Estimation bayésienne non paramétrique

Vincent Rivoirard

UNIVERSITÉ PARIS 7 - DENIS DIDEROT
UFR de Mathématiques

THÈSE

pour l'obtention du Diplôme de

DOCTEUR DE L'UNIVERSITÉ PARIS 7

Spécialité : MATHÉMATIQUES APPLIQUÉES

présentée par

Vincent RIVOIRARD

Titre :
ESTIMATION BAYÉSIENNE NON PARAMÉTRIQUE

Directeur de thèse : **Dominique PICARD**

Soutenue publiquement le **13 Décembre 2002**, devant le jury composé de

M. Lucien BIRGÉ, Université Paris 6
Mme Laure ÉLIE, Université Paris 7
M. Youri GOLUBEV, Université Aix-Marseille 1
M. Gérard KERKYACHARIAN, Université Paris 10
M. Oleg LEPSKI, Université Aix-Marseille 1
Mme Dominique PICARD, Université Paris 7
M. Alexandre TSYBAKOV, Université Paris 6
Mme Sara van de GEER, Université de Leiden

au vu des rapports de **M. Youri GOLUBEV** et **Mme Sara van de GEER**.

Cette thèse n'aurait pu voir le jour sans le concours de nombreuses personnes. Aussi, je tiens à remercier ici tous ceux qui, d'une manière ou d'une autre, m'ont permis de mener à terme ce projet.

En tout premier lieu, je tiens à exprimer ma gratitude à l'égard de Dominique Picard qui a accepté d'accompagner mes premiers pas dans la recherche en dirigeant mes travaux. Elle a su me faire profiter de sa profonde connaissance de tous les sujets abordés. Sa confiance sans cesse renouvelée et sa rigueur bienveillante ont joué un grand rôle dans l'aboutissement de cette thèse.

J'adresse ma plus vive reconnaissance à Sara van de Geer et Youri Golubev qui ont bien voulu consacrer une partie de leur temps à rapporter cette thèse.

Je tiens à remercier chaleureusement Lucien Birgé, Laure Élie, Gérard Kerkyacharian, Oleg Lepski et Alexandre Tsybakov qui m'ont fait l'honneur d'accepter de faire partie du jury. Leur présence aujourd'hui me réjouit.

Marc Hoffmann, Gérard Kerkyacharian, Jacqueline Mac Aleese, Thierry Meyre et Yves Rozenholc ont fait preuve à mon égard de beaucoup de disponibilité. Leur compétence et leurs conseils toujours avisés m'ont été précieux. Qu'ils en soient ici remerciés.

Ce fut un vrai plaisir d'effectuer mon service de moniteur sous la responsabilité de Gabrielle Viennet, qui a su mêler pendant ces trois années son professionnalisme d'enseignante et sa décontraction naturelle.

Je suis également très reconnaissant envers toutes les personnes qui, autour de Laure Élie, m'ont permis de bénéficier d'excellentes conditions de travail. Un merci tout spécial à Michèle Wasse.

Le bureau des thésards 5B1 a été pour moi un lieu de travail stimulant et chaleureux. Je veux remercier chacun de ses membres (auxquels il me faut associer Aurélie, Hélène, José, Karine et Tristan) pour tout le bonheur qu'ils ont apporté lors de nos petites "Pauses :->" quotidiennes. J'ai une pensée particulière pour Caroline, Florent et Véronique avec qui j'ai partagé des moments extraordinaires.

Bien évidemment, je n'oublie pas d'associer à ces remerciements mes proches : famille, amis (et notamment "les Cachanais"). Qu'il me soit pardonné de ne point nommer chacun d'entre eux.

Enfin, je voudrais adresser mes plus tendres remerciements à Alice qui a su être présente à mes côtés durant toutes ces années. *Ton soutien sans faille m'a été indispensable.*

A mes parents

Table des matières

1 Introduction **11**

1.1 L'approche bayésienne 11

1.1.1 Contexte 11

1.1.2 Ondelettes 12

1.1.3 Modélisation 14

1.2 Le point de vue maxiset 15

1.2.1 Minimax contre maxiset 15

1.2.2 Espaces de Lorentz et espaces maximaux 17

1.2.3 Espaces de Besov faibles 20

1.3 Interactions entre approches bayésienne et déterministe 23

1.4 Principaux résultats 24

1.4.1 Estimation minimax sur les espaces de Besov faibles et applications 24

1.4.2 Maxisets pour les procédures linéaires et bayésiennes 29

1.5 Perspectives 32

2 Préliminaires et compléments **35**

2.1 Construction des bases d'ondelettes et applications 35

2.1.1 Analyse multirésolution 35

2.1.2 Espaces de Besov forts 37

2.2 Modèles statistiques 39

2.2.1 Modèle de régression et transformée en ondelettes discrète 39

2.2.2 Modèle de bruit blanc gaussien 40

2.2.3 Problèmes inverses et modèle hétéroscédastique 41

2.3 Quelques résultats antérieurs de la théorie maxiset 43

2.3.1 Modèle de l'estimation d'une densité 43

2.3.2 Modèle de bruit blanc gaussien homoscédastique 44

2.3.3 Modèle hétéroscédastique. Bases inconditionnelles. Superconcentration . . 46

3 Non linear estimation over weak Besov spaces 51

3.1 Introduction . . . 51
3.2 Weak Besov spaces and model . . . 54
3.2.1 Overview of strong Besov spaces . . . 54
3.2.2 Definition of weak Besov spaces . . . 55
3.2.3 The statistical model . . . 57
3.3 The Bayesian approach . . . 58
3.3.1 Asymptotically least favorable priors . . . 58
3.3.2 Minimax Bayes risk . . . 58
3.4 Results, discussions and simulations . . . 60
3.4.1 Notations and technical tools . . . 60
3.4.2 Main results and comments . . . 62
3.4.3 Thresholding rules via a Bayesian framework . . . 64
3.4.4 Comparison of $\mathcal{B}_{s,p,q}(C)$ with $\mathcal{WB}_{s,p,q}(C)$ from the statistical point of view 65
3.4.5 Realizations from asymptotically least favorable priors . . . 65
3.5 Proof of Theorem 3.1 . . . 68
3.5.1 Upper bound . . . 69
3.5.2 The prior π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$ . . . 75
3.6 Appendix : Proof of Theorem 3.3 . . . 81

4 Bayesian thresholding with priors based on Pareto distributions 87

4.1 Introduction . . . 87
4.2 Estimation over weak Besov spaces . . . 90
4.2.1 Wavelet series representation . . . 90
4.2.2 Sparsity and definition of weak Besov spaces . . . 91
4.2.3 Minimax risk and least favorable priors . . . 93
4.3 Construction of functions typical of weak Besov spaces . . . 96
4.3.1 The prior model . . . 96
4.3.2 The main result and simulations . . . 97
4.4 Thresholding rules . . . 100
4.4.1 Model and discrete wavelet transform . . . 100
4.4.2 Choice of the threshold . . . 101
4.4.3 Examples and discussion . . . 105

4.5 Appendix : Proof of Theorem 4.1 108

5 Maxisets for linear procedures 117

5.1 Introduction 117
5.2 Model and function spaces 119
5.2.1 Heteroscedastic white noise model 119
5.2.2 Function spaces 120
5.3 Maxisets and linear estimates 122
5.3.1 Our main result 122
5.3.2 Maxisets for the L_p-risk and with projection weights 129
5.4 Appendix : Proof of Propositions 5.1 and 5.2 132

6 Bayesian modelization of sparse sequences and maxisets for Bayes rules 137

6.1 Introduction 137
6.2 Model and maxiset point of view 140
6.2.1 Heteroscedastic white noise model 140
6.2.2 Sequence spaces and sparsity 141
6.2.3 Thresholding rules 142
6.3 Bayesian procedures 143
6.3.1 The prior model 143
6.3.2 Bayes rules 144
6.3.3 Maxisets for the posterior median 145
6.3.4 Maxisets for the posterior mean 148
6.3.5 L_p-risk for Bayesian procedures $(1 < p < \infty)$ 153
6.4 Relationships between (M_1) and $wl_{p,q}(\sigma)$ spaces 160
6.5 Appendix : Proof of Propositions 6.3 and 6.4 163

Bibliographie 169

Chapitre 1

Introduction

1.1 L'approche bayésienne

1.1.1 Contexte

Les travaux de cette thèse ont pour objet l'étude de quelques aspects modernes de statistique non paramétrique. Nous supposerons données n observations $X_1, \ldots, X_n$ issues d'un modèle statistique très général noté $(\mathcal{X}_n, \mathcal{A}_n, \mathbb{P}_f^{(n)}, f \in \mathcal{F}(\mathbb{R}^d, \mathbb{R}^{d'}))$, où $\mathcal{F}(\mathbb{R}^d, \mathbb{R}^{d'})$ désigne l'ensemble des applications de $\mathbb{R}^d$ dans $\mathbb{R}^{d'}$. La plupart du temps, le statisticien dispose d'informations sur la "régularité" du paramètre f à estimer, qui lui permettent de supposer que f appartient à une classe d'espaces fonctionnels connue. C'est sous cet angle déterministe que sont souvent envisagés les problèmes d'estimation fonctionnelle. Cependant, puisque la connaissance, qui peut être très vague, de la "régularité" de f revient à disposer d'une information *a priori*, il semble naturel d'envisager l'estimation de cette fonction dans un cadre bayésien. Dans ce cas, f est vue comme la réalisation d'un processus stochastique. L'idée d'utiliser une approche bayésienne en estimation fonctionnelle n'est pas neuve puisqu'on la trouve déjà dans un article de Whittle (1958). Elle fut reprise, entre autres, par Kimeldorf et Wahba (1970) et Leonard (1978) qui supposèrent que la distribution a priori de f est celle d'un processus gaussien. Notons que Kimeldorf et Wahba (1970) exhibèrent des estimateurs bayésiens qui peuvent être vus comme des 'L-spline'. Wahba (1978) illustra l'existence de liens entre méthodes bayésiennes et de pénalisation en prouvant que pour le modèle de régression (voir section 2.2.1), si la distribution a priori de f est la même que celle d'un processus X_ξ pouvant s'écrire :

$$X_\xi(t) = \sum_{j=1}^{m} \theta_j \frac{t^{j-1}}{(j-1)!} + \text{Const} \int_0^t \frac{(t-u)^{m-1}}{(m-1)!} dW(u), \quad t \in [0,1], \tag{1.1}$$

où W est le processus de Wiener et $(\theta_1, \ldots, \theta_m)' \sim \mathcal{N}(0, \xi I_m)$, le M-estimateur pénalisé pour la perte l_2 sur la classe

$$W_m = \{f : \quad f, f^{(1)}, \ldots, f^{(m-1)} \text{ sont absolument continues }, f^{(m)} \in L_2[0,1]\}$$

associé à la pénalité $\int_0^1 (f^{(m)}(u))^2 du$ pouvait être vu comme la limite de la moyenne de la loi a posteriori quand $\xi \to +\infty$. Nous verrons par la suite qu'il est très fréquent de chercher à construire des modèles bayésiens de façon à ce que les estimateurs associés soient de "type connu". En fait, pour traiter du problème de l'estimation fonctionnelle, il semblait tout naturel (et l'écriture (1.1) l'encourageait) de supposer que f admettait une décomposition sur une famille, finie ou pas, de polynômes $(\phi_j)_{j \in J}$:

$$f = \sum_{j \in J} \theta_j \phi_j,$$

et de placer un modèle bayésien sur la suite $\theta = (\theta_j)_{j \in J}$. Plusieurs auteurs envisagèrent cette approche. Citons par exemple Young (1977), Wahba (1981), Silverman (1985) et Steinberg (1990) qui utilisèrent des décompositions sur, respectivement, des polynômes de Legendre, des polynômes trigonométriques, des B-spline et des polynômes de Hermite. En général, les θ_j étaient munis d'une loi a priori de type gaussien et étaient estimés par la moyenne de la loi a posteriori. Même si les polynômes trigonométriques offraient l'avantage de constituer une base orthonormée de $L_2(\mathbb{R})$, on ne retrouvait pas une telle unanimité pour décider du choix idéal de la famille $(\phi_j)_{j \in J}$, ce que confirme les exemples précédemment cités. Dans une large mesure, l'apparition des ondelettes en statistique au début des années 1990 bouleversa la donne.

1.1.2 Ondelettes

L'idée de la construction des bases d'ondelettes trouve sa source dans la volonté d'exhiber des bases orthonormées dont les atomes sont à la fois localisés en temps et en fréquence. Car si la base de Fourier est localisée en fréquence, elle ne l'est pas en temps, et des changements de faible amplitude autour d'une fréquence donnée engendrent en général des changements sur la totalité du domaine temporel. Par ailleurs, les bases d'ondelettes se construisent par translations et dilatations dyadiques de deux fonctions, ϕ (la fonction d'échelle) et ψ (l'ondelette mère), ce qui permet de disposer d'une structure algorithmique simple. Il est de plus possible de construire les fonctions ϕ et ψ suffisamment régulières pour générer des bases inconditionnelles pour différentes classes d'espaces fonctionnels tels que les espaces de Besov ou de Triebel (voir Meyer (1992)). Les ondelettes à support compact exhibées par Daubechies (1988) fournissent des bases inconditionnelles pour les espaces $L_p(D)$, $1 < p < \infty$, $D = \mathbb{R}^d$ ou $D = [0,1]^d$. Ainsi, si on dispose d'une base inconditionnelle d'ondelettes, notée $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$, sur un espace normé $(X, \| \, \|)$, alors toute fonction f de X peut s'écrire de manière unique :

$$f = \sum_{j\geq -1} \sum_k \beta_{jk}\psi_{jk},$$

et si $T = (T_{jk})_{jk}$ est un opérateur de contraction (i.e., $\forall\, x \in \mathbb{R}, \quad |T_{jk}(x)| \leq |x|$), alors

$$\| \sum_{j\geq -1} \sum_k T_{jk}(\beta_{jk})\psi_{jk} \| \leq K \| \sum_{j\geq -1} \sum_k \beta_{jk}\psi_{jk} \|,$$

où K est une constante indépendante des β_{jk}. L'opérateur T apparaît donc comme un opérateur de régularisation. Comme les propriétés visuelles d'une procédure d'estimation ne se trouvent pas nécessairement traduites par une distance hilbertienne, il semble naturel de privilégier des procédures qui soient efficaces pour toute une gamme de distance. Ceci motive donc l'emploi de procédures qui allient les décompositions sur une base d'ondelettes et l'utilisation d'estimateurs par contraction. L'analyse temps-fréquence offre l'avantage de fournir des décompositions qui contiennent une grande majorité de petits coefficients et quelques grands coefficients qui portent l'essentiel de l'information contenue dans le signal. Un signal dont la proportion de coefficients non négligeables est faible, est dit *sparse*. Cette notion, pour laquelle nous fournirons un cadre mathématique, reviendra de manière récurrente dans ces travaux. Pour extraire les coefficients significatifs d'un signal, Donoho et Johnstone (1994a) ont introduit des estimateurs par contraction qui étaient aussi des estimateurs de type seuillage : si la taille du coefficient d'ondelette bruité est inférieure à un certain seuil t, le coefficient est estimé par zéro. Plus particulièrement, Donoho et Johnstone (1994a) exhibèrent deux estimateurs de type seuillage qui se révélèrent très performants d'un point de vue pratique :

- l'estimateur par seuillage dur : $d_t^h(x) = x\mathbb{1}_{|x|>t}$,
- l'estimateur par seuillage doux : $d_t^s(x) = \text{sign}(x)(|x| - t)_+$.

Le choix du seuil t est alors essentiel et a fait l'objet de nombreux travaux. Citons Donoho et Johnstone (1994a, 1995), Nason (1996), Abramovich et Benjamini (1995), Ogden et Parzen (1996a, 1996b), Jansen, Malfait et Bultheel (1997) et les diverses approches bayésiennes citées en section 1.1.3. D'un point de vue théorique, les procédures mêlant à la fois "décomposition sur une base d'ondelettes" et "estimateurs par seuillage", s'avérèrent très efficaces comme l'attestent les articles de Donoho, Johnstone, Kerkyacharian et Picard (Donoho et Johnstone (1994a, 1995), Donoho, Johnstone, Kerkyacharian et Picard (1995, 1996, 1997)).
La section 2.1.1 rappelle quelles sont les bases d'ondelettes choisies aux différents stades de notre étude. Mais, jusqu'à la fin de la section 1.4.1, nous supposerons donnée une base orthonormée d'ondelettes de $L_2(\mathbb{R})$, que nous noterons $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$.

