer f en réalisant la transformée en ondelettes associée au plan équidistant de $\tilde{Y}_1 \dots \tilde{Y}_{[n/2]}$, en appliquant le seuillage, puis la transformée en ondelettes inverse.

Remarque : La raison pour laquelle Kerkyacharian et Picard (2004) choisissent de séparer l'échantillon en deux parties égales est que sinon, ceci entraînerait un biais dans l'estimation des coefficients : en estimant la fonction de répartition G avec n/q observations, nous obtiendrions

$$\hat{\beta}_{j,k,n} = \frac{1}{[n/q]} \sum_{i=[n/q]+1}^{n} Y_i \psi_{j,k} o \hat{G}_{[n/q]}(T_i),$$

soit l'estimateur cherché avec un facteur multiplicatif $\frac{n-[n/q]}{[n/q]} \sim q/q'$ avec $1/q + 1/q' = 1$.

Nous reviendrons sur la mise en œuvre de cette procédure au Chapitre 5. Nous verrons alors les problèmes rencontrés lors de l'exécution de ces algorithmes : si la théorie assure les bonnes propriétés asymptotiques de ce schéma d'estimation, celles-ci ne sont pas assurées à distance finie.

Conclusion

Ce chapitre a rappelé brièvement les techniques d'estimation par ondelettes que nous souhaitons appliquer dans un modèle partiellement linéaire. Les outils (Analyse multirésolution, espaces de Besov) définis, l'estimation par ondelettes, associée à des techniques de seuillage, est particulièrement attractive. Nous avons vu que dans le cadre d'un plan d'observation équidistribué, nous pouvions mettre en œuvre une procédure de débruitage quasi-minimax et adaptative relativement à la régularité de la fonction. Nous verrons au chapitre suivant comment ce schéma peut se transcrire dans un modèle partiellement linéaire. Lorsque le plan d'observation est aléatoire, nous avons choisi de focaliser notre étude sur la méthode d'estimation de Kerkyacharian et Picard (2004). Cette approche, quasi-minimax pour une large gamme de densité, est particulièrement adaptée à la transcription dans un modèle partiellement linéaire. Ceci sera étudié au Chapitre 7.

CHAPITRE 3

DU SEUILLAGE À L'ESTIMATION ROBUSTE

Sommaire

3.1 Identifiabilité et transformée en ondelettes 46

3.2 Critère des moindres carrés penalisés 48

3.2.1 Conditions du premier ordre 48

3.2.2 Seuillage doux et estimateur de Huber 49

3.2.3 Vers d'autres M-estimateurs 52

Nous considérons dans ce chapitre le modèle partiellement linéaire suivant :

$$y_i = X_i^T \beta + f(t_i) + u_i \qquad i = 1, \dots, n, \tag{3.1}$$

où les X_i sont des vecteurs de covariables de taille p connus, β est un vecteur de paramètres inconnu et les t_i, $i = 1, \dots, n$, forment un plan équidistribué de $[0;1]$, $t_i = \frac{i}{n}$. Les termes de bruit $(u_i)_{i=1\dots n}$ seront par la suite supposés indépendants et gaussiens de moyenne nulle et de variance σ^2. Nous supposons de plus que nous disposons de $n = 2^J$ observations. Étant donnée l'observation de $(y_i, X_i)_{i=1\dots n}$, notre but est d'estimer le vecteur des paramètres de régression β et la fonction f à partir de ces données.

Comme cela a été précisé au premier chapitre, une technique d'estimation reposant sur la représentation en coefficients d'ondelettes du modèle est développée, avec l'idée que l'introduction d'un seuillage sur les coefficients doit permettre d'obtenir une estimation adaptative de la partie fonctionnelle avec une vitesse quasi-minimax sans restreindre la régularité. Dans un modèle partiellement linéaire, l'estimation non linéaire par ondelettes a déjà été réalisée par F. Meyer (2003), Chang et Qu (2004) et Fadili et Bullmore (2005). Nous comparerons principalement nos résultats à Fadili et Bullmore (2005) dont le cadre d'étude est relativement similaire, mais ceci fera l'objet des chapitres suivants.

L'originalité de la procédure d'estimation présentée est d'établir un lien entre le seuillage des coefficients d'ondelettes de la partie fonctionnelle et des M-estimateurs usuels. Effectuer un seuillage doux, un seuillage fort ou encore un seuillage SCAD sur les coefficients d'ondelettes de f revient à estimer le vecteur des paramètres de régression β respectivement par l'estimateur de Huber, la moyenne tronquée ou l'estimateur de Hampel.

Le chapitre est organisé de la façon suivante : dans un premier temps, nous explicitons la transformée en ondelettes du modèle et introduisons les notations utilisées par la suite. Dans la section suivante, nous mettons en évidence le lien entre l'estimation par seuillage et les M-estimateurs usuels et détaillons le schéma d'estimation.

3.1 Identifiabilité et transformée en ondelettes

Sous forme matricielle le modèle (3.1) s'écrit

$$Y = X\beta + F + U, \tag{3.2}$$

avec $Y = \left(y_1 \dots y_n\right)^T$ vecteur des observations, $X^T = \left(X_1 \dots X_n\right)$ matrice de taille $p \times n$ et $F = \left(f(t_1) \dots f(t_n)\right)^T$. Le bruit $U = \left(u_1 \dots u_n\right)^T$ est un vecteur gaussien de moyenne nulle et de variance $\sigma^2 I_n$.

Nous supposerons de plus que la fonction f appartient à un espace de Besov $\mathcal{B}^s_{\pi,r}([0;1])$ avec $s + 1/\pi - 1/2 > 0$, afin que sa représentation par des coefficients d'ondelettes soit creuse. Afin d'assurer que le modèle soit asymptotiquement identifiable, nous introduisons deux hypothèses :

(A1) Le vecteur $\frac{1}{n}X^T F$ tend vers 0 quand n tend vers l'infini.

(A2) La matrice X est de plein rang colonnes, c'est-à-dire $\frac{1}{n}X^T X$ converge vers une matrice inversible.

L'expression du vecteur des paramètres en fonction des données du problème,

$$\beta = \left(\frac{1}{n}X^T X\right)^{-1} X^T (Y - F - U),$$

permet de vérifier aisément que (A1) et (A2) sont suffisantes pour que le modèle soit asymptotiquement identifiable. Aucune de ces deux hypothèses n'est contraignante. Cela sera étudié en Section 4.1.

Soit $(V_j)_{j\in\mathbb{N}}$ une Analyse Multi Résolution sur $[0;1]$ engendrée par la fonction d'échelle φ. On note $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$. Alors $V_j = Vect(\varphi_{j,k}, k = 0 \dots 2^j - 1)$. Soit ψ l'ondelette associée. On utilise des notations similaires pour $\psi_{j,k}$ et les espaces engendrés par $\{\psi_{j,k}\}_k$ sont notés W_j. Notons W la matrice de projection sur l'espace $V_J = V_{j_0} \oplus_{j=j_0}^{J-1} W_j$. Nous renvoyons à la Section 2.2.1 pour la forme de la matrice W.

Notons $Z = WY$, $A = WX$, $\theta = WF$ et $\underline{\varepsilon} = WU$. Alors la transformée en ondelettes du modèle (3.2) aboutit au modèle

$$Z = A\beta + \theta + \underline{\varepsilon}\,; \tag{3.3}$$

l'orthogonalité de la transformation W assure que $\underline{\varepsilon}$ est un bruit gaussien centré de variance $\sigma^2 I_n$. La représentation du modèle en coefficients d'ondelettes permet de conserver la structure linéaire de la partie linéaire tout en donnant une représentation creuse de la partie fonctionnelle. L'intérêt de l'estimation du paramètre de régression dans le modèle projeté est que le biais introduit par la présence de la partie non paramétrique du modèle sera réduit par ce caractère creux, comme l'illustre la Figure 3.1 ci-dessous.

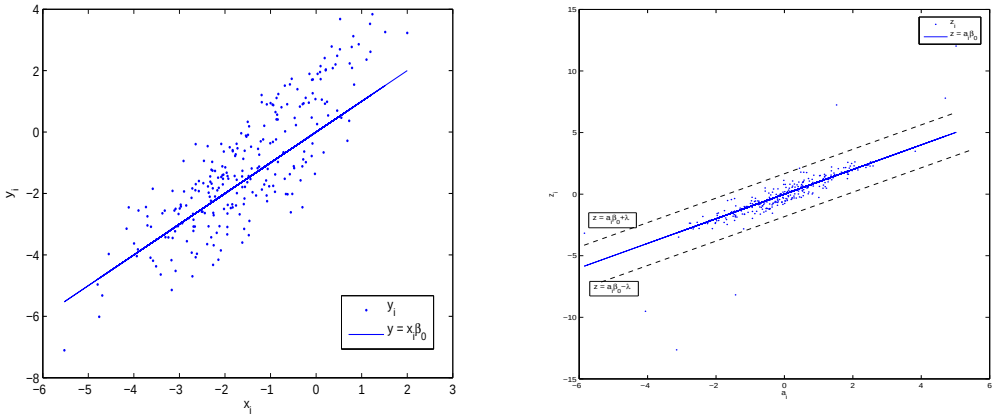

FIG. 3.1: La figure (a) représente les observations y_i en fonction des covariables X_i dans un cas de modèle partiellement linéaire (avec $p = 1$). La droite est la droite d'équation $y_i = X_i\beta_0$. La figure (b) représente le même modèle après application d'une transformée en ondelettes, c'est-à-dire les points de coorconnées (A_i, z_i). La droite en traits pleins est la droite d'équation $z_i = A_i\beta_0$. Les droites en tirets sont les droites d'équations respectives $z_i = A_i\beta_0 \pm \lambda$.

Cette figure montre que la transformée en ondelette devrait mener à une meilleure estimation des paramètres de régression, étant donné que la cohérence entre la droite d'équa-

tion $y = x\beta_0$ (i.e. la partie linéaire) et le nuage de points des observations est fortement augmentée. De plus, sur le modèle projeté, nous avons représenté les droites d'équations $y = x\beta_0 \pm \lambda$. L'idée est que les points en dehors de l'espace délimité par ces deux droites ont une influence moindre dans le calcul d'estimateurs robustes du paramètre de régression.

3.2 Critère des moindres carrés penalisés

Nous proposons d'estimer les vecteurs de paramètres β et θ du modèle (3.3) par les moindres carrés pénalisés. Le problème que nous considérons est le suivant

$$(\hat{\beta}_n, \hat{\theta}_n) = \underset{(\beta,\theta)}{argmin} \left\{ J_n(\beta,\theta) = \sum_{i=1}^{n} \frac{1}{2}(z_i - A_i^T\beta - \theta_i)^2 + \lambda \sum_{i=i_0}^{n} Pen(|\theta_i|) \right\}, \tag{3.4}$$

avec $i_0 = 2^{j_0} + 1$. Le terme de pénalité $\sum_{i=i_0}^{n} Pen(|\theta_i|)$ ne porte que sur les coefficients d'ondelettes empiriques de la partie fonctionnelle du modèle, et ne porte pas en particulier sur les coefficients d'échelle. L'introduction d'une pénalité sur les coefficients d'ondelettes permet dans le cadre non paramétrique d'introduire un seuillage sur les coefficients, comme cela a été vu en Section 2.3.2.

3.2.1 Conditions du premier ordre

Nous nous intéresserons par la suite de manière plus spécifique à certaines pénalités, mais rappelons d'abord ce résultat de Fadili et Bullmore (2005) qui établit que les conditions du premier ordre sont nécessaires pour un large éventail de pénalités :

Théorème 3.1. *Fadili et Bullmore (05)*
Supposons que la pénalité vérifie :

(i) *Pen est paire, à valeurs positive, non décroissante sur* $[0, +\infty)$ *et* $Pen(0) = 0$.

(ii) *Pen est deux fois différentiable sur* $\mathbb{R}\backslash\{0\}$.

(iii) *Pen est continue sur* $\mathbb{R}$ *et admet une dérivée à droite* $Pen'_+(0)$ *strictement positive en* 0, *finie ou infinie.*

Supposons de plus que le vecteur des observations Y *appartient à* $l^2(n)$ *et que* $\mathbb{1}_n$ *n'est pas généré par les vecteurs de* X. *Alors, les solutions (éventuellement locales)* $(\hat{\theta}_n, \hat{\beta}_n)$ *du problème de mini-*

misation (3.4) vérifient les conditions du premier ordre du problème.

$$\begin{aligned}
\hat{\beta}_n &= (A^T A)^{-1} A^T (Z - \hat{\theta}_n) \\
\hat{\theta}_{i,n} &= \begin{cases} z_i - A_i^T \hat{\beta}_n & si\ i < i_0 \\ 0 & si\ i \geqslant i_0\ et\ |z_i - A_i^T \hat{\beta}_n| \leqslant \lambda Pen'_+(0)\ ,\ i = 1 \dots n, \\ z_i - Pen'(\hat{\theta}_{i,n}) & si\ i \geqslant i_0\ et\ |z_i - A_i^T \hat{\beta}_n| < \lambda Pen'_+(0) \end{cases}
\end{aligned}$$

L'intérêt du résultat de Fadili et Bullmore (2005) est qu'il englobe une vaste catégorie de pénalités, notamment les pénalités associées aux seuillages dur et SCAD, comme nous le verrons par la suite. Il est donc valable pour une gamme de seuillage plus importante que les résultats que nous établissons ici, qui ont l'inconvénient de ne pas avoir été exprimés sous une forme plus générale. Ceci serait une perspective intéressante de prolongation. Notons enfin que Chang et Qu (2004) dans le cadre de la pénalité l^1 avaient établi que non seulement les conditions du premier ordre sont nécessaires, mais qu'elles sont de plus, pour cette pénalité, suffisantes.

3.2.2 Seuillage doux et estimateur de Huber

Nous choisissons une pénalité l^1 sur les coefficients d'ondelettes. C'est-à-dire que les estimateurs sont solutions du problème d'optimisation :

$$(\hat{\beta}_n, \hat{\theta}_n) = \underset{(\beta,\theta)}{argmin} \left\{ J_n(\beta, \theta) = \sum_{i=1}^{n} \frac{1}{2}(z_i - A_i^T \beta - \theta_i)^2 + \lambda \sum_{i=i_0}^{n} |\theta_i| \right\}. \tag{3.5}$$

La pénalité l^1 choisie ici correspond à une estimation par ondelettes avec seuillage doux dans un modèle non paramétrique (voir Donoho (1992)). La fonction de seuillage associée dans un modèle non paramétrique est $\gamma_S(u; \lambda) = signe(u)(|u| - \lambda)_+$.

Afin d'établir le lien avec l'estimateur de Huber, essayons de résoudre le problème de minimisation de J_n. Pour β fixé, le critère $J_n(\beta, .)$ est minimal en

$$\tilde{\theta}_i(\beta) = \begin{cases} z_i - A_i^T \beta & \text{si } i < i_0 \\ \text{signe}(z_i - A_i^T \beta) \left(|z_i - A_i^T \beta| - \lambda \right)_+ & \text{si } i \geqslant i_0 \end{cases}. \tag{3.6}$$

Ainsi, trouver $\hat{\beta}_n$ solution du problème (3.5) est équivalent à trouver $\hat{\beta}_n$ qui minimise le critère $J_n(\tilde{\theta}(\beta), \beta)$. Or,

$$J_n(\tilde{\theta}(\beta), \beta) = \sum_{i=i_0}^{n} \rho_\lambda(z_i - A_i^T \beta) \tag{3.7}$$

où ρ_λ est la fonction de coût de Huber de seuil λ, définie par :

$$\rho_\lambda(u) = \begin{cases} u^2/2 & \text{si } |u| \leqslant \lambda \\ \lambda|u| - \lambda^2/2 & \text{si } |u| > \lambda \end{cases} \tag{3.8}$$

Démonstration. Soit $i \geqslant i_0$. Minimiser le critère (3.5) en θ_i est équivalent à minimiser $j(\theta_i) := \frac{1}{2}(z_i - A_i^T\beta - \theta_i)^2 + \lambda|\theta_i|$. Les conditions du premier ordre sont : $j'(\theta_i) = \theta_i - (z_i - A_i^T\beta) + \text{signe}(\theta_i)\lambda = 0$.

- Si $\theta_i \geqslant 0$, alors $j'(\theta_i) = 0$ si et seulement si $\theta_i = z_i - A_i^T\beta - \lambda$ Donc si $z_i - A_i^T\beta \leqslant \lambda$, $\theta_i = 0$ et sinon $\theta_i = z_i - A_i^T\beta - \lambda$.
- Si $\theta_i \leqslant 0$, $j'(\theta_i)$ s'annule si et seulement si $\theta_i = z_i - A_i^T\beta + \lambda$ Donc si $z_i - A_i^T\beta \geqslant -\lambda$, $\theta_i = 0$ et sinon $\theta_i = z_i - A_i^T\beta + \lambda$.

Ceci prouve qu'à β fixé, le problème (3.4) est minimal en $\tilde{\theta}(\beta)$ donné en (3.6).

Remplaçons dans le critère : $J_n(\beta, \tilde{\theta}(\beta)) = \frac{1}{2}\sum_{i=i_0}^{n}\left((z_i - A_i^T\beta - \tilde{\theta}_i)^2 + \lambda|\tilde{\theta}_i|\right)$ car $\tilde{\theta}_i = z_i - A_i^T\beta$ pour $i < i_0$. Ensuite, si I désigne l'ensemble $I := \left\{j = i_0 \dots n, \quad |z_j - A_j\beta| < \lambda\right\}$, alors $J_n(\beta, \tilde{\theta}(\beta))$ s'écrit $J_n(\beta, \tilde{\theta}(\beta)) = \frac{1}{2}\sum_I (z_i - A_i^T\beta)^2 + \frac{1}{2}\sum_{I^C}\lambda^2 + \lambda\sum_{I^C}\left(|z_i - A_i^T\beta| - \lambda\right)$ en remplaçant $\tilde{\theta}_i$ par la valeur (3.6). Nous retrouvons bien la fonction de Huber. □

Nous obtenons ainsi la proposition suivante :

Proposition 3.2.
Si $\hat{\beta}_n$ et $\hat{\theta}_n$ sont solutions du problème d'optimisation (3.4), alors ils vérifient

$$\hat{\beta}_n = \underset{\beta}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - A_i^T\beta), \tag{3.9}$$

$$\hat{\theta}_{i,n} = \begin{cases} z_i - A_i^T\hat{\beta}_n & si\ i < i_0 \\ \gamma_{soft,\lambda}(z_i - A_i^T\hat{\beta}_n) & si\ i \geqslant i_0 \end{cases}, \ i = 1 \dots n, \tag{3.10}$$

avec ρ_λ fonction de coût de Huber définie en (3.8) et $\gamma_{soft,\lambda}$ fonction de seuillage doux de seuil λ ; $\gamma_{soft,\lambda}(u) = signe(u)\,(|u| - \lambda)_+$.

Ce résultat permet de calculer les estimateurs $\hat{\beta}_n$ et $\hat{\theta}_n$ de manière non itérative. Nous pouvons estimer le paramètre β directement à partir des observations, sans tenir compte de la partie non paramétrique du modèle à l'aide de l'équation (3.9), puis obtenir $\hat{\theta}_n$, donc $\hat{f}_n$, à l'aide de l'équation (3.10).

Nous avons ainsi mis en évidence une méthode d'estimation directe des paramètres du modèle, qui offre la possibilité d'étudier les propriétés asymptotiques des estimateurs. Nous verrons dans le Chapitre 5 que cette écriture donne aussi la possibilité d'appliquer des algorithmes issus de l'estimation robuste. Mais surtout, nous avons une interprétation de ce que représentent ces estimateurs ; à partir des observations (X, Y) la procédure d'estimation proposée est la suivante :

1. Nous appliquons la transformée en ondelettes d'ordre $J = \log_2(n)$ sur X et Y afin d'obtenir leur représentations respectives A et Z en coefficients d'ondelettes.
2. L'estimation du paramètre β est donnée par l'estimateur de Huber sans tenir compte de la présence d'une partie fonctionnelle dans le modèle :
$$\hat{\beta}_n = \underset{\beta}{argmin} \sum_{i=1}^{n} \rho_\lambda(z_i - A_i^T \beta).$$
Cela revient à considérer que nous avons un modèle linéaire $z_i = A_i^T \beta + e_i$ de bruit $e_i = \theta_i + \varepsilon_i$.
3. La représentation en coefficients d'ondelettes θ de la fonction f est estimée par seuillage doux de $Z - A\hat{\beta}_n$.
$$\hat{\theta}_{i,n} = \begin{cases} z_i - A_i^T \hat{\beta}_n & \text{si } i < i_0 \\ \text{signe}(z_i - A_i^T \hat{\beta}_n) \left(|z_i - A_i^T \hat{\beta}_n| - \lambda \right)_+ & \text{si } i \geqslant i_0 \end{cases}.$$
L'estimation de f est ensuite obtenue en appliquant la transformée en ondelettes inverse. Nous avons en fait l'estimation par transformée en ondelettes avec seuillage doux dans le modèle non paramétrique :
$$y_i - X_i^T \hat{\beta}_n = f(t_i) + v_i, \; i = 1 \dots n,$$
avec $v_i = X_i^T(\beta_0 - \hat{\beta}_n) + u_i$.

Remarque : Comme évoqué en Section 1.2.2, Speckman (1988) fait remarquer qu'il est préférable d'estimer préalablement la partie linéaire puis d'en déduire une estimation de la partie fonctionnelle plutôt que de prendre le schéma inverse. Le principe de l'estimation peut ici être décrit de manière similaire à celui préconisé par Speckman (1988) :

1. Si β était connu, nous estimerions f par $\tilde{f} = TW(Y - X\beta)$. La différence avec la démarche de Speckman réside dans la présence de l'opérateur de seuillage T.
2. β est alors estimé par la méthode des moindres carrés dans le modèle $Y - \tilde{f} = X\beta + U$, ce qui donne $\hat{\beta} = (X^T S^T S X)^{-1} S Y$ avec $S = (I - T)W$.
3. On remplace dans $\tilde{f}$ afin d'obtenir un estimateur de f : $\hat{f} = (I - S)(Y - X\hat{\beta})$.

La forme des estimateurs résultante est identique à celle de Speckman (1988) donnée au Chapitre 1, à ceci près que S n'est pas un opérateur linéaire.

3.2.3 Vers d'autres M-estimateurs

Nous avons établi un lien entre l'introduction d'une pénalité menant au seuillage doux des coefficients d'ondelettes et le M-estimateur de Huber. Il s'avère que pour d'autres seuillages, nous obtenons une procédure d'estimation équivalente, basée sur des M-estimateurs usuels. Si γ_λ désigne la fonction de seuillage obtenue, alors nous pouvons montrer de manière analogue que les estimateurs vérifient

$$\begin{aligned}
\hat{\beta}_n &= \underset{\beta}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - A_i^T\beta), \\
\hat{\theta}_{i,n} &= \begin{cases} z_i - A_i^T\beta & \text{si } i < i_0 \\ \gamma_\lambda(z_i - A_i^T\beta) & \text{si } i \geqslant i_0 \end{cases}, \; i = 1 \dots n,
\end{aligned}$$

avec ρ_λ primitive de $u \mapsto u - \gamma_\lambda(u)$. Nous présentons dans cette section quelques exemples de fonctions de coût obtenues sur ce principe, mais un vaste éventail peut être obtenu en utilisant le lien entre seuillage et pénalité sur les coefficients, décrit en Section 2.3.2 et reposant sur les travaux de Antoniadis et Fan (2001).

Seuillage fort

Pour introduire le seuillage fort, nous pouvons prendre la pénalité de comptage ; elle mène au critère :

$$(\hat{\theta}_n, \hat{\beta}_n) = \underset{\theta,\beta}{argmin} \left\{ J_n(\theta,\beta) = \frac{1}{2}\sum_{i=1}^{n}(z_i - A_i^T\beta - \theta_i)^2 + \lambda^2/2 \sum_{i=i_0}^{n} \mathbb{1}_{\{\theta_i \neq 0\}} \right\}. \tag{3.11}$$

Alors les estimateurs $\hat{\theta}_n, \hat{\beta}_n$ sont solutions des conditions du premier ordre et nous pouvons montrer que celles-ci s'écrivent

$$\hat{\beta}_n = \underset{\beta}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - A_i^T\beta) \tag{3.12}$$

$$\hat{\theta}_{i,n} = \begin{cases} Z_i - A_i^T\hat{\beta}_n & \text{si } i < i_0 \\ (Z_i - A_i^T\hat{\beta}_n)\mathbb{1}_{|Z_i - A_i^T\hat{\beta}_n| \geqslant \lambda} & \text{si } i \geqslant i_0 \end{cases} \tag{3.13}$$

où ρ_λ est la fonction de coût de la moyenne tronquée de seuil λ, définie par :

$$\rho_\lambda(u) = \begin{cases} u^2/2 & \text{si } |u| \leqslant \lambda \\ \lambda^2/2 & \text{si } |u| > \lambda \end{cases}$$

Remarque : D'autres pénalités mènent au seuillage dur. Nous avons opté pour celle-ci mais nous pouvons prendre de manière équivalente $\lambda^2/2 - \sum_{i=i_0}^{n}(|\theta_i| - \lambda)^2 \mathbb{1}_{\{|\theta_i| \leqslant \lambda\}}$ ou encore $\sum_{i=i_0}^{n} \lambda|\theta_i| \mathbb{1}_{\{|\theta_i|<\lambda\}} + \lambda^2/2 \mathbb{1}_{\{|\theta_i| \geqslant \lambda\}}$ (cf Antoniadis et Fan (2001) –nos pénalités diffèrent légèrement afin d'aboutir ensuite à une fonction de coût continue dans l'estimation du paramètre β–).

Il en résulte que le seuillage dur est associé à la fonction de coût de la moyenne tronquée.

Seuillage SCAD

Une autre forme de seuillage présentée en Section 2.3.2 est le seuillage SCAD. La pénalité correspondante est

$$Pen_\lambda(\theta) = \mathbb{1}_{|\theta| \leqslant \lambda} + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} \mathbb{1}_{|\theta| > \lambda}.$$

Ce seuillage est donc associé au M-estimateur de Hampel de paramètres $(\lambda, 2\lambda, a\lambda)$.

La Figure 3.2 ci-après donne les représentations des trois M-estimateurs usuels obtenus par la procédure des moindres carrés pénalisés pour les seuillages doux, fort et SCAD.

Pénalisation par une norme de Besov

Soient $s_{j_0,\cdot}$ et $w_{j,\cdot}$ respectivement les coefficients d'échelle et d'ondelette de f ; nous avons $\theta = (s_{j_0\cdot} w_{j_0,\cdot} \dots w_{J-1,\cdot})^T$. Lorsque la pénalité est la norme associée à l'espace de Besov $\mathcal{B}^s_{\pi,r}$,

$$Pen_\lambda(\theta) = \lambda \left(\|s_{j_0,\cdot}\|_{l^\pi} + \left(\sum_{j=j_0}^{J-1} 2^{jq(s+1/2-1/\pi)} \|w_{j,\cdot}\|^q_{l^\pi} \right)^{1/q} \right),$$

alors nous obtenons le seuillage

$$\gamma_{j,k,s}(u, \lambda) = \frac{u}{1 + \lambda 2^{2js}}.$$

L'étude des propriétés asymptotiques réalisée par Fadili et Bullmore (2005) se fait dans le cadre d'un seuillage très proche. Les auteurs considèrent la pénalité associée à la norme de Sobolev sur l'espace W^s :

$$Pen_\lambda(\theta) = \lambda \left(\sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} 2^{2js} w^2_{j,\cdot} \right),$$

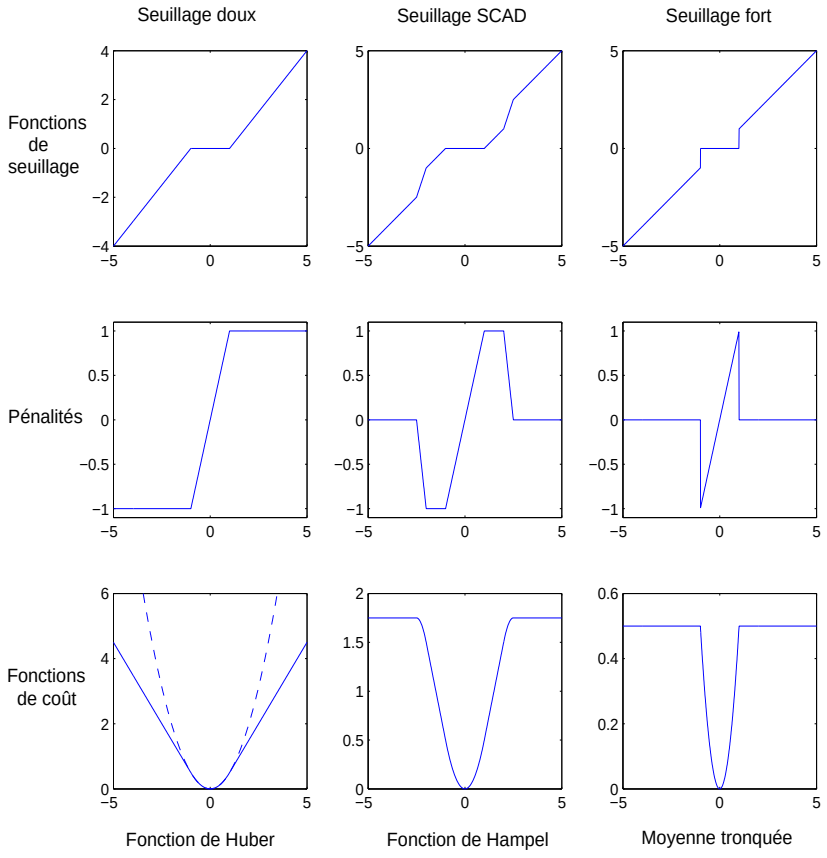

FIG. 3.2: Fonctions de seillages, pénalités correspondantes et fonctions de coût associées pour divers seuillages. En pointillés est tracée la fonction de coût des moindres carrés.

qui mène au seuillage

$$\gamma_{j,k,s}(u,\lambda) = \frac{u}{1+2\lambda 2^{2js}}.$$

Le M-estimateur résultant dans l'estimation de la partie linéaire a pour fonction de coût :

$$\rho_{j,k,\lambda}(u) = \frac{\lambda 2^{2js}}{1+2\lambda 2^{2js}} u^2,$$

ce qui correspond à une procédure des moindres carrés pondérés. Notons que cet estimateur présente la particularité intéressante de pouvoir s'écrire de manière explicite. En effet, nous avons :

$$\hat{\beta}_n = (A^T T A)^{-1} A^T T z,$$

avec $T = diag\left(2\lambda 2^{2js}/(1+2\lambda 2^{2js})\right)_{j=j_0,\ldots,J-1,\,k=0,\ldots,2^j-1}$ lorsque $(A^TTA)^{-1}$ existe.

Le seuil dépend ici du degré de résolution. Cette procédure présente l'inconvénient majeur de ne pas être adaptative, le paramètre de régularité de la fonction, s, intervenant dans la pénalité, donc dans le seuillage et la fonction de coût résultante. Afin d'y remédier, une procédure de validation croisée est nécessaire. C'est ce pour quoi optent Fadili et Bullmore (2005).

Conclusion

Le parrallèle établi entre l'estimation par moindres carrés pénalisés et les M-estimateurs usuels dans une régression linéaire apporte de nombreuses ouvertures à l'inférence statistique dans un modèle partiellement linéaire. En effet, cette propriété permet d'élaborer des estimations en deux étapes : l'estimation du paramètre de régression linéaire et consécutivement l'estimation de la partie fonctionnelle. Nous verrons au Chapitre 5 que ceci offre la possibilité d'algorithmes efficaces. Ensuite, l'étude des propriétés asymptotiques est grandement facilitée par cette approche.

L'estimateur de Huber fut élaboré dans le but de "robustifier" la méthode des moindres carrés, trop sensible aux valeurs aberrantes. L'introduction de la fonction de coût de Huber permet de donner une importance moindre aux grands écarts, de sorte à moins prendre en compte d'éventuelles valeurs aberrantes. –A titre de remarque, l'estimateur de Huber est robuste au sens de Huber (1964) mais n'est pas robuste si l'on considère la définition actuelle de la robustesse qui repose sur la notion des points de rupture (cf. Maronna *et al.* (1979)).– D'après Hampel *et al.* (1986) cet estimateur est peu sensible aux écarts en ordonnée. Les M-estimateurs obtenus par les seuillages Hard et SCAD appartiennent quant à eux à une catégorie de M-estimateurs considérée comme plus robustes que l'estimateur de Huber : les *Hard-Redescending* M-estimateurs (voir par exemple Müller (2004)). Ces derniers ont en fait la particularité d'avoir une fonction de coût constante à partir d'un certain seuil.

La présence de la partie non paramétrique pouvant être considérée en quelque sorte comme un écart en ordonnée relativement au modèle $z_i = A_i^T\beta + \varepsilon_i$, l'idée est que sa présence affectera peu le comportement de l'estimateur. Les coefficients d'ondelettes non négligeables pouvant en effet être considérés comme des valeurs aberrantes, la robustesse des estimateurs obtenus permettra de minimiser le biais induit. L'approfondissement de cette idée fait l'objet du Chapitre 4.

CHAPITRE 4

PROPRIÉTÉS ASYMPTOTIQUES DES ESTIMATEURS

Sommaire

4.1 Hypothèses **57**
4.2 Propriétés asymptotiques **61**
4.2.1 Cas du seuillage doux et de l'estimateur de Huber 61
4.2.2 Exemple issu de Fadili et Bullmore (2005) 63
4.3 Estimation de la variance **65**
4.4 Preuves **67**
4.4.1 Consistance de $\hat{\beta}_n$ avec le seuillage doux 68
4.4.2 Consistance de $\hat{\beta}_n$ avec la pénalité quadratique 82
4.4.3 Estimation de la partie fonctionnelle 85

Ce chapitre donne les propriétés asymptotiques des estimateurs dans les contextes de la pénalité l^1, qui correspondait à la fonction de coût de Huber, et d'une pénalité de Sobolev. Dans chacun de ces cas, nous obtenons des vitesses de convergence minimax ou quasi-minimax et pour la partie linéaire du modèle et pour la partie non paramétrique lorsque la structure des covariables de régression est de la forme suggérée par Rice (1986).

4.1 Hypothèses

Rappelons les deux hypothèses introduites au Chapitre 3 afin d'assurer l'identifiabilité du modèle :

(A1) Le vecteur $\frac{1}{n}X^T F$ tend vers 0 quand n tend vers l'infini.

(A2) La matrice X est de plein rang colonnes, c'est-à-dire $\frac{1}{n}X^TX$ converge vers une matrice inversible.