1.1.3 Modélisation

Les nombreuses propriétés des ondelettes rappelées dans la section précédente expliquent l'impact généré par l'apparition de ces bases en statistique non paramétrique. Il semble que ce soit Müller et Vidakovic (1995) les premiers qui exploitèrent ces nouveaux outils dans un cadre bayésien. Cet article fut suivi de nombreux travaux qui s'attachèrent à montrer, essentiellement d'un point de vue pratique, l'efficacité des procédures bayésiennes alliées aux bases d'ondelettes. Le traitement de données par l'analyse par ondelettes est rendu possible grâce à l'existence de la transformée en ondelettes discrète (voir section 2.2.1). Citons, entre autres, les travaux de Chipman, Kolaczyk et McCulloch (1997), Abramovich, Sapatinas et Silverman (1998), Clyde et George (1998, 2000), Clyde, Parmigiani et Vidakovic (1998), Crouse, Nowak et Baraniuk (1998), Johnstone et Silverman (1998), Vidakovic (1998), George et Foster (2000), Huang et Cressie (2000), Abramovich, Besbeas et Sapatinas (2002) et le recueil d'articles édités par Müller et Vidakovic (1999). La plupart d'entre eux s'appuie sur le caractère a priori sparse du signal f à estimer décomposé sur la base d'ondelettes $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$:

$$f = \sum_{j\geq -1} \sum_{k} \beta_{jk}\psi_{jk}.$$

Cette propriété se modélise de manière très naturelle en statistique bayésienne, notamment à travers le modèle suivant que beaucoup des auteurs précédemment cités envisagèrent :

$$\beta_{jk} = w_{jk}\gamma^{l}_{jk} + (1 - w_{jk})\gamma^{s}_{jk},$$

avec w_{jk} suivant une loi de Bernoulli. La loi de γ^{l}_{jk} représente les grands coefficients (la plupart du temps, une loi normale), la loi de γ^{s}_{jk} représente les coefficients négligeables (une loi normale de faible variance ou la mesure de Dirac au point 0). En général, les β_{jk} sont supposés indépendants. Cette hypothèse discutée par Brown, Vannucci et Fearn (1998) trouve sa justification dans les propriétés de décorrélation de la transformée en ondelettes discrète. Sous le modèle statistique de bruit blanc gaussien (voir section 2.2.2),

$$x_{jk} = \beta_{jk} + \varepsilon\xi_{jk}, \quad \xi_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad j \geq -1,\ k \in \mathbb{Z},$$

Chipman, Kolaczyk et McCulloch (1997) et Clyde, Parmigiani et Vidakovic (1998) considérèrent l'estimateur le plus classique : la moyenne de la loi a posteriori. Ils obtinrent ainsi un estimateur par contraction. Abramovich, Sapatinas et Silverman (1998) préférèrent utiliser la médiane, qui s'avère être un estimateur par seuillage pour leur modèle. C'est également ce type d'estimateurs que considéra Vidakovic (1998) à partir d'un modèle bayésien construit sur des lois de Student.

Citons également les travaux de Clyde et George (1998), Crouse, Nowak et Baraniuk (1998) et Johnstone et Silverman (1998) qui discutèrent de la construction d'estimateurs de type seuillage à travers une approche bayésienne. La comparaison de ces diverses procédures ne fut envisagée que du point de vue pratique. Ce n'est que très récemment que Johnstone et Silverman (2002a, 2002b) et Zhang (2002) établirent des résultats de convergence pour des procédures bayésiennes. Aussi, un des objectifs de ces travaux sera de montrer comment la modélisation bayésienne d'un signal influe sur les performances théoriques des estimateurs bayésiens classiques. Pour cela, nous utiliserons *le point de vue maxiset*.

1.2 Le point de vue maxiset

1.2.1 Minimax contre maxiset

Laissons provisoirement de côté le cadre bayésien pour décrire l'approche maxiset. Mais avant cela, rappelons brièvement le point de vue théorique classique pour mesurer la performance d'une procédure d'estimation : le point de vue minimax. Considérons toujours le problème statistique décrit en section 1.1.1 de l'estimation d'une fonction f à l'aide d'un estimateur $\hat{f}_n$ construit à partir de n observations, et le risque de $\hat{f}_n$ associé à une perte ρ :

$$R^n_\rho(\hat{f}_n, f) = \mathbb{E}_f\, \rho(\hat{f}_n, f).$$

Les fonctions de perte usuelles dérivent des normes L_p ou des normes associées à des espaces de Sobolev, de Hölder ou de Besov. Sans hypothèses de régularité sur f, on ne peut pas, en général, obtenir des résultats de convergence pour

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(\mathbb{R}^d, \mathbb{R}^{d'})} R^n_\rho(\hat{f}_n, f)$$

(voir par exemple Farrell (1967)). Aussi, nous devons choisir un espace fonctionnel V auquel appartiendra f. Le risque minimax pour V est alors défini par

$$\mathcal{R}^n_\rho(V) = \inf_{\hat{f}_n} \sup_{f \in V} \mathbb{E}_f\, \rho(\hat{f}_n, f),$$

où l'infimum est pris sur l'ensemble de tous les estimateurs. Si r_n est une suite qui tend vers 0 telle que

$$C_1 r_n \leq \mathcal{R}^n_\rho(V) \leq C_2 r_n,$$

avec C_1 et C_2 deux constantes, alors on dit que r_n est la vitesse minimax pour l'espace V associée à la perte ρ. Un estimateur $\hat{f}_n$ sera dit optimal au sens minimax s'il atteint la vitesse

r_n, c'est-à-dire si

$$\sup_{f\in V} \mathbb{E}_f\, \rho(\hat{f}_n, f) \leq C_3 r_n,$$

où C_3 est une constante. Les vitesses minimax ont été calculées pour différents modèles statistiques et pour différentes classes fonctionnelles (voir Bretagnolle et Huber (1979), Ibragimov et Khasminskii (1981), Stone (1982), Birgé (1983, 1985), Nemirovskii (1986), Kerkyacharian et Picard (1992)). Nous aussi, nous utiliserons l'approche minimax pour une classe particulière d'espaces fonctionnels : les espaces de Besov faibles introduits en section 1.2.3.
Indéniablement, l'approche minimax présente au moins deux inconvénients. Ayant pour but de rechercher des estimateurs qui minimisent "le risque maximum", elle semble trop pessimiste pour fournir une stratégie de décision similaire à celle que l'on pourrait envisager d'un point de vue pratique. D'autre part, le choix de l'espace fonctionnel V, qui ne fait pas l'objet d'un consensus parmi la communauté statistique, est très subjectif. Ce sont, entre autres, les raisons pour lesquelles Cohen, DeVore, Kerkyacharian et Picard (2001) ou Kerkyacharian et Picard (2000, 2002) envisagèrent une alternative au point de vue minimax : l'approche maxiset qui consiste à chercher l'espace maximal sur lequel une procédure d'estimation atteint une vitesse de convergence donnée. Par exemple (voir section 2.3.2), ces auteurs appliquèrent cette théorie pour deux procédures d'estimation efficaces : le seuillage des coefficients d'ondelette et la sélection locale du pas d'un noyau (voir Lepskii (1991)). Sous certaines conditions, puisque le maxiset de la première procédure est inclus dans le maxiset de la deuxième, ils purent conclure que la procédure de Lepskii était au moins aussi performante que la procédure par seuillage. On peut noter que cette approche est moins pessimiste et surtout, elle fournit des espaces fonctionnels directement liés à la procédure d'estimation choisie. En fait, si l'approche maxiset a été récemment conceptualisée en statistique à partir d'une approche similaire de la théorie de l'approximation (voir Cohen, DeVore et Hochmuth (2000)), elle était sous-jacente dans les résultats de Kerkyacharian et Picard (1993). En étudiant uniquement des procédures linéaires pour estimer une densité f, ces auteurs montrèrent que pour la perte L_p, $(p \geq 2)$, l'espace maximal sur lequel la vitesse $n^{-s/(2s+1)}$ est atteinte est une boule de l'espace de Besov fort $\mathcal{B}_{s,p,\infty}$ (voir Théorème 2.2). Un des objectifs de ces travaux est le suivant : pour un modèle statistique donné, on souhaite exhiber les espaces maximaux associés respectivement à des procédures bayésiennes et linéaires. Plus précisément, on considérera le modèle suivant :

$$x_k = \theta_k + \varepsilon \sigma_k \xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k \in \mathbb{N}^*. \tag{1.2}$$

où $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$ est une suite connue de réels strictement positifs et $\theta = (\theta_k)_{k\in\mathbb{N}^*}$ est la suite à estimer. Ce modèle hétéroscédastique fournit un cadre séquentiel idéal pour modéliser un grand nombre de problèmes statistiques inverses. Il semble également approprié pour approcher le modèle de régression non paramétrique considéré avec des erreurs corrélées et à condition que la taille de l'échantillon soit grande. Le modèle (1.2) est présenté de manière détaillée en section 2.2.3. C'est sous ce modèle que Kerkyacharian et Picard (2000) exhibèrent les maxisets pour la procédure de type seuillage associé au seuil universel

$$t_{k,\varepsilon} = \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}$$

(voir Théorème 2.5). Nous réaliserons le même objectif pour :

- les estimateurs linéaires de la forme

$$\hat{\theta}^{\lambda}(x) = (\lambda_k x_k)_{k \in \mathbb{N}^*}, \tag{1.3}$$

  souvent considérés dans la littérature des problèmes inverses,

- les estimateurs bayésiens classiques (moyenne, médiane) d'une loi a posteriori de $\theta = (\theta_k)_{k\in\mathbb{N}^*}$ conditionnellement à $x = (x_k)_{k\in\mathbb{N}^*}$.

Ceci nous permettra de comparer chacune de ces procédures à celle considérée par Kerkyacharian et Picard (2000) en comparant les espaces maximaux.

1.2.2 Espaces de Lorentz et espaces maximaux

Introduisons à présent une classe d'espaces fonctionnels qui fourniront la description des espaces maximaux d'un grand nombre de procédures statistiques. Pour cela, rappelons la définition des espaces de Lorentz, aussi appelés espaces L_p faibles, ou espaces de Marcinkiewicz (voir Lorentz (1950, 1966), DeVore et Lorentz (1993)).

Définition 1.1. *Si Ω est un espace muni d'une mesure positive μ, pour tout $0 < p < \infty$, l'espace de Lorentz $L_{p,\infty}(\Omega,\mu)$ est l'ensemble des fonctions $f : \Omega \longrightarrow \mathbb{R}$ μ-mesurables telles que*

$$\sup_{\lambda>0} \lambda^p \mu(|f| > \lambda) = \|f\|^p_{L_{p,\infty}(\Omega,\mu)} < \infty.$$

Si $\Omega = \mathbb{N}^*$ et si μ est une mesure sur $\mathbb{N}^*$, on notera $wl_p(\mu) = L_{p,\infty}(\mathbb{N}^*,\mu)$ et $wl_p = wl_p(\mu^*)$ si μ^* est la mesure de comptage sur $\mathbb{N}^*$.

On peut immédiatement remarquer que

$$wl_p = \left\{ \theta = (\theta_n)_{n\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^p \sum_n \mathbb{1}_{\,|\theta_n|>\lambda} < \infty \right\}$$

peut être identifié avec l'ensemble des suites $\theta = (\theta_n)_{n\in\mathbb{N}^*}$ telles que

$$\sup_{n\in\mathbb{N}^*} n^{\frac{1}{p}} |\theta|_{(n)} < \infty, \tag{1.4}$$

où

$$|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(n)} \ldots,$$

est le réarrangement de θ dans l'ordre décroissant. La notation pour l'espace wl_p n'a pas été choisie de manière anodine puisque cet espace séquentiel est fortement lié à l'espace l_p, notamment par les inclusions strictes suivantes :

$$l_p \subset wl_p \subset l_{p+\delta}, \quad \delta > 0.$$

Ainsi wl_p peut être vu comme une version faible de l'espace plus classique l_p. La majoration (1.4) fournit un contrôle polynomial de la suite $(|\theta|_{(n)})_{n\in\mathbb{N}^*}$, et donc un contrôle de la proportion des grandes composantes de θ, qui varie en fonction de p. Ainsi, les espaces wl_p semblent constituer une classe idéale pour mesurer le caractère sparse d'une suite. Il en est de même pour les espaces $wl_p(\mu)$, et avec un bon choix de μ, on peut introduire en plus une mesure de la régularité des suites considérées. Ces propriétés vont permettre à ces espaces séquentiels de jouer un rôle éminent en codage via la notion d'entropie métrique (voir Donoho (1996)), en théorie de l'approximation et en statistique. Pour illustrer cette affirmation, supposons données une base inconditionnelle $\mathcal{E} = (\phi_n)_{n\in\mathbb{N}^*}$ d'un espace Ω et une mesure μ sur $\mathbb{N}^*$, et posons $\forall\ 0 < p < \infty$,

$$\mathcal{L}_{p,\mu}(\mathcal{E}) = \left\{ f = \sum_n \theta_n \phi_n : \quad \theta = (\theta_n)_{n\in\mathbb{N}^*} \in wl_p(\mu) \right\}.$$

Kerkyacharian et Picard (2001) montrèrent que sous certaines conditions, les espaces $\mathcal{L}_{p,\mu}(\mathcal{E})$ sont caractérisés par le comportement asymptotique de leur entropie métrique. En théorie de l'approximation, l'importance des espaces $\mathcal{L}_{p,\mu}(\mathcal{E})$ a été révélée par Cohen, DeVore et Hochmuth (2000) à travers l'approche maxiset. Ils montrèrent que sous certaines conditions, les espaces $\mathcal{L}_{p,\mu}(\mathcal{E})$ apparaissent comme des espaces maximaux associés aux procédures de type seuillage des coefficients d'ondelette. On trouvera d'autres résultats mêlant espaces de Lorentz et théorie de l'approximation dans les travaux de DeVore (1989), DeVore et Lorentz (1993), DeVore, Konyagin et Temlyakov (1998), Temlyakov (1999) ou Cohen (2000).
La théorie de l'approximation étant un champ d'étude fortement lié à l'estimation fonctionnelle, il est naturel de voir certains espaces de Lorentz jouer un rôle important en statistique non paramétrique, notamment à travers l'utilisation des ondelettes. Il semble que les travaux pionniers relatifs à cette étude soient dus à Donoho (1993), Johnstone (1994) et Donoho et Johnstone

(1996). On retrouve certains de ces espaces dans l'approche maxiset (voir Cohen, DeVore, Kerkyacharian et Picard (2001), Kerkyacharian et Picard (2000, 2002) ou voir section 2.3). Citons à titre d'exemple, le théorème obtenu par Kerkyacharian et Picard (2000) pour les procédures de type seuillage :

Théorème 1.1. *Soit* $1 < r < \infty$ *et* $\alpha \in (0,1)$. *On suppose donnée une fonction*

$$f = \sum_{j \geq -1} \sum_{k} \beta_{jk} \psi_{jk} \in L_r([0,1]).$$

Sous le modèle de bruit blanc gaussien

$$x_{jk} = \beta_{jk} + \varepsilon \xi_{jk}, \quad \xi_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad j \geq -1,\ k \in \mathbb{Z},$$

on considère l'estimateur suivant :

$$\hat{f}_\varepsilon^T = \sum_{j=-1}^{j_\varepsilon} \sum_{k} x_{jk} \mathbb{1}_{\ |x_{jk}| > \kappa t_\varepsilon} \psi_{jk},$$

avec

- $t_\varepsilon = \varepsilon \sqrt{\log(\varepsilon^{-1})}$,
- $2^{-j_\varepsilon} \leq \varepsilon^2 \log(\varepsilon^{-1}) < 2^{-j_\varepsilon + 1}$,
- κ *une constante assez grande.*

On a $\forall\, \varepsilon$,

$$\mathbb{E}\|\hat{f}_\varepsilon^T - f\|_r^r \leq K \left(\varepsilon \sqrt{\log(\varepsilon^{-1})}\right)^{\alpha r},$$

où K *est une constante si et seulement si* f *appartient à la fois à l'espace de Besov fort* $\mathcal{B}_{\frac{\alpha}{2},r,\infty}$ *et à* $\mathcal{W}^*(r,(1-\alpha)r)$, *où* $\forall\ 0 < r,p < \infty$,

$$\mathcal{W}^*(r,p) = \left\{ f = \sum_{j \geq -1} \sum_{k} \beta_{jk} \psi_{jk} : \quad \sup_{\lambda > 0} \lambda^p \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} \sum_{k} \mathbb{1}_{\ |\beta_{jk}| > \lambda} < \infty \right\}.$$

Les espaces $\mathcal{W}^*(r,p)$ constituant une sous-classe des espaces $\mathcal{L}_{p,\mu}(\mathcal{E})$, la norme associée à chacun d'entre eux mesure deux choses : la régularité et le caractère sparse d'une fonction. Il est évident que la performance d'une procédure d'estimation dépend de la régularité de la fonction estimée. Mais à travers le résultat précédent, l'approche maxiset montre d'un point de vue théorique l'assertion suivante souvent observée en pratique : "La procédure statistique qui consiste à seuiller les coefficients d'ondelette fonctionne bien à condition que le signal soit sparse." Pour les procédures consistant à sélectionner localement le pas d'un noyau (voir Lepskii (1991)), des espaces de même type que ceux précédemment cités entrent en jeu dans la caractérisation des espaces maximaux.

1.2.3 Espaces de Besov faibles

L'approche maxiset nous conduit naturellement à chercher à exhiber les fonctions les plus difficiles à estimer associées à une procédure statistique. Compte tenu du résultat précédent, il est naturel de chercher "les ennemis typiques" associés à la procédure qui consiste à seuiller les coefficients d'ondelette, à l'intérieur des boules des espaces $\mathcal{W}^*(r,p)$. Pour atteindre cet objectif, notre approche sera de nature bayésienne et inspirée par celle de Pinsker (1980) (voir ci-dessous) : pour chaque boule $\mathcal{W}^*(r,p)(C)$, et en reprenant les notations de la section 1.2.1, nous souhaitons exhiber une mesure de probabilité π dont le support est inclus dans $\mathcal{W}^*(r,p)(C)$ et telle que le risque bayésien $B(\pi)$ associé à cette distribution

$$B(\pi) = \inf_{\hat{f}_n} \mathbb{E}_\pi \mathbb{E}_f \, \rho(\hat{f}_n, f)$$

soit égal au risque minimax sur $\mathcal{W}^*(r,p)(C)$. Les simulations de ces distributions, appelées *distributions les plus défavorables*, fourniront alors de bonnes représentations des fonctions les plus difficiles à estimer des boules $\mathcal{W}^*(r,p)(C)$. Au préalable, il nous faut donc évaluer le risque minimax sur chaque boule $\mathcal{W}^*(r,p)(C)$.