Nous introduisons ensuite l'hypothèse $(A3)$:

(A3) La suite de matrices (K_n) définie par $K_n := \frac{1}{n}\sum_{i=1}^n A_iA_i^T\rho''_\lambda(\theta_{0,i}+\varepsilon_i)$ converge en norme L^2 vers une matrice K_0 inversible.

L'orthogonalité de W assure que $A^TA = X^TX$. Ainsi sous (A2) la matrice A^TA est inversible à partir d'un certain rang. Si nous définissons alors la matrice de projection sur les vecteurs colonnes de A, $H = A(A^TA)^{-1}A^T$, H est de rang p. Si, pour tout $i = 1, \dots, n$, la quantité h_i désigne le $i^{\text{ème}}$ terme diagonal de H, nous avons $\sum h_i = p$.

Nous faisons aussi une hypothèse sur la structure de la matrice de régression. Usuelle en régression robuste –elle permet d'appliquer le critère de Lindeberg-Feller–, elle est transcrite ici pour la représentation en coefficients d'ondelettes de la matrice X.

(A4) La quantité $h := \max_{i=1,\dots,n} A_i^T(A^TA)^{-1}A_i$ tend vers 0 quand n tend vers l'infini.

La quantité h est en fait le terme diagonal maximal de la matrice H.

Discussion des hypotheses

Afin d'étudier la cohérence des hypothèses, nous pouvons supposer, de manière similaire à Rice (1986) –qui a été reprise entre autres par Speckman (1988), Chen et Shiau (1991) et Fadili et Bullmore (2005)– que les éléments de X peuvent s'écrire sous la forme $x_{i,j} = g_j(t_i) + \xi_{i,j}$ avec $(g_j)_{j=1,\dots,p}$ telles que $\int fg_j = 0$ pour tout j et $\xi_{i,j}$ réalisations de variables ξ_j aléatoires i.i.d. centrées, de variance finie non nulle ; nous notons G et E les matrices de termes généraux respectifs $g_{i,j} = g_j(t_i)$ et $\xi_{i,j}$. Nous avons $\frac{1}{n}E^TE \to V > 0$. Avec X sous cette forme, les hypothèses (A1), (A2) et (A4) deviennent :

(A1) Nous avons $\|\frac{1}{n}X^TF_0\|^2 = \sum_{j=1}^p \left(\frac{1}{n}\sum_{j=1}^n g_i(t_j)f(t_j) + \frac{1}{n}\sum_{j=1}^n \xi_{i,j}f(t_j)\right)^2$.
La loi des grands nombres assure que le second terme tend presque sûrement vers 0 car les variables ξ_i sont centrées.
Le premier terme converge aussi vers 0 grâce à l'hypothèse que $\int fg_i = 0$ pour tout $i = 1, \dots, p$.
Remarquons que le fait de supposer $\forall i$, $\int fg_i = 0$ implique que soit f est d'intégrale nulle soit aucune colonne de X n'est un vecteur constant. Ce constat est usuel pour avoir l'identifiabilité d'un modèle partiellement linéaire (voir par exemple Chen (1988) ou Donald et Newey (1994)).

(A2) Nous avons $\frac{1}{n}X^TX = \frac{1}{n}G^TG + \frac{1}{n}E^TE + \frac{1}{n}G^TE + \frac{1}{n}E^TG$
Etudions la convergence du premier terme :

$$\left(n^{-1}G^TG\right)_{i,j} = \frac{1}{n}\sum_{l=1}^{n} g_i(t_l)g_j(t_l) \to_{n\to\infty} \int g_ig_j.$$

Notons $V(g)$ la matrice de terme général $\int g_ig_j$.
Le deuxième terme, $\frac{1}{n}E^TE$ tend vers V définie positive.
Ensuite, $\|n^{-1/2}G^TE\|^2 = \sum_{i,j=1,\dots,p}\left(\frac{1}{n}\sum_{l=1}^{n} g_i(t_l)\xi_{l,j}\right)^2$. Les variables ξ_i étant centrées, la loi des grands nombres assure que $\|n^{-1/2}G^TE\|$ tend vers 0 en probabilité.
Ainsi, $\frac{1}{n}X^TX$ converge en probabilité vers $V(g)+V$. Si nous supposons que la famille de fonctions $(g_i)_{i=1,\dots,n}$ est orthogonale dans $\mathbb{L}^2$, la matrice $V(g)+V$ est diagonale et ses termes diagonaux sont positifs, donc elle est inversible.

(A4) Montrons dans un premier temps qu'il est équivalent de montrer que $\frac{1}{n}\|A_i\|^2$ tend vers 0 pour avoir (A4). La quantité qui nous intéresse est

$$A_i^T(A^TA)^{-1}A_i = \frac{1}{n}\|(\frac{1}{n}A^TA)^{-1/2}A_i\|^2.$$

Nous avons

$$s((\frac{1}{n}A^TA)^{-1})\|A_i\|^2 \leqslant \|(\frac{1}{n}A^TA)^{-1/2}A_i\|^2 \leqslant \|\frac{1}{n}A^TA\|^2\|A_i\|^2$$

avec $s((\frac{1}{n}A^TA)^{-1})$ plus petite valeur propre de $(\frac{1}{n}A^TA)^{-1}$. Sous l'hypothèse (A2), il suffit donc d'avoir $\frac{1}{n}\|A_i\|^2 \to 0$ pour que (A4) soit vérifiée.
La quantité $\frac{1}{n}\|A_i\|^2$ vaut $n^{-1}A_i^TA_i = \sum_{l=1}^{p}\left[\frac{1}{n}\sum_{j=1}^{n}\psi_i(t_j)X_{j,l}\right]^2$. En reprenant les notations précédentes, $x_{j,l} = g_l(t_j) + \xi_{j,l}$, la loi des grands nombres implique que $n^{-1}A_i^TA_i$ converge presque sûrement vers $\sum_{l=1}^{p}\left(\int \psi_ig_l\right)^2$. Autrement dit, $n^{-1}A_i^TA_i$ tend vers $\sum_{l=1}^{p}(w_i^l)^2$ avec w_i^l coefficient d'ondelettes continu d'ordre i de g_l.
Si nous supposons que pour tout $l = 1,\dots,p$, g_l est une fonction polynômiale de degré inférieur à N, avec N le nombre de moments nuls de l'ondelette, alors cette condition est vérifiée.

L'hypothèse (A3) ne semble pas particulièrement contraignante même s'il semble dur de montrer qu'elle est en effet bien vérifiée dans tous les cas. Nous l'étudions ici dans le cas du seuillage quadratique :

(A3) Dans le cas d'une pénalité quadratique, les matrices K_n s'écrivent $K_n = A^T TA$ avec $T = diag\left(\frac{2\lambda 2^{2[\log_2(i)]s}}{1+2\lambda 2^{2[\log_2(i)]s}}\right)_{i=1,\dots,n}$. La preuve qui suit peut être trouvée dans Fadili et Bullmore (2005). Le terme (i,j) de la matrice $\frac{1}{n}A^T TA$ s'écrit

$$\begin{aligned}(A^T TA)_{i,j} &= \frac{1}{n}\sum_{l=1}^{n} t_l w_l^i w_l^j + \frac{1}{n}\sum_{l=1}^{n} t_l w_l^i \eta_l^j + \frac{1}{n}\sum_{l=1}^{n} t_l w_l^j \eta_l^i \\ &\quad + \frac{1}{n}\sum_{l=1}^{n}(1-t_l)\eta_l^i\eta_l^j + \frac{1}{n}\sum_{l=1}^{n}\eta_l^i\eta_l^j, \\ &= T_1 + T_2 + T_3 + T_4 + T_5,\end{aligned}$$

t_l désignant $t_l := \frac{2\lambda 2^{2[\log_2 l]s}}{1+2\lambda 2^{2[\log_2 l]s}}$.
Par Cauchy-Schwarz, le premier terme est majoré par

$$T_1 \leqslant \left(\frac{1}{n}\sum_{l=1}^{n} t_l^2 w_l^{i2}\right)^{1/2}\left(\frac{1}{n}\sum_{l=1}^{n} w_l^{j2}\right)^{1/2},$$

donc nous avons $T_1 \leqslant \lambda^{1/2}\|g_i\|_{W^s}\|g_j\|$ par l'égalité de Parseval. Si g_i appartient à W^s et g_j appartient à L^2, alors le terme T_1 tend vers 0.
Les termes T_2 et T_3 sont d'espérance nulle. Leur variance est majorée respectivement par $\sigma^2\frac{1}{n^2}\sum_{l=1}^{n} t_l^2 w_l^k$ avec $k = i$ ou $k = j$, donc par $\frac{\lambda}{n}\|g_k\|_{W^s}$. Nous en déduisons que ces deux termes tendent en probabilité vers 0 dès lors que g_i et g_j appartiennent à W^s.
Si $i = j$, le terme T_4 est d'espérance nulle et si $i \neq j$, son espérance vaut $\frac{1}{n}trace(I - T)$. D'après Fadili et Bullmore (2005), $trace(I - T) = \bigcirc(\lambda^{-1/(2s)})$. Donc pour une régularité $s > 1/4$ l'espérance tend vers 0. La variance, à une constante près, peut être majorée par $\frac{1}{n^2}\sum_{l=1}^{n} t_l^2 = trace((I-T)^T(I-T))$. Cette quantité est de l'ordre de $\lambda^{-1/2s}$ d'après Fadili et Bullmore (2005). Nous en déduisons que le terme T_4 tend en probabilité vers 0.
Quant au terme T_5, il est immédiat qu'il converge vers V.
Il en résulte que la suite de matrices $\frac{1}{n}A^T TA$ converge en probabilité vers la matrice V. Il est suffisant pour cela que les fonctions g_i, $i = 1,\dots,n$ appartiennent à un espace de Sobolev W^s (ce qui est vérifié pour les fonctions polynomiales) et soient dans L^2.

En résumé, si les observations X_i, $i = 1,\dots,n$ sont de la forme $X_i = g(t_i) + \xi_i$ avec $g = (g_1 \dots g_p)^T$, g_j fonctions polynômiales de degré inférieur à N, où N est le nombre de moments nuls de l'ondelette et ξ_i variables centrées indépendantes de variances finies non nulles, si de plus la famille $(g_j)_{j=1,\dots,p}$ est orthonormée dans $\mathbb{L}^2$ et si $\int fg_j = 0$ pour tout

$j = 1, \ldots, p$, alors les conditions (A1), (A2) et (A4) sont vérifiées. Ces hypothèses semblent peu restrictives et offrent donc un large champ d'application des résultats présentés.

Nous retrouvons une condition quant à la forme des fonctions g_j similaire à celle de Rice (1986), citée au Chapitre 1, dans la mesure où ces fonctions sont supposées polynomiales afin que les hypothèses soient vérifiées. Notre hypothèse est toutefois légèrement plus faible dans la mesure où le degré maximal dépend du nombre de moments nuls de l'ondelette et non de la régularité de la fonction étudiée. Cette restriction est ainsi cohérente avec les résultats précédemment établis. Cependant, contrairement à Rice (1986), nous n'établissons pas que cette hypothèse est nécessaire à la bonne convergence, uniquement qu'elle est suffisante. Nous n'avons pas considéré le cas de fonctions plus complexes, mais une procédure par étapes comme suggéré par Chen et Shiau (1991) (cf. Chapitre 1) pourrait être envisagée.

4.2 Propriétés asymptotiques

Dans cette section, nous donnons les principaux résultats quant aux propriétés asymptotiques du schéma d'estimation proposé. Nous considérons deux des cas présentés au Chapitre 3 : la pénalité l^1 et la pénalisation par une norme de Sobolev. Les preuves des résultats exposés dans cette section se trouvent en Section 4.4.

En s'inspirant des travaux de Rice (1986), l'étude des comportements asymptotiques distingue selon qu'une hypothèse quant à la structure des variables est émise ou pas. cette structure sera considérée sous une forme similaire à celle de Rice le cas échéant.

4.2.1 Cas du seuillage doux et de l'estimateur de Huber

Considérons dans un premier temps le cas de la pénalité l^1. Rappelons que celle-ci mène à une estimation de la partie linéaire du modèle par le M-estimateur de Huber et à un seuillage doux des coefficients de la partie fonctionnelle. Les vitesses de convergences obtenues sont les suivantes :

Théorème 4.1.
Considérons le seuillage universel $\lambda = \sigma\sqrt{2\log(n)}$. Supposons les hypothèses (A1) à (A4) vérifiées.

Alors,

$$\begin{aligned} \hat{\beta}_n - \beta_0 &= \bigcirc_{\mathbb{P}}\left(\left(\frac{\log(n)}{n}\right)^{s/(1+2s)}\right), \\ \text{et} \qquad \sqrt{n}(\hat{\beta}_n - \beta_0) &= K_0^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho'_\lambda(\theta_{0,i}+\varepsilon_i)A_i\right) + o_{\mathbb{P}}(\log(n)^{s/(1+2s)}n^{1/(2(1+2s))}). \end{aligned}$$

Soit N le nombre de moments nuls de l'ondelette ψ.
Si de plus, pour tout $i = 1, \ldots, n$, $j = 1, \ldots, p$, nous considérons que $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j fonctions polynomiales de degré au plus égal au nombre N de moments nuls de l'ondelette et $\xi_{i,j}$ issu d'un vecteur de variables ξ_j de loi centrée de matrice de variance-covariance V finie, alors

$$\begin{aligned} \hat{\beta}_n - \beta_0 &= \bigcirc_{\mathbb{P}}\left(\sqrt{\frac{\log(n)}{n}}\right), \\ \text{et} \qquad \sqrt{n}(\hat{\beta}_n - \beta_0) &= K_0^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho'_\lambda(\theta_{0,i}+\varepsilon_i)A_i\right) + o_{\mathbb{P}}\left(\sqrt{\log(n)}\right). \end{aligned}$$

Supposons que la fonction d'échelle φ et l'ondelette ψ appartiennent à $\mathcal{C}^R$.
Si f appartient à l'espace de Besov $\mathcal{B}^s_{\pi,r}$ avec $0 < s < min(R, N)$ et $s + 1/2 - 1/\pi > 0$, alors

$$\|\hat{f}_n - f\|_2 = \bigcirc_{\mathbb{P}}\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{1+2s}}\right),$$

avec $\|\hat{f}_n - f\|_2^2 = \int_0^1 (\hat{f}_n - f)^2$.

La preuve de ce résultat est donnée en fin de chapitre.

La vitesse de l'estimateur de la partie fonctionnelle est quasi-minimax. Cependant, la vitesse d'estimation pour le vecteur des paramètres de régression n'est en général pas optimale. En effet, en l'absence d'hypothèses (autres que (A1) à (A4)) quant à la structure de la matrice de régression, l'estimateur élaboré admet une vitesse non paramétrique.

Lorsque nous supposons que les covariables vérifient une relation de type $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j fonction polynomiale de degré au plus égal au nombre N de moments nuls de l'ondelette et $\xi_{i,j}$ issu d'une loi centrée de variance finie, alors nous obtenons une vitesse quasi-minimax dans l'estimation de la partie linéaire. L'estimation est presque optimale dans le sens où la vitesse ne diffère que d'un terme logarithmique de la vitesse minimax.

Nous perdons en effet un terme $\sqrt{\log(n)}$ dans l'estimation des paramètres de régression linéaire.

L'explication de la non-optimalité de la vitesse pour la partie paramétrique réside dans le choix du paramètre de seuillage λ. Si nous pouvions prendre λ tendant vers 0, alors nous obtiendrions un résultat minimax pour l'estimateur $\hat{\beta}_n$. Toutefois, ce choix de paramètre est exclu dans la mesure où il ne permet pas d'aboutir à une bonne qualité d'estimation de la partie non paramétrique. Nous retrouvons le compromis de Rice (1986), entre l'optimalité dans la partie linéaire et le surlissage de la fonction ou une perte de la qualité d'estimation de la régression mais un lissage adapté de la partie fonctionnelle.

De même que Rice (1986), nous trouvons que lorsque la structure des covariables de régression est suffisamment "décorrélée" de la partie fonctionnelle (voir plus précisément au Chapitre 1) alors les estimateurs de la partie linéaire et de la partie fonctionnelle du modèle sont quasi-minimax. La condition quant à la structure des covariables est de plus très similaire à celle de Rice (1986). Elle est relativement moins contraignante dans la mesure où le degré maximal des fonctions polynomiales intervenant dans la partie linéaire dépend de la régularité de l'ondelette considérée et non de celle de la partie fonctionnelle du modèle.

L'intérêt de notre procédure, outre l'extension à des fonctions appartenant à des espaces de Besov, est qu'elle est adaptative par rapport à la régularité de la fonction.

Ainsi, nous retrouvons le compromis de Rice (1986) quant à la structure des covariables. Lorsque celles-ci sont de la forme $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j fonction polynomiale de degré au plus égal au nombre N de moments nuls de l'ondelette et $\xi_{i,j}$ issu d'une loi centrée de variance finie, nous obtenons des estimateurs quasi-minimax. Nous avons de plus une écriture asymptotique au second ordre de $\hat{\beta}_n$ qui permettra probablement ultérieurement d'élaborer des tests sur ces paramètres. Nous avons par ailleurs une estimation adaptative de la partie fonctionnelle.

4.2.2 Exemple issu de Fadili et Bullmore (2005)

Nous étudions ici le cadre d'une pénalisation par une norme de Sobolev. Les résultats sont résumés dans le théorème suivant :

Théorème 4.2.
Supposons que les hypothèses (A1) à (A4) soient vérifiées. Supposons que $\lambda = \bigcirc(n^{-2s/(1+2s)})$

et que f appartient à un espace de Hölder $\Lambda^s(M)$ avec $s > 1/4$.

Soit φ une fonction d'échelle admettant N moments nuls, avec $N > s+1$.
Les estimateurs $\hat{\beta}_n$ et $\hat{f}_n$ vérifient :

$$\begin{aligned} \hat{\beta}_n - \beta_0 &= \bigcirc_{\mathbb{P}}\left(n^{-s/(1+2s)}\right), \\ \|\hat{f}_n - f\|_2 &= \bigcirc_{\mathbb{P}}\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{1+2s}}\right) \end{aligned}$$

Si de plus, pour tout $i = 1, \ldots, n$, $j = 1, \ldots, p$, nous considérons que $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j appartenant à $\Lambda^s(M_j)$ et $\xi_{i,j}$ issu d'une loi centrée de variance finie, alors

$$\hat{\beta}_n - \beta_0 = \bigcirc_{\mathbb{P}}\left(n^{-2s/(1+2s)}\right).$$

Lorsque pour tout $j = 1, \ldots, p$, g_j est une fonction polynomiale de degré inférieur ou égal au nombre de moments nuls de l'ondelette, alors

$$\hat{\beta}_n - \beta_0 = \bigcirc_{\mathbb{P}}\left(n^{-1/2}\right).$$

De même que pour le seuillage précédemment étudié, nous obtenons une vitesse quasi-minimax pour l'estimation de la partie fonctionnelle, mais la vitesse de convergence de l'estimateur du paramètre de régression n'est pas satisfaisante dans le cas général. Nous retrouvons le résultat de Fadili et Bullmore (2005) lorsqu'aucune hypothèse autre que les hypothèses (A1) à (A3) n'est faite. La vitesse est améliorée sous l'hypothèse de structure des covariables émise dans le théorème, et devient minimax lorsque nous avons une forme similaire à celle des théorèmes précédents. Dans ce dernier cas, nous avons une procédure minimax pour l'estimation de la partie fonctionnelle et de la partie linéaire. Nous avons donc un résultat similaire à celui de Rice (1986).

La majeure différence avec le seuillage doux est que le schéma d'estimation n'est pas adaptatif dans la mesure où il est nécessaire de connaître la régularité s de la fonction f. Afin de résoudre ce point, une procédure de validation croisée peut être appliquée. Un exemple de validation croisée dans ce contexte est développée en Section 6.2.1. L'un des inconvénients de cette procédure de seuillage en pratique est le temps de calcul résultant de la détermination par validation croisée des paramètres. Celui-ci incite à choisir les schémas adaptatifs, tels que les seuillages fort et doux. Mais surtout, comme le soulignent Rice (1986) et Speckman (1988), la procédure de validation croisée risque d'entraîner des biais importants : lorsque la condition de structure de la partie linéaire $X_{i,j} = g_j(t_i) + \xi_{i,j}$

avec g_j fonction polynomiale de degré au plus égal à N n'est pas vérifier le paramètre λ doit converger vers 0 plus vite que la valeur optimale pour l'estimation de la partie fonctionnelle pour améliorer la qualité d'estimation de la partie linéaire. Dans ce contexte, les conséquences d'une procédure de validation croisée seront alors un sous-lissage dans l'estimation de la partie fonctionnelle.

4.3 Estimation de la variance

Le schéma d'estimation que nous proposons nécessite de connaître la variance du bruit, σ^2. En effet, le paramètre de seuillage λ dépend de cette variance ; le seuillage universel que nous avons considéré pour les seuillages doux et fort s'écrit $\lambda = \sigma\sqrt{2\log(n)}$. Afin d'estimer σ^2, une première approche possible est de le calculer de manière itérative, c'est-à-dire en actualisant notre estimation à l'aide de l'estimation du paramètre de régression β_0 puis en réestimant β_0 avec l'estimation de σ obtenue. Cela est par exemple proposé dans Fadili et Bullmore (2005). Néanmoins, une méthode directe, plus facile à justifier du point de vue théorique, est préférable. L'implémentation confirme par ailleurs l'utilisation de la méthode présentée ci-après.

Dans un modèle non paramétrique, l'étude de Donoho *et al.* (1995) incite à estimer la variance du bruit par l'estimateur MAD des coefficients d'ondelettes z_i de résolution $J = \log_2(n)$ (cf. Section 2.3.2). Néanmoins, contrairement à ce qu'affirment Fadili et Bullmore (2005), cet estimateur n'est pas adapté au modèle partiellement linéaire car la représentation de la partie linéaire du modèle n'est pas creuse. En pratique, ceci est confirmé par des estimations très éloignées des valeurs réelles.

L'idée est ici d'effectuer une décomposition QR de la matrice de régression A. Cependant nous n'appliquons pas cette décomposition sur toutes les colonnes de A mais sur les colonnes correspondant aux coefficients d'ondelettes d'échelle maximale $J = \log_2(n)$. Notons A_J cette matrice. A_J est une matrice de taille $n_J \times p$ avec $n_J = n/2 > p$. La décomposition QR nous dit qu'il existe une matrice orthogonale Q et une matrice triangulaire supérieure R telles que

$$A_J = Q \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

(Q est de taille $n_J \times n_J$, R de taille $p \times p$.)

Notons Z_J, θ_J et $\underline{\varepsilon}_J$ les coefficients d'ondelettes correspondant au degré de résolution J

respectivement de Y, F et U. En multipliant par Q^T le modèle projeté, il en résulte que

$$Q^T z_J = \begin{pmatrix} R \\ 0 \end{pmatrix} \beta_0 + Q^T \theta_J + Q^T \underline{\varepsilon}_J.$$

L'estimateur MAD appliqué non pas aux dernières composantes de z mais de $Q^T z_J$ va alors donner une bonne estimation de σ. En effet, les composantes de la partie linéaire sont nulles donc nous sommes ramenés au cadre d'un modèle non paramétrique. La raison pour laquelle nous nous sommes restreints à la décomposition QR de A_J^T au lieu de celle de A est que sinon les coefficients d'ondelettes de faible résolution de la partie fonctionnelle interviennent dans l'estimation et engendrent un biais important. Le fait que la représentation de la partie fonctionnelle soit creuse et que le bruit est bien gaussien centré et de variance σ^2 assure de bonnes propriétés à l'estimateur (cf Donoho *et al.* (1995)).

Nous avons en effet le résultat suivant :

Proposition 4.3.
Si $\hat{\sigma}_n$ désigne l'estimateur défini ci-dessus, alors

$$\inf_f \mathbb{P}(\sigma < \hat{\sigma}_n) \underset{n\to\infty}{\longrightarrow} 1.$$

De plus, si l'ondelette ψ est $\mathcal{C}^r$ et admet N moments nuls, alors pour $1/\pi < s < \min(R, N) - (1/2 - 1/\pi)_+$,

$$\sup_{f \in \mathcal{B}^s_{\pi,r}} \mathbb{P}(\hat{\sigma}_n > 1.01\,\sigma) \underset{n\to\infty}{\longrightarrow} 1.$$

Démonstration. Soit $\kappa \in \{p+1, \ldots, n_J\}$. Introduisons $\tilde{\theta}_\kappa$ la $\kappa^{\text{ième}}$ composante de $Q^T\theta_J$. Alors $\hat{\sigma}_n$ est l'estimateur MAD défini par Donoho *et al.* (1995) dans le modèle non paramétrique :

$$y_i = g(i/n) + \sigma U_i$$

avec U_i indépendants de loi normale centrée réduite et $g = \sum_{k=p+1}^{n_J} \tilde{\theta}_{J,k} \psi_{J,k}$.

Rappelons le résultat de Donoho *et al.* (1995) (Section 5.1. page 325) dans un tel modèle :

Proposition 4.4. *(Donho, Johnstone, Kerkyacharian et Picard (1995))*

$$\inf_g \mathbb{P}(\sigma < \hat{\sigma}_n) \underset{n\to\infty}{\longrightarrow} 1.$$

Si l'ondelette ψ est $\mathcal{C}^R$ et admet N moments nuls, alors, pour tout $1/\pi < s < \min(R, D)$,

$$\sup_{g \in \mathcal{B}^s_{\pi,r}} \mathbb{P}(\hat{\sigma}_n > 1.01\,\sigma) \underset{n\to\infty}{\longrightarrow} 1.$$

Le premier résultat de la Proposition 4.3 est donc une application directe de celui de Donoho *et al.* (1995).

Quant au second il résulte du fait que la matrice Q^T étant orthogonale, $\|Q^T\|_2 = 1$, donc $\|Q^T\|_\infty \leqslant \sqrt{n_J}$. D'où

$$\|\underline{\tilde{\theta}}\|_\infty \leqslant \sqrt{n_J}\| \begin{pmatrix} \tilde{\theta}_{p+1} & \dots & \tilde{\theta}_{n_J} \end{pmatrix} \|_\infty \leqslant 2^{(J-1)/2}\|\theta_J\|_\infty. \tag{4.1}$$

Si f appartient à un espace de Besov $\mathcal{B}^s_{\pi,r}$ avec $s > 1/\pi$, alors $f \in \mathcal{B}^{s-1/\pi}_{\infty,r}$, donc

$$\|\theta_J\|_\infty = C2^{(J-1)(s-1/\pi+1/2)}$$

avec C constante. D'où

$$\|\underline{\tilde{\theta}}\|_\infty \leqslant C2^{(J-1)(s-1/\pi+1)}.$$

Il en résulte que la fonction $g = \sum_{\kappa=p+1}^{n_J} \tilde{\theta}_\kappa \psi_\kappa$ appartient à l'espace de Besov $\mathcal{B}^{s+1/2-1/\pi}_{\infty,r}$ qui est bien défini car $s + 1/2 - 1/\pi > 0$. Sous les conditions de la Proposition 4.3, cet espace vérifie les conditions du second résultat de Donoho *et al.* (1995), ce qui permet donc de conclure la preuve. □

4.4 Preuves

Cette section est constituée des preuves des Théorèmes 4.1 et 4.2.

D'autres voies de démonstration que celle présentée ici ont été explorées, notamment l'utilisation des pseudo-variables (-voir Oh *et al.* (2007)- qui ne donnait pas semble-t'il de résultats concluants), le schéma de Mammen et Van der Geer (1997), reposant sur l'entropie des boules de Besov donnée par Loubes et Van der Geer (2002) (qui ne distingue pas les parties paramétrique et non paramétrique ; son intérêt est à première vue d'offrir un cadre plus vaste pour le choix des pénalités). Dans le cadre de *Redescending* M-estimateurs telle la moyenne tronquée, une preuve de consistance dans le modèle linéaire peut être trouvée dans Davies (1990). L'idée est d'utiliser les lemmes de symétrisation de Pollard (1990) en approximant la fonction de coût par des fonctions constantes par morceaux. La présence de la partie non paramétrique empêchait à première vue de reprendre la preuve citée.

Nous établissons d'abord la consistance de l'estimateur $\hat{\beta}_n$ dans chacun des contextes étudiés avant de nous intéresser à la convergence de la partie fonctionnelle. En ce qui

concerne la consistance de $\hat{\beta}_n$, celle-ci est donnée dans un premier temps dans le cas du seuillage doux, facilement approximable par une fonction plus lisse. Lorsque la pénalité choisie est celle de Fadili et Bullmore (2005), le même schéma peut être appliqué mais l'écriture explicite de l'estimateur offre une alternative plus rapide. Nous terminons cette section par la preuve de la convergence de la partie fonctionnelle.

Mais donnons d'abord ce lemme utile pour la suite :

Lemme 4.5. *Sous (A2) et (A3),*

(i) $n^{-1/2} \sup_{i=1,\dots,n} \|A_i\| \to 0$

(ii) $n^{-1} \sum_{i=1,\dots,n} \|A_i\|^2 = \bigcirc(1)$

Démonstration. Il suffit de remarquer que

$$\|A_i\|^2 = A_i^T (A^T A)^{1/2} (A^T A)^{-1/2} A_i \leqslant \|A_i\| \|(A^T A)^{1/2}\| h_i^{1/2},$$

donc $\|A_i\| \leqslant n^{1/2} \|(\frac{1}{n} A^T A)^{1/2}\| h_i^{1/2}$. □

4.4.1 Consistance de $\hat{\beta}_n$ avec le seuillage doux

La preuve est construite de la façon suivante : dans un premier temps nous définissons un changement de variables permettant de développer l'argumentation dans un modèle équivalent. Ensuite, nous rappelons des théorèmes relatifs à la convergence de fonctions convexes et donnons une approximation de la fonction de coût par une fonction plus lisse. L'établissement des propriétés de l'estimateur est enfin réalisé en deux étapes : l'étude du comportement asymptotique du critère minimisé et la consistance qui en découle.

Elaboration d'un modèle équivalent par changement de variable

Le changement de variable réalisé repose sur le résultat suivant :

Proposition 4.6.
Si les hypothèses (A2) et (A3) sont vérifiées, et $\lambda = \sigma\sqrt{2\log(n)}$, alors :

$$\frac{1}{n} \sum_{i=i_0}^{n} \rho'_\lambda(\theta_{0i} + \varepsilon_i) A_i = \bigcirc_{\mathbb{P}}((\log(n)/n)^{s/(1+2s)}).$$

Si de plus la matrice de régression X vérifie : pour tout $i = 1, \dots, n$, pour tout $j = 1, \dots, n$,

$$x_{i,j} = g_j(t_i) + \xi_{i,j},$$

avec g_j fonctions polynomiales de degré au plus égal au nombre de moments nuls de l'ondelette, et $\xi_{i,j}$ réalisations du vecteur de variables aléatoires $(\xi_1, \xi_2, \dots, \xi_p)$ centré et de matrice de variance-covariance convergeant vers une matrice V positive, alors :

$$\frac{1}{\sqrt{n}} \sum_{i=i_0}^{n} \rho'_\lambda(\theta_{0i} + \varepsilon_i) A_i = \bigcirc_{\mathbb{P}}(\lambda).$$

Démonstration. Notons

$$B := \frac{1}{n} \sum_{i=i_0}^{n} \rho'_\lambda(\theta_i + \varepsilon_i) A_i.$$

Nous avons :

$$B = \frac{1}{n} \sum_{i=i_0}^{n} A_i(\theta_i + \varepsilon_i - \gamma(\theta_i + \varepsilon_i, \lambda)),$$

avec $\gamma(., \lambda)$ fonction de seuillage doux de seuil λ.

En appliquant l'inégalité de Cauchy-Schwarz, la norme de

$$B_2 := \frac{1}{n} \sum_{i=i_0}^{n} A_i(\theta_i - \gamma(\theta_i + \varepsilon_i, \lambda)).$$

peut être majorée par

$$\left(\frac{1}{n} \sum_{i=i_0}^{n} \|A_i\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=i_0}^{n} (\theta_i - \gamma(\theta_i + \varepsilon_i, \lambda))^2 \right)^{1/2}.$$

D'après le Lemme 4.5, $\frac{1}{n} \sum_{i=i_0}^{n} \|A_i\|^2$ est borné. Soit

$$B_2 := \frac{1}{n} \sum_{i=i_0}^{n} (\theta_i - \gamma(\theta_i + \varepsilon_i, \lambda))^2.$$

Par l'égalité de Parseval, nous avons $B_2 = \|f - \tilde{f}\|^2$ avec $\tilde{f}$ estimation de f dans le modèle non paramétrique $y_i = f(i/n) + \varepsilon_i$. D'après Donoho et Johnstone (1994), ce terme est d'ordre $(n/\log(n))^{-2s/(1+2s)}$. Ce qui achève la démonstration du premier point de la proposition.

Reprenons maintenant l'étude du comportement asymptotique du terme B lorsque la matrice de régression a la forme suggérée dans le cadre du Théorème 4.1 : pour tout

couple (i,j), $i = 1,\dots,n$ et $j = 1,\dots,p$, nous considérons que $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec $\xi_{i,j}$ issu d'une loi centrée de matrice de variance-covariance V. Supposons que pour tout j, g_j est une fonction polynomiale de degré inférieur ou égal au nombre de moments nuls de l'ondelettes. Soit $j \in \{1,\dots,p\}$ fixé. Alors, les coefficients d'ondelettes des fonctions g_j sont nuls. D'où :

$$B = \frac{1}{n}\sum_{i=i_0}^{n} \eta_i \rho'_\lambda(\theta_i + \varepsilon_i),$$

où les vecteurs $(\eta_i)_{i=1,\dots,n}$ sont définis par $\eta = W\zeta$. Par orthogonalité de la transformée en ondelettes W, les vecteurs $(\eta_i)_{i=1,\dots,n}$ sont issus d'une loi centrée de matrice de variance-covariance V. L'espérance de B_2 est donc nulle. De plus nous avons :

$$\mathbb{E}B^2 \leqslant \frac{\lambda^2}{n^2}\mathbb{E}\left(\sum_{i=i_0}^{n} \eta_i^2\right).$$

Nous en déduisons que $B = \bigcirc_{\mathbb{P}}(\lambda n^{-1/2})$.

□

Remarque : La vitesse de convergence du terme $\frac{1}{\sqrt{n}}\sum_{i=i_0}^{n} \rho'_\lambda(\theta_{0i} + \varepsilon_i)A_i$ est déterminante dans la vitesse de l'estimation du vecteur β. Si ce terme est borné, la vitesse de convergence en probabilité de l'estimateur $\hat{\beta}_n$ est d'ordre $n^{-1/2}$, donc optimale, et s'il est d'ordre v_n, alors $\hat{\beta}_n$ est d'ordre $v_n n^{-1/2}$.

Nous sommes amenés à étudier le modèle

$$z_i = A_i^T\beta_0 + e_i,\ i = 1,\dots,n, \quad \text{sous (A2)-(A4)} \tag{4.2}$$

avec $e_i = \theta_i + \varepsilon_i$. Nous ne prenons pas en compte ici l'hypothèse (A1) dans la mesure où elle n'intervient pas dans les preuves ; cette hypothèse est introduite uniquement pour assurer que le modèle est identifiable.

De manière similaire à Huber (1981), Bai *et al.* (1992) ou encore Wu (2004), nous pouvons établir les résultats pour un modèle équivalent en effectuant des changements de variables. L'idée de ce changement de variable est de se ramener à une matrice de régression R telle que $R^TR = I_p$.

Les changements de variable suivants sont effectués :

$$\begin{aligned} R &= A(A^TA)^{-1/2} \sim_{n\to\infty} n^{-1/2}A\Sigma^{-1/2}, \\ \zeta &= (A^TA)^{1/2}\beta. \end{aligned}$$

Nous obtenons alors le modèle

$$z_i = R_i^T \zeta_0 + e_i \quad \text{sous (A2')-(A4')} \tag{4.3}$$

avec

(A2') $R^T R = I_p$.

(A3') $\max_{i=i_0,\dots,n} R_i^T R_i$ tend vers 0.

(A4') $\sum_{i=i_0}^{n} R_i R_i^T \mathbb{E}\left[\rho''_\lambda(e_i)\right]$ converge vers une matrice définie positive notée K'_0.

Nous souhaitons dans un second temps nous ramener à une hypothèse de la forme $\sum_{i=i_0}^{n} R_i \rho'_\lambda(e_i) = \bigcirc_{\mathbb{P}}(1)$. Pour déterminer un changement de variables permettant de se ramener à un modèle vérifiant cette hypothèse, nous utilisons les propriétés de changement d'échelle de la fonction ρ.

La fonction de coût de Huber vérifie pour tout $v > 0$:

$$\rho_\lambda(u) = v^2 \rho_{\lambda/v}(u/v). \tag{4.4}$$

Alors, $\hat{\zeta}_n = \underset{\zeta}{argmin} \sum_{i=i_0}^{n} \rho_\lambda(z_i - R_i^T \zeta)$ est aussi solution du problème de minimisation

$$\hat{\zeta}_n = \underset{\zeta}{argmin} \sum_{i=i_0}^{n} \rho_{\lambda/v_n}\left(\frac{1}{v_n}(z_i - R_i^T \zeta)\right).$$

En remplaçant z_i par la valeur donnée par le modèle, nous avons $\frac{1}{v_n}(z_i - R_i^T \zeta) = e_i - R_i^T \frac{1}{v_n}(\zeta - \zeta_0)$. En posant $\alpha = \frac{1}{v_n}(\zeta - \zeta_0)$, nous obtenons que

$$\hat{\alpha}_n = \underset{\alpha}{argmin} \sum_{i=1}^{n} \rho_{\lambda/v_n}(d_i - R_i^T \alpha),$$

avec $d_i = \frac{1}{v_n} e_i$.

Remarquons que $\rho'_\lambda(u) = v_n \rho'_{\lambda/v_n}(u/v_n)$ et $\rho''_\lambda(u) = \rho''_{\lambda/v_n}(u/v_n)$. Les résultats seront ainsi établis pour le modèle auxiliaire :

$$z_i = R_i^T \alpha_0 + d_i \quad \text{sous (A2'')-(A4'');} \tag{4.5}$$

(A2'') $R^T R = I_p$.

(A3″) $h = \max_{i=i_0,\dots,n} R_i^T R_i$ tend vers 0.

(A4″) $K_n'' := \sum_{i=i_0}^n R_i R_i^T \mathbb{E}\left[\rho''_{\lambda/v_n}(d_i)\right]$ tend vers K_0'' inversible.

La suite v_n est prise de telle sorte que

$$\frac{1}{v_n}\frac{1}{\sqrt{n}}\sum_{i=1}^n A_i \rho'_\lambda(\theta_i + \varepsilon_i) = \bigcirc_{\mathbb{P}}(1).$$

Les résultats asymptotiques établis pour ce modèle sont alors transcrits au modèle (3.2) en effectuant les transformations :

$$\begin{aligned} R &\sim n^{-1/2} A \Sigma^{-1/2}, \\ \alpha &\sim \sqrt{n}\Sigma^{1/2}\left(\frac{1}{v_n}(\beta - \beta_0)\right). \end{aligned}$$

Nous avons asymptotiquement $K_0'' \sim \Sigma^{-1} K_0$. Notons de plus que la Proposition 4.6 devient dans le modèle (4.5) :

$$\sum_{i=i_0}^n \rho'_{\lambda/v_n}(d_i) R_i = \bigcirc_{\mathbb{P}}(1).$$

Montrer la consistance dans le modèle (4.5) est alors équivalent à montrer la $\frac{\sqrt{n}}{v_n}$-consistance dans le modèle (3.2) ; et supposer que $\hat{\alpha}_n$ est borné revient à supposer que $\hat{\beta}_n$ est d'ordre $\frac{n^{1/2}}{v_n}$. Pour toutes les preuves, nous considérerons le modèle (4.5) et nous transcrirons ensuite les résultats à l'aide des relations données ci-dessus.

Outils

Comme cela a été expliqué précédemment, nous reprenons le schéma de preuve de Bai *et al.* (1992) dans le modèle (4.5). Cette preuve s'appuie essentiellement sur la convexité de ρ, notamment sur les théorèmes de Rockafellar (1970), que nous traduisons pour la convergence en probabilité afin de pouvoir les appliquer au contexte présent.

Proposition 4.7.
Soit C un compact de $\mathbb{R}^m$. Soit f_n une famille de fonctions aléatoires convexes et soit f une fonction aléatoire convexe définies sur C à valeur dans un espace de probabilité (Ω, P, μ) donné. Si pour tout $u \in C$, $f_n(u) - f(u)$ converge en probabilité vers 0, alors $\sup_{\{u\in C\}} f_n(u) - f(u)$ converge en probabilité vers 0.

Démonstration. Rappelons un théorème donné dans Rockafellar (1970) (Théorème 10.8 page 90) :

Proposition 4.8.
Soit $\mathcal{C}$ un ouvert convexe. Soit f_n une famille de fonctions convexes finies et soit f une fonction convexe. Si f_n converge simplement vers f sur $\mathcal{C}$, alors la convergence est uniforme sur tout compact inclus dans $\mathcal{C}$.

Afin d'obtenir un résultat similaire pour la convergence en probabilité, nous utilisons le lemme suivant :

Lemme 4.9.
Soient $(X_n)_n$ une suite de variables aléatoires et X variable aléatoire.
Alors (X_n) converge en probabilité vers X si et seulement si de toute suite extraite de (X_n) on peut extraire une sous-suite convergeant presque sûrement vers X.

Soit C un compact. Soit f_n une famille de fonctions aléatoires convexes différentiables et soit f une fonction aléatoire convexe différentiable. Supposons que pour tout $u \in C$, $f_n(u)$ converge en probabilité vers $f(u)$. Soit $\nu(n)$ sous-suite de n.

Nous souhaitons montrer dans un premier temps qu'il existe une sous-suite $\eta(n)$ de $\nu(n)$ telle que pour tout $u \in C$, $f_{\eta(n)}(u) - f(u) \xrightarrow{p.s.} 0$.

Le Lemme 4.9 permet tout d'abord de dire que pour tout $u \in C$, il existe $\eta_u(n)$ extraction de $\nu(n)$ telle que $f_{\eta_u(n)}(u) - f(u) \xrightarrow{p.s.} 0$. Considérons $\mathcal{D} = \{u_0, u_1, u_2 \ldots\}$ sous-ensemble dense dénombrable de C. Par procédé diagonal, nous pouvons construire $(\eta(n))$ telle que pour tout $u \in \mathcal{D}$, $f_{\eta(n)}(u) - f(u) \xrightarrow{p.s.} 0$:

- Il existe $\eta^0 = \eta_{u_0}$ tel que $f_{\eta^0(n)}(u_0) - f(u_0) \xrightarrow{p.s.} 0$;
  - La suite $f_{\eta^0(\eta^{u_1}(n))(u_1)} - f(u_1)$ converge vers 0 presque sûrement donc il existe $\eta^1 = \eta^0 o \eta^{u_1}$ tel que $f_{\eta^1(n)}(u_1) - f(u_1) \xrightarrow{p.s.} 0$ et $f_{\eta^1(n)}(u_0) - f(u_0) \xrightarrow{p.s.} 0$;
  - Supposons η^k construit, en prenant $\eta^{k+1} = \eta^0 o \eta^1 o \ldots o \eta^k o \eta_{u_{k+1}}$,nous avons

$$\forall i = 0, \ldots, k+1,\ f_{\eta^{k+1}(n)}(u_i) - f(u_i) \xrightarrow{p.s.} 0.$$

  Ainsi pour tout k, il existe η^{k+1} tel que pour tout $i = 0, \ldots, k$, $f_{\eta^k(n)}(u_i) - f(u_i)$ tende presque sûrement vers 0. En posant $\eta(n) := \eta^n(n)$, nous obtenons le résultat souhaité.

La convergence de $f_{\eta(n)} - f$ sur un sous espace dense de C assure la convergence sur C par continuité de $f_{\eta(n)} - f$.

Ainsi, $f_{\eta(n)}$ est une suite de fonctions convexes différentiables convergeant presque sûrement vers f différentiable sur C. Le théorème de Rockafellar (1970) peut donc être appliqué et permet d'obtenir que $\sup_{u \in C} f_{\eta(n)}(u) - f(u)$ tend presque sûrement vers 0.

En résumé, il a été montré que de toute suite extraite $\sup_{u \in C} f_{\nu(n)}(u) - f(u)$ de $\sup_{u \in C} f_n(u) - f(u)$ on pouvait extraire une sous-suite convergeant presque sûrement vers 0. Le Lemme 4.9 permet alors de conclure que $\sup_{u \in C} f_n(u) - f(u)$ converge en probabilité vers 0. □

Approximation de la fonction de coût par une fonction plus régulière

L'idée est d'approximer la dérivée troisième de ρ par une fonction affine puis d'intégrer cette fonction pour obtenir les approximations de ρ et de sa dérivée première.

Soit $0 < \delta < 1$. Introduisons r_δ^3 définie sur $\mathbb{R}$ par

$$r_\delta^3 : u \mapsto \begin{cases} 0 & \text{si } |u| \leqslant 1 - \delta/2 \\ \frac{6}{\delta^3}(u - (1-\delta/2))(u - (1+\delta/2)) & \text{si } 1 - \delta/2 < |u| < 1 + \delta/2 \\ 0 & \text{si } |u| \geqslant 1 + \delta/2 \end{cases}.$$

Soient r_δ^2 la primitive de r_δ^3 s'annulant en $1 + \delta/2$, r_δ^1 la primitive de r_δ^2 s'annulant en 0 et r_δ, primitive de r_δ^1 s'annulant en 0.

En posant $\tilde{\rho}_1 = r^3_{1/n^2}$, nous avons exhibé une suite de fonctions convexes et $\mathcal{C}^3$, convergeant uniformément vers ρ_1. Nous pouvons de plus montrer que $\int |\tilde{\rho}_1^{(3)}| \leqslant 12$, que

$$n\|\tilde{\rho}_1 - \rho_1\|_\infty \underset{n\to\infty}{\longrightarrow} 0, \tag{4.6}$$

$$n\|\tilde{\rho}_1' - \rho_1'\|_\infty \underset{n\to\infty}{\longrightarrow} 0, \tag{4.7}$$

$$\|\tilde{\rho}_1'' - \rho_1''\|_\infty \leqslant 1, \tag{4.8}$$

et que $\tilde{\rho}_1''$ et ρ_1'' ne diffèrent que sur deux intervalles de longueur respective $1/(n^2)$.

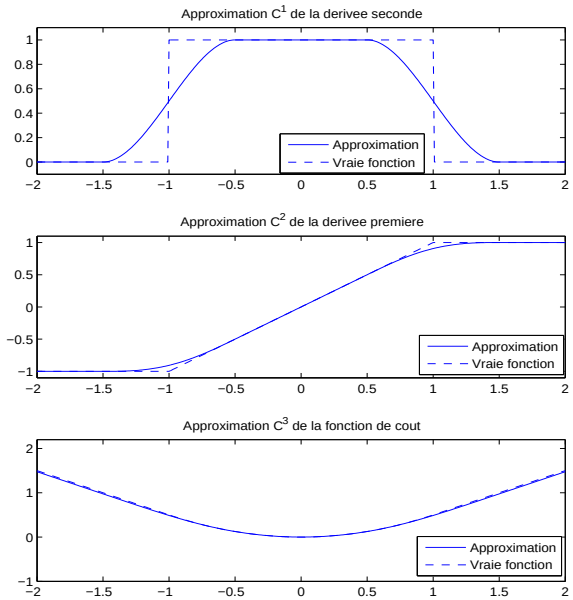

FIG. 4.1: Approximation des dérivées de la fonction de Huber ρ_1 avec $\delta = 1$

Convergence du critère

Nous souhaitons montrer dans le modèle (3.2) la proposition suivante :

Proposition 4.10.
Sous (A1) à (A4), pour tout $c > 0$,

$$\sup_{\{\|\beta-\beta_0\|\leqslant c v_n n^{-1/2}\}} v_n^2 \Big| \sum_{i=i_0}^n \left(\rho_\lambda(\theta_{0i}+\varepsilon_i - A_i^T(\beta-\beta_0)) - \rho_\lambda(\theta_{0i}+\varepsilon_i)\right) + \sum_{i=i_0}^n \rho'_\lambda(\theta_{0i}+\varepsilon_i)A_i^T(\beta-\beta_0) - n\frac{1}{2}(\beta-\beta_0)^T K_0(\beta-\beta_0)\Big| \xrightarrow{\mathbb{P}} 0.$$

Dans le modèle (4.5) la Proposition 4.10 s'écrit :

$$\sup_{\{\|\alpha\|\leqslant c\}} \left| \sum_{i=i_0}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)\right) + \sum_{i=1}^n \rho'_{\lambda/v_n}(d_i)R_i^T\alpha - \frac{1}{2}\alpha^T K_0''\alpha \right| \xrightarrow{\mathbb{P}} 0.$$

• Considérons la fonction :

$$\Delta(\alpha) := \sum_{i=1}^{n} \left(\tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i) + \tilde{\rho}'_{\lambda/v_n}(d_i) R_i^T\alpha \right).$$

A i fixé, la convexité de la fonction de coût $\tilde{\rho}_{\lambda/v_n}$ donne l'inégalité :

$$\left| \tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i) + \tilde{\rho}'_{\lambda/v_n}(d_i) R_i^T\alpha \right| \leqslant |\tilde{\rho}'_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}'_{\lambda/v_n}(d_i)||R_i^T\alpha|. \tag{4.9}$$

Cette inégalité permet de majorer la variance de $\Delta(\alpha)$ par :

$$Var(\Delta(\alpha)) \leqslant \sum_{i=1}^{n} \mathbb{E}\left[\left(\tilde{\rho}'_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}'_{\lambda/v_n}(d_i) \right)^2 \right] |R_i^T\alpha|^2.$$

Supposons que $\|\alpha\| \leqslant c$. Alors $\sup_{i=1,\dots,n} |R_i^T\alpha|$ tend vers 0 quand n tend vers l'infini et $\sum_{i=1}^{n} |R_i^T\alpha|^2$ est borné.

La fonction $\tilde{\rho}'_{\lambda/v_n}$ est 1-lipschitzienne, donc

$$\forall n \in \mathbb{N},\ \forall i = 1, \dots, n,\ \forall u \in \mathbb{R}^+,\ \mathbb{E}\left(\tilde{\rho}'_{\lambda/v_n}(d_i + u) - \tilde{\rho}'_{\lambda/v_n}(d_i) \right)^2 \leqslant u^2.$$

D'où

$$Var(\Delta(\alpha)) \leqslant \sum_{i=1}^{n} |R_i^T\alpha|^4 \leqslant \|\alpha\|^4 \sum_{i=1}^{n} \|R_i\|^4.$$

Comme α est supposé borné et comme $\sum_{i=1}^{n} |R_i|^4 \leqslant h \sum h_i = hp$ tend vers 0, ceci implique que $Var(\Delta(\alpha))$ tend vers 0.

Par l'inégalité de Bienaymé-Tchebychev, il en résulte que $|\Delta(\alpha) - \mathbb{E}\Delta(\alpha)|$ tend vers 0 en probabilité.

• Explicitons $\mathbb{E}\Delta(\alpha)$.

La fonction $\tilde{\rho}$ étant $\mathcal{C}^3$, le développement de Taylor avec reste intégral de $\tilde{\rho}_{\lambda/v_n}$ au voisinage de d_i s'écrit :

$$\begin{aligned} \tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i) + \tilde{\rho}'_{\lambda/v_n}(d_i) R_i^T\alpha - \frac{1}{2}\tilde{\rho}''_{\lambda/v_n}(d_i)\alpha^T R_i R_i^T \alpha \\ = -\int \tilde{\rho}^{(3)}_{\lambda/v_n}(t)(d_i - t)^3 \mathbb{1}_{d_i - R_i^T\alpha \leqslant t \leqslant d_i} dt/6. \end{aligned}$$

Comme $\int \left|\tilde{\rho}^{(3)}_{\lambda/v_n}(t)\right| dt \leqslant 12$,

$$\mathbb{E}\left|\sum_{i=1}^{n}\left(\tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i) + \tilde{\rho}'_{\lambda/v_n}(d_i)R_i^T\alpha - \frac{1}{2}\tilde{\rho}''_{\lambda/v_n}(d_i)\alpha^T R_i R_i^T\alpha\right)\right| \leqslant 2\|\alpha\|^3 \sum_{i=1}^{n}\|R_i\|^3.$$

Notons que $\sum_{i=1}^{n}\|R_i\|^3 \leqslant h^{1/2}\sum h_i = h^{1/2}p \to 0$. Ainsi, pour $\|\alpha\| \leqslant c$,

$$\mathbb{E}\Delta(\alpha) = \frac{1}{2}\alpha^T\tilde{K}''_n\alpha + o(1),$$

avec

$$\tilde{K}''_n = \sum_{i=i_0}^{n} R_i R_i^T \mathbb{E}\left[\tilde{\rho}''_{\lambda/v_n}(d_i)\right].$$

Or, $\tilde{K}''_n$ converge vers K''_0. En effet la différence $\|\tilde{K}''_n - K''_0\|$ peut se décomposer en :

$$\|\tilde{K}''_n - K''_0\| \leqslant \|\tilde{K}''_n - K''_n\| + \|K''_n - K''_0\|.$$

Le second terme tend vers 0 par l'hypothèse (A3). Le premier vaut :

$$\tilde{K}''_n - K''_n = \sum R_i^T \mathbb{E}(\tilde{\rho}''_{\lambda/v_n}(d_i) - \rho''_{\lambda/v_n}(d_i)).$$

Comme $\tilde{\rho}''_1$ et ρ''_1 ne diffèrent que sur deux intervalles de longueur respective $1/(n^2)$, $\mathbb{E}(\tilde{\rho}''_\lambda(d_i) - \rho''_\lambda(d_i)) \leqslant 2/(n^2)\|\tilde{\rho}''_\lambda - \rho''_\lambda\|_\infty\|f_\varepsilon\|_\infty$ avec f_ε densité de ε_i. Ceci nous donne l'inégalité suivante :

$$\|\tilde{K}''_n - K''_n\| \leqslant \frac{1}{n}h^{1/2}C,$$

avec C constante, donc comme h tend vers 0, nous en déduisons que $\tilde{K}''_n$ converge vers K''_0.

D'où : $\mathbb{E}\Delta(\alpha) = \frac{1}{2}\alpha^T K''_0\alpha + \circ(1)$. Donc, sous (H3'), si $\|\alpha\| \leqslant c$, alors la convergence en probabilité de $|\Delta(\alpha) - \mathbb{E}\Delta(\alpha)|$ vers 0 implique :

$$\left|\sum_{i=1}^{n}\left(\tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i) + \tilde{\rho}'_{\lambda/v_n}(d_i)R_i^T\alpha\right) - \frac{1}{2}\alpha^T K''_0\alpha\right| \xrightarrow{\mathbb{P}} 0.$$

Considérons $\tilde{D} := \sum_{i=1}^{n}\tilde{\rho}_{\lambda/v_n}(d_i - R_i^T\alpha) - \tilde{\rho}_{\lambda/v_n}(d_i)$ et $D := \sum_{i=1}^{n}\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)$. Alors $|D - \tilde{D}| \leqslant n\|\tilde{\rho}_{\lambda/v_n} - \rho_{\lambda/v_n}\|_\infty$, donc $D - \tilde{D}$ tend presque sûrement vers 0 d'après (4.6).

De même si $\tilde{B} = \sum_{i=1}^n \tilde{\rho}'_{\lambda/v_n}(d_i) R_i^T \alpha$ et $B = \sum_{i=1}^n \rho'_{\lambda/v_n}(d_i) R_i^T \alpha$, alors nous avons $|B - \tilde{B}| \leqslant n \|\tilde{\rho}'_{\lambda/v_n} - \rho'_{\lambda/v_n}\|_\infty \|\alpha\| h^{1/2}$. Pour $\|\alpha\| \leqslant c$, le point (4.7) implique que $B - \tilde{B}$ tend presque sûrement vers 0. Par conséquent, nous avons :

$$\left| \sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T \alpha) - \rho_{\lambda/v_n}(d_i) + \rho'_{\lambda/v_n}(d_i) R_i^T \alpha \right) - \frac{1}{2} \alpha^T K_0'' \alpha \right| \xrightarrow{\mathbb{P}} 0.$$

• Reste à montrer que la convergence est uniforme sur $\{\|\alpha\| \leqslant c\}$.

Les fonctions qui à α associent

$$\sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T \alpha) - \rho_{\lambda/v_n}(d_i) + \rho'_{\lambda/v_n}(d_i) R_i^T \alpha \right) \text{ et } \frac{1}{2} \alpha^T K_0'' \alpha$$

sont convexes et l'ensemble $\{\|\alpha\| \leqslant c\}$ est convexe, compact et indépendant de n. En appliquant la Proposition 4.7, il vient le résultat final :

$$\sup_{\{\|\alpha\| \leqslant c\}} \left| \sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T \alpha) - \rho_{\lambda/v_n}(d_i) + \rho'_{\lambda/v_n}(d_i) R_i^T \alpha \right) - \frac{1}{2} \alpha^T K_0'' \alpha \right| \xrightarrow{\mathbb{P}} 0. \tag{4.10}$$

Consistance

Nous souhaitons montrer que dans le modèle (3.2), nous avons

Proposition 4.11.

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \bigcirc_{\mathbb{P}}(v_n).$$

Dans le modèle (4.3), ceci équivaut à

$$\hat{\alpha}_n = \bigcirc_{\mathbb{P}}(1).$$

Soit $c_n \to \infty$. Montrons que $\mathbb{P}(\|\hat{\alpha}_n\| > c_n) \to 0$.

• Nous aimerions montrer qu'il existe une suite c'_n telle que $c'_n \to \infty$, $c'_n \leqslant c_n$ et

$$\sup_{\{\|\alpha\| \leqslant c'_n\}} \left| \sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T \alpha) - \rho_{\lambda/v_n}(d_i) + \rho'_{\lambda/v_n}(d_i) R_i^T \alpha \right) - \frac{1}{2} \alpha^T K_0'' \alpha \right| \xrightarrow{\mathbb{P}} 0.$$

Nous noterons $S_n(\alpha) := \left| \sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T \alpha) - \rho_{\lambda/v_n}(d_i) + \rho'_{\lambda/v_n}(d_i) R_i^T \alpha \right) - \frac{1}{2} \alpha^T K_0'' \alpha \right|$.

Soit $\varepsilon > 0$. Alors, il existe n_0 tel que $\forall n \geqslant n_0$, $1/n \leqslant \varepsilon$. En traduisant la convergence de la Proposition 4.10 avec $c = c_n$, il existe $\nu(n) \geqslant n$ tel que

$$\forall k \geqslant \nu(n), \quad \sup_{\{\|\alpha\| \leqslant c_n\}} S_k(\alpha) \leqslant 1/n.$$

Nous construisons alors c'_n de la façon suivante :

$$\forall k = \nu(n), \ldots, \nu(n+1) - 1, \ c'_k = \inf\{c_k, c_n\}$$

en choisissant $\nu(n)$ strictement croissante. Alors, c'_n tend vers l'infini car c_n tend vers l'infini, $c'_n \leqslant c_n$ pour tout n, et enfin,

$$\forall \varepsilon > 0, \exists n_1, \forall n \geqslant n_1, \quad \sup_{\{\|\alpha\| \leqslant c'_n\}} S_n(\alpha) \leqslant \varepsilon,$$

ce qui assure la convergence recherchée.

Comme $\mathbb{P}\left(\|\hat{\alpha}_n\| \geqslant c_n\right) \leqslant \mathbb{P}\left(\|\hat{\alpha}_n\| \geqslant c'_n\right)$, il suffit de montrer que $\mathbb{P}\left(\|\hat{\alpha}_n\| \geqslant c'_n\right) \to 0$.

- Supposons que $\|\alpha\| = c'_n$.

Alors,

$$\sum_{i=1}^{n} \left(\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)\right) = -\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i) R_i^T \alpha + \frac{1}{2}\alpha^T K_0'' \alpha + o_{\mathbb{P}}(1).$$

Or,

$$\|\frac{1}{2}\alpha^T K_0'' \alpha\| \geqslant \frac{1}{2}\underline{s}(K_0'')(c'_n)^2,$$

avec $\underline{s}(K_0'')$ plus petite des valeurs propres de K_0''. La matrice K_0'' étant supposé inversible, $\underline{s}(K_0'') > 0$. Ensuite, la suite v_n a été choisie de telle sorte que $\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i) R_i^T = \bigcirc_{\mathbb{P}}(1)$, il en découle que :

$$\left|\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i) R_i^T \alpha\right| = \bigcirc_{\mathbb{P}}(c'_n).$$

Il en résulte que la probabilité que la quantité

$$\sum_{i=1}^{n} \rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i) = -\sum \rho'_{\lambda/v_n}(d_i) R_i^T \alpha + \frac{1}{2}\alpha^T K_0'' \alpha + o_{\mathbb{P}}(1)$$

soit négative tend vers 0. Ceci étant vrai uniformément pour α tel que $\|\alpha\| = c'_n$, nous en déduisons que

$$\mathbb{P}\left(\inf_{\{\alpha,\ \|\alpha\| = c'_n\}} \sum_{i=1}^{n} \rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i) \leqslant 0\right) \to 0. \tag{4.11}$$

- Soit ensuite α tel que $\|\alpha\| \geqslant c'_n$. En posant $t = \frac{c'_n}{\|\alpha\|} \in]0;1]$ et $\alpha' = t\alpha$ et en écrivant $d - R\alpha' = (1-t)d_i + t(d - R\alpha)$, la convexité de ρ assure que si $\sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)\right) \geqslant 0$ alors

$$0 \leqslant \sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha') - \sum_{i=1}^n \rho_{\lambda/v_n}(d_i) \leqslant t\sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)\right).$$

D'où lorsque $\sum_{i=1}^n \left(\rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i)\right) \geqslant 0$,

$$\sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha') - \sum_{i=1}^n \rho_{\lambda/v_n}(d_i) \leqslant \sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \sum_{i=1}^n \rho_{\lambda/v_n}(d_i).$$

Ainsi, pour tout α tel que $\|\alpha\| \geqslant c'_n$, nous avons avec un probabilité tendant vers 1,

$$\sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i) \geqslant \inf_{\{\alpha',\ \|\alpha'\| = c'_n\}} \sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha') - \rho_{\lambda/v_n}(d_i).$$

Ceci implique que :

$$\mathbb{P}\left(\inf_{\{\alpha,\ \|\alpha\| \geqslant c'_n\}} \sum_{i=1}^n \rho_{\lambda/v_n}(d_i - R_i^T\alpha) - \rho_{\lambda/v_n}(d_i) \leqslant 0\right) \to 0, \tag{4.12}$$

soit :

$$\mathbb{P}\left(\inf_{\{\alpha,\ \|\alpha\| \geqslant c'_n\}} J_n(\alpha) \leqslant J_n(0)\right) \to 0.$$

Etant donné que $\hat{\alpha}_n$ réalise le minimum de J_n, ceci impose

$$\mathbb{P}\left(\|\hat{\alpha}_n\| \geqslant c'_n\right) \to 0,$$

ce qui achève la démonstration.

Ecriture au second ordre

Dans le modele (3.2), nous aimerions montrer la proposition suivante :

Proposition 4.12. *Sous (A2) à (A4),*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = K_0^{-1}\left(\frac{1}{\sqrt{n}} \sum_{i=i_0}^n \rho'_{\lambda/v_n}(d_i)A_i\right) + o_{\mathbb{P}}(v_n).$$

Dans le modèle (4.5), cette écriture au second ordre devient :

$$\hat{\alpha}_n = K_0''^{-1} \left(\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i) R_i \right) + o_{\mathbb{P}}(1).$$

Rappelons tout d'abord le résultat de Rockafellar (1970) :

Proposition 4.13. *Soit $\mathcal{C}$ un ouvert convexe. Soit f_n une famille de fonctions convexes différentiables et soit f une fonction convexe différentiable. Si f_n converge simplement vers f sur $\mathcal{C}$, alors ∇f_n converge simplement vers ∇f sur $\mathcal{C}$, et de plus, la convergence est uniforme sur tout compact inclus dans $\mathcal{C}$.*

De même que pour la Proposition 4.7, ce résultat peut s'appliquer à la convergence en probabilité (en utilisant le Lemme 4.9).

En appliquant la proposition ci-dessus au résultat de la Proposition 4.10, nous obtenons que, pour tout $c > 0$,

$$\sup_{\|\alpha\| \leqslant c} \left| \sum_{i=1}^{n} \left(\rho'_{\lambda/v_n}(d_i - R_i^T \alpha) R_i + \rho'_{\lambda/v_n}(d_i) R_i \right) - K_0'' \alpha \right| \xrightarrow{\mathbb{P}} 0.$$

En particulier, comme $\hat{\alpha}_n = \bigcirc_{\mathbb{P}}(1)$ d'après ce qui précède, il en résulte que

$$\left| \sum_{i=1}^{n} \left(\rho'_{\lambda/v_n}(d_i - R_i^T \hat{\alpha}_n) R_i + \rho'_{\lambda/v_n}(d_i) R_i \right) - K_0'' \hat{\alpha}_n \right| \xrightarrow{\mathbb{P}} 0. \tag{4.13}$$

Par définition de $\hat{\alpha}_n$, $\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i - R_i^T \hat{\alpha}_n) R_i = 0$. Alors la convergence (4.13) permet de conclure que :

$$\hat{\alpha}_n = K_0''^{-1} \left(\sum_{i=1}^{n} \rho'_{\lambda/v_n}(d_i) R_i \right) + o_{\mathbb{P}}(1),$$

ce qui est le résultat annoncé.

La normalité asymptotique sous conditions supplémentaires donnée eu Théorème 4.1 en découle de manière immédiate au vu de la Proposition 4.6.

4.4.2 Consistance de $\hat{\beta}_n$ avec la pénalité quadratique

Le terme déterminant dans la vitesse de l'estimateur, de même que pour le seuillage doux, est

$$\begin{aligned} B: &= \frac{1}{n}\sum_{i=i_0}^{n} \rho'_{\lambda/v_n}(\theta_i+\varepsilon_i)A_i \\ &= \frac{1}{n}\sum_{i=i_0}^{n} A_i(\theta_i+\varepsilon_i)\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}. \end{aligned}$$

Nous avons le contrôle suivant :

Proposition 4.14. *Sous (A1) à (A4), si $\lambda \to 0$ et $\sqrt{n}\lambda \to \infty$, alors,*

$$B = \bigcirc_{\mathbb{P}}(\lambda^{1/2}).$$

Si de plus, pour tout $i = 1, \ldots, n$, $j = 1, \ldots, p$, nous considérons que $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j appartenant à $\Lambda^s(M_j)$ et $\xi_{i,j}$ issu d'une loi centrée de variance finie, alors

$$B = \bigcirc_{\mathbb{P}}(\lambda).$$

Si les fonctions g_j sont des fonctions polynomiales de degré au plus égal au nombre de momants nuls de l'ondelette, alors

$$B = \bigcirc_{\mathbb{P}}(1/\sqrt{n}).$$

Démonstration. Le terme B peut s'écrire :

$$B = \frac{1}{n}\sum_{i=i_0}^{n} \frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}} A_i\theta_i + \frac{1}{n}\sum_{i=i_0}^{n} \frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}} A_i\varepsilon_i.$$

Soit $j \in \{1, \ldots, p\}$ fixé. Nous avons

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=i_0}^{n} \frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}} A_{i,j}\varepsilon_i\right] = 0$$

et

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=i_0}^{n} \frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}} A_{i,j}\varepsilon_i\right)^2 \leqslant \frac{\sigma^2}{n}\left(\frac{1}{n}\sum_{i=i_0}^{n} A_{i,j}^2\right).$$

Nous obtenons donc que $\mathbb{E}\left(\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}A_{i,j}\varepsilon_i\right)^2 = \bigcirc(1/n)$, via le Lemme 4.5. D'après l'inégalité de Bienaymé-Tchebychev, il en résulte que

$$\frac{1}{\sqrt{n}}\sum_{i=i_0}^{n}\frac{1}{1+2\lambda 2^{2\log_2(i)s}}A_{i,j}\varepsilon_i = \bigcirc_{\mathbb{P}}(n^{-1/2}).$$

Afin de majorer le deuxième terme, nous distinguons selon qu'une hypothèse sur la structure des covariable est émise ou pas. Dans le cas général, la norme de

$$\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}A_{i,j}\theta_i$$

peut être majorée à l'aide de l'inégalité de Cauchy-Schwarz par

$$\left(\frac{1}{n}\sum_{i=i_0}^{n}\|A_i\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=i_0}^{n}\frac{4\lambda^2 2^{4\log_2(i)s}}{(1+2\lambda 2^{2\log_2(i)s})^2}\theta_i^2\right)^{1/2}.$$

D'après le Lemme 4.5, $\frac{1}{n}\sum_{i=i_0}^{n}\|A_i\|^2$ est borné. Comme f appartient à un espace $\Lambda^s(M)$, nous en déduisons la majoration, à une constante multiplicative près, par $\lambda^{1/2}$, ce qui était le résultat annoncé.

Supposons maintenant que la matrice de régression a la forme suggérée dans le cadre du Théorème 4.2 : pour tout couple (i,j), $i=1,\dots,n$ et $j=1,\dots,p$, nous considérons que $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j appartenant à $\Lambda^s(M_j)$ et $\xi_{i,j}$ issu d'une loi centrée de variance finie. Soit $j \in \{1,\dots,p\}$ fixé. Alors,

$$\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}A_{i,j}\theta_i = \frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}\eta_i^j\theta_i + \frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}w_i^j\theta_i, \tag{4.14}$$

avec $(\eta_i^j)_{i=1,\dots,n}$ réalisations d'une variable aléatoire centrée de variance finie et w_i^j coefficient d'ondelette de g_j d'ordre (l,k) où $2^l + k = i$.