Avant cela, faisons la remarque suivante : dans la section 1.2.2, nous avions noté que l'espace wl_p apparaissait comme une version faible de l'espace l_p. De la même manière, et en utilisant la caractérisation séquentielle des espaces de Besov forts, on peut remarquer que $\mathcal{W}^*(r,p)$ apparaît comme une version faible de l'espace de Besov fort classique $\mathcal{B}_{s,p,p}$, avec $s = \frac{1}{2}(\frac{r}{p} - 1)$, $r > p$. Aussi, les espaces $\mathcal{W}^*(r,p)$ peuvent ils être nommés espaces de Besov faibles. Dès lors, il semble naturel de chercher à comparer d'un point de vue statistique, les espaces de Besov forts (très largement utilisés de nos jours en estimation fonctionnelle) et les espaces de Besov faibles. Mais pour que cette comparaison soit la plus exhaustive possible, nous élargissons la classe des espaces de Besov faibles et nous adoptons un formalisme différent pour nommer chaque membre de cette classe :

Définition 1.2. *Fixons $0 < s,p,q < \infty$. On dira que la fonction f décomposée sur $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$*

$$f = \sum_{j\geq -1} \sum_k \beta_{jk}\psi_{jk},$$

(ou de manière équivalente, la suite des coefficients d'ondelette $\beta = (\beta_{jk})_{j\geq -1,k\in\mathbb{Z}}$) appartient à l'espace de Besov faible de paramètres s,p,q, noté $\mathcal{WB}_{s,p,q}$ si

$$\sup_{\lambda>0} \lambda^q \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_k \mathbb{1}_{\{|\beta_{jk}|>\lambda\}}\right)^{\frac{q}{p}} < \infty.$$

A chaque espace de Besov faible $\mathcal{WB}_{s,p,q}$, on associe les boules :

$$\mathcal{WB}_{s,p,q}(C) = \left\{ f = \sum_{j\geq -1} \sum_{k} \beta_{jk}\psi_{jk} : \quad \sup_{\lambda>0} \lambda^q \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_k \mathbb{1}_{\ |\beta_{jk}|>\lambda}\right)^{\frac{q}{p}} \leq C^q \right\}.$$

Avec ce formalisme, il est facile de voir que $\mathcal{B}_{s,p,q}(C) \subset \mathcal{WB}_{s,p,q}(C)$ et d'identifier $W^*(r,p)$ et $\mathcal{WB}_{\frac{r}{2p}-\frac{1}{2},p,p}$.
Nous chercherons donc à comparer les vitesses minimax et les distributions les plus défavorables associées respectivement aux boules de Besov faibles $\mathcal{WB}_{s,p,q}(C)$ et de Besov forts $\mathcal{B}_{s,p,q}(C)$. Nous nous placerons sous le modèle de bruit blanc gaussien et la fonction de perte du risque minimax sera celle associée à la norme de l'espace de Besov fort $\mathcal{B}_{s',p',p'}$ ($0 \leq s' < \infty$, $1 \leq p' < \infty$). Ce choix s'appuie sur le fait que la classe des espaces de Besov forts modélise une large gamme d'inhomogénéité spatiale. D'autre part, le calcul du risque minimax pour cette norme constitue la première étape nécessaire pour espérer évaluer le risque minimax associé à une perte $L_{p'}$. Cette évaluation ne sera pas envisagée dans cette thèse.
Ainsi, nous allons généraliser le résultat de Johnstone (1994) qui obtint le comportement asymptotique du risque minimax sur les boules des espaces wl_p pour la perte l_2 et dans le cadre du modèle de bruit blanc gaussien. Pour obtenir ce résultat, Johnstone exploita l'approche bayésienne de Pinsker (1980) (mentionnée ci-dessus) qui, sous le même modèle statistique, souhaitait évaluer asymptotiquement le risque minimax d'un signal appartenant à un ellipsoïde Θ d'un espace de Hilbert. A cet effet, Pinsker montra qu'il existe des distributions a priori gaussiennes presque concentrées sur Θ et telles que leur risque bayésien est asymptotiquement égal au risque minimax. L'article de Pinsker inspira une littérature considérable. Mentionnons les travaux d'autres auteurs qui utilisèrent cette méthode, appelée *méthode minimax bayésienne*, pour résoudre des problèmes paramétriques ou non paramétriques : Casella et Strawderman (1981) et Bickel (1981) pour le problème de l'estimation de la moyenne d'une loi normale sachant qu'elle appartient à un intervalle compact, ou Donoho et Johnstone (1994b) et Donoho et Johnstone (1998) qui usèrent de cette méthode respectivement pour l'estimation sur les boules des espaces l_p et de Besov forts. On peut ajouter que la plupart des distributions les plus défavorables produites par ces problèmes sont concentrées sur un nombre fini de points (voir Casella et Strawderman (1981), Donoho et Johnstone (1994b) ou Donoho et Johnstone (1998)). Cependant, Bickel (1981) présenta des distributions ayant des densités basées sur des fonctions trigonométriques. De plus, cette méthode permit à Donoho et Johnstone (1994b) d'exhiber des estimateurs par seuillage atteignant les vitesses minimax. Ajoutons à cette liste le résultat de Johnstone (1994) sur les boules des espaces l_p faibles, mentionné ci-dessus :

Théorème 1.2. *Supposons donné le modèle statistique suivant :* $\forall\, n \in \mathbb{N}^*$,

$$x_k = \theta_k + \varepsilon_n \xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k = 1, \ldots, n, \quad \varepsilon_n \overset{n\to\infty}{\longrightarrow} 0,$$

et R_n *le risque minimax sur*

$$wl_p(r_n) = \left\{ \theta = (\theta_k)_k : \quad k^{1/p} |\theta|_{(k)} \leq r_n, \quad k = 1, \ldots, n \right\}$$

associé à la norme l_2 *:*

$$R_n = \inf_{\hat{\theta}_n} \sup_{\theta \in wl_p(r_n)} \mathbb{E}_\theta \|\hat{\theta}_n - \theta\|_{l_2}^2.$$

Si $p < 2$ *et si* $\eta_n = n^{-1/p}(r_n/\varepsilon_n)$ *vérifie* $\eta_n \overset{n\to\infty}{\longrightarrow} 0$ *et*

$$(\varepsilon_n / r_n)^2 \log(\eta_n^{-p}) = o_n((\log n)^{-6/p}),$$

alors

$$R_n \overset{n\to\infty}{\sim} \frac{2}{2-p} \varepsilon_n^{2-p} r_n^p (2\log(\eta_n^{-p}))^{1-\frac{p}{2}},$$

et l'estimateur construit sur l'estimateur par seuillage doux d^s

$$\hat{\theta}_n = \left(d_{t_n}^s(x_k) \right)_{k=1,\ldots,n}$$

avec $t_n = \varepsilon_n \sqrt{2\log(\eta_n^{-p})}$, *est asymptotiquement minimax.*

De plus, Johnstone exhiba une suite $(\pi_n)_n$ de *Distributions Asymptotiquement les Plus Défavorables* (notées DAPD), c'est-à-dire, qui vérifient :

$$\forall\, n \geq 1, \quad C_1 B(\pi_n) \leq R_n \leq C_2 B(\pi_n),$$

où C_1 et C_2 sont deux constantes et

$$\forall\, n \geq 1, \quad \exists\, \gamma_n > 1, \text{ tel que } \gamma_n \overset{n\to\infty}{\longrightarrow} 1, \text{ et } \mathbb{P}_{\pi_n}(\theta \in wl_p(\gamma_n r_n)) \overset{n\to\infty}{\longrightarrow} 1.$$

Ces distributions sont construites à partir des lois de Pareto de paramètre p.
Très naturellement, nous utiliserons comme Johnstone (1994) la méthode minimax bayésienne. Elle générera des estimateurs de type seuillage qui atteignent les vitesses minimax sur les boules de Besov faibles $\mathcal{WB}_{s,p,q}(C)$. Ces résultats serviront de base à la construction d'une procédure d'estimation adaptative qui sera testée et comparée à d'autres procédures plus classiques.

1.3 Interactions entre approches bayésienne et déterministe

Bien entendu, chacune de ces deux approches (bayésienne et déterministe) qui coexistent en estimation fonctionnelle ne vit pas dans un domaine complètement hermétique à l'autre. On a vu en section 1.1.1 que des estimateurs obtenus par la méthode de pénalisation pouvaient être vus comme des estimateurs bayésiens. Et la section 1.2.3 montre que la méthode minimax bayésienne permet d'exhiber des distributions simulant les fonctions les plus difficiles à estimer d'une classe fonctionnelle donnée.

Abramovich, Sapatinas et Silverman (1998) établirent l'existence d'autres interactions entre ces deux approches en utilisant un point de vue différent. Ils montrèrent comment la connaissance de la régularité d'une fonction f peut être déterminante pour choisir la modélisation bayésienne des coefficients d'ondelette de f. Plus précisément, supposons donnée une fonction périodique de période 1 décomposée sur une base d'ondelettes périodiques $(\tilde{\psi}_{jk})_{j\geq -1, 0\leq k<2^j}$:

$$f = \sum_{j\geq -1} \sum_{0\leq k<2^j} \beta_{jk}\tilde{\psi}_{jk},$$

dont le caractère sparse est mesuré par le modèle :

$$\beta_{jk} \sim w_j\mathcal{N}(0, \tau_j^2) + (1 - w_j)\delta_0, \quad j \geq 0, \quad 0 \leq k < 2^j,$$

avec les paramètres τ_j et w_j de la forme

$$\tau_j^2 = c_1 2^{-\alpha j}, \quad w_j = \min(1, c_2 2^{-\beta j}), \quad c_1, c_2, \alpha \geq 0, \quad 0 \leq \beta \leq 1.$$

Sous certaines hypothèses sur la base d'ondelettes, et si $1 \leq p \leq \infty$, $1 \leq q < \infty$, $s > \max(0, 1/p - 1/2)$, et pour toute valeur fixée de β_{-10},

$$f \in \mathcal{B}_{s,p,q} \quad p.s. \iff s + \frac{1}{2} - \frac{\beta}{p} - \frac{\alpha}{2} < 0.$$

Ils utilisèrent ce résultat pour donner une interprétation des paramètres s et p de l'espace de Besov.

Nous chercherons à établir une équivalence du même type, mais avec un modèle bayésien s'inspirant très fortement des DAPD pour les boules de Besov faibles, et avec les espaces $W^*(r, p)$ à la place des espaces de Besov $\mathcal{B}_{s,p,q}$. Ce choix semble plus naturel puisqu'on a vu que les espaces de Besov faibles modélisent naturellement le caractère sparse d'un signal. Nous montrerons l'existence d'une telle équivalence dans le chapitre 4. Ce résultat nous permettra de donner une interprétation des paramètres r et p. De plus, il confirmera les liens directs qui peuvent exister entre une classe fonctionnelle donnée et un modèle bayésien. Ceci nous encouragera à envisager la problématique suivante : quelles sont les distributions qui doivent entrer en jeu pour qu'un modèle bayésien décrive au mieux le caractère sparse d'une suite ? Pour cela, nous considérerons une suite $(\theta_k)_{k\in\mathbb{N}^*}$ aléatoire dont les composantes sont deux à deux indépendantes et telle que loi a priori de θ_k soit de la forme :

$$\theta_k \sim w_k\gamma_k(\theta_k) + (1-w_k)\delta_0(\theta_k),$$

et nous chercherons à déterminer le comportement des queues des densités γ_k pour que la suite $(\theta_k)_{k\in\mathbb{N}^*}$ appartienne presque sûrement à un espace $wl_p(\mu)$ donné.

1.4 Principaux résultats

Après avoir rappelé quelques notions sur la construction des bases d'ondelettes et les espaces de Besov forts, introduit les modèles statistiques utilisés dans ces travaux, et détaillé les résultats antérieurs de la théorie maxiset, tout ceci dans le chapitre 2, les chapitres 3, 4, 5 et 6 contiennent l'ensemble des résultats et leur preuve que nous souhaitions présenter dans cette thèse. Nous décrivons dans cette section les plus importants d'entre eux. Plus précisément, la section 1.4.1 contient les résultats majeurs des chapitres 3 et 4. Les théorèmes essentiels des chapitres 5 et 6 sont présentés en section 1.4.2.

1.4.1 Estimation minimax sur les espaces de Besov faibles et applications

Dans le chapitre 3, nous évaluons, sous le modèle de bruit blanc gaussien de niveau de bruit ε, le risque minimax associé à la norme $\mathcal{B}_{s',p',p'}$ $(0 \leq s' < \infty,\ 1 \leq p' < \infty)$ pour chaque boule de Besov faible $\mathcal{WB}_{s,p,q}(C)$ $(s' < s < \infty,\ 0 < p, q < \infty)$:

$$\mathcal{R}_\varepsilon = \inf_{\hat\beta_\varepsilon} \sup_{\beta\in\mathcal{WB}_{s,p,q}(C)} \mathbb{E}_\beta \|\hat\beta_\varepsilon - \beta\|^{p'}_{\mathcal{B}_{s',p',p'}}$$

(voir Théorème 3.1). Trois cas distincts émergent naturellement de cette étude. Si on suppose que $p'(s'+\frac{1}{2}) < p(s+\frac{1}{2})$ (cas régulier) alors nous montrons que la vitesse minimax r_ε est polynomiale : $r_\varepsilon = \varepsilon^{p'\nu}$ avec $\nu = (s-s')/(s+\frac{1}{2})$. Dans le cas où $p'(s'+\frac{1}{2}) = p(s+\frac{1}{2})$ (cas critique), la vitesse minimax devient $r_\varepsilon = \varepsilon^{p'\nu}\log\left(\frac{C}{\varepsilon}\right)^{\frac{p'\nu}{2}+(1-\frac{p}{q})_+}$. Il faut noter que pour ce cas, la majoration n'est possible que si nous imposons en plus que $\beta\in\mathcal{B}_{\eta,p',\infty}$ $(\eta > s')$. Cette hypothèse de régularité minimale déjà nécessaire pour le Théorème 1.1 permet de contrôler la taille des coefficients d'ondelette lorsque le niveau de résolution est élevé. La constante de majoration devant r_ε dépend alors de η et explose quand $\eta \to s'$. Bien sûr, avec $p' = 2$, $s' = 0$, $q = p$ et

$s = \frac{1}{p} - \frac{1}{2}$, on retrouve les résultats de Johnstone (1994) relatifs aux calculs des vitesses minimax sur les boules des espaces wl_p. A ce stade, il faut noter, grâce aux travaux de Donoho, Johnstone, Kerkyacharian et Picard (1997), que si $p'(s' + \frac{1}{2}) \leq p(s + \frac{1}{2})$, alors les vitesses minimax associées aux boules $\mathcal{WB}_{s,p,q}(C)$ et $\mathcal{B}_{s,p,q}(C)$ sont les mêmes. La preuve de ces résultats s'appuie sur la méthode minimax bayésienne (évoquée en section 1.2.3) qui consiste à majorer le risque minimax sur $\mathcal{WB}_{s,p,q}(C)$ par le risque minimax bayésien associé à un espace M de mesures de probabilité convenablement choisi et qui doit être convexe et compact. Pour cette étude, M sera l'enveloppe convexe généralisée de

$$m_{s,p,q}(C) = \left\{ \pi(d\beta) : \quad \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left[\sum_k \mathbb{E}_\pi (\mathbb{1}_{\ |\beta_{jk}|>\lambda}) \right]^{\frac{q}{p}} \leq \left(\frac{C}{\lambda} \right)^q, \quad \forall\, \lambda > 0 \right\}$$

et le risque minimax bayésien pour M sera donc

$$B(M, \varepsilon) = \inf_{\hat{\beta}_\varepsilon} \sup_{\pi \in M} \mathbb{E}_\pi \mathbb{E}_\beta \|\hat{\beta}_\varepsilon - \beta\|^{p'}_{\mathcal{B}_{s',p',p'}}.$$

On peut alors appliquer le théorème minimax et inverser l'infimum et le supremum. Puis, on se restreint à une classe particulière d'estimateurs, notée $\mathcal{C}$. Pour un bon choix de $\mathcal{C}$, il est possible d'exhiber une distribution $\tilde{\pi_\varepsilon} \in M$ et un estimateur $\hat{\beta}_\varepsilon \in \mathcal{C}$ tels que

$$B(M, \varepsilon) \approx B(\tilde{\pi_\varepsilon}) \approx \mathbb{E}_{\tilde{\pi_\varepsilon}} \mathbb{E}_\beta \|\hat{\beta}_\varepsilon - \beta\|^{p'}_{\mathcal{B}_{s',p',p'}}.$$