Avec cette écriture, nous pouvons montrer de même que précédemment que :

$$\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}\eta_i^j\theta_i = \bigcirc_{\mathbb{P}}(n^{-1/2}). \tag{4.15}$$

Quant au terme

$$\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}w_i^j\theta_i,$$

sa norme peut être majorée à l'aide de l'inégalité de Cauchy-Schwarz par

$$\left(\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}w_i^{j2}\right)^{1/2}\left(\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}\theta_i^2\right)^{1/2}.$$

En utilisant l'hypothèse que les fonctions f et $(g_j)_{j=1,\dots,p}$ appartiennent à des espaces de Hölder de régularité $s > 1/4$, nous en déduisons que

$$\left\|\frac{1}{n}\sum_{i=i_0}^{n}\frac{2\lambda 2^{2\log_2(i)s}}{1+2\lambda 2^{2\log_2(i)s}}w_i^j\theta_i\right\| \leqslant \lambda,$$

ce qui achève la démonstration.

Le dernier cas considéré est celui où les covariables vérifient une relation de la forme $X_{i,j} = g_j(t_i) + \xi_{i,j}$ avec g_j appartenant dorénavant à l'ensemble des fonctions polynomiales de degré inférieur ou égale au nombre de moments nuls de l'ondelette. Les variables $\xi_{i,j}$ étant toujours supposées issues d'une loi centrée de variance finie. Alors, en conservant les notations précédentes, les coefficients d'ondelettes w_i^j des fonctions g_j sont nuls. D'où le résultat énoncé dans le théorème en utilisant les équations (4.14) et (4.15).

□

Nous pouvons effectuer des transformées de variables similaires à celles décrites pour le seuillage doux –en normalisant non par $v_n = \lambda^{1/2}\sqrt{n}$, $v_n = \lambda\sqrt{n}$ ou $v_n = 1$ selon les cas– puis reprendre intégralement la preuve pour le seuillage doux (sans besoin de passer par l'approximation). Ceci est possible car les fonctions de coût associées à la pénalité de Fadili et Bullmore (2005) sont convexes, trois fois continûment dérivables et 1-lipschitziennes. Mais nous avons ici le résultat de manière immédiate étant donné que l'estimateur s'écrit de manière explicite.

L'estimateur $\hat{\beta}_n$ s'écrit

$$\hat{\beta}_n = (A^T TA)^{-1}A^T Tz$$

avec $T = diag\left(\frac{2\lambda 2^{2[\log_2(i)]s}}{1+2\lambda 2^{2\log_2(i)s}}\right)_{i=1,\dots,n}$. Par conséquent,

$$\hat{\beta}_n - \beta_0 = \left(\frac{1}{n}A^T TA\right)^{-1} B.$$

Or par hypothèse, $(A^T TA)^{-1}$ converge vers une matrice inversible, donc nous avons le résultat énoncé dans le Théorème 4.2.

4.4.3 Estimation de la partie fonctionnelle

Nous considérons ici le modèle initial (3.2) et non les modèles (4.3) ou (4.5) utilisés pour les preuves précédentes.

D'après Parseval, $\|\hat{f}_n - f\|_2 \sim \frac{1}{n}\|\hat{\theta}_n - \theta_0\|$. Nous pouvons ensuite décomposer cette quantité : $\frac{1}{n}\|\hat{\theta}_n - \theta_0\| \leqslant \frac{1}{n}\|\hat{\theta}_n - \tilde{\theta}_n\| + \frac{1}{n}\|\tilde{\theta}_n - \theta_0\|$ avec

$$\tilde{\theta}_{i,n} = \begin{cases} z_i - A_i^T\beta_0 & \text{si } i < i_0 \\ \gamma(z_i - A_i^T\beta_0, \lambda) & \text{si } i \geqslant i_0 \end{cases},$$

où $\gamma(., \lambda)$ désigne la fonction de seuillage de seuil λ. D'après les résultats Donoho (1992), il existe une constante C telle que $\mathbb{E}\frac{1}{n}\|\tilde{\theta}_n - \theta_0\| \leqslant C\left(\frac{\log(n)}{n}\right)^{\frac{s}{1+2s}}$. Pour le seuillage associé à la pénalité quadratique, cette inégalité est donnée dans Antoniadis (1996). Ayant la convergence L^2, il en découle la convergence en probabilité. Montrons ensuite que le terme $\frac{1}{n}\|\hat{\theta}_n - \tilde{\theta}_n\|$ est négligeable, ce qui achèvera la démonstration.

Nous avons la majoration $\frac{1}{n}\|\hat{\theta}_n - \tilde{\theta}_n\| \leqslant \frac{1}{n}\|A\|\|\hat{\beta}_n - \hat{\beta}_0\| + 2\frac{\lambda}{n}$. Les hypothèses (A2) et (A3) assurent que $\frac{1}{\sqrt{n}}\|A\| = \left(\frac{1}{n}\sum\|A_i\|^2\right)^{1/2}$ est borné. Pour le seuillage doux, nous avons $\|\hat{\beta}_n - \hat{\beta}_0\| = \bigcirc_{\mathbb{P}}(v_n)$, avec $v_n \leqslant (\log(n)/n)^{s/(1+2s)}$. D'où, $\frac{1}{n}\|\hat{\theta}_n - \tilde{\theta}_n\| = \bigcirc_{\mathbb{P}}(\frac{v_n}{\sqrt{n}}) = \bigcirc_{\mathbb{P}}((\frac{\log(n)}{n})^{s/(1+2s)})$. Nous obtenons bien le résultat souhaité.

Pour la pénalité quadratique, sous les hypothèses du Théorème 4.2, l'estimateur $\hat{\beta}_n$ vérifie $\|\hat{\beta}_n - \hat{\beta}_0\| = \bigcirc_{\mathbb{P}}(\lambda^{1/2})$. D'où, $\frac{1}{n}\|\hat{\theta}_n - \tilde{\theta}_n\| = \bigcirc_{\mathbb{P}}((\lambda/n)^{1/2})$. Avec $\lambda = n^{-2s/(1+2s)}$, ce terme est bien négligeable.

Conclusion

Ce chapitre présente les résultats de convergence obtenus pour deux procédures décrites au Chapitre 3. Pour le schéma de pénalisation menant au seuillage doux dans l'estimation par ondelettes de la partie fonctionnelle, les estimateurs de la partie fonctionnelle et de la partie paramétrique du modèle s'avèrent avoir une vitesse quasi-minimax en probabilité. Cependant, pour obtenir ces résultats nous avons supposé que la structure des covariables était similaire à celle suggérée par Rice (1986). Une écriture asymptotique au second ordre de l'estimateur du paramètre de régression est par ailleurs donnée. Le

seuillage associé à une norme de Sobolev vérifie l'optimalité au sens minimax sous cette hypothèse de structure des covariables. Mais contrairement à la procédure du seuillage doux, cette démarche n'est pas adaptative par rapport à la régularité de la fonction. Ceci risque alors d'entraîner des biais dans la sélection du seuil par validation croisée. Cette étude des comportements théoriques est complétée en Chapitre 6 d'une comparaison numérique sur des simulations, et est généralisée au cas d'un plan non équidistant pour le seuillage fort au Chapitre 7. Notons enfin qu'un schéma d'estimation de la variance du bruit est aussi proposé dans ce chapitre.

CHAPITRE 5

QUELQUES ALGORITHMES

Sommaire

5.1 Estimation conjointe des paramètres **87**
5.1.1 Backfitting 88
5.1.2 Deux algorithmes pour le seuillage doux 89
5.2 Approche semi-quadratique **91**
5.2.1 ARTUR, ou les poids modifiés 93
5.2.2 LEGEND, ou les résidus modifiés 94

Grâce à la procédure d'estimation présentée, nous pouvons estimer le vecteur des paramètres de régression β_0 indépendamment de la partie fonctionnelle, puis en déduire une estimation de la partie fonctionnelle. Cependant, l'estimateur $\hat{\beta}_n$ ne peut être donné explicitement, mais doit être obtenu par des algorithmes itératifs. L'avantage d'aboutir à un M-estimateur usuel est que des techniques de programmation de cet estimateur existent. La première partie du chapitre propose des algorithmes de calcul estimant conjointement les parties linéaire et non paramétrique du modèle, notamment le *Backfitting*. Dans la seconde section, nous présenterons les algorithmes résultant de cette estimation en deux étapes.

5.1 Estimation conjointe des paramètres

Dans cette section sont présentés trois algorithmes. Ces algorithmes calculent simultanément les estimations du vecteur des paramètres et de la fonction dans un modèle partiellement linéaire, les estimateurs étant définis par minimisation d'un critère des

moindres carrés pénalisés tel que cela a été vu au Chapitre 3. Le premier algorithme, le *Backfitting* est un algorithme usuel dans un modèle partiellement linéaire. Une version adaptée au contexte de notre étude est proposée. Ensuite, nous exposons deux algorithmes qui sont propres au seuillages doux.

5.1.1 Backfitting

Chang et Qu (2004) pour la pénalité l^1 et Fadili et Bullmore (2005) pour une large gamme de pénalités proposent d'obtenir les vecteurs de paramètres $\hat{\beta}_n$ et $\hat{\theta}_n$ à l'aide d'un algorithme de *Backfitting*. L'idée d'un tel algorithme est de résoudre itérativement les équations du premier ordre du problème de minimisation conjointement aux deux vecteurs. Avec les notations que nous avons introduites, l'estimation par *Backfitting* consiste à appliquer dans un premier temps la transformée en ondelettes aux observations, à l'aide de l'algorithme pyramidal de Mallat (1989), puis à appliquer l'algorithme suivant :

1. Initialisation des paramètres : $m = 0$ et $\theta_i^{(0)} = \begin{cases} z_i & \text{si } i < i_0, \\ 0 & \text{si } i \geqslant i_0. \end{cases}$

2. Tant que $m < M$ ou $\frac{\|\beta^{(m+1)} - \beta^{(m)}\|_2}{\|\beta^{(m)}\|_2} < \delta$,

$$\begin{aligned} \hat{\beta}^{(m+1)} &= (A^T A)^{-1} A^T (z - \hat{\theta}^{(m)}) \\ \hat{\theta}_i^{(m+1)} &= \begin{cases} z_i - A_i^T \hat{\beta}^{(m+1)} & \text{si } i < i_0, \\ \gamma_\lambda(z_i - A_i^T \hat{\beta}^{(m+1)}) & \text{si } i \geqslant i_0, \end{cases} \end{aligned}$$

   avec γ_λ fonction de seuillage des coefficients d'ondelettes de seuil λ associée à la pénalité donnée dans les moindres carrés pénalisés, tel que cela fut décrit au Chapitre 3.

L'estimation $\hat{f}_n$ de la partie fonctionnelle est ensuite obtenue en appliquant la transformée en ondelettes inverse au vecteur $\hat{\theta}_n$. Un nombre maximal d'itérations M est fixé ici afin de s'assurer de l'arrêt de l'algorithme. De plus, nous considérons que l'algorithme a convergé lorsque la vitesse de convergence relative entre deux itérations successives dans le calcul de $\hat{\beta}_n$ est inférieure à un seuil δ donné.

Nous renvoyons au livre de Hardle *et al.* (2000) pour des références et des détails supplémentaires sur le *Backfitting*.

Fadili et Bullmore (2005) appliquent un algorithme légérement différent de celui proposé ici. En effet, les auteurs font l'hypothèse que la matrice de covariables X admet

une représentation en fonction d'échelle creuse. Par conséquent, les itérations ci-dessus deviennent :

$$\begin{aligned}\hat{\beta}^{(m+1)} &= (A_w^T A_w)^{-1} A_w^T (z_w - \hat{\theta}_w^{(m)}) \\ \hat{\theta}_{w\ i}^{(m+1)} &= \gamma_\lambda (z_{w\ i} - A_{w\ i}^T \hat{\beta}^{(m+1)}),\end{aligned}$$

où l'indice w désigne le coefficient d'ondelettes. Les coefficients d'échelle $\hat{\theta}_i$ valent z_i. En l'absence de critères d'arrêt, Fadili et Bullmore (2005) montrent que cet algorithme est convergent lorsque la pénalité Pen vérifie les hypothèses du Théorème 3.1 et lorsque de plus la fonction qui à θ associe $\theta + \lambda Pen'(\theta)$ est unimodale sur $(0, +\infty)$.

La raison pour laquelle nous avons considéré l'algorithme de Fadili et Bullmore (2005) est que le temps de calcul est réduit lorsque seuls les coefficients d'ondelettes sont considérés. En pratique les covariables de régression seront centrées avant d'exécuter cet algorithme.

5.1.2 Deux algorithmes pour le seuillage doux

Nous présentons deux algorithmes pouvant être appliqués dans le cadre du seuillage doux. Le premier est issu d'une première version de l'article de Fadili et Bullmore (2005) et le deuxième est proposé par Fuchs (1999).

Algorithme *BCR*

Pour une pénalité l^1 sur les coefficients d'ondelettes, Fadili et Bullmore proposent un autre algorithme. Les auteurs s'inspirant de l'algorithme BCR développé dans le cadre non paramétrique, nous l'appelerons BCR par la suite.

1. Initialisation des paramètres : $\beta^{(0)} = (A^T A)^{-1} A^T z$ et $l^{(0)} = \max_{i=1,\dots,n} |\theta_i|$.
2. Pour $m = 0, \dots, M-1$, les étapes a et b ci-dessous sont répétées K fois :
   a) Calculer une estimation des coefficients d'ondelettes de la partie fonctionnelle de la manière suivante :

$$\hat{\theta}_i^{(m+k+1)} = \begin{cases} z_i - A_i^T \hat{\beta}^{(m+k)} & \text{si } i < i_0, \\ \gamma_{l^{(m)}}(z_i - A_i^T \hat{\beta}^{(m+k)}) & \text{si } i \geqslant i_0, \end{cases}$$

   avec γ_l fonction de seuillage doux de seuil l.

b) Estimer β par $\hat{\beta} =^{(m+k+1)} (A^T A)^{-1} A^T (z - \hat{\theta}^{(m+k)})$.

Actualiser le seuil $l^{(m)}$: $l^{(m+1)} = l^{(m)} - (l^{(m)} - \lambda)/M$.

Algorithme quadratique de Fuchs

Une autre méthode d'estimation conjointe des paramètres a été implémentée dans le cas d'une pénalité l^1, correpondant au seuillage doux. Il s'agit du programme quadratique proposé par Fuchs (1999). Ce dernier remarque que la minimisation du critère (3.5) est équivalent à la résolution du programme :

$$\begin{array}{lll} \max\limits_{\beta,\theta^+,\theta^-} & \frac{1}{2}\sum_{i=1}^n (z_i - A_i\beta + \theta_i^- - \theta_i^+)^2 + \sum_{i=i_0}^n (\theta_i^+ + \theta_i^-) & \\ s.c. & \theta_i^+ \geqslant 0, & \text{pour } i = 1, \ldots, n, \\ & \theta_i^- \geqslant 0, & \text{pour } i = 1, \ldots, n, \end{array}$$

avec $\theta_i^+ = \max(\theta_i, 0)$ et $\theta_i^- = \max(-\theta_i, 0)$. Ce programme peut s'écrire sous la forme :

$$\begin{array}{ll} \max\limits_{\beta,T^+,T^-} & \frac{1}{2}\begin{pmatrix}\beta^T & T^{+T} & T^{-T}\end{pmatrix} H \begin{pmatrix}\beta \\ T^+ \\ T^-\end{pmatrix} + h^T \begin{pmatrix}\beta \\ T^+ \\ T^-\end{pmatrix} \\ s.c. & -T^+ \leqslant 0, \\ & -T^- \leqslant -z, \end{array}$$

avec $T_i^+ = \theta_i^+$ et $T_i^- = \theta_i^- + z_i$,

$$H = \begin{pmatrix} A^T A & A^T & -A^T \\ A^T & I_n & -I_n \\ -A^T & -I_n & I_n \end{pmatrix}$$

et $h^T = \begin{pmatrix}0_p & 0_{i_0-1} & \mathbb{1}_{n-i_0-1} & 0_{i_0-1} & \mathbb{1}_{n-i_0-1}\end{pmatrix}$. Pour le résoudre, nous pouvons utiliser les fonctions de résolution de programmes quadratiques sous contraintes linéaires accessibles sous les logiciels de programmation. Notamment, nous avons utilisé la fonction *quadprog* de MATLAB. Les valeurs initiales des paramètres dans l'algorithme sont fixées à $\beta^{(0)} = (A^T A)^{-1} A^T Z$, soit l'estimateur des moindres carrés, $T^+ = 0$ et $T^- = 0$.

Remarquons aussi que Fuchs (1999) établit l'équivalence entre l'estimation par critère des moindres carrés pénalisés avec une pénalité l^1 dans un modèle linéaire avec présence de valeurs aberrantes et l'estimation par le critère de Huber et le seuillage doux dans

un modèle linéaire. Il propose alors trois options de programmation : le *Backfitting*, la résolution du programme présenté ici ou IRLS (ou LEGEND), qui est présenté en section suivante.

5.2 Approche semi-quadratique

Résumons tout d'abord les étapes de l'estimation telle qu'elle fut présentée au Chapitre 3 :

1. Transformée discrète en ondelettes à l'aide de l'algorithme pyramidal de Mallat (1989) ;
2. Estimation de β_0 par résolution d'un problème de minimisation de la forme :

$$\hat{\beta}_n = \underset{\beta}{argmin} \sum_{i=2^{j_0}+1}^{n} \rho_\lambda(z_i - A_i\beta);$$

3. Estimation de θ_0 par seuillage de $Z - A\hat{\beta}_n$:

$$\hat{\theta}_{i,n} = \begin{cases} z_i - A_i\hat{\beta}_n & \text{si } i \leqslant 2^{j_0} \\ \gamma(z_i - A_i\hat{\beta}_n, \lambda) & \text{sinon,} \end{cases}$$

   avec $\gamma(., \lambda)$ fonction de seuillage de seuil λ.
4. Transformée inverse en ondelettes (algorithme pyramidal de Mallat (1989)) de $\hat{\theta}_n$ pour obtenir $\hat{f}_n$.

Le but de cette section est de proposer un algorithme permettant de résoudre l'étape 2 ci-dessus. L'estimateur $\hat{\beta}_n$ s'obtient par minimisation d'un critère de la forme

$$J(\beta) = \sum_{i=1}^{n} \rho\left(\frac{y_i - X_i^T\beta}{\sigma}\right).$$

Le critère que nous souhaitons minimiser n'étant pas quadratique, l'idée est de se ramener à une forme quadratique du problème, car alors des techniques de recherche linéaires de minima existent. Deux méthodes furent explorées pour se ramener à un problème d'optimisation proche d'un problème quadratique.

La première approche est celle proposée par Huber (1981) qui consiste à approximer le critère à l'aide d'une famille de polynômes. Les deux algorithmes, ARTUR et LEGEND,

mis en œuvre dans ce travail peuvent être obtenus par ce principe (cf. Huber (1981)) mais nous verrons leur fondement du point de vue de la seconde approche qu'est la théorie semi-quadratique.

Cette méthode consiste à introduire un nouveau critère K ayant le même minimum global que le critère initial. Le principe est de faire en sorte que la fonction K, fonction de β et d'une variable auxiliaire, soit telle que J et K admettent le même minimum et que ce minimum soit atteint pour une même valeur de β. Dans le domaine de la théorie semi-quadratique, les algorithmes ARTUR et LEGEND peuvent être notamment trouvés dans Dahyot et Kokaram (2004), Dahyot *et al.* (2000), Dahyot *et al.* (2004) ou Vik (2004). Ce dernier montre le lien entre ces algorithmes et l'approche de Huber (1981).

Soit $J(\beta) = \sum_{i=1}^{n} \rho\left(\frac{y_i - X_i^T\beta}{\sigma}\right)$ le critère que l'on souhaite minimiser. Nous construisons un critère K, appelé énergie augmentée, vérifiant :

(i) il existe c tel que $\forall\beta,\ J(\beta) = \min_c K(\beta, c)$,

(ii) il existe $\breve{\rho}$ tel que $\forall\beta,\ \forall c,\ K(\beta, c) = \sum_{i=1}^{n} \breve{\rho}(y_i, X_i, \beta, c)$.

Le principe est alors de résoudre le problème de minimisation de l'énergie augmentée de manière itérative. Si $\beta^{(m)}$ et $c^{(m)}$ sont les valeurs obtenues après m itérations, l'étape $m+1$ de l'algorithme consiste alors à actualiser les valeurs par :

$$\beta^{(m+1)} = \underset{\beta}{argmin}\, K(\beta, c^{(m)}) \tag{5.1}$$

$$c^{(m+1)} = \underset{c}{argmin}\, K(\beta^{(m+1)}, c) \tag{5.2}$$

Toute la difficulté de la théorie semi-quadratique réside dans la capacité à trouver une énergie augmentée K facile à minimiser.

La paternité de la construction des deux algorithmes présentés ci-après est attribuée à Geman et Reynolds (1992) et Geman et Yang (1995). L'idée est de chercher l'énergie auxiliaire K décrite ci-dessus sous la forme

$$K(\beta, c) = \sum_{i=1}^{n} Q(g_i^T\beta, c_i) + \psi(c_i),$$

avec Q forme quadratique. Deux formes de Q ont été proposées : multiplicative $Q(t, s) = t^2 s/2$ –qui mène à ARTUR– ou additive $Q(t, s) = (t - s)^2$ –qui mène à LEGEND–.

Remarque : Le principe des algorithmes semi-quadratiques peut s'appliquer dans un problème des moindres carrés avec une pénalité convexe :

$$\min_{\beta} J(\beta) = \|Y - X\beta\|^2 + \gamma \sum_{i=1}^{n} \varphi(g_i^T \beta).$$

Nous renvoyons à Nikolova et Ng (2005) pour une analyse des algorithmes dans ce contexte.

5.2.1 ARTUR, ou les poids modifiés

L'algorithme exposé ici se rencontre sous les termes d'ARTUR dans la littérature semi-quadratique ou de *Iterative Reweighted Least Squares* (IRLS) dans la théorie robuste. Le théorème de Geman et Reynolds (1992) permet de construire une énergie augmentée de la forme

$$K(\beta, c) = \sum_{i=1}^{n} c_i (y_i - X_i^T \beta)^2 + \Psi(c).$$

Théorème 5.1.
Supposons que la fonction de coût ρ soit symétrique, croissante sur $\mathbb{R}^+$ et que la fonction c définie sur $\mathbb{R}$ par $c(t) = \rho'(t)/(2t)$ existe, soit décroissante sur $\mathbb{R}^+$ et bornée. Supposons de plus que $\lim_{t\to\infty} b(t) = 0$ et que $\lim_{t\to 0} b(t) = 1$.
Alors il existe une fonction $\psi : [0,1] \mapsto [0,1]$ strictement convexe et décroissante telle que pour tout $t \in \mathbb{R}$, $\rho(t) = \inf_c ct^2 + \psi(c)$. De plus, pour $t \leqslant 0$, l'infimum est atteint en une valeur unique $c = \rho'(t)/(2t)$.

Dans le critère $K(\beta, c) = \sum_{i=1}^{n} c_i (y_i - X_i^T \beta)^2 + \Psi(c)$ obtenu, la variable auxiliaire c correspond à une pondération sur les résidus, d'où l'appellation IRLS. Intuitivement, les pondérations sur les résidus élevés tendent à les exclure de l'estimation.

D'après le théorème ci-dessus, à β fixé le minimum est atteint pour $c_i = \frac{\rho'(r_i)}{r_i}$ où r_i est le $i^{ème}$ résidu $r_i = y_i - X_i^T \beta$, et en ce point la fonction Ψ vaut $\rho(r_i) - \rho'(r_i) r_i / 2$. Il en résulte que ARTUR consiste à réévaluer les valeurs à l'étape $m+1$ de la façon suivante :

$$\begin{cases} r_i^{(m)} & = & z_i - A_i^T \beta^{(m)} \\ c_i^{(m+1)} & = & \frac{\rho'_\lambda(2r_i^{(m)})}{2r_i^{(m)}}, \qquad \forall i \in \{1, \ldots, n\} \\ \beta^{(m+1)} & = & (A^T c^{(m+1)} A)^{-1} A^T c^{(m+1)} Z \end{cases}$$

5.2.2 LEGEND, ou les résidus modifiés

LEGEND, ou *Iterative Modified Residuals* (IMR), est légèrement différent ; la variable auxiliaire ne joue pas le rôle d'une pondération mais réduit les résidus élevés par soustraction. L'existence de l'énergie augmentée est démontrée à l'aide du deuxième théorème de Geman et Reynolds (1992).

Théorème 5.2.
Sous les conditions du Théorème 5.1, il existe une fonction $\xi : [0,\infty) \mapsto [0,\infty)$ *telle que pour tout* $t \in \mathbb{R}$, $\rho(t) = \inf_c (c-t)^2 + \xi(c)$. *De plus, pour* $t \leqslant 0$, *l'infimum est atteint en une valeur unique* $c = t(1 - \rho'(t)/(2t))$.

Le critère obtenu est de la forme

$$K(\beta, c) = \sum_{i=1}^{n} (y_i - X_i^T \beta - c_i)^2 + \xi(c).$$

A β fixé, le minimum est atteint pour $c_i = r_i \left(1 - \frac{\rho'(r_i)}{2r_i}\right)$ où r_i est le $i^{ème}$ résidu $r_i = y_i - X_i^T \beta$, et qu'en ce point la fonction ξ vaut $\rho(r_i) - \rho'(r_i)^2/4$.

Avec des notations similaires à la sous-section précédente, l'itération $m+1$ est effectuée par :

$$\left\{ \begin{array}{rcl} r^{(m)} & = & Z - A\beta^{(m)} \\ c_i^{(m+1)} & = & r_i^{(m)} \left(1 - \frac{\rho'_\lambda(2r_i^{(m)})}{2r_i^{(m)}}\right) \qquad \forall i \in \{1, \ldots, n\} \\ \beta^{(m+1)} & = & (A^T A)^{-1} A^T (Z - c^{(m+1)}) \end{array} \right.$$

L'un des intérêts de cette méthode est que les itérations (5.1) peuvent être données explicitement. D'où la facilité de programmation d'ARTUR et LEGEND. Nikolova et Ng (2005) prouvent que le risque obtenu par une forme multiplicative est toujours plus faible que celui obtenu par une forme additive, mais l'analyse des coûts des algorithmes montre que la forme additive permet un coût moindre. La cause de cette différence de coût réside dans la nécessité d'inverser une matrice à chaque itération dans l'algorithme résultant de la forme multiplicative. En pratique, Vik (2004) et Dahyot et Kokaram (2004) observent en effet que si ARTUR converge en un nombre d'itérations plus faible que LEGEND, il est néanmoins plus lent car chaque itération est plus coûteuse.

Remarque : Huber (1981) réalise un parallèle entre les fonctions de coûts et l'a priori que l'on a sur les résidus. Dahyot *et al.* (2004) et Vik (2004) reprennent cette optique en

établissant quelles sont les densités a priori que l'on attribue aux résidus lorsque l'on choisit les algorithmes ARTUR ou LEGEND.

De même que précédemment, afin de réduire les temps de calcul, un critère d'arrêt a par ailleurs été introduit dans ARTUR et LEGEND : plutôt que de fixer arbitrairement le nombre d'itérations, nous considèrons que les algorithmes sont stabilisés dès lors que la différence relative entre entre deux itérations est inférieure à un seuil δ donné. Plus précisément, l'algorithme est arrêté si $\frac{\|b^{(m+1)}-b^{(m)}\|_2}{\|b^{(m)}\|_2} < \delta$. Un nombre maximal d'itérations est aussi donné afin d'assurer l'arrêt des algorithmes.

Conclusion

Ce chapitre propose deux types d'algorithmes permettant de calculer les estimations dans un modèle partiellement linéaire. La première catégorie consiste à estimer conjointement les parties linéaire et fonctionnelle du modèle. La seconde est issue du constat établi au Chapitre 3 qui permet d'estimer successivement le vecteur de régression et la fonction du modèle. Le lien avec la théorie de l'estimation robuste vu au Chapitre 3 donne accès à des algorithmes semi-quadratiques. Cette estimation en deux étapes est à notre connaissance novatrice dans la mesure où l'estimation conjointe des parties du modèle semble être l'unique approche considérée auparavant. Le chapitre suivant compare ces algorithmes d'estimation sur des données simulées avant de les appliquer sur des données réelles.

CHAPITRE 6

SIMULATIONS ET UN EXEMPLE SUR DONNÉES RÉELLES

Sommaire

6.1 Comparaison des algorithmes pour le seuillage doux 98
6.1.1 Exemple 1 : Fonction sinusoïdale 99
6.1.2 Exemple 2 : Fonction linéaire par morceaux 104
6.1.3 Exemple 3 : Dimension 5 106
6.2 Comparaison des différents seuillages 109
6.2.1 Validation croisée pour la pénalité quadratique 109
6.2.2 Exemple 1 110
6.2.3 Exemple 2 112
6.2.4 Exemple 3 115
6.3 Application à l'étude de données réelles issues d'une IRM fonctionnelle 118
6.3.1 Description des données 118
6.3.2 Modélisation par un modèle partiellement linéaire 119

Dans ce chapitre, nous comparons les algorithmes présentés dans le chapitre précédent sur des données simulées. Nous nous intéressons ensuite aux comportements des différentes procédures de seuillage étudiées au Chapitre 4 sur des simulations. La dernière section de ce chapitre consiste en l'étude de données réelles, provenant d'Imagerie à Résonance magnétique fonctionnelle.

Les calculs ont été réalisés sous MATLAB 7.4 sur un environnement Unix. Pour la décomposition en ondelettes (DWT) nous avons utilisé la *toolbox* WaveLab développée par Donoho et ses collaborateurs (disponible sur *http://www-stat.stanford.edu/∼wavelab*). Résumons tout d'abord les procédures mises en œuvre pour chacun des exemples :

1. Application de la transformée en ondelette discrète (DWT) aux observations et aux covariables à l'aide de l'algorithme pyramidal de Mallat (1989) ;
2. Estimation de la variance σ^2 au moyen d'une décomposition QR de la matrice des coefficients d'ondelettes de résolution maximale de la matrice de régression, suivie d'une estimation MAD, tel que cela est décrit en Section 4.3 ;
3. Estimation du vecteur de régression β et du vecteur des coefficients de la partie fonctionnelle θ par les différents algorithmes :
   - Estimation conjointe des deux vecteurs par *Backfitting*.
   - Estimations successives des deux vecteurs à l'aide de la relation établie au Chapitre 3 :
     (a) Estimation de β à l'aide de ARTUR ou LEGEND ;
     (b) Estimation de θ par application de la fonction de seuillage à $Z - A\hat{\beta}_n$;
4. Et enfin estimation de f en appliquant la transformée en ondelette inverse à $\hat{\theta}_n$.

Dans le cadre d'un seuillage doux, nous comparerons aussi à l'algorithme *BCR* de Fadili et Bullmore ainsi qu'à l'algorithme quadratique suggéré par Fuchs.

Nous remercions chaleureusement M. Fadili de nous avoir transmis les algorithmes du *Backfitting* et *BCR*.

6.1 Comparaison des algorithmes pour le seuillage doux

Nous nous intéressons dans un premier temps à l'estimation mettant en œuvre un seuillage doux des coefficients d'ondelettes.

Pour illustrer les procédures ci-dessus, nous les avons évalué sur les trois problèmes "test" suivants. Deux fonctions différentes ont été utilisées pour la partie non paramétrique f, une fonction sinusoïdale et une constante par morceaux. Les covariables ont été générées selon l'expression $X_i = g(i/n) + \eta_i$ avec des fonctions g polynomiales et un échantillon $(\eta_i)_{i=1,\ldots,n}$ généré indépendamment depuis une loi centrée de variance finie, comme cela fut expliqué lors de la discussion des hypothèses au Chapitre 4. Nous avons considéré $n = 2^8$ observations et répété l'estimation avec 500 réalisations différentes de U pour chaque expérience. La variance du bruit a été choisie de telle sorte que le rapport signal-sur-bruit respectivement de la partie non paramétrique et de la partie paramétrique valent 5 and 8.88. Ce choix nous a semblé raisonnable. Les données ainsi

simulées, nous avons appliqué les algorithmes présentés ci-dessus. L'application de la décomposition DWT a été réalisée à partir de Symmlets de Daubechies ayant 8 moments nuls. Afin de mieux comparer les algorithmes, ce filtre a été appliqué à chacun d'eux et seul le seuillage doux a été mis en œuvre dans un premier temps. Le seuil choisi dans le seuillage des coefficients d'ondelettes est le seuil universel. Quant aux critères d'arrêt, le seuil de tolérance δ a été fixé à 10^{-10} pour ARTUR, LEGEND et le *Backfitting*. Afin d'éviter des délais de calcul trop longs, nous avons enfin limité le nombre d'itérations à 2000. En ce qui concerne l'algorithme *BCR*, le nombre d'itérations a aussi été fixé à 2000.

D'autres exemples de simulations peuvent être trouvés dans Gannaz (2007). La raison pour laquelle les temps de calcul sont nettement réduits pour le *Backfitting* comparativement à l'article, est que nous n'avons pas ici utilisé l'algorithme de Fadili et Bullmore (2005) gracieusement fourni par les auteurs. Nous avons en particulier calculé la matrice $(A^T A)^{-1} A^T$ nécessaire à chaque étape de l'algorithme plutôt que d'utliser la fonction de Matlab "$A\backslash$", ce qui permet un gain de temps considérable.

6.1.1 Exemple 1 : Fonction sinusoïdale

Dans les exemples 1 and 2, la covariable fut générée avec une fonction polynomiale $g(t) = 15t^4 - 2(t-0.5)^2 + t$ et un bruit η indépendant de loi $N(0,1)$. Nous avons par ailleurs effectué aussi des simulations avec des fonctions g générant le plan de régression plus complexes telles que $g(t) = 2^t$, $g(t) = e^{-t^2}$ ou encore $g(t) = cos(t)$ afin d'étudier le comportement des estimateurs lorsque l'hypothèse (A4) était affaiblie. Les résultats étant similaires, cette hypothèse semble donc pouvoir être réduite pour assurer le bon comportement des estimateurs.

Nous considérons dans un premier temps le cas où la partie non paramétrique du modèle est constituée d'une fonction sinusoïdale. Dans un tel contexte, l'estimation semiparamétrique basée sur des techniques de splines est accessible, mais il est intéressant d'observer le comportement de la procédure par ondelettes. La Figure 6.1 représente la transformée en ondelette des observations et de la matrice de régression. Nous pouvons constater que la représentation creuse de la partie non paramétrique génère une réduction importante du biais entre les données et un modèle linéaire. Les lignes en pointillés de la Figure 6.1 représentent respectivement les droites d'équations $y = X\beta - \lambda$ et $y = X\beta + \lambda$. Les observations hors de ces lignes ont un poids moindre dans l'estimation de β.

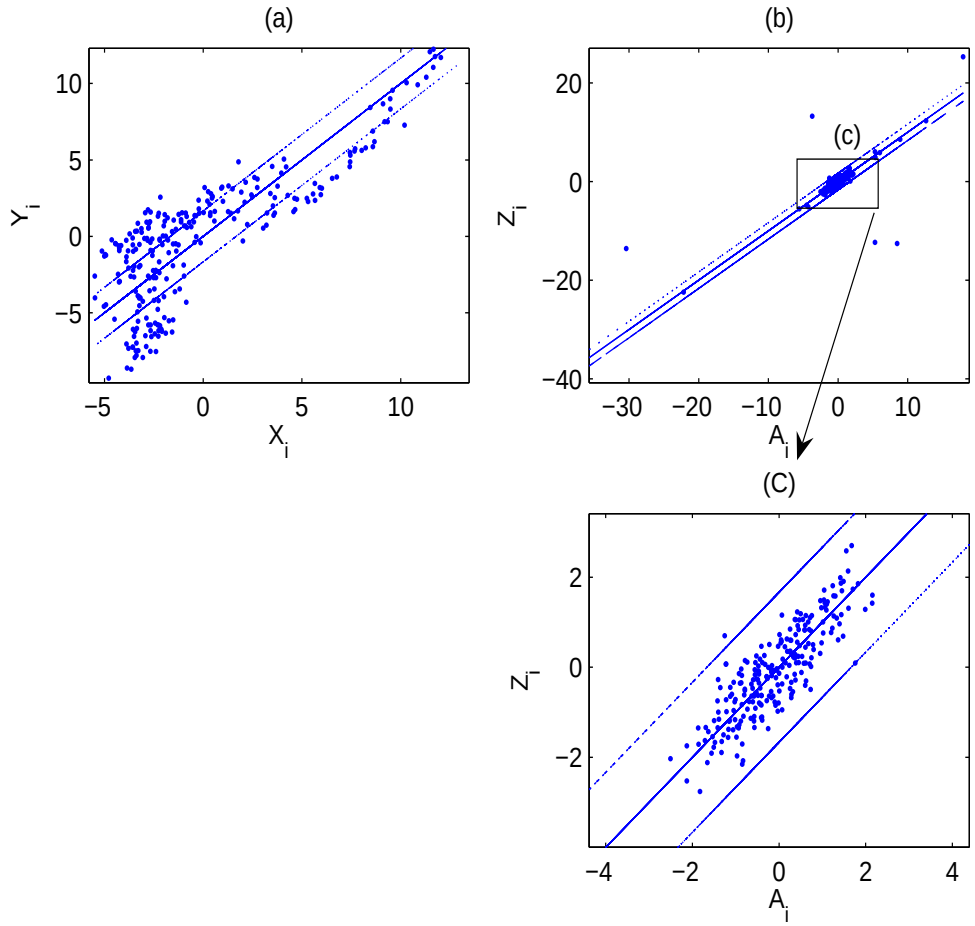

FIG. 6.1: Transformée en ondelettes des données. La figure (a) représente le nuage des observations y_i en fonction des covariables X_i. La droite correspond à la partie linéaire du modèle, d'équation $y_i = X_i\beta$. La figure (b) représente le nuage de points de la figure (a) après la Transformée en Ondelettes Discrète, c'est-à-dire les coefficients z_i en fonction de A_i. La ligne pleine correspond à la partie linéaire du modèle (d'équation $z_i = A_i\beta$) et les lignes pointillées sont les droites d'équations $z = A_i\beta \pm \lambda$. La figure (c) est un agrandissement de la figure (b), correspondant au rectangle tracé sur celle-ci.

Nous étudions maintenant l'influence de la décomposition QR sur l'estimation de la variance du bruit, puis nous comparons les temps d'exécution de chacun des algorithmes sur 500 répliques des données.

Estimation de σ par MAD		
Vraie valeur	sans décomposition QR	avec décomposition QR
0.5	1.1144(0.090578)	0.50085(0.050492)

TAB. 6.1: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 1 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

La Table 6.1 donne un bon aperçu de l'effet d'une décomposition QR sur l'estimation de la variance du bruit : la présence de la partie linéaire introduit un fort biais dans l'estimation MAD. La décomposition QR permet de réduire considérablement ce biais. Ceci peut expliquer pourquoi dans leur comparaison des divers estimateurs, Fadili et Bullmore (2005) obtiennent souvent des estimateurs surlissés, étant donné que leur seuil est sur-évalué par l'intermédiaire d'une sur-estimation de la variance du bruit. Pour que la comparaison des méthodes soit plus juste, nous avons appliqué pour tous les algotithmes, y compris le *Backfitting*, le seuil universel $\lambda = \sigma\sqrt{2\log(n)}$ avec σ donné par l'estimateur MAD après une décomposition QR.

Estimation de β			
Vraie valeur	Fuchs	*BCR*	*Backfitting*
1	0.85283(0.0065982)	0.8678(0.0098)	0.9285(0.0148)
Temps moyen de calcul	391.5878	0.8212	0.0326

Estimation de β		
Vraie valeur	ARTUR	LEGEND
1	0.9285(0.0148)	0.9285(0.0148)
Temps moyen de calcul	0.0885	0.0480

TAB. 6.2: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 1 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.2687 pour le *Backfitting*, ARTUR et LEGEND, 2.4318 pour le *BCR* et 17.5383 pour l'algorithme de Fuchs.

L'algorithme quadratique de Fuchs a été exécuté avec un degré de tolérance égal à 10^{-2} et un nombre maximal de 10 itérations dans la fonction *quadprog* de Matlab. Cette tolérance très forte et ce très faible nombre d'itérations expliquent en grande partie les résultats obtenus dans la Table 6.2. Néanmoins, la dernière ligne de la Table fait apparaître un temps

de calcul considérablement plus long que les autres algorithmes. Ainsi le coût de calcul que nécessiterait cet algorithme pour obtenir des estimations de qualités équivalentes aux autres procédures semble excessif. Par conséquent, nous ne considérerons plus cet algorithme par la suite.

La dernière ligne de la Table 6.2 met en évidence que l'algorithme *BCR* est beaucoup plus lent que les autres procédures et que de plus la qualité de l'estimation obtenue est nettement inférieure. Notons que la différence de qualité d'estimation de l'algorithme *BCR* est peut-être due au choix du nombre d'itérations. Une augmentation du nombre d'itérations permettrait probablement d'améliorer la qualité de l'estimation, mais au détriment du temps de calcul.

La Table 6.2 montre aussi que le *Backfitting* est plus rapide que les procédures semi-quadratiques (ARTUR et LEGEND), pour des qualités d'estimation comparables.

Rappelons que dans les schémas d'estimation semi-quadratiques, après avoir obtenu une estimation du paramètre β, la partie non paramétrique est estimée à partir des résidus associés, $y_i - X_i\hat{\beta}_n$, à l'aide d'une estimation basée sur des ondelettes. La Figure 6.2 représente un exemple typique de ces résidus ainsi que de l'estimation correspondant à ceux-ci. Ces résidus sont issus de l'une des simulations, à laquelle l'algorithme LEGEND a été appliqué.

Etant donnée la valeur du rapport signal-sur-bruit choisie pour la partie non paramétrique ($SNR_f = 5$) l'estimation ne détecte pas la discontinuité de la fonction. Nous pouvons aussi observer des "décalages" au niveau des extrêmes de la fonction les plus proches des bords de l'intervalle ; ce phénomène usuel peut s'expliquer par le surlissage lié au choix du seuil universel et par les problèmes de bord rencontrés dans une estimation par ondelettes. Cependant, la qualité de l'estimation est identique à celle qu'on obtiendrait par des techniques d'ondelettes dans un modèle usuel de régression fonctionnelle (c'est-à-dire sans partie linéaire) avec un même SNR. Ceci étaye notre propos du Chapitre 4 qui stipulait que la présence de la partie paramétrique dans le modèle n'altère pas la qualité de l'estimation de la partie fonctionnelle.

Lors de l'application de ARTUR et LEGEND, Vik (2004) et Dahyot et Kokaram (2004) arrivent tout deux à la conclusion que LEGEND converge plus rapidement, ce qui confirme les résultats théoriques de Nikolova et Ng (2005). Afin de vérifier nous aussi ce point, nous avons comparé les deux algorithmes sur un nombre plus élevé d'observations. Avec

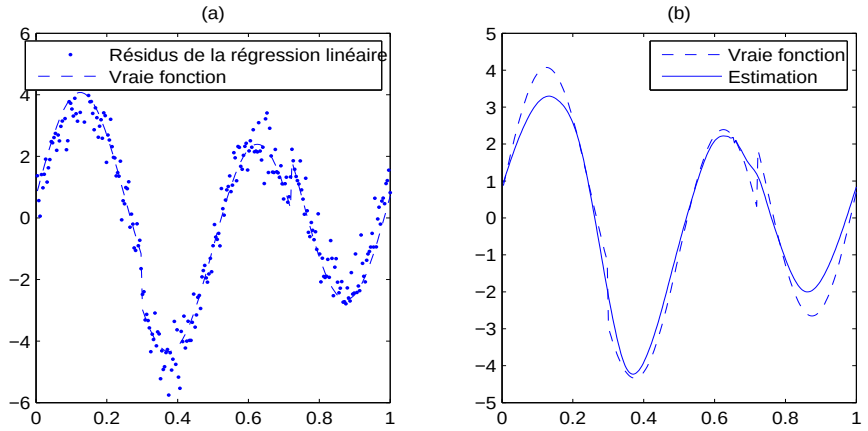

FIG. 6.2: Estimation de la partie non paramétrique dans l'Exemple 1. La figure (a) représente les résidus obtenus après estimation de la partie linéaire du modèle, soit $z_i - A_i\hat{\beta}_n$; la vraie fonction y est représentée en pointillés. La figure (b) donne l'estimation (en traits pleins) qui en résulte ainsi que la vraie fonction (en pointillés).

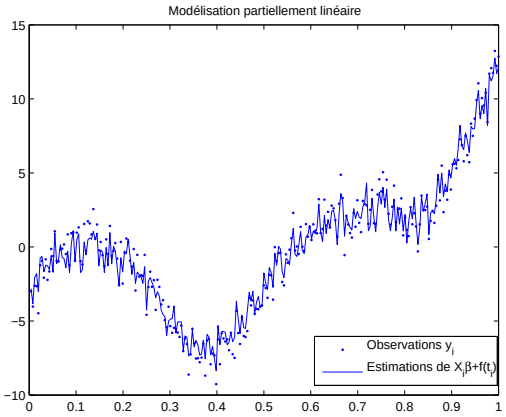

FIG. 6.3: Exemple de modélisation par un modèle partiellement linéaire dans l'Exemple 1. Le nuage de points des observations y est représenté, ainsi que l'estimation partiellement linéaire correspondante (en traits pleins) pour un plan d'expérience.

$n = 2^{10}$ observations et un rapport signal-sur-bruit inchangé, une différence très nette apparaît en terme de temps de calcul pour des qualités d'estimation équivalentes, comme cela est mis en évidence par la Table 6.3. Cette Table montre aussi que le *Backfitting* reste

plus rapide que ARTUR et LEGEND pour une qualité équivalente.

Estimation de β			
Vraie valeur	*Backfitting*	ARTUR	LEGEND
1	0.9814(0.0048)	0.9814(0.0048)	0.9814(0.0048)
Temps moyen de calcul	0.0178	0.6590	0.0328
Nombre indicatif d'itérations	31	14	71

TAB. 6.3: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 1 (les écarts-types sont donnés entre parenthèses) avec $n = 2^{10}$.

6.1.2 Exemple 2 : Fonction linéaire par morceaux

Le but est maintenant de comparer les procédures lorsque la partie non paramétrique du modèle présente de fortes irrégularités. C'est pourquoi nous considérons ici une fonction f linéaire par morceaux. Il est évident que pour une telle fonction l'estimation à base de techniques d'ondelettes est plus adaptée qu'une estimation par splines. Les autres caractéristiques du modèle sont identiques à celles de l'Exemple 1.

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.50371(0.052899)

TAB. 6.4: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 2 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

Les résultats donnés dans la Table 6.5 confirment de nouveau que le *Backfitting* est l'algorithme le plus rapide parmi ceux appliqués. Les procédures semi-quadratiques ont toutefois des temps comparables au *Backfitting*. On peut de plus remarquer que le manque de régularité de la partie fonctionnelle ne semble pas affecter la qualité de l'estimation du paramètre de régression pour ces trois procédures. En ce qui concerne l'algorithme *BCR*, même si la qualité de l'estimation est meilleure que dans l'Exemple 1 pour la partie

Estimation de β				
Vraie valeur	*BCR*	*Backfitting*	ARTUR	LEGEND
1	1.0012(0.0107)	0.9976(0.0148)	0.9976(0.0148)	0.9976(0.0148)
Temps moyen de calcul	0.8092	0.0270	0.0679	0.0412

TAB. 6.5: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 2 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.2216 pour le *Backfitting*, ARTUR et LEGEND et 2.3824 pour l'algorithme *BCR*.

paramétrique, elle reste médiocre pour la partie non paramétrique. De plus, le temps de calcul est nettement supérieur aux procédures ARTUR, LEGEND et *Backfitting*.

De même que dans l'Exemple 1, les Tables 6.5 et 6.6 montrent que LEGEND converge plus rapidement qu'ARTUR, et que cet écart dans les temps de calcul augmente avec le nombre d'observations n. Le *Backfitting* reste l'algorithme le plus rapide.

Estimation de β			
Vraie valeur	*Backfitting*	ARTUR	LEGEND
1	0.9563(0.0068)	0.9563(0.0068)	0.9563(0.0068)
Temps moyen de calcul	0.0261	0.8488	0.0463

TAB. 6.6: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 2 (les écarts-types sont donnés entre parenthèses) avec $n = 2^{10}$.

Au vu de la Figure 6.4, l'estimation de la partie fonctionnelle détecte le saut de la fonction. Lorsque nous comparons l'estimation usuelle par ondelettes dans un modèle de regression non paramétrique et l'estimation obtenue dans un modèle partiellement linéaire, nous obtenons des qualités similaires.

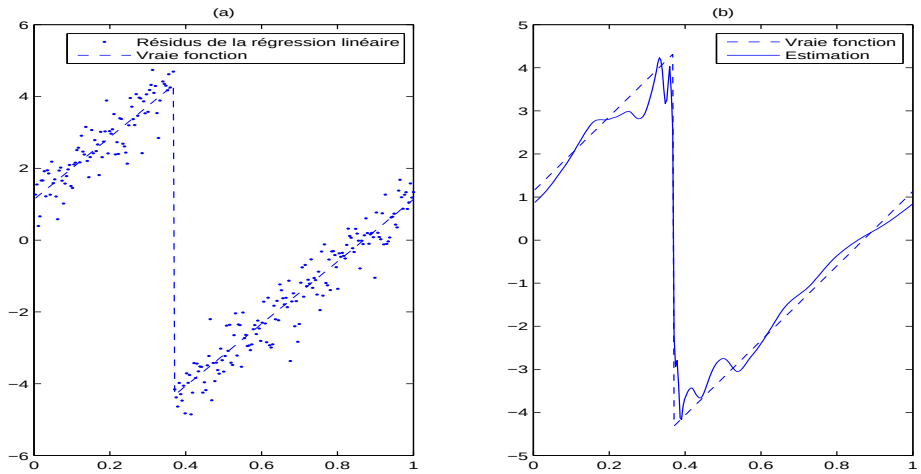

FIG. 6.4: Estimation de la partie non paramétrique dans l'Exemple 2. La figure (a) représente les résidus obtenus après estimation de la partie linéaire du modèle, soit $z_i - A_i\hat{\beta}_n$; la vraie fonction y est représentée en pointillés. La figure (b) donne l'estimation (en traits pleins) qui en résulte ainsi que la vraie fonction (en pointillés).

6.1.3 Exemple 3 : Dimension 5

Nous nous intéressons dans cet exemple au cas où le vecteur des paramètres de régression β appartient à $\mathbb{R}^5$ (La dimension de la matrice de covariables X étant alors $n \times 5$). La partie non paramétrique f est la même que dans l'Exemple 2, c'est-à-dire que la fonction présente de fortes discontinuités. De même que pour les exemples précédents, le SNR a été fixé à 5 pour la partie non paramétrique. Le SNR de la partie linéaire du modèle vaut 5 également. Les Tables ci-après résument les résultats obtenus dans ces simulations.

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.50922(0.051209)

TAB. 6.7: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 3 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

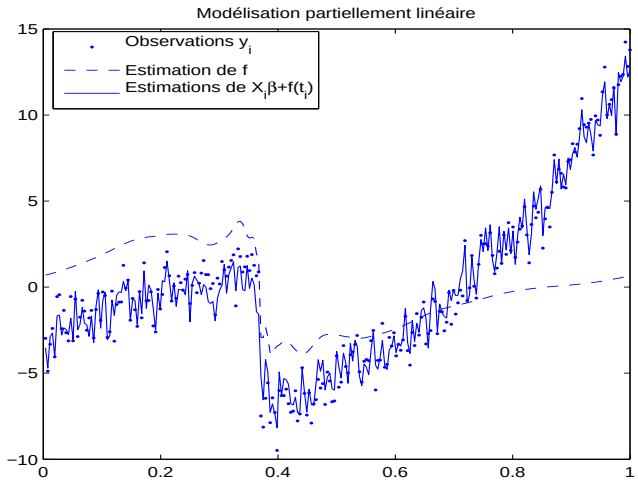

FIG. 6.5: Exemple de modélisation par un modèle partiellement linéaire dans l'Exemple 2. Le nuage de points des observations y est représenté, ainsi que la partie fonctionnelle du modèle estimée (en traits pointillés) et l'estimation partiellement linéaire correspondante (en traits pleins) pour un plan d'expérience.

Nous retrouvons dans la Table 6.8 des résultats similaires à ceux des Exemples 1 et 2. Le *Backfitting* est l'algorithme le plus rapide et l'algorithme *BCR* le plus lent malgré sa qualité peu satisfaisante d'estimation.

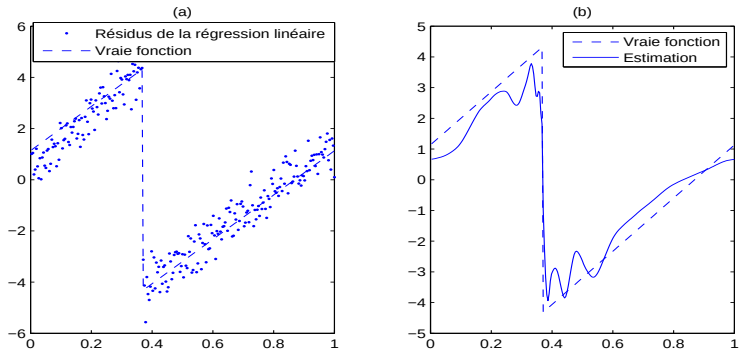

FIG. 6.6: Estimation de la partie non paramétrique dans l'Exemple 3. La figure (a) représente les résidus obtenus après estimation de la partie linéaire du modèle, soit $z_i - A_i\hat{\beta}_n$; la vraie fonction y est représentée en pointillés. La figure (b) donne l'estimation (en traits pleins) qui en résulte ainsi que la vraie fonction (en pointillés).

	Estimation de β			
Vraie valeur	*BCR*	*Backfitting*	ARTUR	LEGEND
-1	1.4160(0.2519)	-0.5755(0.2411)	-0.5755(0.2411)	-0.5755(0.2411)
5	4.6518(0.0915)	4.9183(0.0912)	4.9183(0.0912)	4.9183(0.0912)
0	0.3450(0.3002)	0.4859(0.3078)	0.4859(0.3078)	0.4859(0.3078)
-2	-5.0459(0.3229)	-2.9461(0.2349)	-2.9461(0.2349)	-2.9461(0.2349)
10	10.9563(0.2076)	10.4691(0.1963)	10.4691(0.1963)	10.4691(0.1963)
Erreur quadratique moyenne	4.0717	1.3390	1.3390	1.3390
Temps moyen de calcul	0.6916	0.0337	0.0931	0.0444

TAB. 6.8: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 3 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.2798 pour le *Backfitting*, LEGEND et ARTUR et 2.1359 pour le *BCR*.

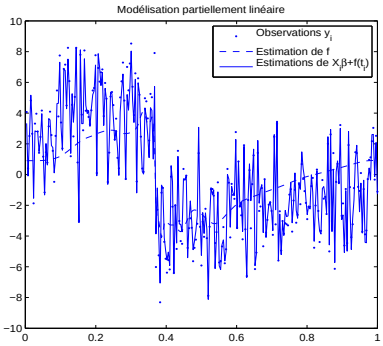

FIG. 6.7: Exemple de modélisation par un modèle partiellement linéaire dans l'Exemple 3. Le nuage de points des observations y est représenté, ainsi que la partie fonctionnelle du modèle estimée (en traits pointillés) et l'estimation partiellement linéaire correspondante (en traits pleins) pour un plan d'expérience.

La particularité de cet exemple comparativement aux deux exemples précédents est aussi que la partie fonctionnelle du modèle a des valeurs du même ordre de grandeur que la partie paramétrique du modèle. Nous pouvons voir que ceci n'altère pas la qualité de l'estimation pour la fonction (à SNR égal).

6.2 Comparaison des différents seuillages

Nous reprenons les trois exemples précédents avec l'objectif de comparer les trois procédures de seuillage étudiées dans ce travail : le seuillage doux, le seuillage associé à une pénalité quadratique et le seuillage fort. Le seuillage doux et le seuillage fort permettent de définir des procédures adaptatives. Nous avons pris pour ces deux procédures le seuil universel $\lambda = \sigma\sqrt{2\log(n)}$. Dans le cas du seuillage correspondant à une pénalisation par la norme de Sobolev, le schéma d'estimation obtenu n'est pas adaptatif. Par conséquent nous mettons en œuvre la procédure de validation croisée décrite en début de section. Les autres paramètres de l'estimation sont choisis de manière identique à la section précédente. Au vu des résultats de la section précédente, seul le *Backfitting* a été appliqué.

Signalons que les résultats présentés peuvent sembler éloignés de ceux obtenus dans Fadili et Bullmore (2005). La raison principale est que la notion de SNR diffère. En effet, Les auteurs définissent le rapport signal sur bruit comme $\|f\|_\infty/\sigma$, tandis que dans ce travail, celui-ci vaut $SNR_f = \|f\|_2/(\sqrt{n}\sigma)$.

6.2.1 Validation croisée pour la pénalité quadratique

La procédure de seuillage par une pénalité de Sobolev, issue de Fadili et Bullmore (2005), aboutit à un estimateur ayant une forme similaire à celui de Speckman (1988). Nous pouvons donc appliquer la procédure de validation croisée définie par celui-ci. Celui-ci consiste à choisir les paramètres s et λ minimisant le critère

$$GCV = \frac{n^{-1}\|(I - M(\lambda, s))y\|^2}{(1 - n^{-1}trace(M(\lambda, s)))^2},$$

avec $M(\lambda, s)$ tel que $y - \hat{y} = (I - M(\lambda, s))y$.

Fadili et Bullmore (2005) optent aussi pour ce schéma, mais en ne considérant que les coefficients d'ondelettes. Dans le cas de l'estimateur $\hat{\beta}_n$ obtenu, nous obtenons :

$$GCV(\lambda, s) = \frac{(n - 2^{i_0})^{-1}\|(I - M(\lambda, s))y\|^2}{(1 - (n - 2^{i_0})^{-1}trace(M(\lambda, s)))^2},$$

avec $M(\lambda, s) = I - T + TA(A^TTA)^{-1}A^TT$, la matrice T ayant pour terme général

$$T_{i,j} = \frac{2\lambda 2^{2[\log_2(i)]s}}{1 + 2\lambda 2^{2[\log_2(i)]s}}\delta_{i,j}.$$

Sous les conditions du Théorème 4.2, à s fixé, le critère vérifie

$$\mathbb{E}(GCV(\lambda, s)) - \mathbb{E}(R(\lambda, s)) \to \sigma^2$$

avec

$$R(\lambda, s) = (n - 2^{i_0})^{-1} \|X\beta + F - X\hat{\beta}_n - \hat{F}_n\|^2.$$

La preuve de cette convergence peut être trouvée dans Fadili et Bullmore (2005).

Remarquons qu'il serait intéressant de mettre en œuvre cette procédure de validation croisée pour les seuillages doux et fort afin d'améliorer la comparaison, mais n'ayant pas d'écriture explicite des estimateurs du paramètre de régression dans ces contextes, ceci n'est pas possible.

6.2.2 Exemple 1

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.50025(0.052127))

TAB. 6.9: Valeur moyenne et écart-type des estimations de σ sur 500 simulations de l'Exemple 1 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

L'estimation de l'écart-type du bruit, donnée en Table 6.9, ne dépend pas du seuillage choisi.

Au vu de la dernière ligne de la Table 6.10, le temps de calcul est plus élevé pour la pénalité de Sobolev. Ceci est dû à la procédure de validation croisée nécessaire pour déterminer les paramètres de lissage λ et s. Le temps de calcul de la procédure de validation croisée dépend de la précision demandée pour ces paramètres. Le temps peut être réduit au risque de diminuer la qualité d'estimation.

La Table 6.10 montre que le seuillage associé à la pénalité de Sobolev aboutit à une moins bonne qualité d'estimation. Le biais de l'estimation du paramètre de régression β ainsi que l'erreur quadratique moyenne de l'estimation de la partie fonctionnelle du modèle sont nettement plus élevés qu'avec les schémas des seuillages doux et fort. Nous avons vu au Chapitre 3 qu'en théorie nous obtenions des vitesses plus faibles pour l'estimation du paramètre de régression. Cependant la mauvaise qualité observée ici semble être liée

Estimation de β			
Vraie valeur	Seuillage fort	Seuillage doux	Seuillage de Sobolev
1	0.9838(0.0189)	0.9288(0.0158)	0.8637(0.0076)
Temps moyen de calcul	0.0375	0.0398	6.5199

TAB. 6.10: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 1 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.0695 pour le seuillage fort, 0.2473 pour le seuillage doux et 0.6003 pour le seuillage de Sobolev. Le seuillage de Sobolev a été mis en œuvre avec une procédure de validation croisée.

à la procédure de validation croisée appliquée. En effet, en théorie le biais résultant de l'estimation de la partie linéaire n'affecte pas la vitesse d'estimation de la partie fonctionnelle. Or d'après la Table 6.10, la qualité d'estimation de f est peu satisfaisante, par comparaison aux autres seuillages.

Nous avons appliqué une procédure de validation croisée similaire dans un modèle non paramétrique. La fonction choisie est celle de l'Exemple 1, avec un même SNR. La Table 6.11 donne les résultats des simulations réalisées dans ce modèle. Nous pouvons constater que la qualité de l'estimation par pénalisation de Sobolev est améliorée, comparativement à ce que nous avons pu observer dans un modèle partiellement linéaire. Ceci confirme, comme il a été dit au Chapitre 4, que les paramètres obtenus par moyen de la procédure de validation croisée dans un modèle partiellement linéaire entraînent un sous-lissage de la partie fonctionnelle.

Estimation de f dans un modèle non paramétrique			
	Seuillage fort	Seuillage doux	Seuillage de Sobolev
MISE	0.0597	0.1380	0.2501

TAB. 6.11: MISE de l'estimation dans un modèle non paramétrique sur 500 simulations aec la fonction et le SNR de l'Exemple 1 et $n = 2^8$. Le seuillage de Sobolev a été mis en œuvre avec une procédure de validation croisée.

La Table 6.10 montre aussi un biais plus important dans l'estimation de la partie linéaire du modèle avec le seuillage de Sobolev qu'avec les autres seuillages.

Nous pouvons constater au vu de la Table 6.10 que le seuillage fort est le seuillage correspondant à la meilleure qualité (selon le critère de l'erreur quadratique moyenne) dans l'estimation de la partie fonctionnelle. Ce constat est similaire à ce qu'on observe dans un modèle non paramétrique et est donc cohérent. Concernant l'estimation de la partie linéaire, la qualité de l'estimation est aussi meilleure avec le seuillage fort. De manière générale, la classe des *Redescending M-estimateurs*, à laquelle appartient la moyenne tronquée, associée au seuillage fort, est moins sensible aux valeurs aberrantes. D'où probablement le biais moindre dans le cadre d'un seuillage fort que dans celui du seuillage doux, les coefficients d'ondelettes de la partie linéaire ayant eu moins d'influence dans l'estimation.

Les Figures 6.8 et 6.9 confirment le sous-lissage important de la fonction dans le cas du seuillage de Sobolev, en raison de la validation croisée.

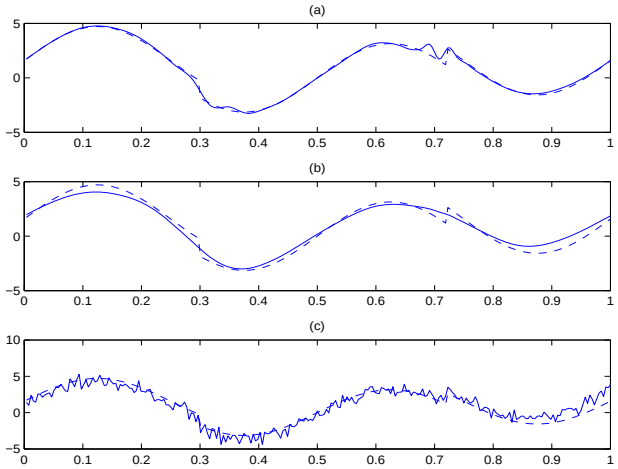

FIG. 6.8: Un exemple d'estimation de la partie non paramétrique dans l'Exemple 1. Sont représentées en traits continus l'estimation par seuillage fort (figure (a)) l'estimation par seuillage doux (figure (b)) l'estimation par seuillage quadratique (figure (c)). La vraie fonction est représentée en traits discontinus.

6.2.3 Exemple 2

De même qu'auparavant, l'estimation de l'écart-type du bruit, donnée en Table 6.12, ne dépend pas du seuillage considéré.

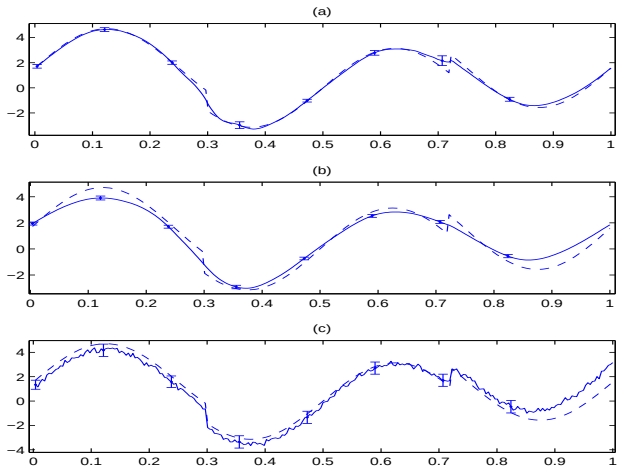

FIG. 6.9: Moyennes des estimations de la partie non paramétrique sur 500 simulations dans l'Exemple 1. Sont représentées en traits continus les moyennes des estimations par seuillage fort (figure (a)) par seuillage doux (figure (b)) et par seuillage quadratique (figure (c)). Les barres verticales sont de taille proportionnelle à l'écart-type de l'estimation. La vraie fonction est représentée en traits discontinus.

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.5078(0.05087)

TAB. 6.12: Valeur moyenne et écart-type des estimations de σ sur 500 simulations de l'Exemple 2 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

Les observations de la Table 6.13 aboutissent aux même conclusions que dans l'exemple précédent. En effet, nous constatons que l'erreur quadratique moyenne de l'estimation de la partie fonctionnelle et le biais de l'estimation du vecteur de régression sont plus élevés dans le cadre du seuillage de Sobolev. Ce résultat ainsi que le temps de calcul sont probablement dus à la procédure de validation croisée, comme expliqué précédemment. Mais la qualité d'estimation par seuillage de Sobolev est nettement meilleure que dans l'Exemple 1. Ensuite, le seuillage fort donne de meilleurs résultats en ce qui concerne l'estimation de la partie fonctionnelle, mais le seuillage doux est plus performant pour l'estimation de la partie linéaire du modèle.

	Estimation de β		
Vraie valeur	Seuillage fort	Seuillage doux	Seuillage de Sobolev
1	0.9800(0.0149)	0.9969(0.0145)	0.9558(0.0074)
Temps moyen de calcul	0.0287	0.0294	5.7920

TAB. 6.13: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 2 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.0743 pour le seuillage fort, 0.2130 pour le seuillage doux et 0.2849 pour le seuillage de Sobolev. Le seuillage de Sobolev a été mis en œuvre avec une procédure de validation Croisée.

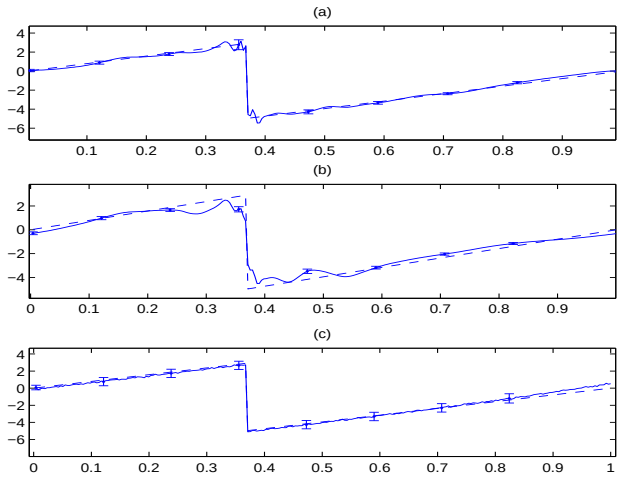

FIG. 6.10: Moyennes des estimations de la partie non paramétrique sur 500 simulations dans l'Exemple 2. Sont représentées en traits continus les moyennes des estimations par seuillage fort (figure (a)) par seuillage doux (figure (b)) et par seuillage quadratique (figure (c)). Les barres verticales sont de taille proportionnelle à l'écart-type de l'estimation. La vraie fonction est représentée en traits discontinus.

En moyenne, l'estimation de la fonction est meilleure pour le seuillage de Sobolev, comme l'illustre la Figure 6.10. Mais la variance de l'estimation est plus forte que pour les autres seuillages. Au vu de la Figure 6.11, ces observations peuvent s'expliquer par un sous-lissage de l'estimation, résultant a priori de la validation croisée.