En calculant le terme de droite, on obtient ainsi une majoration du risque minimax. Cette technique de majoration s'applique particulièrement bien à l'estimation sur les espaces $\mathcal{WB}_{s,p,q}(C)$ à condition que $p'(s' + \frac{1}{2}) \leq p(s + \frac{1}{2})$. Ce n'est pas le cas pour le cas logarithmique : $p'(s' + \frac{1}{2}) > p(s + \frac{1}{2})$. Les fonctions typiques du cas logarithmique ont en effet une structure sparse qui ne peut pas être traduite par les mesures de probabilité $\tilde{\pi}_\varepsilon$ que l'on exhibe. Ceci rend la méthode bayésienne inadaptée pour ce cas. Dans le chapitre 3 et pour les cas réguliers et critiques, $\mathcal{C}$ sera la classe des estimateurs qui estiment chaque coefficient d'ondelette en utilisant l'estimateur par seuillage doux

$$d_t^s(x) = \text{sign}(x)(|x| - t)_+.$$

On exploitera notamment les propriétés de monotonie du risque de d_t^s (Proposition 3.4). La minoration du risque minimax s'obtient en utilisant la minoration du risque pour $\mathcal{B}_{s,p,q}(C)$ (voir Donoho, Johnstone, Kerkyacharian et Picard (1997)) et l'inclusion $\mathcal{B}_{s,p,q}(C) \subset \mathcal{WB}_{s,p,q}(C)$. C'est naturellement en s'inspirant des mesures $\tilde{\pi_\varepsilon}$ que sont bâties des DAPD, notées π_ε, dont le support pour chacune d'elles est asymptotiquement inclus dans $\mathcal{WB}_{s,p,q}(C) \backslash \mathcal{B}_{s,p,q}(C)$. Pour montrer cette propriété (Propositions 3.6 et 3.7), on s'aide de l'inégalité DKW. La difficulté de l'évaluation de $B(\pi_\varepsilon)$ réside dans le choix de la norme $\mathcal{B}_{s',p',p'}$ pour laquelle il n'est pas possible en général d'exhiber l'estimateur bayésien. Nous surmontons cette difficulté en étudiant le cas $p' = 1$ qui nous fournit une idée du résultat que nous généralisons ensuite au cas où p' est choisi de manière quelconque (voir Théorème 3.3). Sous π_ε, la distribution des coefficients β_{jk} a la forme suivante (voir Théorème 3.1) : les β_{jk} sont indépendants, leur distribution est symétrique par rapport à 0 et $|\beta_{jk}|$ peut s'écrire :

$$|\beta_{jk}| = \begin{cases} \varepsilon\alpha_j & \text{si } j < j_* \\ \varepsilon \min(\alpha_j X_{jk}, \mu_j) & \text{sinon,} \end{cases} \tag{1.5}$$

où X_{jk} est une variable de Pareto de paramètre p, $j_* \in \mathbb{N}$, α_j et μ_j sont des réels non aléatoires positifs. Nos DAPD ont donc une structure très différente des DAPD associées aux espaces de Besov forts construites par Johnstone (1994) à partir de variables aléatoires, qui sont gaussiennes pour le cas $p = q = 2$, ou dont la masse est répartie en deux ou trois points pour le cas $p < 2$.
L'estimateur $\hat{\beta}_\varepsilon$ de type seuillage que l'on exhibe en utilisant la méthode minimax bayésienne atteint les vitesses minimax. Il n'est pas adaptatif car il dépend des paramètres de l'espace de Besov faible. En fait, à chaque niveau de résolution j, chaque coefficient d'ondelette est estimé grâce à l'estimateur par seuillage doux $d^s_{t_j(\varepsilon)}$ avec $t_j(\varepsilon)$ de la forme

$$t_j(\varepsilon) = \begin{cases} \varepsilon\sqrt{-2\log(\alpha_j^p)} & \text{si } j \geq j_*, \\ 0 & \text{sinon,} \end{cases} \tag{1.6}$$

où ε est le niveau de bruit et p, α_j et j_* apparaissent dans l'égalité (1.5).
Pour mesurer la performance de cet estimateur d'un point de vue pratique, nous avons procédé de la manière suivante : sous le modèle de régression

$$g_i = f(\frac{i}{n}) + \sigma\varepsilon_i, \quad 1 \leq i \leq n = 1024, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1),$$

où σ est supposé connu, nous supposons que la fonction f à estimer fait partie de la classe des fonctions les plus difficiles à estimer d'une boule de Besov faible de paramètres inconnus. Nous appliquons alors la transformée en ondelettes discrète pour les différents vecteurs introduits précédemment et nous obtenons le modèle statistique suivant :

$$y_{jk} = d_{jk} + \sigma z_{jk}, \quad z_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad -1 \leq j \leq N-1, 0 \leq k < 2^j,$$

où $y_{jk} = (\mathcal{W}g)_{jk}$, $d_{jk} = (\mathcal{W}f^0)_{jk}$, $f^0 = (f(\frac{i}{n}), \quad 1 \leq i \leq n)^T$ et $z_{jk} = (\mathcal{W}\varepsilon)_{jk}$. En prenant en compte les propriétés qui lient les d_{jk} et les coefficients d'ondelette ordinaires de la fonction f (voir section 2.2.1) et l'hypothèse que l'on fait sur f, il est alors naturel de supposer que les coefficients d_{jk} peuvent être munis d'un modèle bayésien s'inspirant fortement des DAPD. Aussi, nous supposons que les d_{jk} sont deux à deux indépendants et que $\sigma^{-1}d_{jk} \sim F_j^{\alpha_j,\mu_j,p}$, avec

- $F_j^{\alpha_j,\mu_j,p} = \frac{1}{2}(F_j^+ + F_j^-)$,
- F_j^- est la distribution obtenue par symétrie de F_j^+ par rapport à 0,
- F_j^+ est la distribution de $\min(\alpha_j X_j - \alpha_j, \mu_j)$, où X_j est une variable de Pareto de paramètre p,
- α_j et μ_j sont deux nombres positifs.

Nous apportons dans le chapitre 4 des justifications plus précises des raisons qui nous conduisent à envisager ce modèle bayésien. Il est alors assez naturel d'estimer chaque d_{jk} par l'estimateur de type seuillage doux associé au seuil $t_j(\sigma)$ défini en (1.6) puis d'appliquer la transformée en ondelettes discrète inverse. Pour cela, il faut connaître les valeurs de α_j, j_* et p. Nous décrivons en section 4.4.2 la méthode choisie pour estimer j_* et chaque α_j.
Toutes ces étapes permettent donc de construire une procédure (appelée *ParetoThresh*) dépendant uniquement de p. Une fois le paramètre p fixé, nous obtenons une méthode pour décider niveau par niveau si les coefficients bruités y_{jk} doivent être seuillés et avec quel seuil. La procédure ParetoThresh est appliquée pour différentes valeurs de ce paramètre autour de la valeur $p = 1.3$ aux quatre fonctions "test" classiques de Donoho et Johnstone ("Blocks", "Bumps", "Heavisine", "Doppler"). Le Tableau 1.1 compare ParetoThresh aux procédures déterministes classiques (VisuShrink et SureShrink) en fournissant la moyenne de l'erreur en moyenne quadratique (AMSE) calculée avec 100 applications de chaque procédure, et pour différents rapports signal/niveau de bruit (RSNR). Les résultats obtenus (voir section 4.4.3 pour plus de détails) confirment l'efficacité des méthodes bayésiennes du point de vue pratique. Néanmoins, si ParetoThresh s'avère assez performante pour les fonctions "Blocks", "Bumps" et "Doppler", ce n'est plus le cas pour la fonction "Heavisine". Ce phénomène peut s'expliquer par le fait que "Heavisine" étant assez "lisse", elle entre difficilement dans le cadre de notre modèle bayésien censé modéliser les fonctions les plus difficiles à estimer d'une certaine classe fonctionnelle. Si les bonnes performances de ParetoThresh semblent avérées à travers l'utilisation de l'erreur en moyenne quadratique, il faut noter l'apparition d'artefacts (voir par exemple Figures 4.2, 4.3, 4.4 et 4.5). Il est possible de faire disparaître ces artefacts en diminuant la valeur de p. Malheureusement, ceci a un coût : en général l'erreur en moyenne quadratique augmente. Quand nous considérons des valeurs de p plus grandes que 1.3, l'erreur en moyenne quadratique augmente et nous obtenons également un plus grand nombre d'artefacts. Ce dernier phénomène s'explique

RSNR	Signal	VisuShrink	SureShrink	ParetoThresh ($p = 1.3$)
RSNR=3	Blocks	3.3143	1.7850	*1.4559*
	Bumps	5.6100	2.0378	*1.8025*
	Heavisine	0.3136	*0.3042*	0.3059
	Doppler	2.1588	1.0911	*0.9221*
RSNR=5	Blocks	1.8624	0.7645	*0.6928*
	Bumps	2.7345	0.8523	*0.8128*
	Heavisine	0.1946	*0.1816*	0.1851
	Doppler	1.0358	0.4378	*0.4310*
RSNR=8	Blocks	0.9745	0.3449	*0.3248*
	Bumps	1.3139	*0.3032*	0.3714
	Heavisine	0.1312	0.1028	*0.0818*
	Doppler	0.5374	0.2434	*0.2140*

TAB. 1.1 – AMSEs pour VisuShrink, SureShrink et ParetoThresh ($p = 1.3$) avec quatre fonctions "test" et différentes valeurs du rapport signal/bruit.

très bien à l'aide des conclusions tirées à la fin de cette section.
Un des buts mentionnés en section 1.2.3 étaient de représenter les "ennemis typiques" pour les procédures de type seuillage et on a vu qu'il était naturel de les chercher à l'intérieur des espaces de Besov faibles $W^*(r,p)$. Encore une fois, l'outil bayésien va nous permettre de réaliser cet objectif à travers deux approches distinctes et complémentaires.

- La première consiste tout simplement à construire des simulations des DAPD d'une boule de Besov faible de rayon fixé égal à 1. On a vu que ces distributions sont typiques des espaces de Besov faibles et étrangères aux espaces de Besov forts.

- La seconde s'inspire de l'approche d'Abramovich, Sapatinas et Silverman (1998). On considère $\forall\, j \geq 0$, la mesure de probabilité $F_j^{\alpha_j,\mu_j,p}$ définie précédemment et dont la définition s'inspire des DAPD. En s'aidant de cette distribution, on établit la CNS suivante :

Théorème 1.3. *On considère une fonction f périodique de période 1 décomposée sur une base d'ondelettes périodiques $(\tilde{\psi}_{jk})_{j\geq -1, 0\leq k<2^j}$:*

$$f(t) = \sum_{j=-1}^{\infty} \sum_{0\leq k<2^j} \beta_{jk}\tilde{\psi}_{jk}(t).$$

Soient $0 < p < \infty$ *et* $0 < r < \infty$. *Étant donnés trois nombres strictement positifs* δ, C *et* T, *on définit pour tout* $j \geq 0$, $\alpha_j = C2^{-j\delta}$, *et* $\mu_j = \sqrt{\max\left(T^2, -2\log(\alpha_j^p)\right)}$. *On suppose que les coefficients d'ondelette* β_{jk} *de* f *sont indépendants et pour* $j \geq 0$ *et* $0 \leq k < 2^j$, $F_j^{\alpha_j,\mu_j,p}$ *est la distribution de* β_{jk}. *Pour toute valeur fixée de* β_{-10},

$$f \in \mathcal{W}^*(r,p) \ p.s. \quad \Longleftrightarrow \quad \frac{r}{2} < \delta p.$$

Cette approche fournit des fonctions appartenant effectivement aux espaces $W^*(r,p)$, ce qui n'était pas le cas des DAPD, éléments de l'espace M défini plus haut. Cependant, cette approche ne contrôle pas le rayon de la boule de Besov faible qui contient la fonction f simulée, ce qui peut altérer notre perception de l'inhomogénéité de la fonction f.
Les deux approches montrent les phénomènes suivants (voir figures 3.1 et 4.1). Pour chaque espace $W^*(r,p)$, le paramètre r influe sur la régularité des fonctions de cet espace, le paramètre p sur le caractère sparse : quand p diminue, le nombre de coefficients négligeables augmente mais les quelques rares coefficients non négligeables peuvent être très grands.

1.4.2 Maxisets pour les procédures linéaires et bayésiennes

L'objet des chapitres 5 et 6 consiste essentiellement à étudier les espaces maximaux associés aux estimateurs linéaires et bayésiens classiques. Rappelons que le modèle statistique considéré est le modèle hétéroscédastique (1.2).
Dans le chapitre 5 nous considérons les estimateurs de la forme donnée en (1.3). Le risque est évalué pour une norme l_p pondérée, notée $l_p(\mu)$, qui peut être la norme l_p, une norme Sobolev ou une norme associée à un espace de Besov fort. Le Théorème 5.2 montre que sous des hypothèses très générales portant sur la suite de poids $(\lambda_k)_{k\in\mathbb{N}^*}$, et sous un contrôle de la norme $l_p(\mu)$ de la suite $(\sigma_k\lambda_k)_{k\in\mathbb{N}^*}$, les espaces maximaux pour $\hat{\theta}^\lambda(x) = (\lambda_k x_k)_{k\in\mathbb{N}^*}$ sont de la forme

$$B^\eta_{p,\infty}(\mu) = \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^{p\eta} \sum_{k\geq\lambda} \mu_k|\theta_k|^p < \infty \right\},$$

où η dépend de la vitesse de convergence donnée. Nous montrons que les poids classiques de la littérature des problèmes inverses (les poids de projection, de Pinsker ou de Tikhonov-Phillips) vérifient les hypothèses du Théorème 5.2 (voir Propositions 5.1 et 5.2). En s'aidant du résultat établi par Kerkyacharian et Picard (2000) (voir Théorème 5.1), il est alors possible de comparer les performances des procédures linéaires et de type seuillage du point de vue maxiset, en considérant la norme l_p et la vitesse de convergence $(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(1+2s)}$, $0 < s < \infty$. Si $\sum_k \sigma_k^p < \infty$ ou si la suite $(\sigma_k)_{k\in\mathbb{N}^*}$ a une croissance/décroissance polynomiale, alors l'inclusion stricte des espaces maximaux des procédures linéaires dans ceux des procédures de type seuillage (lemme 5.1) établit la supériorité des secondes (qui sont de plus adaptatives) sur les premières. On retrouve là un résultat classique de la théorie minimax. Nous étendons ensuite ce résultat en étudiant les maxisets pour l'estimation de fonctions décomposées sur des bases inconditionnelles $(\psi_k)_{k\in\mathbb{N}^*}$ de L_p $(1 < p < \infty)$ et pour les estimateurs par projection. Nous supposons de plus que la suite $(\sigma_k\psi_k)_{k\in\mathbb{N}^*}$ vérifie une inégalité de superconcentration (voir l'inégalité (5.10)). Une fois encore, pour le risque associé à la norme L_p, nous montrons la supériorité des estimateurs de type seuillage sur les estimateurs par projection. De plus, sous certaines conditions portant sur la base $(\psi_k)_{k\in\mathbb{N}^*}$, les espaces maximaux sont exactement des espaces de Besov forts de la forme $\mathcal{B}_{s,p,\infty}$ comme pour le modèle de l'estimation de densité. Le chapitre 5 montre donc une grande stabilité de la nature des maxisets pour les procédures linéaires quand le modèle statistique varie.

Dans le chapitre 6, nous nous plaçons à nouveau dans un cadre bayésien et nous considérons le modèle suivant : les θ_k sont supposés indépendants et $\forall\, k \in \mathbb{N}^*$,

$$\theta_k \sim w_{k,\varepsilon}\gamma_{k,\varepsilon}(\theta_k) + (1 - w_{k,\varepsilon})\delta_0(\theta_k), \qquad (M_1)$$

où $w_{k,\varepsilon} \in (0,1)$. On suppose que la densité $\gamma_{k,\varepsilon}$ s'écrit

$$\forall\, \theta \in \mathbb{R}, \quad \gamma_{k,\varepsilon}(\theta) = (\varepsilon\sigma_k)^{-1}\gamma((\varepsilon\sigma_k)^{-1}\theta),$$

où γ est symétrique, positive et unimodale, et que ses queues ont une décroissance exponentielle ou plus lente. Dans un premier temps, nous supposons que $w_{k,\varepsilon}$ ne dépend que de ε et tend vers 0 quand ε tend vers 0. Il est alors aisé d'exhiber le comportement asymptotique de la médiane et de la moyenne de la loi a posteriori de θ_k conditionnellement à x_k (Propositions 6.3 et 6.4). De plus, nous montrons que la médiane est un estimateur de type seuillage. Pour le risque l_p, nous exhibons les maxisets pour chacune des deux procédures bayésiennes associées respectivement à la médiane et à la moyenne de la loi a posteriori, mais qui imposent néanmoins que θ_k soit estimé par 0 pour les grandes valeurs de k (Théorèmes 6.2, 6.3, 6.4 et 6.5). On note que l'hypothèse sur les queues de γ est essentielle pour obtenir des espaces maximaux aussi grands que possibles. Pour chacune de ces deux procédures, les maxisets sont constitués de l'intersection de deux espaces : un espace de type "l_q faible pondéré" noté $wl_{p,q}(\sigma)$ $(q < p)$ avec

$$wl_{p,q}(\sigma) = \left\{\theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^q \sum_k \mathbb{1}_{\ |\theta_k|>\lambda\sigma_k}\sigma_k^p < \infty\right\}$$

et $B^{\eta}_{p,\infty}(\mu^*)$ où μ^* est la mesure de comptage sur $\mathbb{N}^*$ (q et η dépendent bien sûr de la vitesse de convergence donnée). Le résultat essentiel de ce chapitre est alors le suivant : du point de vue maxiset et pratiquement sous les mêmes hypothèses, chacune de nos deux procédures bayésiennes atteint exactement la même performance que la procédure de type seuillage associée au seuil $t_{k,\varepsilon} = \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}$, pour la vitesse de convergence $(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(1+2s)}$, $0 < s < \infty$. Chacune d'entre elles est donc préférable aux procédures linéaires. Ces résultats et ces conclusions sont étendus aux procédures analogues envisagées sur les espaces L_p $(1 < p < \infty)$, mais la caractérisation des espaces maximaux fait intervenir une classe particulière d'espaces de Lorentz (voir Théorème 6.6).

Un autre objectif de ce chapitre était de s'interroger sur la meilleure façon de modéliser le caractère sparse d'une suite à travers une approche bayésienne. Nous avons vu que cette propriété était bien mesurée par les espaces l_q faibles pondérés et il n'est donc pas très étonnant de voir ce type d'espaces jouer un rôle si important dans la caractérisation des espaces maximaux de nos procédures bayésiennes. L'approche maxiset relie ainsi notre modèle bayésien (M_1) et les espaces $wl_{p,q}(\sigma)$. Mais pour choisir au mieux le modèle bayésien et notamment la densité γ, nous avons naturellement cherché à établir des relations plus directes entre (M_1) et les espaces $wl_{p,q}(\sigma)$. Nous montrons le résultat suivant :

Théorème 1.4. *Supposons donnés $1 \leq p < \infty$ et $0 < q < p$. On considère toujours le modèle (M_1) avec $\varepsilon = 1$. On note $w_k = w_{k,1}$ et $\forall\, \lambda \geq 0$, $\tilde{F}(\lambda) = 2\int_\lambda^{+\infty} \gamma(x)dx$. S'il existe une constante C telle que*

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{\,|\theta_k|>\sigma_k\lambda} \leq C^q \quad p.s.,$$

alors

$$\sup_{\lambda>0} \lambda^q \tilde{F}(\lambda) \sum_k w_k \sigma_k^p \leq C^q.$$

Réciproquement, s'il existe une constante C telle que

$$\sup_{\lambda>0} \lambda^q \tilde{F}(\lambda) \sum_k w_k \sigma_k^p \leq C^q, \tag{1.7}$$

alors

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{\,|\theta_k|>\sigma_k\lambda} < \infty \quad p.s.$$

Ce résultat illustre les liens très forts qui semblent exister entre les densités de Pareto de paramètre q et les espaces l_q faibles pondérés. Ces relations furent déjà observés à travers l'approche fournie par les DAPD.

1.5 Perspectives

Tout au long de ces travaux, nous supposons, notamment pour les aspects pratiques, que le niveau de bruit pour chaque modèle statistique est connu. Bien entendu, c'est une hypothèse dont nous pourrions nous affranchir en cherchant à estimer ce niveau de bruit à travers une approche bayésienne par exemple, comme l'ont fait Clyde, Parmigiani et Vidakovic (1998), Vidakovic (1998) ou Vidakovic et Ruggeri (2001). Naturellement, les lois a posteriori ne seraient pas les mêmes et les résultats sensiblement différents. Ce thème peut constituer un axe de recherche possible.

Cette thèse s'appuie constamment sur l'outil des bases d'ondelettes, mais nous avons choisi de ne pas nous interroger sur la façon de choisir cette base. Si nous exploitons abondamment les propriétés essentielles de certaines bases d'ondelettes (l'orthonormalité, l'analyse temps-fréquence, leur régularité, la compacité de leur support, le fait que pour certains espaces fonctionnels ce sont des bases inconditionnelles vérifiant une inégalité de concentration), il y en a une que nous laissons de côté : le caractère unique d'une décomposition une fois la base choisie. De cette constatation, surgit naturellement l'idée d'utiliser "des familles génératrices surabondantes". Davis, Mallat et Zhang (1994) et Chen, Donoho et Saunders (1998) ont montré les avantages de ces familles qui peuvent offrir des décompositions plus adaptatives. L'outil des ondelettes suggère naturellement d'utiliser la famille $(\psi_{a,b})_{a\in[1,\infty),b\in\mathbb{R}_+}$ avec

$$\psi_{a,b}(t) = a^{\frac{1}{2}}\psi(at-b)$$

et de décomposer les fonctions à estimer sur cette famille. Il incombe alors au statisticien de choisir la meilleure décomposition. Une fois encore, l'approche bayésienne peut s'avérer un outil efficace comme le montrèrent Abramovich, Sapatinas et Silverman (2000b) en imposant une loi a priori sur les indices (a,b). Il serait intéressant d'associer cette idée avec les modèles bayésiens décrits dans ces travaux et d'en explorer les résultats.

Nos résultats s'appuient toujours sur l'indépendance des composantes d'une suite à estimer. Cette forte hypothèse peut se justifier lorsqu'on cherche à estimer des coefficients issus de la transformée en ondelettes discrète. Pas toujours très réaliste, cette hypothèse offre malgré tout l'avantage de simplifier beaucoup de situations. Néanmoins, quelques auteurs (Müller et Vidakovic (1995), Crouse, Nowak et Baraniuk (1998), Huang et Cressie (2000) ou Vannucci et Corradi (1999)) envisagèrent des approches bayésiennes pour modéliser la dépendance de coefficients. Un axe de recherche possible serait de chercher à prolonger ces travaux qui semblent s'appliquer naturellement au débruitage de signaux très corrélés (images,...).

Si un grand nombre d'articles ont montré les bonnes performances d'un point de vue pratique des estimateurs bayésiens, nous avons noté que leurs performances théoriques avaient été très peu étudiées. Ceci est sans doute dû au fait qu'il est souvent difficile d'expliciter l'expression de ces estimateurs. Pourtant les articles de Johnstone et Silverman (2002a, 2002b), de Zhang (2002) et les travaux de cette thèse, considérant respectivement, des approches minimax, oracle et maxiset révèlent l'optimalité de certaines procédures bayésiennes. Nous avons rappelé dans la section 1.2.1 les inconvénients de l'approche minimax soulignant ainsi les avantages de la théorie maxiset, qui elle aussi, n'est pourtant pas sans inconvénients. Par exemple, et même si nous n'avons pas eu à subir ce désagrément, il aurait pu se produire que la nature des espaces maximaux obtenus ne permettent pas de les comparer. Néanmoins, cette approche s'est révélée prometteuse, notamment parce qu'elle confirmait des phénomènes souvent observés en pratique. Un axe de recherche naturel serait d'utiliser ce point de vue, complémentaire des approches minimax et oracle, pour mesurer les performances d'autres estimateurs (les estimateurs CART, les estimateurs obtenus par maximisation d'une loi a posteriori ou par pénalisation,...) ou pour d'autres problèmes statistiques (estimation ponctuelle, change point models, ...).

Chapitre 2

Préliminaires et compléments

2.1 Construction des bases d'ondelettes et applications

2.1.1 Analyse multirésolution

L'objet de cette section est de rappeler les principaux éléments de la construction des bases d'ondelettes. Pour plus de détails on se référera aux ouvrages de Meyer (1992), Daubechies (1992) et Mallat (1998). La définition des bases d'ondelettes repose sur la notion fondamentale d'analyse multirésolution :

Définition 2.1. *Une analyse multirésolution de $L_2(\mathbb{R})$ est une suite croissante de sous espaces fermés de $L_2(\mathbb{R})$, notée $(V_j)_{j\in\mathbb{Z}}$, possédant les propriétés suivantes :*

- $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$,
- $\bigcup_{j\in\mathbb{Z}} V_j$ *est dense dans* $L_2(\mathbb{R})$,
- $\forall\, f \in L_2(\mathbb{R}), \quad \forall\, j \in \mathbb{Z}, \quad f(x) \in V_j \iff f(2x) \in V_{j+1}$,
- $\forall\, f \in L_2(\mathbb{R}), \quad \forall\, k \in \mathbb{Z}, \quad f(x) \in V_0 \iff f(x-k) \in V_0$,
- *il existe une fonction $\phi \in V_0$, appelée fonction d'échelle de l'analyse multirésolution, telle que $\{\phi(x-k) : \quad k \in \mathbb{Z}\}$ soit une base orthonormée de V_0.*

Ainsi, à chaque niveau de résolution j, l'espace V_j d'une analyse multirésolution $(V_j)_{j\in\mathbb{Z}}$ possède une base orthonormée obtenue par translations et dilatations de la fonction d'échelle ϕ : $\{\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) : \quad k \in \mathbb{Z}\}$. Citons l'exemple classique de l'analyse multirésolution engendrée par la fonction de Haar $\phi(x) = 1_{x\in[0,1]}$ où $\forall\, j \in \mathbb{Z}$,

$$V_j = \overline{\text{vect}\{\phi_{jk} : \quad k \in \mathbb{Z}\}}.$$

L'approximation au niveau de résolution j d'une fonction de $L_2(\mathbb{R})$ s'obtient par projection sur l'espace V_j. Notons P_j l'opérateur de projection de $L_2(\mathbb{R})$ sur l'espace V_j. Si $f \in L_2(\mathbb{R})$, la différence d'approximation $P_{j+1}f - P_jf$ entre les niveaux j et $j+1$ s'obtient par projection de f sur W_j, espace supplémentaire orthogonal de V_j dans V_{j+1}. On peut construire une fonction ψ, appelée *ondelette*, de façon à ce que $\{\psi_{jk}(x) = 2^{j/2}\psi(2^jx - k) : \quad k \in \mathbb{Z}\}$ soit une base orthonormée de W_j. Toute fonction f de $L_2(\mathbb{R})$ peut alors être reconstruite par raffinements successifs en ajoutant à un terme d'approximation des termes de détail : pour tout $j_0 \in \mathbb{Z}$,

$$f = \sum_{k\in\mathbb{Z}} \alpha_{j_0k}\phi_{j_0k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk},$$

où les coefficients d'ondelette sont définis par

$$\alpha_{j_0k} = \int f(x)\phi_{j_0k}(x)dx$$

et

$$\beta_{jk} = \int f(x)\psi_{jk}(x)dx.$$

Tout au long de ces travaux, nous avons privilégié (sans perte de généralité) le niveau de résolution $j_0 = 0$ et avons noté $\forall\, k \in \mathbb{Z}$, $\forall\, x \in \mathbb{R}$, $\psi_{-1k}(x) = \phi_{0k}(x)$, $\beta_{-1k} = \alpha_{0k}$. L'outil de l'analyse multirésolution nous permet donc de disposer de bases orthonormées dont les atomes sont à la fois localisés en temps et en fréquence et construits par translations et dilatations dyadiques d'un système "fonction d'échelle/ondelette" (ϕ, ψ). Le système de Haar déjà mentionné, où $\phi(x) = 1_{x\in[0,1]}$ et $\psi(x) = 1_{x\in[0,1/2]} - 1_{x\in]1/2,1]}$, possède ces propriétés mais les fonctions associées présentent le défaut d'être irrégulières et peu oscillantes. C'est pour pallier cet inconvénient que Daubechies (1988) exhiba des fonctions ϕ et ψ de ***régularité*** r, c'est-à-dire : ϕ et ψ sont de classe C^r, à décroissance rapide ainsi que chacune de leurs dérivées jusqu'à l'ordre r. Ainsi, ces propriétés vont permettre d'analyser non seulement l'espace des fonctions de $L_2(\mathbb{R})$, comme le fait déjà l'analyse de Fourier, mais également d'autres classes fonctionnelles comme on le verra en section 2.1.2. Daubechies montra également que les fonctions ϕ et ψ pouvaient être choisies à support compact fournissant ainsi des bases inconditionnelles des espaces $L_p(\mathbb{R})$, $1 < p < \infty$ (pour plus de détails, voir section 2.3.3). Des exemples de systèmes "fonction d'échelle/ondelette" (ϕ, ψ) possédant ces diverses propriétés sont donnés par Härdle, Kerkyacharian, Picard et Tsybakov (1998).

Terminons cette section en rappelant qu'en adaptant certains aspects de la théorie décrite ci dessus, on peut exhiber des bases orthonormées d'ondelettes pour l'espace $L_2([0,1])$ (voir Meyer (1992)). L'analyse par ondelettes s'applique également aux fonctions périodiques. Dans ce cas, on utilise des formes périodisées des fonctions ϕ et ψ et la base orthonormée correspondante peut alors s'écrire sous la forme $(\tilde{\psi}_{jk})_{j\geq -1, 0\leq k<2^j}$ si la période est égale à 1. Enfin, il est possible de construire des bases orthonormées d'ondelettes pour $L_2(\mathbb{R}^d)$, $d \in \mathbb{N}^*$, par la méthode du produit tensoriel. Cette méthode est brièvement décrite en section 5.3.1.

2.1.2 Espaces de Besov forts

Nous avons défini en introduction des espaces fonctionnels fondamentaux pour notre étude : les espaces de Lorentz et plus particulièrement la classe des espaces de Besov faibles. Nous avons illustré les fortes connexions qui existent entre les espaces de Besov faibles et forts sans rappeler la définition de ces derniers de manière précise. L'objet de cette section est de combler ce vide et de présenter succinctement quelques propriétés des espaces de Besov forts notamment dans le cadre de l'analyse par ondelettes. Pour plus de détails, on se référera aux travaux de Bergh et Löfström (1976), Peetre (1976), Meyer (1992) ou DeVore et Lorentz (1993).
Commençons par donner la définition de la classe des espaces de Besov forts en termes de module de continuité. Soit $0 < s < 1$, $1 \leq p \leq \infty$, $1 \leq q < \infty$, et posons $\forall\ (x,h) \in \mathbb{R}^2$, $\tau_h f(x) = f(x-h)$. On définit

$$\gamma_{spq}(f) = \left(\int_{\mathbb{R}} \left(\frac{\|\tau_h f - f\|_{L_p}}{|h|^s}\right)^q \frac{dh}{|h|}\right)^{1/q},$$

et

$$\gamma_{sp\infty}(f) = \sup_{h\in\mathbb{R}} \frac{\|\tau_h f - f\|_{L_p}}{|h|^s}.$$

Si $s = 1$, on pose,

$$\gamma_{1pq}(f) = \left(\int_{\mathbb{R}} \left(\frac{\|\tau_h f + \tau_{-h} f - 2f\|_{L_p}}{|h|}\right)^q \frac{dh}{|h|}\right)^{1/q},$$

et

$$\gamma_{1p\infty}(f) = \sup_{h\in\mathbb{R}} \frac{\|\tau_h f + \tau_{-h} f - 2f\|_{L_p}}{|h|}.$$

En s'aidant de ces définitions, on définit $\forall\ 0 < s \leq 1$ et $\forall\ 1 \leq p, q \leq \infty$, l'espace de Besov fort de paramètres s, p, et q

$$\mathcal{B}_{s,p,q} = \{f \in L_p(\mathbb{R}) : \quad \gamma_{spq}(f) < \infty\},$$

muni de la norme

$$\|f\|_{\mathcal{B}_{s,p,q}} = \|f\|_{L_p} + \gamma_{spq}(f).$$

Si $s = N + \alpha$, avec $N \in \mathbb{N}$ et $0 < \alpha \leq 1$, on dit que $f \in \mathcal{B}_{s,p,q}$ si et seulement si $f^{(m)} \in \mathcal{B}_{\alpha,p,q}$, pour tout $m \leq N$. Cet espace est muni de la norme :

$$\|f\|_{\mathcal{B}_{s,p,q}} = \|f\|_{L_p} + \sum_{m \leq N} \gamma_{\alpha pq}(f^{(m)}).$$

Rappelons que l'espace de Sobolev H^s correspond à l'espace de Besov $\mathcal{B}_{s,2,2}$, l'espace de Hölder C^s (avec $0 < s \notin \mathbb{N}$) à $\mathcal{B}_{s,\infty,\infty}$.
Une caractérisation essentielle des espaces de Besov forts repose sur la notion de vitesse d'approximation. On a le résultat suivant :

Théorème 2.1. *Supposons donné un système "fonction d'échelle/ondelette" (ϕ, ψ) de régularité r. Considérons les opérateurs de projection P_j, $j \geq 0$, sur les espaces V_j de l'analyse multirésolution engendré par (ϕ, ψ). Alors, pour tous $0 < s < r$, $1 \leq p, q \leq \infty$, f appartient à l'espace de Besov fort $\mathcal{B}_{s,p,q}$ si et seulement si $f \in L_p(\mathbb{R})$ et il existe une suite de nombres positifs $(\varepsilon_j)_{j \in \mathbb{N}} \in l_q(\mathbb{N})$ telle que*

$$\forall\, j \in \mathbb{N}, \quad \|f - P_j f\|_{L_p} \leq 2^{-js} \varepsilon_j.$$

Enfin, rappelons que les espaces de Besov forts admettent une caractérisation en termes de coefficients d'ondelette. Sous les hypothèses du théorème précédent, pour toute fonction $f \in L_p(\mathbb{R})$, on peut définir les coefficients :

$$\alpha_{0k} = \int f(x)\phi_{0k}(x)dx, \quad \beta_{jk} = \int f(x)\psi_{jk}(x)dx,$$

et f appartient à l'espace de Besov fort $\mathcal{B}_{s,p,q}$ si et seulement si

$$J_{s,p,q}(f) = \|\alpha_{0.}\|_{l_p} + \left(\sum_{j \geq 0} 2^{jq(s+1/2-1/p)} \|\beta_{j.}\|_{l_p}^q \right)^{1/q} < \infty, \quad \text{si } q < \infty,$$

et

$$J_{s,p,\infty}(f) = \|\alpha_{0.}\|_{l_p} + \sup_{j \geq 0} 2^{j(s+1/2-1/p)} \|\beta_{j.}\|_{l_p} < \infty \quad \text{si } q = \infty.$$

De plus, la norme $\|.\|_{\mathcal{B}_{s,p,q}}$ est équivalente à la norme $J_{s,p,q}$. On en déduit aisément les inclusions suivantes :

$$\mathcal{B}_{s',p,q'} \subset \mathcal{B}_{s,p,q}, \text{ pour } s' > s \text{ ou pour } s' = s \text{ et } q' \leq q,$$

$$\mathcal{B}_{s,p,q} \subset \mathcal{B}_{s',p',q}, \text{ pour } p' > p, \ s' = s - (1/p - 1/p').$$

En particulier, pour $s > 1/p$, $q > 1$, $\mathcal{B}_{s,p,q} \subset \mathcal{B}_{s',\infty,\infty}$ est inclus dans l'espace des fonctions continues et bornées. On a aussi :

$$\mathcal{B}_{0,p',\min(p',2)} \subset L_{p'}, \quad p' \geq 1$$

où $\mathcal{B}_{0,p',\min(p',2)}$ est défini à travers la norme $J_{s,p,q}$ avec $s = 0$.
Jusqu'à la fin du chapitre 2, nous considérons une base d'ondelettes $(\psi_{jk})_{j\geq -1,k\in\mathbb{Z}}$ de régularité plus grande que 1.

2.2 Modèles statistiques

L'objet de cette section est de décrire, d'une part, les modèles statistiques qui sont à la base des résultats de nos travaux, mais surtout les raisons qui nous ont conduits à considérer de telles modélisations.