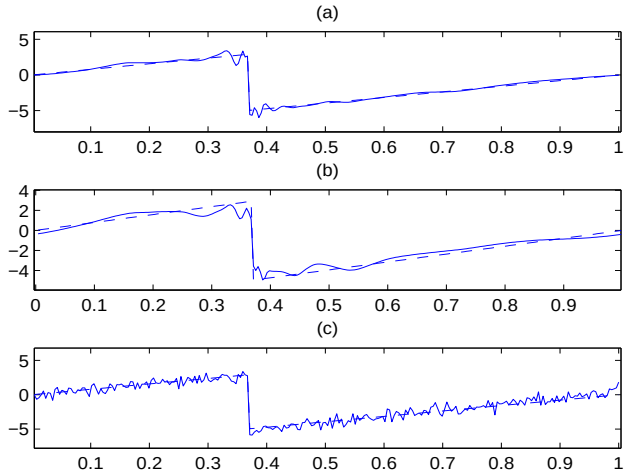

FIG. 6.11: Un exemple d'estimation de la partie non paramétrique dans l'Exemple 2. Sont représentées en traits continus les moyennes des estimations par seuillage fort (figure (a)) par seuillage doux (figure (b)) et par seuillage quadratique (figure (c)). La vraie fonction est représentée en traits discontinus.

6.2.4 Exemple 3

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.50289(0.052125)

TAB. 6.14: Valeur moyenne et écart-type des estimations de σ sur 500 simulations de l'Exemple 3 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

Les résultats donnés en Table 6.15 concordent avec ceux observés dans les Exemples 1 et 2. Le seuillage fort mène à l'estimation du vecteur des paramètres β ayant l'erreur quadratique moyenne la plus faible. La qualité de l'estimation fonctionnelle en terme de risque quadratique est aussi plus faible pour le seuillage fort. Le seuillage de Sobolev apparaît nettement moins performant que les seuillages doux et fort.

De plus, comparativement à l'Exemple 2, la qualité de l'estimation par seuillage de Sobolev est fortement diminuée. En effet le MISE de l'estimation de la partie fonctionnelle vaut près du double de celui de l'Exemple 2. Ceci s'explique probablement par le fait que, comme cela fut précisé en section précédente, l'Exemple 3 admet une partie linéaire

	Estimation de β		
Vraie valeur	Seuillage fort	Seuillage doux	Seuillage de Sobolev
-1	-1.2512(0.2311)	-0.5732(0.2362)	1.6523(0.1334)
5	4.9989(0.0891)	4.9243(0.0874)	4.4646(0.0833)
0	0.3013(0.3275)	0.4813(0.2939)	-0.8198(0.2807)
-2	-2.0768(0.2421)	-2.9468(0.2465)	-7.4036(0.1847)
10	10.2048(0.2342)	10.4883(0.2117)	12.0784(0.1963)
Erreur quadratique Moyenne	0.6956	1.3453	6.4566
Temps moyen de calcul	0.0382	0.0361	6.6344

TAB. 6.15: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 3 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.0806 pour le seuillage fort, 0.2769 pour le seuillage doux et 1.4281 pour le seuillage de Sobolev. Le seuillage de Sobolev a été mis en œuvre avec une procédure de validation Croisée.

de modalités comparables à la partie fonctionnelle, tandis que dans l'Exemple 2 la partie fonctionnelle "prédomine" dans le modèle. Or, le biais de l'estimation de la partie linéaire est non négligeable, au vu des résultats établis au Chapitre 4. Les modalités de la matrice de régresion étant plus élevées, le biais sera donc plus impotrant dans l'Exemple 3 que dans l'Exemple 2. Ce biais implique une moins bonne estimation de la partie fonctionnelle.

Les Figures 6.13 et 6.12 mènent à des conclusions similaires aux deux exemples précédents pour les seuillages fort et doux. En ce qui concerne le seuillage issu d'un pénalisation par une norme de Sobolev, la qualité visuelle de la moyenne des estimations est dégradée. Ceci en raison probablement de la remarque ci-dessus : la partie linéaire a une importance plus grande que dans les exemples précédents et le biais dans l'estimation de cette partie se répércute dans l'estimation fonctionnelle.

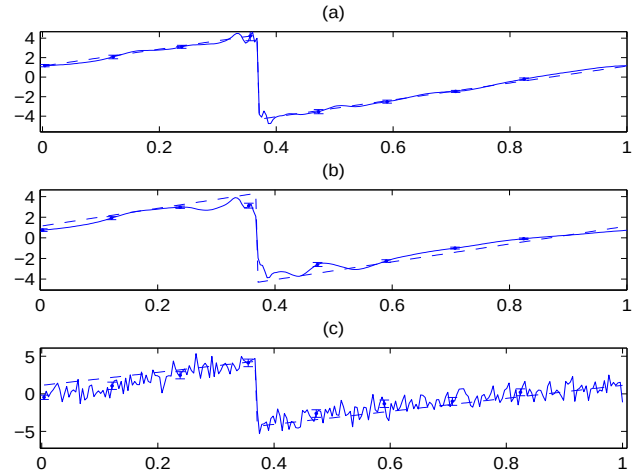

FIG. 6.12: Moyennes des estimations de la partie non paramétrique sur 500 simulations dans l'Exemple 3. Sont représentées en traits continus les moyennes des estimations par seuillage fort (figure (a)) par seuillage doux (figure (b)) et par seuillage quadratique (figure (c)). Les barres verticales sont de taille proportionnelle à l'écart-type de l'estimation. La vraie fonction est représentée en traits discontinus.

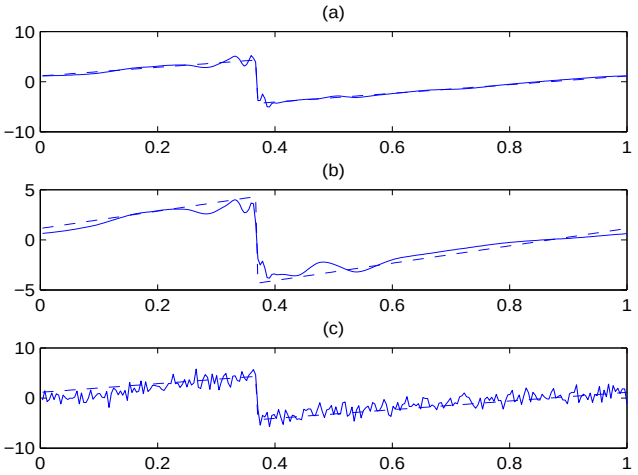

FIG. 6.13: Un exemple d'estimation de la partie non paramétrique dans l'Exemple 3. Sont représentées en traits continus les moyennes des estimations par seuillage fort (figure (a)) par seuillage doux (figure (b)) et par seuillage quadratique (figure (c)). La vraie fonction est représentée en traits discontinus.

6.3 Application à l'étude de données réelles issues d'une IRM fonctionnelle

Dans cette section, nous nous intéressons à la modélisation de signaux issus d'Imagerie à Résonance Magnétique fonctionnelle par un modèle partiellement linéaire. Nous remercions grandement M. Fadili qui nous a procuré les données étudiées ici.

6.3.1 Description des données

Nous étudions trois signaux issus d'une Imagerie à Résonance Magnétique fonctionnelle (IRMf). Le principe de l'IRM fonctionnelle est de détecter les flux sanguins dans le cerveau à l'aide d'un champ magnétique de forte intensité. Lors de l'activation d'une zone du cerveau, les neurones consomment plus d'oxygène, ce qui se traduit par un afflux sanguin. L'IRMf, en mesurant la densité d'hydrogène, va permettre de détecter cet activation. Plusieurs "coupes" en deux dimensions du cerveau sont effectuées, et ces coupes sont répétées à des intervalles de temps rapprochés. La particularité de l'IRMf est de fournir des images de bonne résolution du cerveau à des intervalles de temps réduits (comparativement par exemple aux méthodes de tomographie d'émission de positrons (PET)). Dans les coupes obtenues, plus l'intensité enregistrée est grande, plus l'afflux sanguin dans la zone du cerveau correspondante est important. L'idée est alors d'observer quelles zones sont activées lors d'un stimulus et comment évolue l'activité d'une zone précise au cours d'une expérience.

Parmi les trois signaux, deux sont des signaux de référence, où aucun stimuli n'a été réalisé durant la durée de l'enregistrement. Ces signaux ont été enregistrés au CHU de Caen, initialement sur 6 individus, mais seuls deux des enregistrements sont conservés dans notre étude. Les personnes sont allongées, les yeux fermés, sans perturbation extérieure. Les durées d'enregistrement sont de 8 minutes 32 secondes et 128 mesures ont été réalisées au cours de ce délai, toutes les 4 secondes. Les images obtenues contiennent 64×64 pixels et ont une résolution de 3.75×3.75 mm.

Le troisième signal étudié provient d'un enregistrement réalisé à Cambridge. Seul un signal fut enregistré. Une personne a été placée dans l'IRM fonctionnelle et des stimuli ont été réalisés toutes les 16.5 secondes. Les stimuli étaient des stimuli visuels d'une durée 500 millisecondes, et à chaque stimulus le sujet devait appuyer sur un bouton. La durée de l'enregistrement réalisé est de 4 minutes et 28 secondes, et 128 mesures ont été ef-

fectuées, toutes les 2 secondes. Les images obtenues contiennent 64×64 pixels et ont une résolution de 1.5×3 mm. L'objectif est de mesurer l'activité résultante dans une zone précise du cortex.

Le trois signaux que nous allons étudiés sont issus d'un pixel situé dans la zone du cortex correspondant à la perception visuelle que nous pensons activée par les stimuli. D'où le fait que nous étudions trois jeux de 128 données unidimensionnelles. La restriction à des signaux issus d'un pixel a pour but d'analyser dans un premier temps la validité du modèle proposé sur ces signaux, afin éventuellement de prolonger l'étude aux données bi-dimensionnelles. L'idée serait alors de déterminer les zones activées en extrayant les pixels activés, c'est-à-dire dont la modélisation du signal présente la fréquence des stimuli.

L'idée du modèle partiellement linéaire est de régresser les signaux par rapport à des covariables "de nuisance" et d'extraire une partie fonctionnelle correspondant à l'évolution au cours du temps caractérisant l'activité de la zone du cerveau observée. Les mouvements de têtes, les problèmes d'instabilité des instruments de mesure ou encore l'oscillation naturelle des neurones impliquent une perturbation du signal, correspondant en général à des oscillations basses fréquences. Le fait de régresser sur ces covariables est alors intéressant pour étudier un signal sans biais, dont la tendance rendra mieux compte des phénomènes propres à l'expérience réalisée.

6.3.2 Modélisation par un modèle partiellement linéaire

Nous avons modélisé ces trois signaux par un modèle partiellement linéaire. La partie linéaire du modèle a pour matrice de régression la matrice des trois covariables décrite ci-dessus. Nous utilisons les ondelettes *Symmlet* à 8 moments nuls. Le seuillage considéré est le seuillage fort et le seuil choisi est le seuil universel corrigé :

$$\lambda = \sigma\sqrt{2\log(n) - \log(\log(n))}.$$

Par la suite, nous désignerons respectivement les trois signaux par "Signal 1" et "Signal 2" pour les deux signaux sans stimuli et "Signal 3" pour le signal issu de l'expérience avec stimuli.

Nous avons estimé l'écart-type du bruit dans chacun des signaux par la procédure MAD après une décomposition QR de la matrice, comme décrit en Section 4.3. La première remarque est que l'écart-type du bruit semble élevé comparativement aux valeurs associées

à la partie linéaire du modèle. Pour s'en assurer, nous avons calculé le SNR de la partie linéaire, soit $\frac{\|X\beta\|}{\sqrt{n}\sigma}$, avec les valeurs estimées. Ces SNR sont résumés dans la Table 6.16.

Estimation du SNR		
Signal 1	Signal 2	Signal 3
0.3359	0.1334	0.5052

TAB. 6.16: Estimation du rapport signal sur bruit de la partie linéaire : $S\hat{N}R = \|X\hat{\beta}_n\|/(\sqrt{n}\hat{\sigma}_n)$, pour les 3 signaux d'IRMf.

Les SNR obtenus sont particulièrement faibles. Par exemple, pour l'expérience avec activation, le SNR pour la partie linéaire vaut 0.5052. Les valeurs de SNR ne laissent pas présager une bonne qualité d'estimations du vecteur des paramètres de régression. Cette observation remet en cause la validité d'une modélisation par le modèle partiellement linéaire.

Les estimations des parties fonctionnelles obtenues sur les trois signaux étudiés sont données en Figure 6.14. Si nous effectuons l'estimation par un modèle non paramétrique, les fonctions sont en effet très différentes. Ceci montre que le fait de considérer un modèle partiellement linéaire a un impact important. De manière générale, les fonctions estimées dans un modèle non paramétrique présentent plus d'oscillations. Un exemple est donné en Figure 6.15.

Les courbes obtenues dans les deux cas sans stimuli sont fort semblables à celles obtenues par Fadili et Bullmore (2005). Néanmoins, pour le cas avec stimuli, nous obtenons une courbe très différente. En faisant varier le paramètre d'échelle j_0 et la constante du seuil, nous obtenons des résultats similaires, mais nous ne nous sommes pas autorisés ces changements dans notre travail.

Afin de mieux analyser les estimations, un spectre peut être réalisé, de même que dans Fadili et Bullmore (2005) afin de voir si la fréquence des stimuli est présente dans l'expérience correspondante. Ces spectres sont donnés en Figure 6.16.

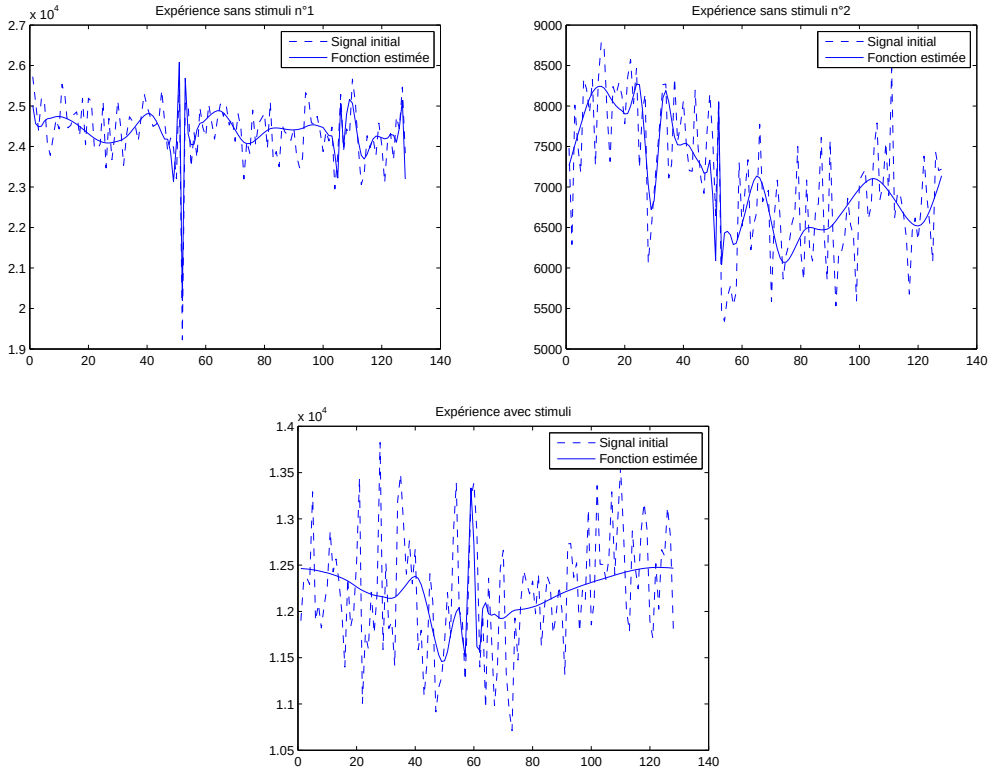

FIG. 6.14: Signaux initiaux (tirets) et fontions estimées dans un modèle partiellement linéaire (continues). Les deux premières figures représentent des cas sans stimuli, la troisième un cas avec stimuli réguliers.

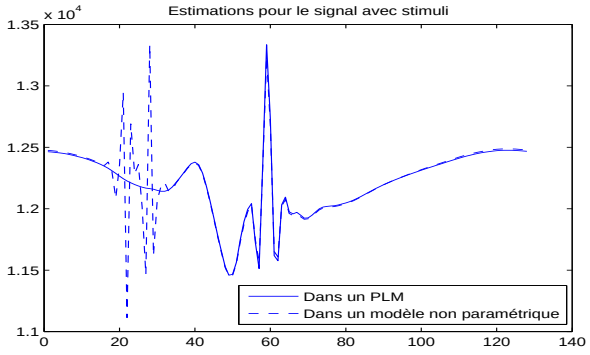

FIG. 6.15: Fontions estimées pour le signal de l'expérience avec stimuli, dans un modèle partiellement linéaire (continues) et dans un modèle non paramétrique (tirets).

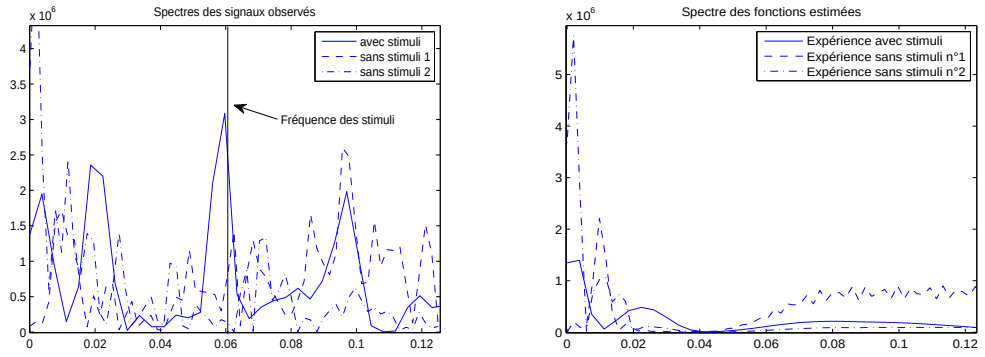

FIG. 6.16: Spectres des signaux initiaux et des fonctions estimées dans un PLM.

L'étude des spectres montre la présence d'un pic très net à la fréquence des stimuli pour le signal issu de l'expérience avec activation, pic qui ne se retrouve pas dans les deux autres expériences. Cependant, ce pic n'apparaît pas dans le spectre de la fonction estimée.

Une autre remarque est que sur chaque enregistrement nous pouvons observer une forte variation de la partie fonctionnelle, très irrégulière et localisée, et ce à des instants identiques proportionnellement aux temps de mesures. N'y aurait-il pas un pré-traitement des données qui créerait cette irrégularité ? Ou un phénomène autre, par exemple une covariable qui n'a pas été prise en compte ? Le fait que les signaux proviennent d'enregistrements dont la fréquence de mesure n'est pas identique, alors que nous considérons une unique matrice de régression (donc les même fréquences dans la partie linéaire) semble par ailleurs peu adapté. Il serait plus justifié de prendre comme signaux de référence des signaux enregistrés dans les mêmes conditions que celui avec stimuli.

En conclusion, la modélisation par un modèle partiellement linéaire dans ce contexte est peu convaincante. Outre que la fréquence des stimuli n'est pas mise en évidence par la partie fonctionnelle du modèle, c'est surtout l'importance du niveau de bruit qui rend ce modèle peu adapté. L'écart-type du bruit semble trop élevé pour pouvoir obtenir une estimation satisfaisante du vecteur de regression dans la partie linéaire du modèle. Cette étude ne remet pas en cause la modélisation par un modèle partiellement linéaire de signaux d'IRMf, uniquement son application aux données étudiées dans ce chapitre.

Conclusion

Lors de la mise en œuvre des estimateurs sur des simulations, il est apparu que les qualités des estimations obtenues étaient satisfaisantes, tant pour le seuillage doux que pour le seuillage fort. La procédure construite permet de détecter des irrégularités dans la partie fonctionnelle, où une estimation par splines pourrait ne pas être performante. De plus le temps de calcul est faible. Le *Backfitting* s'est révélé l'algorithme le plus rapide, mais LEGEND admet un temps de calcul comparable. Cependant, pour le seuillage issu d'une pénalisation quadratique, la procédure de validation croisée semble entraîner un sous-lissage de la partie fonctionnelle. Ce seuillage paraît par conséquent moins adapté au modèle partiellement linéaire que les seuillages fort et doux. Enfin, la dernière section de ce chapitre appliquait la procédure d'estimation développée à des données réelles. Ces données, issues d'Imagerie à Résonance Magnétique fonctionnelle semblent fortement bruitées. Ce qui explique probablement que les résultats obtenus n'ont pas étés satisfaisants.

CHAPITRE 7

CAS DE PLANS D'OBSERVATIONS ALÉATOIRES

Sommaire

7.1 Principe d'estimation **125**
7.2 Simulations avec un plan d'observations aléatoire **128**
7.2.1 Modèle non paramétrique 128
7.2.2 Modèle partiellement linéaire 140

L'étude développée jusqu'à présent supposait que les observations de la partie fonctionnelle étaient réalisées sur une grille équidistante de l'intervalle unité. L'objet de ce chapitre est de généraliser ce travail à des plans d'observations aléatoires de la partie fonctionnelle. Nous souhaitons pour cela appliquer l'estimation développée par Kerkyacharian et Picard (2004) et présentée dans le Chapitre 2.

7.1 Principe d'estimation

Rappelons que ce modèle s'écrit, sous forme matricielle :

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta_0} + \underset{n\times 1}{F(T)} + \underset{n\times 1}{U}$$

avec $F(T) = (f(T_1)\dots f(T_n))^T$ et U bruit blanc gaussien centré de variance $\sigma^2 I_n$. Nous considérons dans ce chapitre que les $(X_i, T_i)_{i=1\dots n}$ sont des variables aléatoires i.i.d. Nous supposons que les variables T_i admettent une fonction de répartition G et une densité g.

Etudions dans un premier temps l'identifiabilité de ce modèle. De même que pour l'étude dans le cadre d'un plan d'observation équidistribué, lorsque $\frac{1}{n}X^TX$ est inversible, nous pouvons écrire le paramètre β_0 sous la forme

$$\beta_0 = (\frac{1}{n}X^TX)^{-1}\frac{1}{n}X^T(Y - F(T) - U).$$

Les variables (X_i, T_i) étant indépendantes et identiquement distribuées, la Loi des Grands Nombres assure que $\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n X_i^TX_i$ tend presque sûrement vers $\mathbb{E}[X^TX]$ et que $\frac{1}{n}X^TF_0(T) = \frac{1}{n}\sum_{i=1}^n X_i^Tf_0(T_i)$ tend presque sûrement vers $\mathbb{E}[X^Tf(T)]$. Ainsi l'hypothèse

$$(A1) \qquad \mathbb{E}[X^TX] > 0 \ \ et \ \ \mathbb{E}[X^Tf(T)] = 0$$

est suffisante pour avoir l'identifiabilité asymptotique presque sûre du modèle.

La transformation en ondelettes appliquée diffère du cas du plan équidistant. En s'appuyant sur les travaux de Kerkyacharian et Picard (2004), nous définssons W_T la transformation associée au plan $T_1 \dots T_n$. Une approximation du terme général de la transformée en ondelettes W est donnée par :

$$W_{T,i,j} \approx \frac{1}{\sqrt{n}}\psi_i \circ G(T_j).$$

Ainsi si e est un vecteur de $\mathbb{R}^n$, sa transformée en ondelette est approximativement

$$W_Te \approx \left(\frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_1 \circ G(T_j)e_j \quad \dots \frac{1}{\sqrt{n}}\sum_{j=1}^n \psi_n \circ G(T_j)e_j.\right)$$

Nous pouvons aisément vérifier qu'asymptotiquement, W_T est bien une matrice orthogonale :

Démonstration. Nous avons $(W_T^TW_T)_{i,i'} = \frac{1}{n}\sum_{j=1}^n \psi_i \circ G(T_j)\psi_{i'} \circ G(T_j)$. Les variables $\psi_i \circ G(T_j)\psi_{i'} \circ G(T_j)$ sont i.i.d. et

$$\mathbb{E}\psi_i \circ G(T_j)\psi_{i'} \circ G(T_j) = \int \psi_i \circ G(t)\psi_{i'} \circ G(t)g(t)dt = \int \psi_i\psi_{i'} = \delta_{i,i'}.$$

Nous pouvons alors appliquer la Loi forte des Grands Nombres, qui donne

$$(W_T^TW_T)_{i,i'} \xrightarrow{p.s.} \delta_{i,i'}.$$

Ainsi, La suite de matrices $W_T^TW_T$ converge vers la matrice unité I_n. □

Nous considérerons par la suite que $W_T^T W_T = W_T W_T^T = I_n$, cette égalité étant vraie asymptotiquement presque sûrement.

Après transformation, nous obtenons le modèle suivant :

$$\underset{n\times 1}{Z} = \underset{n\times p}{A}\ \underset{p\times 1}{\beta_0} + \underset{n\times 1}{\theta_0} + \underset{n\times 1}{\underline{\varepsilon}}$$

avec $Z = W_T Y$, $A = W_T X$ et $\underline{\varepsilon} = W_T U$, bruit blanc gaussien centré de variance $\sigma^2 I_n$ par orthogonalité de la matrice W_T.

Lorsque la fonction de répartition n'est pas connue, alors nous remplaçons G par son estimation $\hat{G}_n$, obtenue indépendamment des données utilisées pour l'estimation dans le cadre du modèle partiellement linéaire. Le terme général de la transformée W_T résultante est approximativement

$$W_{T,i,j} \approx \frac{1}{\sqrt{n}} \psi_i \circ \hat{G}_n(T_j).$$

Nous sommes alors ramené à un cadre d'étude similaire à celui du Chapitre 3. Nous pouvons alors reprendre la procédure d'estimation décrite dans ce chapitre. C'est-à-dire que les estimateurs $(\hat{\beta}_n, \hat{\theta}_n)$ sont définis par minimsation d'un critère des moindres carrés pénalisés :

$$(\hat{\theta}_n, \hat{\beta}_n) = \underset{\theta,\beta}{argmin} \left\{ J_n(\theta,\beta) = \frac{1}{2}\sum_{i=1}^{n}(z_i - A_i^T\beta - \theta_i)^2 + \lambda^2/2 \sum_{i=i_0}^{n} \mathbb{1}_{\{\theta_i \neq 0\}} \right\}. \tag{7.1}$$

Nous considérons la pénalité de comptage afin d'obtenir un seuillage fort des coefficients et donc d'obtenir un estimateur de la partie non paramétrique du modèle semblable à celui de Kerkyacharian et Picard (2004). De même qu'au chapitre 3, les estimateurs résultent d'une procédure en deux étapes :

1. L'estimateur du paramètre β est un M-estimateur obtenu en intégrant la partie linéaire dans le bruit :
$$\hat{\beta}_n = \underset{\beta}{argmin} \sum_{i=1}^{n} \rho_\lambda(z_i - A_i^T\beta),$$
avec $\rho_\lambda(u) = u^2/2\mathbb{1}_{|u|\leqslant\lambda} + 1/2\mathbb{1}_{|u|>\lambda}$.
2. Le vecteur des coefficients d'ondelettes θ est estimé par seuillage fort de $Z - A\hat{\beta}_n$.
$$\hat{\theta}_{i,n} = \begin{cases} z_i - A_i^T\hat{\beta}_n & \text{si } i < i_0 \\ \text{signe}(z_i - A_i^T\hat{\beta}_n)\mathbb{1}_{|z_i - A_i^T\hat{\beta}_n| \geqslant \lambda} & \text{si } i \geqslant i_0 \end{cases}.$$

L'estimation de f est ensuite obtenue en appliquant la transformée en ondelettes inverse associée au plan $T_1, \ldots, T_n$.

7.2 Simulations avec un plan d'observations aléatoire

Dans cette section, nous mettons en œuvre la procédure d'estimation dans le cadre d'un plan d'observation non équidistant développée au début du chapitre. En particulier, l'estimation de la partie fonctionnelle repose sur la méthodologie développée dans l'article de Kerkyacharian et Picard (2004) présentée au Chapitre 2. Avant d'appliquer notre procédure dans un modèle partiellement linéaire, nous étudions dans un premier temps les qualités de l'estimation fonctionnelle décrite par Kerkyacharian et Picard (2004) dans un modèle non paramétrique.

7.2.1 Modèle non paramétrique

Rappelons tout d'abord la procédure d'estimation décrite au Chapitre 2 pour un modèle non paramétrique. Le modèle considéré est de la forme :

$$y_i = f(T_i) + v_i,\ i = 1, \dots, n,$$

avec v_i variables i.i.d. de loi normale centrée de variance σ_v^2 et T_i variables i.i.d., indépendantes des v_i, de densité g et de fonction de répartition G. Nous distinguons deux cas :

– Soit la fonction de répartition du plan d'observations est connue, auquel cas nous caculons l'estimateur f_n^* par :

$$f^* = \alpha_{0,0,n}^* \varphi \circ G + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{j,k,n}^* \mathbb{1}_{\{|\beta_{j,k,n}^*| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ G,$$

avec κ constante, $2^{j_1} \sim \sqrt{\frac{n}{\log(n)}}$, et

$$\beta_{j,k,n}^* = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_{j,k} \circ G(T_i).$$

– Soit la fonction de répartition est inconnue et alors nous procédons en deux étapes :

1. Nous estimons la fonction de répartition G par $\hat{G}_n$ calculé sur les observations $(T_i)_{i=1,\dots,n/2}$. Kerkyacharian et Picard (2004) proposent d'utiliser l'estimateur empirique mais d'autres estimateurs plus lisses peuvent être considérés.
2. L'estimation de la fonction f est donnée par $\hat{f}_n$:

$$\hat{f}_n = \hat{\alpha}_{0,0,n}^{\varphi} \circ \hat{G}_{[n/q]} + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \hat{\beta}_{j,k,n} \mathbb{1}_{\{|\hat{\beta}_{j,k,n}| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ \hat{G}_{[n/q]},$$

avec $2^{j_1} \sim \sqrt{\frac{n}{\log(n)}}$ et

$$\hat{\beta}_{j,k,n} = \frac{1}{n - [n/2]} \sum_{i=[n/2]+1}^{n} Y_i \psi_{j,k} \circ \hat{G}_{[n/2]}(T_i).$$

L'objectif de cette sous-section est d'implémenter ces estimateurs. Nous remercions sincèrement M. Willer de nous avoir aimablement donné les programmes relatifs à l'article Chesneau et Willer (2005).

Problèmes à distance finie

Dans les estimateurs ci-dessus, le degré maximal de résolution est $2^{j_1} = \sqrt{n/\log(n)}$. Pour des valeurs usuelles de n, cette grandeur ne permet pas une bonne approximation de la fonction. Ainsi si par exemple nous avons $n = 2^9$ observations, seuls 7 degrés de résolution sont utilisés dans l'approximation. Le choix d'un paramètre de lissage j_1 nécessaire à l'établissement des propriétés asymptotiques de l'estimateur mais ne permettant pas une bonne estimation à distance finie est usuel dans le cadre de l'estimation de la densité par techniques d'ondelettes (voir Donoho *et al.* (1995)). Nous choisissons de prendre $2^{j_1} = n$. Cette modification du paramètre j_1 ne porte pas à conséquence sur la qualité de l'estimation.

La qualité des estimateurs décrits ci-dessus n'est pas satisfaisante, dans le cas où G est connue comme dans celui où G est inconnue. En effet bien que le comportement asymptotique de l'estimateur soit très bon (nous avons un estimateur quasi-minimax), celui-ci ne permet pas une bonne approximation en pratique, i.e. avec un nombre d'observations fini. Afin de mieux comprendre ce phénomène, détaillons la procédure d'estimation.

Considérons le cas où la fonction de répartition G est connue. La difficulté de l'estimateur de Kerkyacharian et Picard (2004) est qu'il nécessite le calcul des ondelettes aux points $G(T_i)$, $i = 1, \ldots, n$. Nous pourrions utiliser une approximation de ces valeurs en utilisant une grille de points l/N, $l = 1, \ldots, N$ avec N très grand, et en approximant la valeur $G(T_i)$ par la grandeur l/N la plus proche. L'algorithme pyramidal de Mallat (1989) peut alors être appliqué. Les temps de calcul sont raisonnables, et on peut espérer que l'approximation résultante est négligeable. Cependant, nous choisissons de calculer les valeurs des ondelettes aux points précis afin de s'assurer que le comportement de l'estimateur observé n'est pas dû à cette approximation.

Le calcul des valeurs des ondelettes en des points donnés peut être réalisé à l'aide de l'algorithme de Daubechies-Lagarias, mis en œuvre par Vidakovic et décrit dans Vidakovic

(2002). L'algorithme utilisé permet des temps de calculs raisonnables, même si plus longs qu'une DWT classique. L'intérêt d'appliquer cette méthode est que, comme cela a été dit, nous calculons avec une très faible approximation l'estimateur défini par Kerkyacharian et Picard (2004).

L'objectif étant de comprendre le comportement de l'estimateur, nous avons opté pour un bruit très faible. Dans un premier temps, nous vérifions notre procédure d'estimation sur un plan équidistant, en Figure 7.1, en considérant une fonction sinusoïdale.

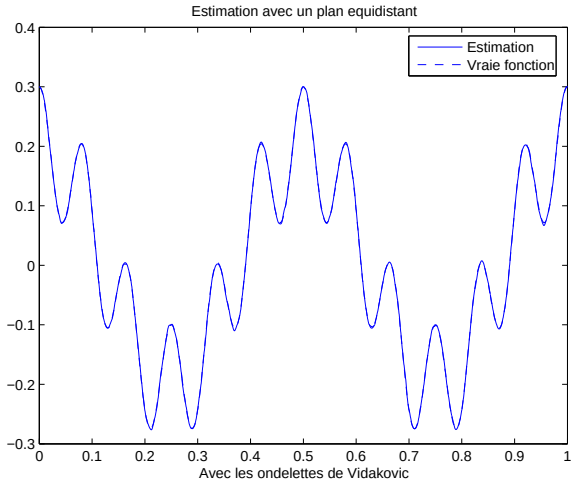

FIG. 7.1: Estimation de fonction avec un plan équidistant. Sont représentées la fonction f à estimer (en tirets) et l'estimation f_n^* obtenue. La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 10 et l'écart-type du bruit σ est supposé connu. Les valeurs des ondelettes aux point i/n, $i = 1, \ldots, n$, ont été calculées à l'aide de l'algorithme de Vidakovic (2006).

La Figure 7.2 représente l'estimation obtenue dans le cas où la densité du plan des observations est uniforme, avec la même fonction qu'en Figure 7.1. Nous pouvons constater que la qualité visuelle de l'estimation n'est pas satisfaisante. En comparaison, nous avons aussi représenté en Figure 7.2 l'estimation à noyau avec seuillage par blocs élaborée par Amato *et al.* (2006).