2.2.1 Modèle de régression et transformée en ondelettes discrète

Un des problèmes statistiques les plus classiques consiste à estimer une fonction à partir des observations bruitées des valeurs de cette fonction calculées en n points répartis de manière équidistante sur un intervalle compact. Ainsi, il est très naturel de considérer le modèle de régression non paramétrique suivant :

$$g_i = f(\frac{i}{n}) + \sigma\varepsilon_i, \quad 1 \leq i \leq n, \tag{2.1}$$

où f est la fonction à estimer à partir des n observations $g_1, \dots, g_n$, et chaque ε_i suit une loi normale centrée réduite. Le niveau de bruit σ sera supposé connu, et en général, les ε_i seront indépendants. Les procédures statistiques décrites dans le chapitre 4 sont construites pour le modèle (2.1). Cette modélisation étant fort utilisée, nous n'avons eu aucune peine à comparer nos procédures à toute une gamme d'autres procédures performantes.
Nous exploitons la modélisation précédente en utilisant les outils de la *transformée en ondelettes discrète* : chaque vecteur de taille dyadique subit une succession de transformations linéaires orthogonales définies à partir de filtres associés à un système fonction d'échelle/ondelette (ϕ, ψ). Si $n = 2^N$, $N \in \mathbb{N}$, on construit ainsi une matrice orthogonale W qui transforme le vecteur $f^0 = (f(\frac{i}{n}), \quad 1 \leq i \leq n)^T$ en un vecteur de même taille noté $d = (d_{jk})_{-1\leq j\leq N-1,k\in\mathcal{I}_j}$, où $\mathcal{I}_j = \{k \in \mathbb{N} : \quad 0 \leq k < 2^j\}$. Le vecteur f^0 est reconstruit en utilisant la formule $f^0 = W^T d$. Mallat (1989) montra que l'ensemble de ces opérations pouvait être effectuées en $O(n)$ opérations. Sous certaines conditions (voir Donoho et Johnstone (1994a)), si $W_{jk,i}$ désigne le coefficient se trouvant à l'intersection de la $([2^j] + 1 + k)$ème ligne et de la $i$ème colonne de W, on a l'approximation suivante :

$$n^{\frac{1}{2}} W_{jk,i} \approx 2^{\frac{j}{2}} \psi(2^j i/n - k).$$

Nous en déduisons :

$$d_{jk} \approx n^{\frac{1}{2}} \beta_{jk}, \tag{2.2}$$

où les β_{jk} désignent les coefficients d'ondelette ordinaires de la fonction f :

$$\beta_{jk} = \int_0^1 f(t)\psi_{jk}(t)dt.$$

Si les ε_i sont indépendants, puisque la transformation W est orthogonale, on obtient donc le modèle suivant :

$$y_{jk} = d_{jk} + \sigma z_{jk}, \quad z_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad -1 \leq j \leq N-1, k \in \mathcal{I}_j,$$

où

$$y_{jk} = (\mathcal{W}g)_{jk},$$

et

$$z_{jk} = (\mathcal{W}\varepsilon)_{jk}.$$

Une présentation détaillée de l'algorithme précédent est donnée par Daubechies (1992) ou par Härdle, Kerkyacharian, Picard et Tsybakov (1998). Parce que cet algorithme utilise une extension périodique du vecteur f^0, il est préférable d'utiliser des fonctions de $[0,1]$ que l'on peut prolonger de manière périodique sur $\mathbb{R}$ et sans perte de régularité, ce que nous ferons tout au long de ces travaux.

Bien que pas toujours des plus fiables, l'approximation (2.2) permet donc de relier un modèle pratique (le modèle (2.1)) et des modèles plus théoriques, comme par exemple, le modèle de bruit blanc gaussien décrit ci-dessous. Ainsi, les résultats théoriques du Théorème 3.1 établis sous ce modèle statistique peuvent s'appliquer de manière concrète, et ceci, grâce à la transformée en ondelettes discrète.

2.2.2 Modèle de bruit blanc gaussien

Dans le chapitre 3, le modèle statistique sous lequel nous calculons le risque minimax pour les boules de Besov faibles est le modèle de bruit blanc gaussien construit à partir d'un processus de Wiener $(W_t)_t$:

$$dY_t = f(t)\, dt + \varepsilon dW_t, \quad t \in [0,1], \tag{2.3}$$

où l'on cherche à reconstruire le signal f quand l'intensité du bruit ε tend vers 0 à partir des observations $\left\{\int_0^1 \phi(t)dY_t : \quad \phi \in L_2([0,1],dt)\right\}$. Ce modèle joue un rôle central en statistique (voir Ibragimov et Khasminskii (1981)). Simple à utiliser, il approche très bien d'autres modèles classiques, comme les modèles de régression non paramétrique, de diffusion ou le modèle de l'estimation d'une densité (voir Brown and Low (1996) ou Nussbaum (1996)). La plupart du temps, on l'utilise de la manière suivante : on suppose donnée une base orthonormée $\mathcal{E} = (\psi_k)_{k\in\mathbb{N}^*}$ de $L_2([0,1])$, et si f est décomposée sur cette base $f(t) = \sum_{k\in\mathbb{N}^*} \theta_k\psi_k(t)$, on estime les coefficients θ_k à partir des observations $x_k = \int \psi_k(t)dY_t$, $k \in \mathbb{N}^*$, qui vérifient :

$$x_k = \theta_k + \varepsilon\xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k \in \mathbb{N}^*. \tag{2.4}$$

Un cas particulier mérite d'être relevé : le cas où la base $\mathcal{E}$ est une base d'ondelettes $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$. Le modèle séquentiel que l'on obtient peut alors s'écrire :

$$y_{jk} = \beta_{jk} + \varepsilon z_{jk}, \quad z_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad j \geq -1,\ k \in \mathbb{Z}.$$

C'est sous cette forme que nous utilisons le modèle de bruit blanc gaussien dans le chapitre 3. Le modèle (2.4) décrit précédemment est en fait un cas particulier d'un modèle séquentiel plus général fréquemment utilisé en statistique des problèmes inverses.

2.2.3 Problèmes inverses et modèle hétéroscédastique

Supposons donné X un espace de Hilbert muni du produit scalaire $(.,.)$ et de la norme associée $\|.\|$. La modélisation statistique d'un problème linéaire inverse avec bruit aléatoire peut s'écrire sous la forme symbolique suivante :

$$Y = Af + \varepsilon\xi, \tag{2.5}$$

où Y est l'observation, A est un opérateur linéaire supposé connu d'un espace $D \subset X$ à valeurs dans X, ξ, aléatoire prend ses valeurs dans X et ε, l'intensité du bruit tend vers 0. Comme pour le modèle (2.3), cette modélisation signifie que $\forall\ u \in X$, $Y(u) = (Af,u) + \varepsilon\xi(u)$ est observable. Nous supposons de plus que $\xi(u)$ est une variable aléatoire gaussienne sur un espace de probabilité $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ de moyenne nulle et de variance $\|u\|^2$. On suppose également que $\forall\ (u,v) \in X^2$, $\mathbb{E}(\xi(u)\xi(v)) = (u,v)$, où $\mathbb{E}$ désigne l'espérance associée à la probabilité $\mathbb{P}$. Les travaux pionniers relatifs à l'étude statistique des problèmes inverses avec bruit aléatoire datent des années 1960 avec les articles de Sudakov et Khalfin (1964) et Bakushinskii (1969). Plus récemment, une multitude de méthodes de résolution furent proposées. Citons par exemple, Wahba (1977, 1990), Vapnik (1982), O'Sullivan (1986), Kress (1989), Johnstone et Silverman (1990), van Rooij et Ruymgaart (1991, 1996), Ruymgaart (1992), Koo (1993), Korostelev et Tsybakov (1993), Donoho (1995), Mair et Ruymgaart (1996), Cavalier (1998), Johnstone (1999), Cavalier, Golubev, Picard et Tsybakov (2000), Cavalier et Tsybakov (2000), Goldenshluger et Pereverzev (2000), Tsybakov (2000) et les travaux cités ci dessous.

Une *décomposition en valeurs singulières* de l'opérateur A permet de construire naturellement une modélisation séquentielle associée au modèle (2.5). Supposons que A est injectif, et que A^*A est un opérateur compact de X, où A^* désigne l'opérateur adjoint de A. Il existe alors une suite strictement positive $(b_k^2)_{k\in\mathbb{N}^*}$ et un système de fonctions orthonormées $(\phi_k)_{k\in\mathbb{N}^*}$ de X telles que $A^*A\phi_k = b_k^2\phi_k$. Posons

$$\psi_k = b_k^{-1}A\phi_k.$$

Le système $(\psi_k)_{k\in\mathbb{N}^*}$ forme également un système orthonormé de X. Les relations

$$A\phi_k = b_k\psi_k, \quad A^*\psi_k = b_k\phi_k$$

constituent une décomposition en valeurs singulières de A. De plus, on peut écrire :

$$\forall\, f \in D, \quad Af = \sum_{k\in\mathbb{N}^*} b_k^{-1}(Af,\psi_k)A\phi_k$$

et comme A est injectif

$$f = \sum_{k\in\mathbb{N}^*} b_k^{-1}(Af,\psi_k)\phi_k = \sum_{k\in\mathbb{N}^*}(f,\phi_k)\phi_k.$$

Le modèle séquentiel associé au modèle (2.5) est alors le suivant :

$$y_k = b_k\theta_k + \varepsilon\xi_k, \quad k \in \mathbb{N}^*,$$

où les $y_k = Y(\psi_k)$ sont les observations, les $\xi_k = \xi(\psi_k)$ sont des variables aléatoires gaussiennes indépendantes centrées réduites et $\forall\, k \in \mathbb{N}^*$, $\theta_k = (f,\phi_k)$. En posant $\forall\, k \in \mathbb{N}^*$, $\sigma_k = b_k^{-1}$, $x_k = \sigma_k y_k$, on obtient de manière équivalente :

$$x_k = \theta_k + \varepsilon\sigma_k\xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k \in \mathbb{N}^*. \tag{2.7}$$

L'opérateur A étant connu, les σ_k le sont également. Le modèle (2.7), appelé *modèle hétéroscédastique de bruit blanc gaussien*, apparaît donc comme une généralisation du modèle homoscédastique (2.4).
L'étude statistique d'un grand nombre de problèmes inverses est réalisée grâce à la modélisation précédente. Elle dépend alors fortement du comportement asymptotique de la suite $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$. Le cas ardu où $\sigma_k \overset{k\to\infty}{\longrightarrow} +\infty$ est fort répandu. Par exemple, les suites σ associées aux opérateurs tels que l'intégration considérés par Ruymgaart (1993), la transformée de Radon (voir Cavalier et Tsybakov (2000)), la convolution pour le cas étudié par Cavalier et Tsybakov (2000) ou les opérateurs associés à certaines équations différentielles elliptiques (voir Mair et Ruymgaart (1996))

ont une croissance polynomiale. Mais les σ_k peuvent croître de manière exponentielle, comme l'ont montré par exemple, Pereverzev et Schock (1999) pour le problème de géodésie satellite, ou Mair et Ruymgaart (1996) et Golubev et Khasminskii (1999) pour les problèmes inverses associés à l'équation de la chaleur.
En exploitant le cadre d'une analyse multirésolution, Johnstone et Silverman (1997) montrèrent que le modèle hétéroscédastique de bruit blanc gaussien peut être aussi utilisé pour représenter des observations ayant une structure corrélée. Plus précisément, supposons donné le modèle de régression non paramétrique suivant :

$$g_i = f(\frac{i}{n}) + e_i, \quad 1 \leq i \leq n,$$

où les variables e_i sont issues d'un processus gaussien stationnaire. En étudiant la fonction d'autocorrelation des erreurs, Johnstone et Silverman (1997) prouvèrent que sous un bon choix de l'intensité de bruit à chaque niveau de résolution j, le modèle (2.7) constitue une bonne approximation du modèle précédent. Pour illustrer cette affirmation, supposons que

$$\lim_{k \to +\infty} k^\alpha \text{cov}(e_i, e_{i+k}) = K < \infty,$$

avec $0 < \alpha < 1$. Dans ce cas, l'intensité du bruit sera proportionnelle à $2^{-j(1-\alpha)/2}$.

2.3 Quelques résultats antérieurs de la théorie maxiset

L'objet de cette section est de rappeler quelques résultats de la théorie maxiset obtenus par Cohen, DeVore, Kerkyacharian et Picard. Nous pourrons alors constater, dans une certaine mesure, le rôle joué par les espaces de Besov forts dans le cadre de cette approche, mais surtout, l'importance primordiale des classes fonctionnelles introduites dans les sections 1.2.2 et 1.2.3 (espaces de Lorentz, de Besov faibles, ...).

2.3.1 Modèle de l'estimation d'une densité

Comme nous le rappelions en introduction, les premiers résultats de type maxiset remontent au début des années 1990 avec l'article de Kerkyacharian et Picard (1993) qui traite du problème de l'estimation d'une densité. Supposons données $X_1, \ldots, X_n$, n variables aléatoires indépendantes ayant chacune pour densité f. Kerkyacharian et Picard (1992) prouvèrent que si $2 \leq p < \infty$, alors les vitesses de convergence minimax pour la perte L_p sur les boules de l'espace de Besov fort $\mathcal{B}_{s,p,\infty}$ pour $0 < s < \infty$ sont atteintes par des estimateurs de la forme

$$\hat{E}_{j_n}(x) = \frac{1}{n} \sum_{i=1}^{n} E_{j_n}(x, X_i) = \frac{1}{n} \sum_{i=1}^{n} 2^{j_n} E(2^{j_n} x, 2^{j_n} X_i),$$

où $E(u,v)$ est un noyau de projection sur un espace V_0 d'une analyse multirésolution et $2^{j_n} = n^{1/(1+2s)}$. Mais Kerkyacharian et Picard (1993) établirent un résultat "inverse" qui prouva les forts liens existant entre les espaces de Besov forts $\mathcal{B}_{s,p,\infty}$ et les estimateurs linéaires. Ils démontrèrent :

Théorème 2.2. *Soit p un réel tel que* $2 \leq p < \infty$, *et* V *un ensemble de densités inclus dans une boule de* L_p. *On suppose qu'il existe* $C_2 > 0$ *telle que,*

$$\forall\, n \geq 1, \quad \inf_{\hat{f} \in F} \sup_{f \in V} \mathbb{E}_f \|\hat{f} - f\|_{L_p}^p \geq C_2 n^{-\frac{ps}{2s+1}},$$

où F *est un ensemble d'estimateurs contenant les estimateurs linéaires, et il existe* E *tel que*

- $E(x+1, y+1) = E(x,y)$,
- $|E(x,y)| \leq \Phi(x-y)$ *où* $\Phi \in L_1(\mathbb{R}) \cap L_\infty(\mathbb{R})$,
- *il existe* $N > s$, *tel que* $x \to E(x,y)$ *est faiblement* N*-dérivable, pour tout* $y \in \mathbb{R}$ *avec* $|D_x^N E(x,y)| \leq \Phi(x-y)$,
- $E \circ E_j = E_j \circ E,\ \forall\, j \in \mathbb{N}$,

et une suite croissante j_n *qui tend vers* $+\infty$ *quand* n *tend vers* $+\infty$, *telle que*

$$\forall\, n \geq 1, \quad \sup_{f \in V} \mathbb{E}_f \|\hat{E}_{j_n} - f\|_{L_p}^p \leq C_1 n^{-\frac{ps}{2s+1}},$$

où C_1 *est une constante strictement positive. Alors,* V *est inclus dans une boule de* $\mathcal{B}_{s,p,\infty}$.

Les conditions requises sur le noyau E sont facilement vérifiées par les noyaux classiques. Ainsi, les espaces de Besov forts $\mathcal{B}_{s,p,\infty}$ apparaissent comme les espaces maximaux associés aux estimateurs linéaires sous le modèle statistique de l'estimation d'une densité.

2.3.2 Modèle de bruit blanc gaussien homoscédastique

Nous avons déjà donné, dans la section 1.2.2 de l'introduction, le résultat suivant concernant la procédure d'estimation par seuillage des coefficients d'ondelette sous le modèle de bruit blanc gaussien :

Théorème 2.3. *Soit* $1 < r < \infty$ *et* $\alpha \in (0,1)$. *On suppose donnée une fonction*

$$f = \sum_{j \geq -1} \sum_k \beta_{jk} \psi_{jk} \in L_r([0,1]).$$

Sous le modèle de bruit blanc gaussien

$$x_{jk} = \beta_{jk} + \varepsilon\xi_{jk}, \quad \xi_{jk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad j \geq -1,\ k \in \mathbb{Z},$$

on considère l'estimateur suivant :

$$\hat{f}_\varepsilon^T = \sum_{j=-1}^{j_\varepsilon} \sum_k x_{jk} \mathbb{1}_{\ |x_{jk}|>\kappa t_\varepsilon} \psi_{jk},$$

avec

- $t_\varepsilon = \varepsilon\sqrt{\log(\varepsilon^{-1})}$,
- $2^{-j_\varepsilon} \leq \varepsilon^2 \log(\varepsilon^{-1}) < 2^{-j_\varepsilon+1}$,
- κ *une constante assez grande.*

On a $\forall\ \varepsilon$,

$$\mathbb{E}\|\hat{f}_\varepsilon^T - f\|_r^r \leq K\left(\varepsilon\sqrt{\log(\varepsilon^{-1})}\right)^{\alpha r},$$

où K *est une constante si et seulement si* f *appartient à la fois à l'espace de Besov fort* $\mathcal{B}_{\frac{\alpha}{2},r,\infty}$ *et à* $\mathcal{W}^*(r,(1-\alpha)r)$, *où* $\forall\ 0 < r,p < \infty$,

$$\mathcal{W}^*(r,p) = \left\{ f = \sum_{j\geq -1}\sum_k \beta_{jk}\psi_{jk} : \quad \sup_{\lambda>0} \lambda^p \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} \sum_k \mathbb{1}_{\ |\beta_{jk}|>\lambda} < \infty \right\}.$$

Sous ce même modèle statistique, Kerkyacharian et Picard (2002) établirent des résultats beaucoup plus généraux. Ils introduisirent la notion suivante :

Définition 2.2. *Soit* $r \geq 1$ *et* F *une fonctionnelle telle que si* f *est une fonction,* $F(f)(.,.)$ *est une fonction positive sur* $\mathbb{N}\times[0,1]$, *et pour presque tout* x, $j \longrightarrow F(f)(j,x)$ *est croissante et* $F(f)(0,x) < \infty$. *On dit que la suite d'estimateurs* $\hat{f}_n$ *satisfait une inégalité oracle locale d'ordre* r *sur un espace fonctionnel* V, *associée à une suite d'opérateurs* E_j, *à la fonctionnelle locale* F, *et à la suite d'arrêts* J_n *tendant vers* $+\infty$, *si les inégalités suivantes sont vraies pour tous* $n \in \mathbb{N}^*$ *et* $f \in V$.

$$\mathbb{E}|\hat{f}_n(x) - f(x)|^r \leq C((2^{j_n^*(x)/2}t_n)^r + |E_{j_n^*(x)}f(x) - f(x)|^r + |E_{J_n}f(x) - f(x)|^r), \quad \forall\ x \in [0,1],$$

$$\|\sup_{j'\geq j}|E_{j'}f - f|\mathbb{1}_{\ j_\lambda^F(f,.)=j}\|_{L_r}^r \leq C'(2^{j/2}\lambda)^r \mu(x : \quad j_\lambda^F(f,x) = j), \quad \forall\ j \geq 0,\ \forall\ \lambda > 0,$$

où

$$j_\lambda^F(f,x) = \inf\{j \in \mathbb{N} : \quad F(f)(j,x) \leq \lambda\},$$

$$t_n = (\log n/n)^{1/2}, \quad j_n^*(x) = j_{t_n}^F(f,x),$$

μ *est la mesure de Lebesgues et* C *et* C' *sont deux constantes.*

Ils montrèrent le résultat suivant qui fait le lien entre les approches oracles et maxiset :

Théorème 2.4. *Soient $T > 0$ et β une constante strictement positive. Si la suite d'estimateurs $\hat{f}_n$ satisfait une inégalité oracle locale d'ordre r sur l'espace fonctionnel $V = \mathcal{B}_{\alpha/\beta,r,\infty}(M)$, associée à une suite d'opérateurs E_j, à la fonctionnelle locale F, et à la suite d'arrêts J_n, et si J_n est telle que $2^{J_n} < t_n^{-\beta} \leq 2^{J_n+1}$, il existe M' telle que*

$$W(F)(r,(1-\alpha)r)(M') \subset \{f \in V : \quad \sup_n \mathbb{E}\|\hat{f}_n - f\|_{L_r}^r (t_n)^{-\alpha r} \leq T\},$$

où $\forall\ 0 < p < r$

$$W(F)(r,p)(M) = \{f \in V : \quad \sup_{\lambda > 0} \lambda^p \sum_j 2^{j\frac{r}{2}} \mu(x : \quad F(f)(j,x) > \lambda) \leq M^p\}.$$

Notons que $W(F)(r,p)(M)$ coïncide avec $\mathcal{B}_{\alpha/\beta,r,\infty}(M) \cap W^*(r,p)(M)$ si

$$F(f)(j,x) = 2^{-j/2} \sum_k |\beta_{jk}| h_{jk}(x),$$

où les β_{jk} sont les coefficients d'ondelettes de f et $h_{jk}(x) = 2^{j/2} \mathbb{1}_{\,2^j x - k \in [0,1]}$. En s'aidant de ce résultat, Kerkyacharian et Picard montrèrent que pour le modèle de bruit blanc gaussien et sous certaines conditions, certains espaces de Lorentz sont inclus dans les espaces maximaux pour la procédure consistant à sélectionner localement le pas d'un noyau. Ceci permit à Kerkyacharian et Picard d'affirmer que cette dernière procédure est au moins aussi performante que la procédure de type seuillage, du point de vue maxiset.

2.3.3 Modèle hétéroscédastique. Bases inconditionnelles. Superconcentration

La section précédente décrit les espaces maximaux obtenus pour les procédures de type seuillage, sous le modèle de bruit blanc gaussien. Il est naturel de chercher à étendre ces résultats pour le modèle de bruit blanc gaussien hétéroscédastique. Pour le risque considéré avec la fonction de perte associée à la norme séquentielle l_p, et sous le modèle statistique (2.7), Kerkyacharian et Picard (2000) établirent le résultat suivant :

Théorème 2.5. *Soit $1 \leq p < \infty$ un réel fixé et $0 < r < \infty$. On pose $\forall\ 0 < \varepsilon \leq \varepsilon_0$ où ε_0 est tel que $\varepsilon_0 \sqrt{\log(1/\varepsilon_0)} = 1$:*

$$\hat{\theta}_k^t(x_k) = \begin{cases} x_k \mathbb{1}_{\,|x_k| \geq \kappa_t \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}} & si\ k < \Lambda_\varepsilon, \\ 0 & sinon. \end{cases}$$

On suppose que

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \Lambda_\varepsilon = \left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{-r},$$

et il existe une constante strictement positive C_t*, telle que* $\forall\, 0 < \varepsilon \leq \varepsilon_0$,

$$\varepsilon^{\frac{\kappa_t^2}{16}} \log(1/\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \leq C_t.$$

Soit ν *un réel fixé tel que* $0 < \nu < \infty$.
Alors, si $\kappa_t \geq \sqrt{2p}$*, il existe une constante strictement positive* C *telle que*

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \left(\mathbb{E}\|\hat{\theta}^t - \theta\|_{l_p}^p\right)^{\frac{1}{p}} \leq C \left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{2\nu/(2\nu+1)}$$

$$\iff \theta \in wl_{p,p/(2\nu+1)}(\sigma) \cap B_{p,\infty}^{\frac{2\nu}{r(2\nu+1)}}(\mu),$$

avec $\forall\, 0 < q < p$, $\forall\, 0 < \eta < \infty$,

$$\begin{aligned} wl_{p,q}(\sigma) &= \left\{\theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^q \sum_k 1\!\!1_{\,|\theta_k|>\lambda\sigma_k} \sigma_k^p < \infty\right\}, \\ B_{p,\infty}^\eta(\mu) &= \left\{\theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^{p\eta} \sum_{k\geq\lambda} \mu_k |\theta_k|^p < \infty\right\}, \end{aligned}$$

et μ *est la mesure de comptage sur* $\mathbb{N}^*$.

Il est bien entendu naturel de chercher à étendre ce résultat à des fonctions de perte L_p, notamment pour l'estimation de fonctions décomposées sur une base $(\psi_k)_{k\in\mathbb{N}^*}$:

$$f = \sum_{k\in\mathbb{N}^*} \theta_k \psi_k.$$

Kerkyacharian et Picard (2000) ont montré que cet objectif pouvait être atteint si la base $(\psi_k)_{k\in\mathbb{N}^*}$ est une base *inconditionnelle* vérifiant *une inégalité de superconcentration.*
Rappelons qu'une base inconditionnelle d'un espace de Banach X muni d'une norme $\|.\|$ est une famille dénombrable $(\phi_k)_{k\in\mathbb{N}^*}$ possédant les deux propriétés suivantes :

- pour tout $g \in X$, il existe une unique suite $\lambda = (\lambda_k)_{k\in\mathbb{N}^*}$ telle que

  $$g = \sum_{k\in\mathbb{N}^*} \lambda_k \phi_k$$

  (la série est convergente pour la norme $\|.\|$),

- il existe une constante absolue notée K telle que si $\forall\, k \in \mathbb{N}^*$, $|\lambda_k| \le |\lambda'_k|$, alors

$$\| \sum_{k\in\mathbb{N}^*} \lambda_k \phi_k \| \le K \| \sum_{k\in\mathbb{N}^*} \lambda'_k \phi_k \|$$

(propriété de contraction).

Dans le cadre de notre étude, nous nous intéressons au cas où $X = L_p(D^d)$, où $d \in \mathbb{N}^*$, $D = [0,1]$ ou $D = \mathbb{R}$ et $1 < p < \infty$, cette dernière restriction étant due au fait qu'en général, il n'existe pas de bases inconditionnelles si $p \notin (1, \infty)$ (voir Lindenstrauss et Tzafriri (1977)). Le théorème suivant prouvé par DeVore, Konyagin et Temlyakov (1998) donne un premier outil pour étendre des résultats établis pour le risque L_2 à l'estimation L_p de fonctions :

Théorème 2.6. *Soit $(\phi_k)_{k\in\mathbb{N}^*}$ une suite de $L_p(D^d)$, avec $1 < p < \infty$. On a l'équivalence des deux assertions suivantes :*

- *$(\phi_k)_{k\in\mathbb{N}^*}$ est une base inconditionnelle de $L_p(D^d)$.*
- *$(\phi_k)_{k\in\mathbb{N}^*}$ est un système total dans $L_p(D^d)$, et il existe une constante $K > 0$, telle que, pour tout ensemble $F \subset \mathbb{N}$, et pour toute suite de coefficients $(\lambda_k)_{k\in\mathbb{N}^*}$,*

$$K^{-1} \left\| \sum_{k\in F} \lambda_k \phi_k \right\|_{L_p} \le \left\| \left(\sum_{k\in F} \lambda_k^2 \phi_k^2 \right)^{1/2} \right\|_{L_p} \le K \left\| \sum_{k\in F} \lambda_k \phi_k \right\|_{L_p}$$

A ce stade, il faut rappeler que les ondelettes à support compact sont des bases inconditionnelles des espaces $L_p(D)$, $1 < p < \infty$, où $D = [0,1]$ ou $D = \mathbb{R}$ (voir Meyer (1992)). Ce résultat peut être étendu aux espaces multidimensionnels grâce aux ondelettes à support compact construites à l'aide de la méthode du produit tensoriel (voir section 5.3.1). On dispose ainsi de bases inconditionnelles pour tous les espaces $L_p(D)$, $1 < p < \infty$ où $D = [0,1]^d$ ou $D = \mathbb{R}^d$, $d \in \mathbb{N}^*$.
Décrivons à présent ce qu'on appelle la propriété de superconcentration :

Définition 2.3. *Soit $(e_i)_{i\in I}$ une suite de fonctions définies sur $\mathbb{R}^d$ ou sur $[0,1]^d$. On dit que cette famille de fonctions satisfait une inégalité de superconcentration si, pour tous r, p avec $0 < p < \infty$ et $0 < r < \infty$, il existe une constante C (dépendant de p, r, et de d) telle que pour tout $F \subset I$,*

$$\left\| \left(\sum_{i\in F} |e_i|^r \right)^{1/r} \right\|_{L_p} \le C \left\| \sup_{i\in F} |e_i| \right\|_{L_p}$$

On en déduit aisément la propriété suivante :

Proposition 2.1. *Si la suite de fonctions* $(e_i)_{i\in I}$ *satisfait une inégalité de superconcentration, alors pour tout* $0 < p < \infty$*, il existe deux constantes* c_p *et* C_p *telles que pour tout* $F \subset I$*, on a*

$$c_p \int \sum_{i\in F} |e_i|^p \leq \int \left(\sum_{i\in F} |e_i|^2\right)^{p/2} \leq C_p \int \sum_{i\in F} |e_i|^p.$$

Les inégalités de la proposition précédente ont été introduites par DeVore (1998) et Temlyakov (1999). Elles entrent en jeu pour traiter du problème de l'estimation fonctionnelle sous la perte L_p. Le théorème suivant démontré par Kerkyacharian et Picard (2001) sera utilisé par les mêmes auteurs pour exhiber les espaces maximaux des procédures de type seuillage des coefficients d'ondelettes sous le modèle hétéroscédastique pour la perte L_p $(1 < p < \infty)$:

Théorème 2.7. *Soit* $(\psi_k)_{k\in\mathbb{N}^*}$ *une base d'ondelettes de* $\mathbb{R}^d$ *ou de* $[0,1]^d$*,* $d \in \mathbb{N}^*$*, construite par la méthode du produit tensoriel, et à support compact . Alors* $(\psi_k)_{k\in\mathbb{N}^*}$ *vérifie une inégalité de superconcentration. Il en est de même pour* $\{\sigma_k\psi_k,\ k \in \mathbb{N}^*\}$ *si les* σ_k *dépendent seulement des niveaux de résolution et si au niveau* j*, l'intensité du bruit est proportionnelle à* 2^{bj}*,avec* $b > -\frac{d}{2}$*.*

Finalement, Kerkyacharian et Picard (2000) établirent le résultat suivant :

Théorème 2.8. *Soit* $1 < p < \infty$ *un réel fixé et* $0 < r < \infty$*. Considérons le modèle statistique (2.7) et l'estimateur* $\hat{\theta}$ *introduit dans le Théorème 2.5. Supposons donnée* $\mathcal{E} = (\psi_k)_{k\in\mathbb{N}^*}$ *une base inconditionnelle de* L_p *telle que* $(\sigma_k\psi_k)_{k\in\mathbb{N}^*}$ *vérifie une inégalité de superconcentration. On suppose que*

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \Lambda_\varepsilon = \left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{-r},$$

où ε_0 *est tel que* $\varepsilon_0\sqrt{\log(1/\varepsilon_0)} = 1$*, et il existe une constante strictement positive* C_t*, telle que* $\forall\, 0 < \varepsilon \leq \varepsilon_0$*,*

$$\varepsilon^{\frac{\kappa_t^2}{16}} \log(1/\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \|\psi_k\|_{L_p}^p \leq C_t.$$

Soit q *un réel strictement positif tel que* $q < p$*. Alors, si* $\kappa_t \geq \sqrt{\frac{4p}{\min(2;p)}}$*, il existe une constante strictement positive* C *telle que* $\forall\, 0 < \varepsilon \leq \varepsilon_0$*,*

$$\left[\mathbb{E}\|\sum_k (\hat{\theta}_k^t(x_k) - \theta_k)\psi_k\|_{L_p}^p\right]^{1/p} \leq C\left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{(1-q/p)}$$
$$\iff f = \sum_k \theta_k\psi_k \in wl_{p,q}(\sigma)(\mathcal{E}) \cap B_{p,\infty}^{\frac{1}{r}(1-q/p)}(\mathcal{E}).$$

avec $\forall\, 0 < q < p$, $\forall\, 0 < \eta < \infty$,

$$B^{\eta}_{p,\infty}(\mathcal{E}) = \left\{ f = \sum_k \theta_k \psi_k : \quad \sup_{\lambda > 0} \lambda^{\eta} \| \sum_{k \geq \lambda} \theta_k \psi_k \|_{L_p} < \infty \right\},$$

$$wl_{p,q}(\sigma)(\mathcal{E}) = \left\{ f = \sum_k \theta_k \psi_k : \quad \sup_{\lambda > 0} \lambda^{q} \sum_k \mathbb{1}_{|\theta_k| > \lambda \sigma_k} \sigma_k^p \|\psi_k\|_{L_p}^p < \infty \right\}.$$

Chapitre 3

Non linear estimation over weak Besov spaces

Weak Besov spaces naturally appear in statistics or in approximation theory to measure the performance of classical procedures like wavelet thresholding. The first goal of this chapter is to build realizations of the worst functions to be estimated for these procedures. For this purpose, we exploit the asymptotically least favorable priors of the weak Besov balls $\mathcal{WB}_{s,p,q}(C)$. The second goal of this chapter is to compare Besov spaces with weak Besov spaces from the statistical point of view. Under suitable conditions, we show that the optimal rate of convergence of the minimax risk for $\mathcal{WB}_{s,p,q}(C)$ is the same as for the strong Besov ball $\mathcal{B}_{s,p,q}(C)$, included into $\mathcal{WB}_{s,p,q}(C)$, but we show that the asymptotically least favorable priors exhibited for $\mathcal{WB}_{s,p,q}(C)$ cannot be asymptotically least favorable priors for $\mathcal{B}_{s,p,q}(C)$. The third goal of this chapter is to explore the Bayes approach to thresholding. For this purpose, we build thresholding rules that attain the minimax rates of convergence, and which level-dependent thresholds are related to the parameters of the asymptotically least favorable priors.

3.1 Introduction

Recently, weak Besov spaces have become more important in approximation theory and statistics :
DeVore (1989), Donoho (1996), Donoho and Johnstone (1996) or Cohen, DeVore and Hochmuth (2000) have pointed out the importance of weak l_p spaces, denoted wl_p. Indeed, these spaces can be viewed as collections of functions on $[0,1]$ that can be approximated in $L_2([0,1])$ at rate $N^{-\sigma}$,

$\sigma = 1/p - 1/2$. Abramovich, Benjamini, Donoho and Johnstone (2000a) used weak l_p spaces to define more precisely the notion of sparsity. For them, a vector is said to be sparse if it has a relatively small proportion of relatively large entries and they introduce a weak l_p constraint to control this proportion. Cohen, DeVore and Hochmuth (2000) have linked the approach of non linear approximation not only to weak l_p spaces but to weak Besov spaces, denoted $W^*(r,p)$, that can be viewed as weighted weak l_p spaces. So, as we shall see in section 3.2.2, weak Besov spaces may appear as natural spaces to capture signals in function of their regularity properties and their sparsity. Cohen, DeVore and Hochmuth (2000) focused on the well known adaptive wavelet approximation that consists in thresholding the coefficients of the function to be estimated. The weak Besov spaces $W^*(r,p)$ appear in the characterization of the L_r-approximation performance of wavelet thresholding.

From the statistical point of view, weak Besov spaces naturally appear when statisticians want to evaluate the performance of a procedure for non parametric estimation. Cohen, DeVore, Kerkyacharian and Picard (2001) and Kerkyacharian and Picard (2002**) have investigated the following problem : rather than taking the minimax point of view for measuring the performance of an estimation procedure, they wondered what is the maximal set over which this procedure attains a given rate of convergence. This view is less pessimistic and is directly connected to the procedure. Kerkyacharian and Picard (2002**) have extended the notion of local oracle inequalities for L_p-norms. Introduced for the L_2-norm by Donoho and Johnstone (1994a), this notion reflects the idea of performing, as if we have an "oracle". Kerkyacharian and Picard roughly proved that if a procedure verifies a local oracle inequality, then a weak Besov space is included into its maxiset. They used this theory for two well known efficient procedures : wavelet thresholding and local bandwidth selection (introduced by Lepskii (1991)).