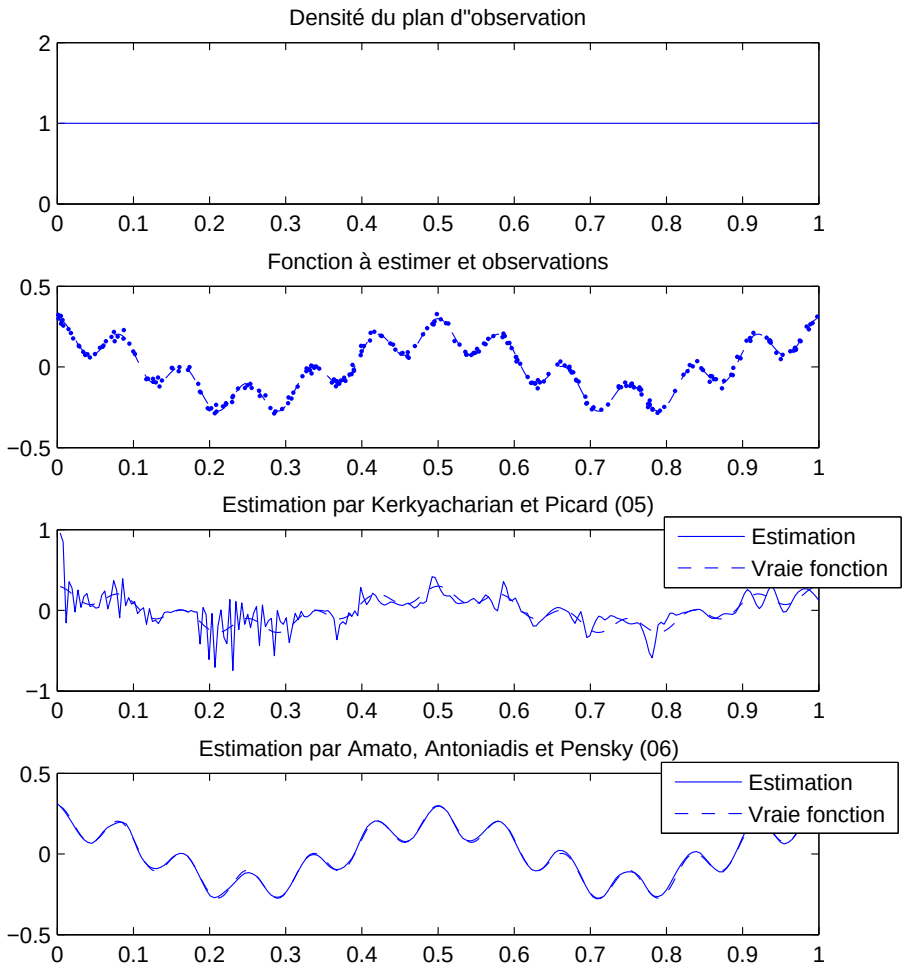

FIG. 7.2: Estimation de fonction avec une densité du plan d'observations uniforme dans le cas où la fonction de répartition est connue. Les figures représentent respectivement la densité g du plan d'observation, la fonction f à estimer et les estimations f_n^* obtenues par les procédures de Kerkyacharian et Picard (2004) et Amato *et al.* (2006). La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 10 et l'écart-type du bruit σ est supposé connu. Les valeurs des ondelettes aux point $G(T_i), i = 1, \ldots, n$, ont été calculées à l'aide de l'algorithme de Vidakovic (2006).

Afin de mieux comprendre la raison pour laquelle l'estimation résultant de la procédure de Kerkyacharian et Picard (2004) est peu satisfaisante (cette observation reste valable pour toutes les densités qui ont été essayées), nous l'appliquons sur une fonction continue par morceaux. La Figure 7.3 représente la fonction que l'on cherche à estimer ainsi que l'estimation obtenue par Kerkyacharian et Picard (2004). Nous pouvons constater que lorsque les valeurs des observations sont proches de 0, l'estimation est correcte, tandis que lorsque les valeurs des observations sont élevées, l'estimation est très irrégulière.

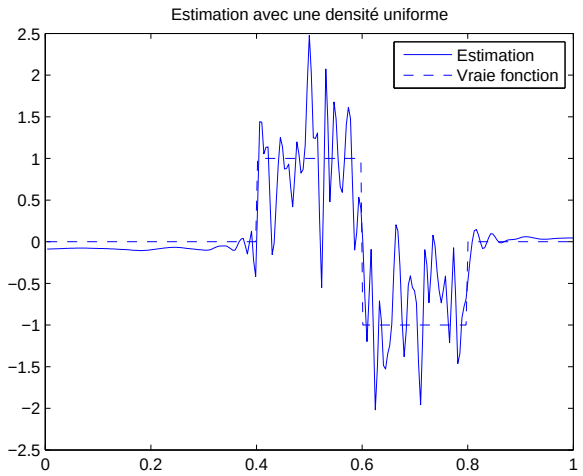

FIG. 7.3: Estimation par Kerkyacharian et Picard (2004) avec une densité du plan d'observations uniforme dans le cas où la fonction de répartition est connue. La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 10 et l'écart-type du bruit σ est supposé connu. Les valeurs des ondelettes aux point $G(T_i), i = 1, \ldots, n$, ont été calculées à l'aide de l'algorithme de Vidakovic (2002).

L'idée est que lorsque nous calculons les coefficients d'ondelettes par la procédure de Kerkyacharian et Picard (2004), le coefficient d'ondelette sera d'autant plus élevé qu'il y a d'observations sur le support de l'ondelette. A l'inverse, les coefficients d'ondelettes sont nuls dès lors que le support de l'ondelette est inclus dans un intervalle de la forme $[G(T_i), G(T_{i+1})]$, $i = 1, \ldots, n/2$. La conséquence est que les estimations des coefficients d'ondelettes sont biaisées lorsque nous sommes dans l'un de ces deux cas.

Pour illustrer ces propos, nous avons représenté en Figure 7.4 l'estimation résultant du schéma de Kerkyacharian et Picard (2004) conjointement aux cumul des observations sur les intervalles $[i/n, (i+1)/n]$, $i = 1, \ldots, n$. Nous pouvons constater que les pics les plus elevés dans l'estimation correspondent bien à des cumul d'observations importants.

Remarque : Nous avons vu en Section 2.4 que lorsque la fonction de répartition est inconnue et estimée par la fonction de répartition empirique, la procédure de Kerkyacharian et Picard (2004) équivaut à sommer les observations sur les intervalles $[\hat{G}_n^{-1}(i/n), \hat{G}_n^{-1}((i+1)/n)]$, $i = 1, \ldots, n$ et à considérer la valeur nulle si ces intervalles sont vides. Les coefficients d'ondelettes associés sont donc d'autant plus élevés que ces intervalles contiennent beaucoup d'observations, et sont nuls lorsqu'ils n'en contiennent pas. Ainsi, l'estimation du coefficient risque d'être biaisée. Le principe lorsque la fonction d'estimation est connue est le même, à ceci près que la valeur prise par l'ondelette dans le calcul du coefficient dépend du point d'observation, ce qui ne permet pas de réécrire la procédure en terme de "données modifiées".

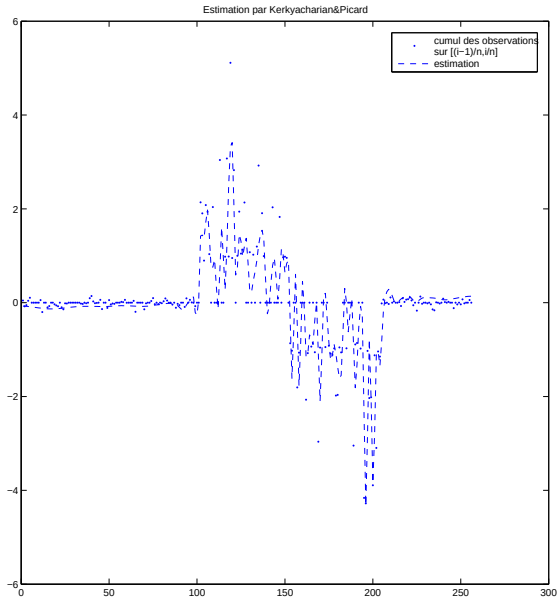

FIG. 7.4: Estimation par Kerkyacharian et Picard (2004) et cumul des observations sur une subdivision de pas $1/n$. La densité du plan des observations est uniforme et la fonction de répartition est supposée connue. La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 10 et l'écart-type du bruit σ est supposé connu. Les valeurs des ondelettes aux point $G(T_i)$, $i = 1, \ldots, n$, ont été calculées à l'aide de l'algorithme de Vidakovic (2002).

Lorsque nous réalisons une décomposition en ondelette d'ordre n, la longueur des support des ondelettes d'ordre maximal est $1/n$. Avec une distribution aléatoire de n points,

il y a nécessairement des supports sans observations. En théorie nous avons une probabilité non nulle qu'il y ait des points, mais en pratique ceci n'est pas vérifié. Si nous réalisons une estimation avec un degré maximal de résolution plus faible, le problème principal sera qu'à l'inverse les supports d'ondelettes contiendront trop d'observations.

Afin d'étayer cette interprétation, nous avons représenté en Figure 7.5 une moyenne des estimations obtenues sur 500 simulations. En effet, les biais des coefficients dépendant de la répartition des données sur l'intervalle, en moyennant les estimations nous devrions retrouver une estimation satisfaisante, ce que la Figure 7.5 confirme.

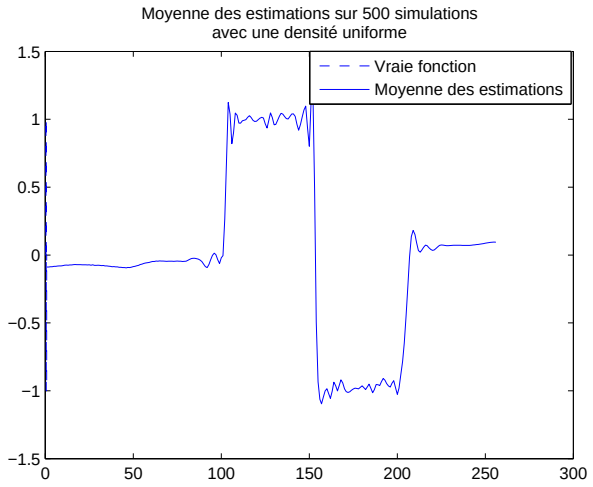

FIG. 7.5: Moyenne des estimations par Kerkyacharian et Picard (2004) sur 500 simulations. La densité du plan d'observations est uniforme et la fonction de répartition est supposée connue. Les simulations ont été réalisées avec $n = 2^8$ observations, un SNR égal à 10 et l'écart-type du bruit σ est supposé connu. Les valeurs des ondelettes aux point $G(T_i)$, $i = 1, \ldots, n$, ont été calculées à l'aide de l'algorithme de Vidakovic (2002).

En résumé, asymptotiquement, tous les intervalles supports des ondelettes contiennent des observations, et en nombre équivalent. Ceci est schématiquement la conséquence de l'hypothèse de régularité (H_2) sur la densité du plan d'observation. C'est pourquoi les propriétés asymptotiques de l'estimation sont satisfaisantes. Mais à distance finie, les observations ne sont par réparties de manière suffisamment homogène pour que cela soit vérifié. Soit la résolution des ondelettes est fine, et alors il existe des intervalles ne contenant pas de points d'observations de la fonction. Sur ces intervalles, le coefficient d'ondelettes correspondant sera nul. Soit la résolution est grossière auquel cas certains intervalles contiennent de nombreuses observations. Ceci a pour conséquence un biais

important dans l'estimation du coefficient d'ondelette correspondant. D'où une estimation qui n'apparaît pas de bonne qualité visuellement lorsque le nombre d'observation est fini.

Alternative choisie

Nous définissons un estimateur très proche de celui utilisé par Chesneau (2006) dans les simulations et par Chesneau et Willer (2005). Le modèle considéré est

$$Y_i = f(X_i) + v_i,\ i = 1, \dots, n,$$

avec X_i variables aléatoires i.i.d. de fonction de répartition G et v_i i.i.d. de loi normale centrée de variance σ_v^2. Il peut aussi s'écrire

$$Y_i = h(U_i) + v_i,\ i = 1, \dots, n, \tag{7.2}$$

avec $h = f \circ G^{-1}$ et $U_i = G(X_i)$ pour tout $i = 1, \dots, n$. Les variables U_i sont i.i.d. de loi uniforme sur $[0; 1]$.

Nous définissons alors l'estimateur suivant :

$$f_n^* = \alpha_{0,0,n}^* \varphi \circ G + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{j,k,n}^* \mathbb{1}_{\{|\beta_{j,k,n}^*| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ G,$$

avec $2^{j_1} = n$, et

$$\beta_{j,k,n}^* = \frac{1}{n+1} \sum_{i=1}^{n} Y_{(i)} \psi_{j,k}(\frac{i}{n+1}),$$

lorsque la fonction de répartition G du plan d'observation est connue. Lorsqu'elle est inconnue, l'estimateur devient

$$\hat{f}_n = \alpha_{0,0,n}^* \varphi \circ \hat{G}_n + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{j,k,n}^* \mathbb{1}_{\{|\beta_{j,k,n}^*| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ \hat{G}_n.$$

Précisons que l'estimation $\hat{G}_n$ de la fonction de répartition utilisée est basée sur les observations T_i, $i = 1, \dots, n$. Par conséquent, $\hat{G}_n(T_{(i)}) = i/n$.

Les estimateurs ainsi définis s'écrivent $f_n^* = \hat{h}_n \circ G$ et $\hat{f}_n = \hat{h}_n \circ \hat{G}_n$, avec $\hat{h}_n$ estimateur par ondelettes défini par Cai et Brown (1998) dans le modèle (7.2). Nous avons donc la propriété suivante :

Proposition 7.1. *Supposons que l'ondelette ψ admet N moments nuls. Si $f \circ G^{-1}$ appartient à l'espace des fonction Hölderiennes $\Lambda^\alpha(M)$ avec $1/2 \leqslant \alpha \leqslant N$ et $0 < M < \infty$, alors*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(f_n^*(T_i) - f(T_i))^2\right] \leqslant C\left(\frac{\log(n)}{n}\right)^{\frac{2\alpha}{1+2\alpha}}.$$

Si $\alpha \geqslant 1$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{f}_n(T_i) - f(T_i)\right)^2\right] \leqslant C\left(\frac{\log(n)}{n}\right)^{\frac{2\alpha}{1+2\alpha}}.$$

Démonstration. Dans le cas où la fonction de répartition G est connue, il suffit de remarquer que $\hat{h}_n(U_i) = f_n^*(T_i)$ et d'appliquer le résultat de Cai et Brown (1998).
Lorsque la fonction G est inconnue, nous avons la décomposition suivante :

$$\hat{f}_n(T_i) - f(T_i) = \hat{h}_n \circ \hat{G}_n(T_i) - h \circ \hat{G}_n(T_i) + h \circ \hat{G}_n(T_i) - h \circ G(T_i).$$

D'où, en remarquant que $\hat{G}_n(T_{(i)}) = i/n$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{f}_n(T_i) - f(T_i)\right)^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{h}_n(\frac{i}{n}) - h(\frac{i}{n})\right)^2\right]}_{T_1} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(h(\frac{i}{n}) - h(U_{(i)})\right)^2\right]}_{T_2}.$$

Le premier terme est borné par $\left(\frac{\log(n)}{n}\right)^{2\alpha/(1+2\alpha)}$, à une constante près, d'après le Théorème de Cai et Brown (1998). Quant au deuxième terme, la fonction h appartenant à l'espace $\Lambda^\alpha(M)$, il peut être majoré par

$$\frac{M}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(U_{(i)} - \frac{i}{n}\right)^{2\min(\alpha,1)}\right].$$

L'inégalité de Jensen permet d'obtenir l'inégalité

$$T_2 \leqslant \frac{M}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left(U_{(i)} - \frac{i}{n}\right)^2\right]^{\min(\alpha,1)}.$$

Les variables $U_{(i)}$ ont pour loi une loi Beta $B(i, n-i+1)$ donc $\mathbb{E}[U_{(i)}] = \frac{i}{n+1}$ et $\mathbb{E}[U_{(i)}^2] = \frac{i(i+1)}{(n+1)(n+2)}$. D'où

$$\mathbb{E}\left[\left(U_{(i)} - \frac{i}{n}\right)^2\right] = \frac{n}{(n+1)(n+2)}\frac{i}{n}(1 - \frac{i}{n}) + \frac{2}{(n+1)(n+2)}i^2n^2.$$

Pour n assez grand, cette quantité est comprise entre 0 et 1. Donc si $\alpha \geqslant 1$,

$$T_2 \leqslant \frac{n}{(n+1)(n+2)} \frac{1}{n} \sum_{i=1}^{n} \frac{i}{n}(1 - \frac{i}{n}) + \frac{2}{(n+1)(n+2)} \frac{1}{n} \sum_{i=1}^{n} \frac{i^2}{n^2}.$$

Nous en déduisons que T_2 est de l'ordre de $1/n$, ce qui achève la démonstration. □

Remarquons que si f appartient à $\Lambda^\alpha(M')$ avec $1/2 \leqslant \alpha \leqslant 1$ et $0 < M < \infty$, alors il suffit que G^{-1} soit lipschitzienne pour qu'existe M telle que $f \circ G^{-1}$ appartienne à l'espace $\Lambda^\alpha(M)$. Ceci est notamment vérifié lorsque g est borné inférieurement par une constante $c > 0$. Cependant, cette condition est plus forte que la condition de Muckenhoupt de Kerkyacharian et Picard (2004). En effet, G^{-1} lipschtzienne équivaut à dire qu'il existe une constante C telle que pour tout intervalle I nous avons

$$\frac{1}{|I|} \int_I \frac{1}{g(G^{-1})} \leqslant C.$$

Or l'hypothèse (H_p) émise par Kerkyacharian et Picard (2004) était qu'il existe une constante C telle que pour tout intervalle I nous avons

$$\frac{1}{|I|} \int_I \frac{1}{g(G^{-1})} \leqslant C \left(\frac{1}{|I|} \int_I g(G^{-1})^{q/p} \right)^{-p/q},$$

avec $1/p + 1/q = 1$. Il est alors immédiat que G^{-1} lipschitzienne implique (H_2).

Ce résultat est donc assez restrictif tant en ce qui concerne la régularité de la fonction étudiée f que celle de la densité du plan d'observation. Néamoins, il permet de justifier en partie les bonnes qualités de l'estimateur observées en pratique. Notons de plus que le fait que l'estimation de la fonction h ne fasse pas intervenir les variables $(T_i)_{i=1,\dots,n}$ permet d'estimer la fonction de répartition G à l'aide de l'intégralité du plan d'observation, ce que ne permet pas la procédure de Kerkyacharian et Picard (2004).

Les Figures 7.6 et 7.7 montrent les bonnes qualités visuelles de l'estimation lorsque la fonction G est connue. Elles mettent aussi en évidence, particulièrement la Figure 7.7, que la composition avec G aboutit à de bien meilleurs estimateurs. Notons de plus que dans cette procédure d'estimation, il n'est plus nécessaire de faire appel à un algorithme calculant les valeurs des ondelettes en des points exacts. L'algorithme pyramidal utilisé dans le cadre équidistant peut être appliqué, ce qui représente un gain en terme de temps de calcul non négligeable.

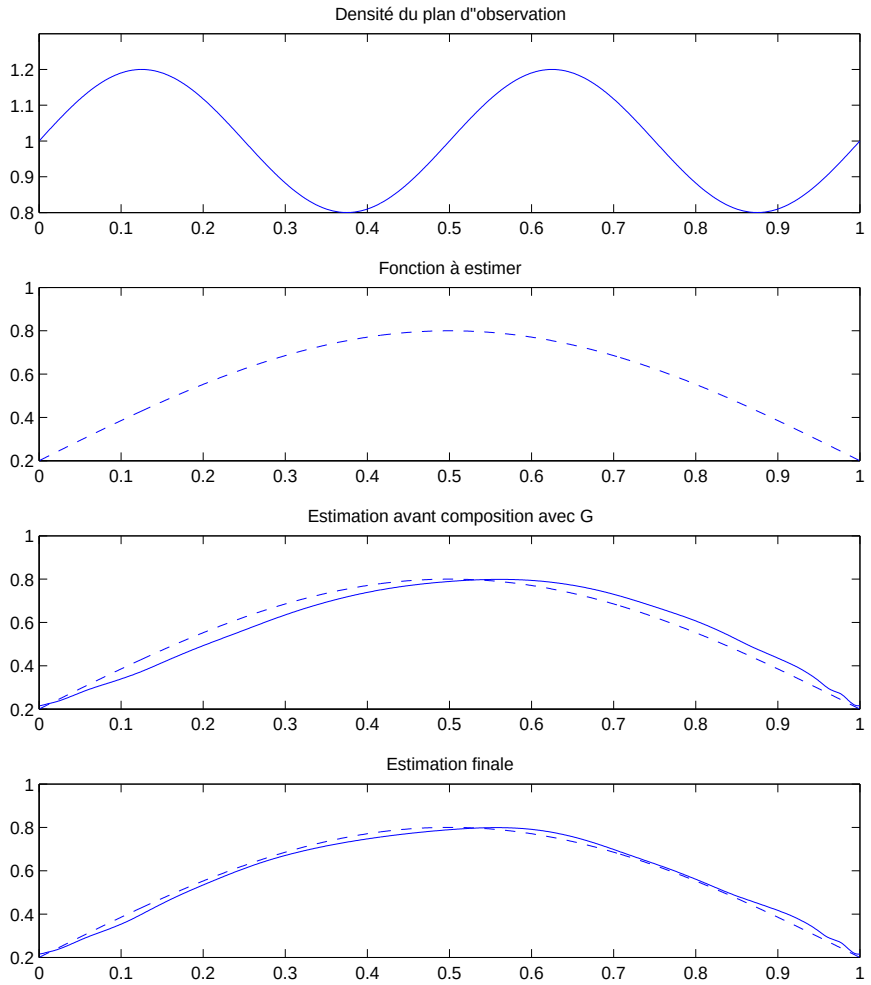

FIG. 7.6: Estimation par l'estimateur de Kerkyacharian et Picard (2004) modifié dans le cas où la fonction de répartition est connue. Sont représentées respectivement la densité g du plan d'observation, la fonction f à estimer, l'estimation f_n obtenue lorsqu'on considère que les observations sont équidistantes et l'estimation $\hat{f}_n$ obtenue en composant avec la fonction de répartition G. La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 9 et l'écart-type du bruit σ est supposé connu.

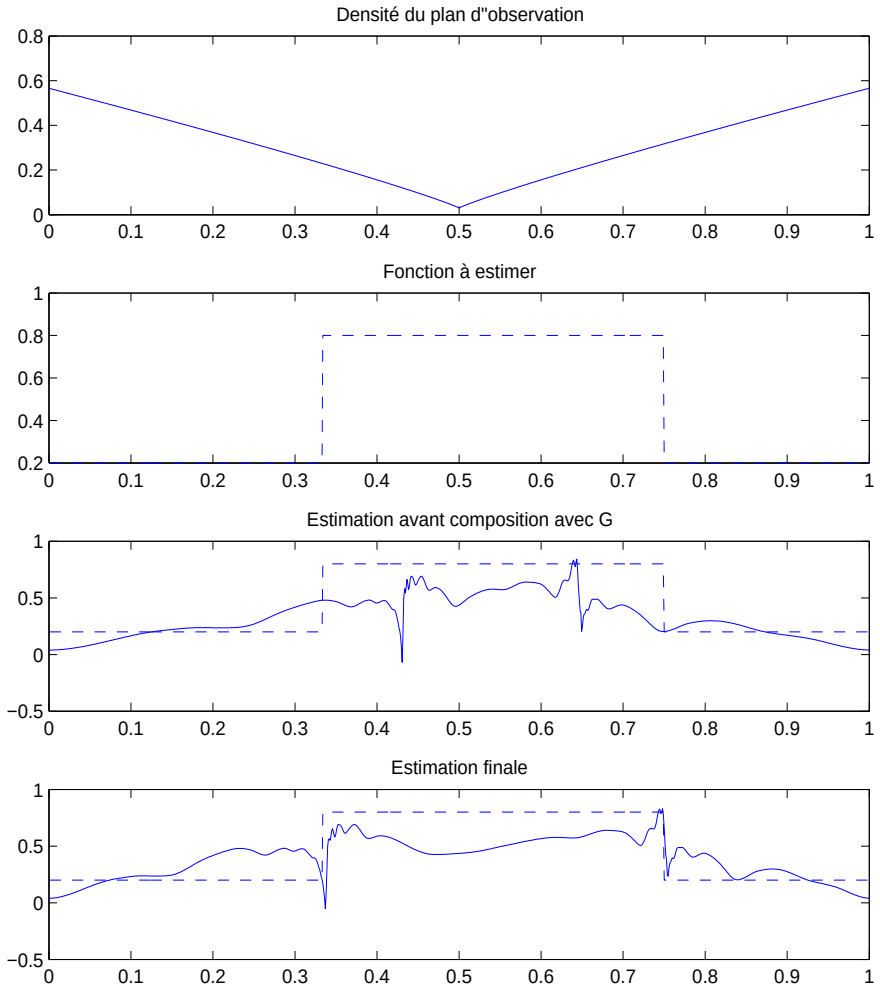

FIG. 7.7: Estimation par l'estimateur de Kerkyacharian et Picard (2004) modifié dans le cas où la fonction de répartition est connue. Sont représentées respectivement la densité g du plan d'observation, la fonction f à estimer, l'estimation f_n obtenue lorsqu'on considère que les observations sont équidistantes et l'estimation $\hat{f}_n$ obtenue en composant avec la fonction de répartition G. La simulation a été réalisée avec $n = 2^8$ observations, un SNR égal à 9 et l'écart-type du bruit σ est supposé connu.

Lorsque la fonction de répartition G est inconnue, l'estimation est toujours satisfaisante. Les figures n'ont pas été mises ici, étant donné qu'il n'est pas possible visuellement de faire la différence avec le cas où la fonction G est connue.

Le problème qui se pose pour l'application de cet estimateur dans un modèle partiellement linéaire est qu'il nécessite de mettre en œuvre une transformée en ondelettes d'ordre $n+1$. Or, il est nécessaire d'avoir une transformée d'ordre n pour que la matrice de régression de la partie linéaire conserve ses propriétés lors de la transformée en ondelettes. C'est pourquoi nous n'appliquons pas les estimateurs ci-dessus dans un modèle partiellement linéaire. Nous prendrons donc les estimateurs suivants :

$$f_n^* = \alpha_{0,0,n}^* \varphi \circ G + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{j,k,n}^* \mathbb{1}_{\{|\beta_{j,k,n}^*| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ G,$$

avec $2^{j_1} = n$, et

$$\beta_{j,k,n}^* = \frac{1}{n} \sum_{i=1}^{n} Y_{(i)} \psi_{j,k}(\frac{i}{n}),$$

lorsque la fonction de répartition G du plan d'observation est connue ; et

$$\hat{f}_n = \alpha_{0,0,n}^* \varphi \circ \hat{G}_n + \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \beta_{j,k,n}^* \mathbb{1}_{\{|\beta_{j,k,n}^*| \geqslant \kappa \sqrt{\frac{\log(n)}{n}}\}} \psi_{j,k} \circ \hat{G}_n,$$

lorsque G est inconnue. Ceci revient en fait à considérer que nous avons un plan d'observations équidistant lors de l'estimation des coefficients d'ondelettes puis à corriger par la fonction G ou son estimation.

Remarquons que ces estimateurs sont très proches de ceux définis ci-dessus. La qualité visuelle des estimateurs est identique et les calculs des MISE sur des simulations aboutissent aux mêmes valeurs. Ces estimateurs furent par ailleurs appliqués par Chesneau (2006) et Chesneau et Willer (2005). Nous renvoyons donc à la thèse de Chesneau pour de plus amples exemples d'applications de ces procédures sur des simulations.

7.2.2 Modèle partiellement linéaire

Mettre en œuvre la procédure d'estimation de la partie fonctionnelle que nous venons de décrire dans un modèle partiellement linéaire consiste à réaliser les étapes suivantes :

1. Application de la transformée en ondelette discrète (DWT) d'ordre n aux observations et aux covariables à l'aide de l'algorithme pyramidal de Mallat (1989) de manière identique au cas équidistant ;

2. Estimation de la variance σ^2 au moyen d'une décomposition QR de la matrice des coefficients d'ondelettes de résolution maximale de la matrice de régression, suivie d'une estimation MAD (bien que ce schéma ne soit plus justifié, n'étant plus assuré du caractère creux des coefficients d'ondelettes) ;
3. Soit estimation conjointe du vecteurs des paramètres de régression β et du vecteur des coefficients d'ondelettes θ de la partie fonctionnelle par *Backfitting*, soit estimation de β par LEGEND puis estimation de θ par application du seuillage à $Z - A\hat{\beta}_n$. Les autres algorithmes ne seront pas considérés au vu des résultats du chapitre précédent ;
4. Et enfin calcul de $\hat{f}_n$ en appliquant la transformée en ondelette inverse à $\hat{\theta}_n$ puis en composant avec l'estimation de la fonction de répartition G.

Reprenons les exemples du chapitre précédent. Nous considérerons trois densités du plan d'observations différentes, que nous nommerons respectivement densité n°1, n°2 et n°3 :

La densité n°1 est la densité uniforme sur $[0, 1]$

La densité n°2 est une densité de forme sinusoïdale, définie à une constante près par $g(t) = 1 + 0.2\sin(4\pi t)$.

La densité n°3 présente une discontinuité. Elle est de la forme $g(t) = |x - a|^{0.9} + 0.03$, avec une constante de normalisation multiplicative. Sauf lorsque cela sera précisé dans l'Exemple 2, le paramètre a, déterminant où se trouve la discontinuité, vaudra 0.5.

Exemple 1

La matrice de régression ainsi que la fonction que nous cherchons à estimer sont celles de l'Exemple 1 du Chapitre 6. La différence est que nous considérons que le plan des observations de la partie fonctionnelle est aléatoirement distribué sur l'intervalle. Nous supposons que le plan d'observation admet une densité g définie sur $[0, 1]$ vérifiant l'hypothèse (H_2). Nous considérons les trois types de densités décrits ci-dessus : une densité uniforme, une densité sinusoïdale et une densité présentant une singularité. Dans chacun de ces trois cas, nous appliquons le schéma d'estimation décrit ci-dessus.

Au vu de la Table 7.1, l'estimation par MAD après une décomposition QR de la matrice de régression aboutit en pratique à de bons résultats. Les estimations obtenues ne dépendent pas de la densité du plan d'observation étant donné que la transformée en ondelette appliquée ne dépend pas du plan d'observation. Comme évoqué plus haut, cette

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.47645(0.042488)

TAB. 7.1: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 1 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

procédure n'a plus le fondement théorique du caractère creux des coeffcicients d'ondelettes de la partie fonctionnelle, étant donné que nous appliquons une DWT pour un plan équidistant à une fonction observée sur un plan aléatoire. Cependant, les bons résultats en pratique de l'estimateur incitent à conserver ce schéma d'estimation de la variance par la suite.

Estimation de β		
Vraie valeur	*Backfitting*	LEGEND
1	0.9976(0.0038)	0.9976(0.0038)
Temps moyen de calcul	0.0058	0.0119

TAB. 7.2: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 1 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$ avec une densité du plan d'observation uniforme. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut 0.2939 pour le *Backfitting* et LEGEND.

La dernière ligne de la Table 7.2 met en évidence que le *Backfitting* reste plus rapide que l'algorithme LEGEND, comme cela avait déjà été observé dans le cadre d'un plan d'observation equidistant. Les ordres de grandeur des temps de calcul sont identiques pour les autres densités et les Exemples 2 et 3, et ne seront donc pas précisés.

Au vu de la Table 7.3 la qualité de l'estimation semble satisfaisante. Le biais ainsi que la variance du bruit sont plus faibles que dans le chapitre précédent en raison du caractère aléatoires des variables de régression. Nous pouvons noter que lorsque le plan des observations est uniforme, la qualité de l'estimation est identique à celle obtenue dans le cadre d'observations équidistantes. Pour les densités n°1 et n°2, nous obtenons des biais et des écart-types un peu plus élevés.

Estimation de β				
Vraie valeur	Estimation			
	Densité n°1	Densité n°2	Densité n°3	Plan équidistant
1	0.9999(0.0038)	1.0071(0.0038)	1.0035(0.0033)	0.9998(0.0039)

TAB. 7.3: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 1 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$ avec un plan d'observation aléatoire. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.5478 pour la densité n°1, 1.0686 pour la densité n°2, 0.7300 pour la densité n°3 et 0.4581 pour le plan équidistant.

Le fait de se placer dans un contexte aléatoire ne semble pas altèrer beaucoup la qualité de l'estimation du paramètre de régression β. Cette conclusion semble logique dans la mesure où la procédure d'estimation est similaire à celle développée dans le cadre d'un plan équidistant.

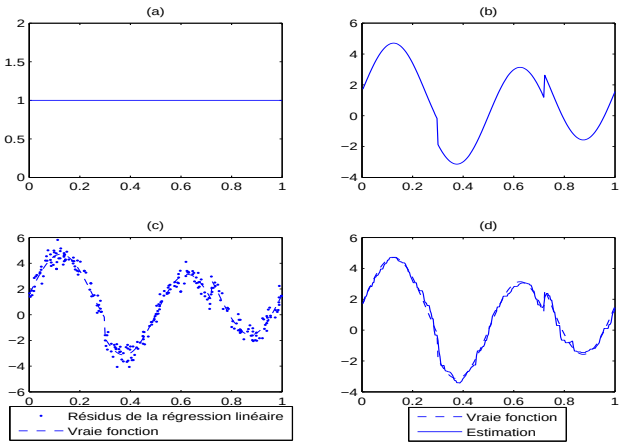

FIG. 7.8: Estimation de la partie non paramétrique dans l'Exemple 1 avec la densité n°1. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

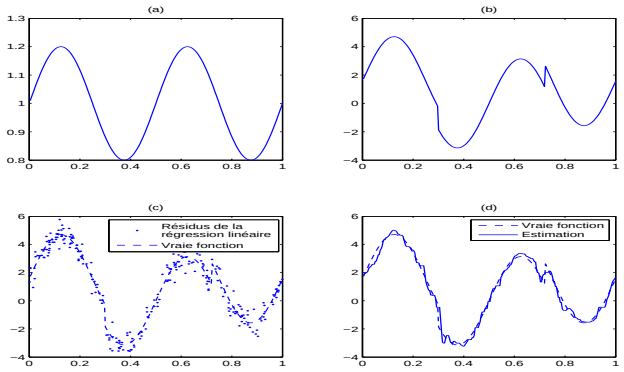

FIG. 7.9: Estimation de la partie non paramétrique dans l'Exemple 1 avec la densité n°2. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

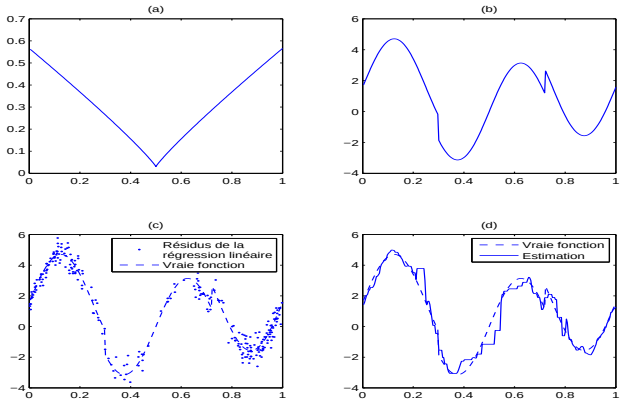

FIG. 7.10: Estimation de la partie non paramétrique dans l'Exemple 1 avec la densité n°3. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

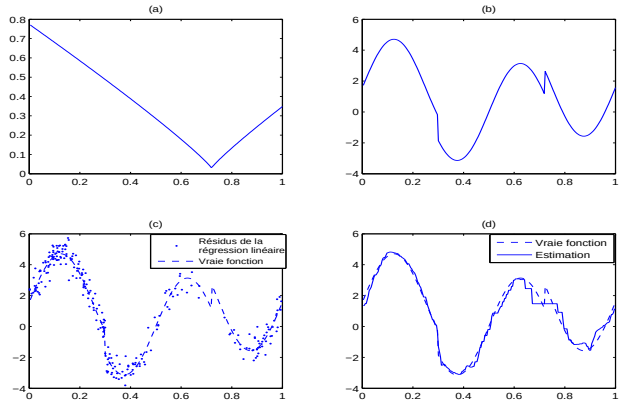

FIG. 7.11: Estimation de la partie non paramétrique dans l'Exemple 1 avec la densité n°3 avec un paramètre $a = 0.7$. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

Les Figures 7.8, 7.9 et 7.10 montrent que les qualités d'estimation de la partie fonctionnelle sont visuellement satisfaisantes. Les estimations sont meilleures aux voisinages des maxima de la densité du plan d'observation (soit là où le nombre d'observations est important) et plus faibles lorsque l'on se situe au voisinage d'un minima de la densité (soit là où le nombre d'observations est moindre). Ceci est particulièrement visible sur les Figures 7.10 et 7.11. Notamment, la discontinuité n'est pas detectée par l'estimation dans le cas où la discontinuité se situe au niveau du minimum de la densité (Figure 7.11), tandis qu'elle est détectée lorsqu'elle ne se situe plus dans le voisinage du minimum (Figure 7.10). A titre de remarque, avec ce SNR, dans le cas d'un plan d'observation équidistant, la discontinuité n'apparaît pas dans l'estimation.

Remarque : L'usage des ondelettes dans la régression fonctionnelle avec un plan d'observation équidistant est encore au stade de la recherche, comme l'a montré le rapide survol bibliographique du Chapitre 2. Dans un tel contexte, les méthodes d'estimations à noyaux s'avèrent plus développées. Leur principal inconvénient réside dans l'incapacité à estimer correctement des fonctions présentant de fortes irrégularités. Les ondelettes permettent d'appréhender des fonctions plus hétérogènes, la régression par ondelettes permettant notamment de détecter des "sauts". Nous avons brièvement comparé

la procédure par ondelettes reposant sur Chesneau et Willer (2005) choisie pour la suite du travail et une technique d'estimation par régression partielle avec noyau (décrite en Section 1.2.2) dans le contexte de l'Exemple 1. Nous avons considéré un noyau gaussien (ce noyau a été pris dans un souci de facilité et n'est pas nécessairement le plus adapté). La procédure de validation croisée appliquée dans l'estimation à noyau est celle de Speckman (1988). La Table 7.4 ainsi que les Figures 7.12, 7.13 et 7.14 illustrent les comportements des deux approches quant à la qualité d'estimation dans les modèles partiellement linéaires sur l'Exemple 1.

Estimation de β				
	Vraie valeur	Estimation		
		Densité n°1	Densité n°2	Densité n°3
Par ondelettes	1	1.0003(0.0032)	1.0003(0.0032)	1.0035(0.0033)
Par noyau	1	1.0134(0.0063)	1.0105(0.0070)	1.0113(0.0058)

Erreur quadratique moyenne de l'estimation de f			
	Estimation		
	Densité n°1	Densité n°2	Densité n°3
Par ondelettes	0.0096	0.0182	0.0098
Par noyau	0.0244	0.0729	0.0334

TAB. 7.4: Valeurs moyennes des estimations et écarts-types pour β (les écarts-types sont donnés entre parenthèses) et erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur 500 simulations dans l'Exemple 1 avec $n = 2^8$ avec un plan d'observation aléatoire.

Il apparaît au vu de la Table 7.4 que la procédure d'estimation par ondelettes semble mener à des estimations de meilleure qualité que l'estimation à noyau mise en œuvre ici. Ce constat est confirmé par les Figures 7.12, 7.13 et 7.14, notamment avec la Densité n°3. La qualité de l'estimation par noyau pourrait être améliorée en considérant une grille plus fine pour le paramètre de lissage dans la procédure de validation croisée. Le choix du noyau de même a probablement une influence sur la qualité de l'estimation. Néanmoins, il semble raisonnable au vu de cette rapide comparaison de penser que la procédure par ondelettes est performante et que les résultats obtenus sont au moins comparables à ceux résultant d'une estimation à noyau.

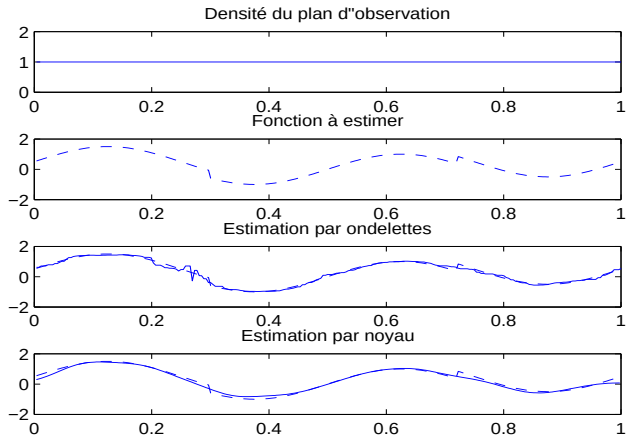

FIG. 7.12: Comparaisons des estimations par ondelettes et par noyau gaussien pour la partie fonctionnelle dans l'Exemple 1 avec un plan d'observation de Densité n°1. La simulation a été réalisée avec $n = 2^8$ observations et un SNR égal à 5.

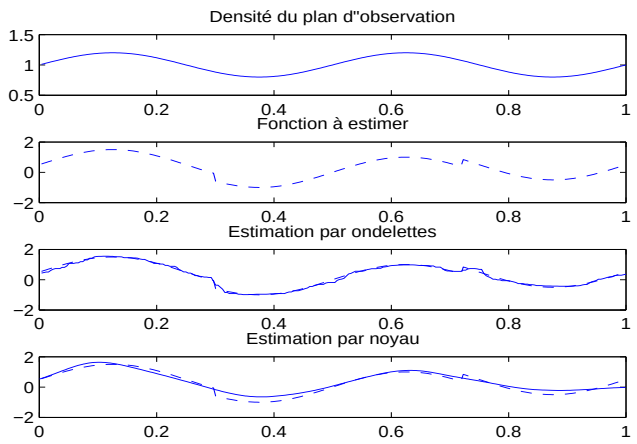

FIG. 7.13: Comparaisons des estimations par ondelettes et par noyau gaussien pour la partie fonctionnelle dans l'Exemple 1 avec un plan d'observation de Densité n°2. La simulation a été réalisée avec $n = 2^8$ observations et un SNR égal à 5.

Dans la Figure 7.15) où nous avons considéré le même modèle mais avec un SNR plus grand, nous pouvons par exemple constater que la procédure par ondelette détecte la singularité de la partie fonctionnelle alors que la procédure à noyau ne la détecte pas.

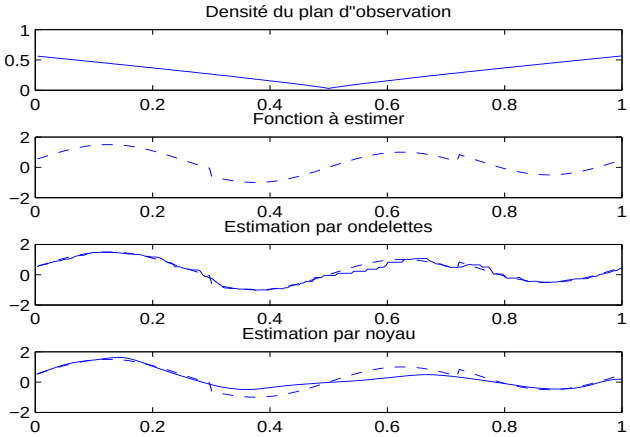

FIG. 7.14: Comparaisons des estimations par ondelettes et par noyau gaussien pour la partie fonctionnelle dans l'Exemple 1 avec un plan d'observation de Densité n°3. La simulation a été réalisée avec $n = 2^8$ observations et un SNR égal à 5.

Cette étude incite donc à développer les procédures par ondelettes dans ce contexte.

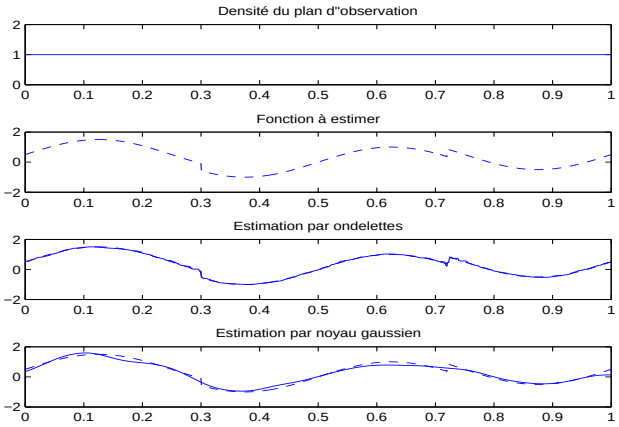

FIG. 7.15: Comparaisons des estimations par ondelettes et par noyau gaussien pour la partie fonctionnelle dans l'Exemple 1 avec un plan d'observation de Densité n°1. La simulation a été réalisée avec $n = 2^8$ observations et un SNR égal à 10.

Exemple 2

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.47751(0.043152)

TAB. 7.5: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 2 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

Estimation de β			
Vraie valeur	Estimation		
	Densité n°1	Densité n°2	Densité n°3
1	1.003(0.0031)	1.0077(0.0029)	1.0019(0.0038)

TAB. 7.6: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 2 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$ avec un plan d'observation aléatoire. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.5858 pour la densité n°1, 0.1012 pour la densité n°2 et 0.7173 pour la densité n°3.

Les qualités d'estimation données en Table 7.6 sont comparables à celles obtenues dans l'Exemple 1.

L'erreur quadratique moyenne de l'estimation de la partie fonctionnelle est plus faible lorsque la densité du plan d'observation est la densité uniforme ou est sinusoïdale.

Visuellement, la qualité de l'estimation est correcte. Nous pouvons toutefois observer un décalage entre le saut de la fonction estimée et celui de la vraie fonction dans le cas où la densité du plan des observations est sinusoïdale (Densité n°2). Ceci s'explique par les faibles valeurs de la densités au voisinage du saut. Pour les autres densités, le saut de la fonction est mieux détecté. Les valeurs prises par les densités du plan d'observation étant relativement éloignées de 0 au voisinage du saut, ceci est cohérent.

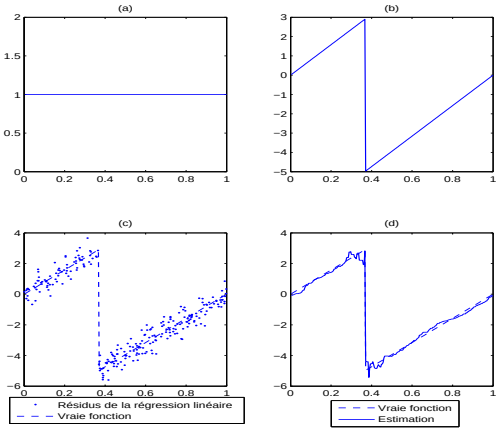

FIG. 7.16: Estimation de la partie non paramétrique dans l'Exemple 2 avec la densité n°1. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) les observations sont donnéess ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

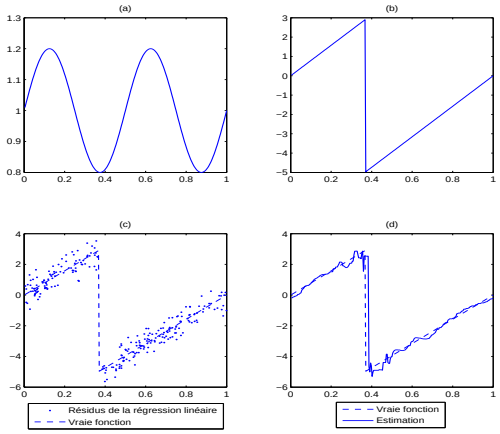

FIG. 7.17: Estimation de la partie non paramétrique dans l'Exemple 2 avec la densité n°2. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

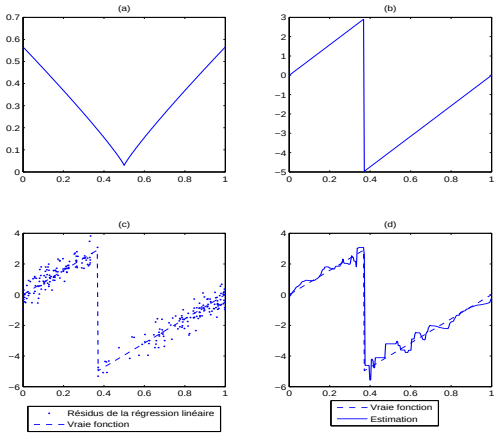

FIG. 7.18: Estimation de la partie non paramétrique dans l'Exemple 2 avec la densité n°3. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

Exemple 3

Nous considérons dans cet exemple le cas où le vecteur des paramètres de régression β appartient à $\mathbb{R}^5$. La partie non paramétrique f est la même que dans l'Exemple 2, c'est-à-dire que la fonction présente de fortes discontinuités.

Estimation de σ par MAD avec décomposition QR	
Vraie valeur	Estimation
0.5	0.53568(0.053989)

TAB. 7.7: Valeurs moyennes et écarts-types des estimations de σ sur 500 simulations de l'Exemple 3 avec $n = 2^8$ observations (l'écart-type est donné entre parenthèses).

Les résultats montrent de même que dans les deux exemples précédents que la qualité d'estimation du paramètre de régression est satisfaisante pour les trois densités considérées. Nous pouvons voir aussi que la meilleure qualité de l'estimation fonctionnelle est obtenue pour la densité uniforme (Densité n°1) et que la moins bonne qualité correspond à la densité présentant une singularité (Densité n°2).

Estimation de β			
Vraie valeur	Estimation		
	Densité n°1	Densité n°2	Densité n°3
-1	-0.3695(0.2626)	-0.3922(0.2447)	-0.4184(0.2590)
5	5.0066(0.0486)	5.0088(0.0478)	5.0070(0.0482)
0	0.4025(0.1660)	0.4095(0.1801)	0.4052(0.1838)
-2	-2.1288(0.1221)	-2.1295(0.1215)	-2.1260(0.1168)
10	9.6128(0.1067)	9.6105(0.1043)	9.6071(0.1097)
Erreur quadratique moyenne	0.9224	0.9088	0.8953

TAB. 7.8: Valeurs moyennes des estimations et écarts-types sur 500 simulations dans l'Exemple 3 (les écarts-types sont donnés entre parenthèses) avec $n = 2^8$ avec un plan d'observation aléatoire. L'erreur quadratique moyenne pour l'estimation de la partie non paramétrique sur ces simulations vaut respectivement 0.5809 pour la densité n°1, 0.6437 pour la densité n°2 et 0.7363 pour la densité n°3.

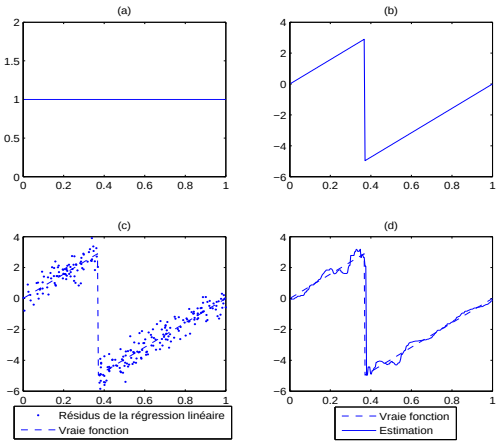

FIG. 7.19: Estimation de la partie non paramétrique dans l'Exemple 3 avec la densité n°1. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

L'ordre de grandeur des erreurs quadratiques moyennes ainsi que les qualités visuelles des estimations sont semblables à celle obtenues dans l'Exemple 2.

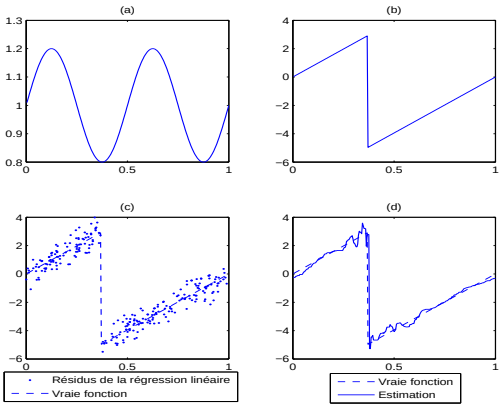

FIG. 7.20: Estimation de la partie non paramétrique dans l'Exemple 3 avec la densité n°2. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

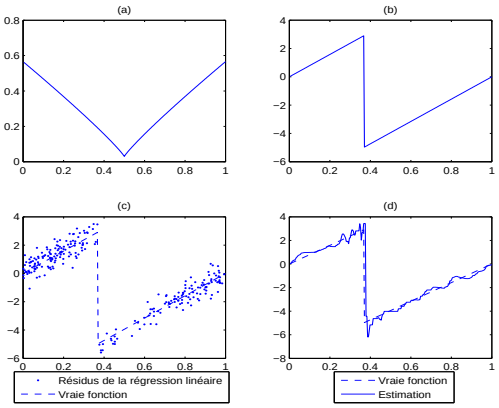

FIG. 7.21: Estimation de la partie non paramétrique dans l'Exemple 3 avec la densité n°3. La figure (a) représente la densité du plan d'observation, la figure (b) la fonction que l'on cherche à estimer. En figure (c) sont donnés les résidus après estimation de la partie linéaire ainsi que la fonction à estimer (en trait discontinu). L'estimation obtenue est représentée en figure (d) (en trait continu) ainsi que la fonction à estimer (en trait discontinu).

Conclusion

Dans ce chapitre, nous avons généralisé le schéma d'estimation au cas où le plan d'observation de la partie fonctionnelle est aléatoire. La procédure envisagée consistait à reprendre l'estimation de la partie fonctionnelle proposée par Kerkyacharian et Picard (2004) et décrite au Chapitre 2.

Lors de la mise en œuvre des schémas d'estimations sur des simulations, il est apparu que la procédure développée par Kerkyacharian et Picard (2004) ne donnait pas une qualité visuelle d'estimation satisfaisante. Nous avons alors opté pour l'alternative de Chesneau et Willer (2005) qui consiste à considérer que le plan est équidistant puis à apporter une correction de l'estimation en composant avec la fonction de répartition des observations. Les résultats obtenus sur les simulations se sont avérés de bonne qualité et ce pour une large gamme de densités.

Conclusion et Perspectives

Cette thèse est consacrée à l'inférence statistique dans un modèle partiellement linéaire. Un tel modèle est constitué de deux parties : une partie paramétrique de régression linéaire et une partie non paramétrique de régression fonctionnelle. L'idée était d'introduire dans ce modèle des techniques d'estimation par ondelettes, notamment des procédures non linéaires permettant d'élaborer une estimation de la partie fonctionnelle qui soit adaptative quant à la régularité de la fonction estimée.

L'écriture des estimateurs comme solutions d'un problème des moindres carrés pénalisés a permis de construire des estimateurs ayant la forme recherchée pour la partie fonctionnelle. Nous établissons l'équivalence entre la minimisation du critère des moindres carrés et une estimation en deux étapes. Nous pouvons estimer le vecteur des paramètres de régression à l'aide d'un M-estimateur usuel, puis appliquer la technique d'estimation par ondelettes avec seuillage sur les résidus résultant. Le lien entre le M-estimateur de la première étape et le seuillage utilisé dans l'estimation fonctionnelle est établi.

Cette caractérisation de l'estimation est intéressante tout d'abord dans la mesure où elle permet une interprétation nouvelle des estimateurs. Elle nous a aussi permis d'établir les propriétés asymptotiques de deux procédures. Notamment, pour la pénalisation menant au seuillage doux, nous obtenons des estimateurs quasi-minimax pour la partie linéaire et pour la partie fonctionnelle du modèle lorsque la structure des covariables est similaire à celle considérée par Rice (1986). Cette procédure présente l'intérêt par rapport aux méthodes existantes d'être adaptative quant à la régularité de la partie fonctionnelle.

En s'appuyant sur cette écriture des estimateurs, nous avons aussi proposé des algorithmes de calcul des estimateurs. Nous les avons exécutés sur des données simulées et comparé à d'autres algorithmes existants. Les performances des algorithmes étaient satisfaisantes. Nous avons aussi appliqué la procédure d'estimation étudiée dans le cadre de données réelles. Celles-ci consistaient en des signaux issus d'Imagerie à Résonance

Magnétique fonctionnelle. Le niveau de bruit relativement élevé ne nous a cependant pas permis de conclure à la validité du modèle.

Les perspectives de ce travail sont nombreuses. Tout d'abord, notre étude s'est focalisée sur deux seuillages. Il serait intéressant d'étendre les propriétés asymptotiques établies à d'autres formes de seuillage, par exemple aux seuillages fort et SCAD. (L'étude des propriétés asymptotiques dans le cadre d'un seuillage fort est actuellement en cours.) Les seuillages décrits dans Antoniadis et Fan (2001) comme résultants d'une procédure de moindres carrés pénalisés sont adaptés. Une forme plus générale des résultats présentés permettrait une meilleure lisibilité de l'application des ondelettes et du seuillage dans un modèle partiellement linéaire. Par ailleurs les résultats énoncés sont donnés pour la convergence en probabilité et l'extension à une convergence presque sûre peut être étudiée.

Ensuite, il serait intéressant de prolonger ce travail en étudiant des modèles partiellement linéaires généralisés.

Nous pouvons aussi complexifier la structure des variables, notamment en introduisant une dépendance au sein des variables de régression. Le fait de considérer un plan d'observation de la partie fonctionnelle avec dépendance nécessitant toutefois une estimation par ondelettes et seuillage avec données dépendantes, ce point est délicat. Nous pourrions étudier la faisabilité d'une estimation en deux étapes, telle que proposée par Chen et Shiau (1991), qui pourrait permettre éventuellement une corrélation plus forte entre les variables des parties linéaire et fonctionnelle du modèle.

Nous pouvons citer une extension vers la MM-estimation, qui consisterait à définir un critère estimant la variance du bruit conjointement au vecteur de paramètre et à la fonction. Ou encore la recherche d'un critère qui permettrait, à partir du schéma d'estimation choisi, de déterminer quelle variable serait la plus appropriée dans la partie fonctionnelle du modèle, de manière similaire à ce qu'ont développé Chen et Chen (1991).

Bibliographie

Amato, U., Antoniadis, A. et Pensky, M. (2006). Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, **16**(1), 37 - 56.

Antoniadis, A. (1992). Taux de convergence pour des modèles de régression à inf-convolution spline et applications. *Statistique et Analyse des Données*, **16**(2), 20–32.

Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, **23**(3), 313–330.

Antoniadis, A.(1997). Wavelets in statistics : a review. Technical Report. Université Joseph Fourier, Grenoble.

Antoniadis, A. et Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96**(455), 939–967.

Antoniadis, A., Grégoire, G. et McKeague, I. (1994). Wavelet methods for curve estimation. *American Statistical Estimation*, **89**(428), 1340–1353.

Antoniadis, A. et Pham, D. (1998). Wavelet regression for random or irregular design. *Computational Statistics and Data Analysis*, **28**(4), 353 - 369.

Bai, Z., Rao, C. et Wu, Y.(1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, 237–254.

Cai, T. et Brown, L. (1998). Wavelet shrinkage for nonequispaced design. *The Annals of Statistics*, **26**(5), 1783–1799.

Cai, T. et Brown, L.(1999). Wavelet estimation for samples with random uniform design. *Statistics and Probability Letters*, **42**, 313–321.

Chang, X.-W. et Qu, L. (2004). Wavelet estimation of partially linear models. *Computationnal Statistics and Data Analysis*.

Chen, H. (1987). Estimation of semiparametric generalized linear models. Technical Report. State University of New York.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, **16**(1), 136–146.

Chen, H. et Chen, K.-W. (1991). Selection of the splined variables and convergence rates in a partial spline model. *The Canadian Journal of Statistics*, **19**(3), 323–339.

Chen, H. et Shiau, J.-J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference*, **27**, 187–201.

Chesneau, C. (2006). Quelques contributions à l'estimation fonctionnelle par méthode d'ondelettes. Ph. D. dissertation, Université Paris 6.

Chesneau, C. (2007). Regression in random design : a minimax study. *Statistics and Probability Letters*, **77**, 40–53.

Chesneau, C. et Willer, T. (2005). Numerical performances of a warped wavelet estimation procedure for regression in random design. Technical Report. Universités Paris 6-Paris 7.

Chui, C. (1992). *Wavelet : a tutorial in theory and applications.* Academic Press, New-York.

Dahyot, R., Charbonnier, P. et Heitz, F. (2000). Robust visual recognition of colour images. *CVPR*, **1**, 685–690.

Dahyot, R., Charbonnier, P. et Heitz, F. (2004). A bayesian approach to object detection using probabilistic appearance-based models. *Pattern Analysis and Application*, **7**, 317–332.

Dahyot, R. et Kokaram, A. (2004). Comparison of two algorithms for robust m-estimation of global motion parameters.

Daubechies, I. (1992). *Ten lectures on wavelets. (Vol. 61)*. SIAM press.

Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics*, **18**(4), 1651–1675.

Delouille, V., Franke, J. et Sachs, R. von. (2001). Nonparametric stochastic regression with design-adapted wavelets. *Sankhya : The Indian Journal of Statistics, Special issue on Wavelets, Series A*, **63**(3), 328–366.

Delouille, V., Simoens, J. et Sachs, R. von. (2001). Smooth design-adapted wavelets for nonparametric stochastic regression. Technical Report No. 0117. Université Catholique de Louvain.

Donald, S. et Newey, W. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, **50**, 30–40.

Donoho, D. (1992). De-noising by soft-thresholding. Technical Report. Department of Statistics, Stanford University.

Donoho, D. et Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Donoho, D. et Johnstone, I. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **26**(3), 879–921.

Donoho, D., Johnstone, I., Kerkyacharian, G. et Picard, D. (1995). Wavelet shrinkage : asymptotia ? *Journal of Royal Statistics Society*, **57**(2), 301–369.

Engle, R., Granger, C., Rice, J. et Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**(394), 310–320.

Fadili, J. et Bullmore, E. (2005). Penalized partially linear models using sparse representation with an application to fMRI time series. *IEEE Transactions on Signal Processing*, **53**(9), 3436–3448.

Fan, J. et Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fuchs, J. (1999). An inverse problem approach to robust regression. *IEEE - ICASSP*, **4**, 1809–1812.

Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, **4**(17), 293–310.

Geman, D. et Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions of Pattern Analysis of Machine Intelligence*, **14**, 367–383.

Geman, D. et Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Transaction on Image Processes*, **4**, 932–946.

Green, P. et Yandell, B. (1985). Semi-parametric generalized linear models. Technical Report No. 2847. University of Wisconsin-Madison.

Hall, P. et Turlach, B. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *The Annals of Statistics*, **25**(5), 1912–1925.

Hamilton, S. et Truong, Y. (1997). Local estimation in partly linear models. *Journal of Multivariate Analysis*, **60**, 1–19.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. et Stahel, W. A. (1986). *Robust statistics : The approach based on influence functions.* Wiley Series in Probability and Mathematical Statistics.

Hardle, W., Kerkyacharian, G., Picard, D. et Tsybakov, A. (1998). *Wavelets approximation and statistical applications. (Vol. 129).* Springer Verlag, Lecture Notes in Statistics.

Hardle, W., Liang, H. et Gao, J. (2000). *Partially linear models.* New-York : Springer-Verlag.

Hardle, W., Liang, H. et Sommerfeld, V. (1999). Bootstrap approximations in a partially linear regression model. (Bootstrap in partially linear model)

Heckman, N. (1986). Spline smoothing in partly linear model. *Journal of Royal Statistical Society*, **48**(2), 244–248.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Huber, P. (1981). *Robust statistics.* Wiley Series in Probability and Mathematical Statistics.

Ibragimov, I. et Has'Minskii, R. (1980). *Statistical estimation : asymptotic theory.* Springer, New York Berlin.

Kerkyacharian, G. et Picard, D. (2004). Regression in random design and warped wavelets. *Bernouilli*, **10**(6), 1053-1105.

Kovak, A. et Silverman, B. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of American Statistics*, **95**, 172–183.

Loubes, J. et Van der Geer, S. (2002). Adaptative estimation with soft thresholding type penalties. *Statistica Neerlandica*, **56**(4), 454–479.

Mallat, S. (1989). A theory for multiresolution signal decomposition : the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

Mallat, S. (1999). *A wavelet tour on signal processing. (2 ed.).* Academic Press.

Mammen, E. et Van der Geer, S. (1997). Penalized quadi-likelihood estimation in partial linear models. *The Annals of Statistics*, **25**(3), 1014–1035.

Maronna, R., Bustos, O. et Yohai, V. (1979). Bias and efficiency robustness of general M-estimators for regression with random carriers. In T. Gasser and M. Rosenblatt eds. (Ed.), Smoothing techniques for curve estimation. (pp. 91–1169). Lecture Notes in Statistics, Springer, Berlin.

Maxim, V. (2003). Restauration de signaux bruités observés sur des plans d'expérience aléatoires. Ph. D. dissertation, Université Grenoble I.

Meyer, F. (2003). wavelet-based estimation of a semiparametric generalized linear model of fmri time-series. *IEEE transactions on medical imaging*, **22**, 315–324.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge University Press.

Müller, C. (2004). Redescending M-estimators in regression analysis, cluster analysis and image analysis. *Discussiones Mathematicae - Probability and Statistics*(24), 59–75.

Nikolova, M. et Ng, M. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SCIAM Journal on Scientific Computing*, **27**(3), 937–966.

Oh, H.-S., Nychka, D. et Lee, T. (2007). The role of pseudo data for robust smoothing with application to wavelet regression. (A paraître)

Pensky, M. et Vidakovic, B. (2001). On non-equally spaced wavelet regression. *Annals of Instute of Statistical Mathematics*.

Pollard, D. (1990). *Empirical processes : theory and applications. (Vol. 2).* NFS-CBMS regional conference series in probability and statistics.

Qu, L. (2005). Bayesian wavelet estimation of partially linear models. *Journal of Statistical Computation and Simulation*.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics ans Probability Letters*, **4**, 203–208.

Rockafellar, R. (1970). *Convex analysis.* Princeton University Press.

Roger-Vial, C. (2003). Deux contributions à l'étude semiparamétrique d'un modèle de régression. Ph. D. dissertation, Université Rennes I.

Sardy, S., Percival, D., Bruce, A., Gao, H. et Stuelze, W. (1998). Wavelet shrinkage for unequally spaced data. Technical Report. MathSoft, Inc.

Schick, A. (1996). Root-n-consistent and efficient estimation in semiparametric additive regression models. *Statistics ans Probability Letters*, **30**, 45–51.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of Royal Statistical Society*, **50**(3), 413–436.

Vidakovic, B. (2002). Pollen bases and daubechies-lagarias algorithm in matlab.

Vik, T. (2004). Modèles statistiques d'apparence non gaussiens. application à la création d'un atlas probabiliste de perfusion cérebrale en imagerie médicale. Ph. D. dissertation, Université Strasbourg 1.

Willer, T. (2006). Estimation non paramétrique et problèmes inverses. Ph. D. dissertation, Université Paris 7.

Wu, Y. (2004). M-estimation of linear models with dependent errors. Technical Report. University of Chicago.

Xue, L. (2003). Rates of random weighting approximation of wavelet estimates in semi-parametric regression model. *Acta Mathematicae Applicatae Sinica*, **26**(1), 1–25.

ESTIMATION PAR ONDELETTES DANS LES MODÈLES PARTIELLEMENT LINÉAIRES

RÉSUMÉ : L'objet de cette thèse est d'apporter une contribution à l'inférence dans les modèles partiellement linéaires en appliquant des méthodes d'estimation adaptative par ondelettes. Ces modèles de régression semi-paramétriques distinguent des relations linéaires et des relations fonctionnelles, non paramétriques. Une procédure des moindres carrés pénalisés permet d'introduire une estimation par ondelettes avec seuillage des coefficients de la partie fonctionnelle. Un parallèle est établi avec une estimation du paramètre de régression par des M-estimateurs usuels dans un modèle linéaire, les coefficients d'ondelettes de la partie fonctionnelle étant considérés comme des valeurs aberrantes. Des résultats relatifs aux propriétés asymptotiques des estimateurs de la partie linéaire et de la partie fonctionnelle sont démontrés. Des simulations permettent d'illustrer les comportements des estimateurs et de les comparer avec d'autres méthodes existantes. Une application sur des données d'IRMf a aussi été réalisée.

MOTS-CLEFS : Modèles semi-paramétriques, modèles partiellement linéaires, ondelettes, estimation fonctionnelle adaptative, régression à pas aléatoires, backfitting, M-estimation, moindres carrés pénalisés.

WAVELET ESTIMATION IN PARTIALLY LINEAR MODELS

ABSTRACT : This dissertation is concerned with the use of wavelet methods in semiparametric partially linear models. These models are composed by a linear component with unknown regression coefficients and an unknown nonparametric function. A wavelet thresholding based procedure is built to estimate the nonparametric part of the model using a penalized least squares criterion. We establish a connection between different thresholding schemes and M-estimators in linear models with outliers. Some asymptotic results of the estimates of both the parametric and the nonparametric part are given. Simulations illustrate the behavior of the proposed methodology and compare it with existing methods. An application to real data from functional IRM is also presented.

KEYWORDS : Semiparametric modelling, partially linear models, wavelet thresholding, regression in random design, backfitting, M-estimation, penalized least squares.