The first goal of this chapter is to get a good representation of 'the typical enemies' for the classical procedures mentioned previously. It is natural to look for them in weak Besov balls. Our approach is the following : from each weak Besov ball $W^*(r,p)(C)$, we wish to exhibit a prior which support is included into $W^*(r,p)(C)$ and such that its Bayes risk is equal to the minimax risk over $W^*(r,p)(C)$. If they exist, these least favorable priors simulate the worst functions to be estimated and belonging to weak Besov balls. Then, to reach this goal, we generalize Johnstone's results who obtained the asymptotic values of the minimax risk for weak l_p balls with the l_2-loss in the framework of the classical white noise model (see Johnstone (1994)) : we evaluate the minimax risk for weak Besov balls and with Besov norms as loss functions. More precisely, we focus on the $\mathcal{B}_{s',p',p'}$-loss, where $0 \leq s' < \infty$, $1 \leq p' < \infty$. As explained in section 3.2.1, the choice of the Besov norms $\mathcal{B}_{s',p',p'}$ as loss functions makes sense since Besov spaces modelize various forms of spatial inhomogeneity. What is more, this result constitutes the first step to get

the evaluation of the minimax risk for $L_{p'}$-norms.
The second goal of this chapter is to compare weak Besov spaces with ordinary Besov spaces (that are nowadays extensively considered by the statistical community) from the statistical point of view. Indeed, using the sequential characterization of Besov spaces recalled in section 2.1, we notice that if $p < r$, $W^*(r,p)$ is very close to $\mathcal{B}_{s,p,p}$, with $s = \frac{1}{2}(\frac{r}{p} - 1)$. So, in section 2.2, we widen the class of weak Besov spaces and we adopt a different formalism to denote each member of this class. This formalism enables us to naturally associate any strong Besov space $\mathcal{B}_{s,p,q}$ with a weak Besov space denoted $\mathcal{WB}_{s,p,q}$, with the inclusion $\mathcal{B}_{s,p,q} \subset \mathcal{WB}_{s,p,q}$. So, it is natural to wonder whether the rates of convergence of the minimax risk and the least favorable priors for a weak Besov ball are the same as for the Besov ball associated with.
The proofs of the results in this chapter exploit the approach used by Pinsker (1980) who wished to evaluate the asymptotic minimax risk of a signal belonging to an ellipsoid Θ in Hilbert space in the framework of the white noise model. For this purpose, he showed that there exist Gaussian priors almost concentrated on Θ and such that their Bayes risk is asymptotically equal to the minimax risk. Pinsker's paper inspired a considerable literature. Let us mention other authors' works that use the minimax Bayes method to solve parametric or non parametric problems : Casella and Strawderman (1981) and Bickel (1981) for the estimation of a bounded normal mean or Donoho and Johnstone (1994b), Johnstone (1994) and Donoho and Johnstone (1998) among others, who used this method respectively for the estimation over l_p balls, weak l_p balls and Besov balls. We can add that most of the least favorable priors produced by these problems are concentrated on a finite number of points (see Casella and Strawderman (1981), Donoho and Johnstone (1994b) or Donoho and Johnstone (1998)). However, Bickel (1981) and Johnstone (1994) present priors respectively based on 'cos' and Pareto distributions. Johnstone (1994) and Donoho and Johnstone (1994b) exhibited minimax thresholding rules by exploiting this Bayes approach. By using exactly the same method as the one developed by these authors, we reach the third goal of this chapter : building optimal thresholding rules which thresholds depend on the parameters of a prior model. Since wavelet thresholding estimators have various important optimality properties, the crucial problem of choosing the threshold, and this in a Bayesian framework, has often been investigated (see section 3.4.3 for a brief review of the methods presented by Abramovich, Sapatinas and Silverman (1998) and Vidakovic (1998)). The results of this chapter keep exploiting the methods developed in the papers cited previously (and more specifically, those of Johnstone (1994) and Donoho and Johnstone (1994b)), therefore, many details of the proofs are omitted.
To present the results concerning the rates of convergence of the minimax risk for the weak Besov

balls $\mathcal{WB}_{s,p,q}(C)$, we refer to two distinct zones :

$$\begin{aligned}\mathcal{R} = &\left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) > p'\left(s'+\frac{1}{2}\right)\right\} \cup \left\{(s,p,q): \quad p' \le p\right\},\\ \mathcal{C} = &\left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) = p'\left(s'+\frac{1}{2}\right)\right\},\end{aligned}$$

which correspond respectively to the regular case and the critical case introduced by Donoho, Johnstone, Kerkyacharian and Picard (1996) for the evaluation of the minimax risk over $\mathcal{B}_{s,p,q}$. In the regular case, the minimax risk attains, up to constants, the rate of convergence $C^{p'(1-\vartheta)}\varepsilon^{p'\vartheta}$, where $\vartheta = (s-s')/(s+\frac{1}{2})$. In the critical case, an additional assumption to bound the wavelet coefficients in the large scales is necessary : we suppose in addition that f lies in $\mathcal{B}_{\eta,p',\infty}$, $(\eta - s' > 0)$. With this additional assumption, the rate of convergence of the minimax risk is $C^{p'(1-\vartheta)}\varepsilon^{p'\vartheta}\left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})_+}$, up to constants depending on η. It is noteworthy that on $\mathcal{R}\cup\mathcal{C}$, for $\mathcal{WB}_{s,p,q}(C)$, the same rates of convergence can be achieved as for $\mathcal{B}_{s,p,q}(C)$. The logarithmic zone

$$\mathcal{L} = \left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) < p'\left(s'+\frac{1}{2}\right)\right\}$$

is very different from the other ones and is omitted in this chapter. For each weak Besov ball $\mathcal{WB}_{s,p,q}(C)$, we present asymptotically least favorable priors derived from Pareto(p) distributions. But these priors are not asymptotically least favorable priors for $\mathcal{B}_{s,p,q}(C)$. In section 3.4.5, we build realizations from these priors. Finally, we exhibit thresholding rules that attain minimax rates of convergence. At large resolution levels j, the threshold is proportional to $\left(-2\log(\alpha_j^p)\right)^{\frac{1}{2}}$, where α_j is a dilation parameter that appears in the definition of the asymptotically least favorable prior π_ε associated with $\mathcal{WB}_{s,p,q}(C)$.
The chapter is organized as follows. Section 2 is devoted to weak Besov spaces and model after an overview of Besov spaces. Section 3 recalls some facts about the Bayesian approach. Section 4 presents the results we get. Finally, in section 5, we give some elements of the proofs.

3.2 Weak Besov spaces and model

3.2.1 Overview of strong Besov spaces

Before defining weak Besov spaces, we briefly recall some relevant aspects for our study of Besov spaces. For a good introduction to these spaces, we refer the reader to Peetre (1976) and DeVore and Lorentz (1993).

Besov spaces are of statistical interest since they modelize important forms of spatial inhomogeneity. The scale of Besov spaces $\{\mathcal{B}_{s,p,q} : 0 < s < \infty, 1 \leq p, q \leq \infty\}$ includes Sobolev spaces ($H^s = \mathcal{B}_{s,2,2}$), and Hölder spaces ($C^s = \mathcal{B}_{s,\infty,\infty}$). The space of functions of bounded variation contains $\mathcal{B}_{1,1,1}$ and is included into $\mathcal{B}_{1,1,\infty}$. Using wavelets, we can connect Besov norms to sequence space norms. Let us suppose that we are given a pair of scaling function and wavelet ϕ and ψ and a function f having the following decomposition :

(3.1)

where $\psi_{jk}(t) = 2^{\frac{j}{2}}\psi(2^j t - k)$ if $j \in \mathbb{N}$, $\psi_{-1k}(t) = \phi(t-k)$, and

$$\beta_{jk} = \int f(t)\psi_{jk}(t)dt.$$

The following facts are true under standard properties of regularity and moment vanishing of ϕ and ψ (see Meyer (1992)) : if we are given $1 \leq p, q \leq \infty$ and $0 < s < \infty$, the function f in (3.1) belongs to $\mathcal{B}_{s,p,q}$, if and only if $\beta = (\beta_{jk})_{jk}$, its wavelet coefficients, verify

$$\|\beta\|_{\mathcal{B}_{s,p,q}} = \left(\sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_k |\beta_{jk}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} < \infty$$

(with the obvious modifications for $p = \infty$, $q = \infty$). In the following, we only use this sequential characterization of Besov spaces, in particular for the evaluation of the minimax risk for $\mathcal{B}_{s',p',p'}$-norms as loss functions and we note $\|f\|_{\mathcal{B}_{s,p,q}} = \|\beta\|_{\mathcal{B}_{s,p,q}}$. This allows to consider the case $s' = 0$ (we can note that when $s' = 0$ and $p' = 2$, $\|f\|_{\mathcal{B}_{s',p',p'}} = \|f\|_{L_2}$). Furthermore, we exploit Daubechies' construction that enables us to suppose in addition and without loss of generality that ϕ and ψ are supported by the interval $[-N, N]$, where $0 < N < \infty$ (see Daubechies (1992)).

3.2.2 Definition of weak Besov spaces

To characterize maxisets for wavelet thresholding, Cohen, DeVore, Kerkyacharian and Picard (2001) introduced weak Besov spaces by using the following definition :

Definition 3.1. *Let us fix $0 < p, r < \infty$. We say that the function f in (3.1) belongs to $W^*(r,p)$ if*

$$\sup_{\lambda>0} \lambda^p \sum_j 2^{j(\frac{r}{2}-1)} N(j,\lambda) < \infty,$$

where $N(j,\lambda)$ is the number of wavelet coefficients at level j greater than λ :

$$N(j,\lambda) = \sum_k \mathbb{1}_{\ |\beta_{jk}|>\lambda}. \tag{3.2}$$

The space $W^*(r,p)$ can be viewed as a weighted weak l_p space (denoted wl_p) considered by Johnstone (1994), Donoho (1996) or Abramovich, Benjamini, Donoho and Johnstone (2000a) :

$$wl_p = \left\{ \theta \in \mathbb{R}^{\mathbb{N}} : \quad \sup_{\lambda>0} \lambda^p \sum_i \mathbb{1}_{\ |\theta_i|>\lambda} < \infty \right\}.$$

The weights penalize the counting of the β_{jk}'s greater than λ for the large scales according to the sign of $r-2$. Obviously, $W^*(2,p)$ can be identified with wl_p. We can note that if we order the components of an infinite vector $\theta \in \mathbb{R}^{\mathbb{N}}$ according to their size :

$$|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(n)} \geq \ldots$$

then

$$\theta \in wl_p \iff \sup_n \ n^{\frac{1}{p}} |\theta|_{(n)} < \infty.$$

For Abramovich, Benjamini, Donoho and Johnstone (2000a), the vector θ is said to be sparse if there is a small proportion of large entries and they use the previous power-law bound to control this proportion. So, weak Besov spaces may appear as natural spaces to capture signals in function of their regularity properties and their sparsity. Using the Markov inequality and the sequential characterization of Besov spaces, it is easy to see that the Besov space $\mathcal{B}_{\frac{r}{2p}-\frac{1}{2},p,p}$ is included into $W^*(r,p)$ for $r>p$. So, to include each Besov space $\mathcal{B}_{s,p,q}$ (for $p<\infty$, $q<\infty$) into a weak one, we set :

Definition 3.2. *Let us fix $0 < s,p,q < \infty$. We say that f in (3.1) (or $\beta = (\beta_{jk})_{j,k}$ its wavelet coefficients) belongs to the weak Besov space of parameters s,p,q, noted $\mathcal{WB}_{s,p,q}$ if*

$$\sup_{\lambda>0} \lambda^q \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \ N(j,\lambda)^{\frac{q}{p}} < \infty,$$

where $N(j,\lambda)$ is defined in (3.2). To each weak Besov space $\mathcal{WB}_{s,p,q}$, we associate the balls :

$$\mathcal{WB}_{s,p,q}(C) = \left\{ f : \quad \sup_{\lambda>0} \lambda^q \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \ N(j,\lambda)^{\frac{q}{p}} \leq C^q \right\}.$$

Now, with this formalism, it is easy to see that $\mathcal{B}_{s,p,q}(C) \subset \mathcal{WB}_{s,p,q}(C)$ and to identify $W^*(r,p)$ with $\mathcal{WB}_{\frac{r}{2p}-\frac{1}{2},p,p}$.
Before going further, we recall the link between maxisets and weak Besov spaces : Cohen, DeVore, Kerkyacharian and Picard (2001) proved the following result that connects the performance of the thresholding rule to the sparsity of the wavelet coefficients of the function to be estimated :

Proposition 3.1. *Let us fix $1 < r < \infty$ and $\alpha \in]0,1[$. Under the white noise model*

$$dY_t = f(t)\,dt + \varepsilon dW_t, \quad t \in [0,1],$$

we consider the following thresholding estimator :

$$\hat{f}_\varepsilon = \sum_{j \le j_\varepsilon} \sum_k \hat{\beta}_{jk} \mathbb{1}_{\ |\hat{\beta}_{jk}| > \kappa t_\varepsilon} \psi_{jk},$$

where

- $\hat{\beta}_{jk} = \int \psi_{jk}(t) dY_t$,
- $2^{-\frac{j_\varepsilon}{2}} = t_\varepsilon = \varepsilon\sqrt{\log(\frac{1}{\varepsilon})}$,
- *κ is a constant large enough.*

We have

$$\mathbb{E}\|\hat{f}_\varepsilon - f\|_{L_r}^r \le K\left(\varepsilon\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right)^{\alpha r} \iff f \in \mathcal{B}_{\frac{\alpha}{2},r,\infty} \cap W^*(r,(1-\alpha)r).$$

3.2.3 The statistical model

In the following, we evaluate the minimax risk for weak Besov balls and for the $\mathcal{B}_{s',p',p'}$-loss with $0 \le s' < \infty$ and $1 \le p' < \infty$. We will restrict our attention to compactly supported functions, namely functions supported by $[0,1]$. So, β_{jk} is non-zero as soon as k is in $\mathcal{I}_j$, with

$$\mathcal{I}_j = \left\{-N+1, \dots, 2^j + N - 1\right\}. \tag{3.4}$$

We consider the white noise model :

$$dY_t = f(t)\,dt + \varepsilon dW_t, \quad t \in [0,1].$$

By integration versus each function ψ_{jk}, the white noise model is translated into the sequence space, and we obtain the following sequence of independent variables :

$$y_{jk} = \beta_{jk} + \varepsilon z_{jk}, \quad z_{jk} \sim \mathcal{N}(0,1), \quad j \ge -1,\ k \in \mathcal{I}_j.$$

The minimax risk is denoted :

$$\begin{aligned} \tilde{R}_\varepsilon &= \inf_{\hat{\beta}} \sup_{\beta \in \mathcal{W}\mathcal{B}_{s,p,q}(C)} \mathbb{E}_\beta \|\hat{\beta} - \beta\|_{\mathcal{B}_{s',p',p'}}^{p'} \\ &= \inf_{\hat{\beta}} \sup_{\beta \in \mathcal{W}\mathcal{B}_{s,p,q}(C)} \mathbb{E}_\beta \sum_j 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^{p'}. \end{aligned}$$

3.3 The Bayesian approach

3.3.1 Asymptotically least favorable priors

To exhibit 'the typical enemies' of the weak Besov ball $\mathcal{WB}_{s,p,q}(C)$ from the statistical point of view, we use π_ε, a prior distribution on the wavelet coefficients, that is required to have the following properties : given $0 \leq s' < \infty$ and $1 \leq p' < \infty$, the support of π_ε must asymptotically be included into $\mathcal{WB}_{s,p,q}(C)$:

$$\forall\, \varepsilon > 0, \quad \exists\, \gamma_\varepsilon > 1, \text{ such that } \gamma_\varepsilon \overset{\varepsilon \to 0}{\longrightarrow} 1 \text{ and under } \pi_\varepsilon, \quad \mathbb{P}(\beta \in \mathcal{WB}_{s,p,q}(\gamma_\varepsilon C)) \overset{\varepsilon \to 0}{\longrightarrow} 1,$$

and the Bayes risk of π_ε defined by

$$B(\pi_\varepsilon) = \inf_{\hat\beta}\ \mathbb{E}_{\pi_\varepsilon}\ \mathbb{E}_\beta \sum_j 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat\beta_{jk} - \beta_{jk}|^{p'}$$

must verify :

$$C_1 B(\pi_\varepsilon) \leq \tilde R_\varepsilon \leq C_2 B(\pi_\varepsilon),$$

where C_1 and C_2 are positive constants depending only on s, p, q, s' and p'. If these properties are checked, we say that π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$.

3.3.2 Minimax Bayes risk

In this chapter, we evaluate the minimax risk over weak Besov balls. For this purpose, we use the Bayes method that is often exploited in the literature (see the references cited in Introduction). That is the reason why most of the details are omitted in this section.
Let us consider $m_{s,p,q}(C)$, the natural set of probability measures associated with $\mathcal{WB}_{s,p,q}(C)$:

$$m_{s,p,q}(C) = \left\{ \pi : \quad \sum_j 2^{jq(s+\frac{1}{2}-\frac{1}{p})} [\mathbb{E}_\pi N(j,\lambda)]^{\frac{q}{p}} \leq \left(\frac{C}{\lambda}\right)^q \quad \forall\, \lambda > 0 \right\},$$

and $M_{s,p,q}(C)$, the closure of its generalized convex hull

$$M_{s,p,q}(C) = \overline{Hu(m_{s,p,q}(C))},$$

where

$$Hu(m_{s,p,q}(C)) = \left\{ \pi = \sum_{l=0}^{+\infty} \mu_l\, \pi_l : \qquad \pi_l \in m_{s,p,q}(C), \quad 0 \leq \mu_l \leq 1, \quad \sum_{l=0}^{+\infty} \mu_l = 1 \right\}.$$

When $q \geq p$, $m_{s,p,q}(C)$ is convex, so $M_{s,p,q}(C) = m_{s,p,q}(C)$. Let us consider the minimax Bayes risk for $M_{s,p,q}(C)$:

$$B(M_{s,p,q}(C), \varepsilon) = \inf_{\hat{\beta}} \sup_{\pi \in M_{s,p,q}(C)} \mathbb{E}_\pi \ \mathbb{E}_\beta \sum_j 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^{p'}.$$

By using standard arguments, we have $\tilde{R}_\varepsilon \leq B(M_{s,p,q}(C), \varepsilon)$ and to get an upper bound for $\tilde{R}_\varepsilon$, we just have to obtain an upper bound for $B(M_{s,p,q}(C), \varepsilon)$. Since $M_{s,p,q}(C)$ is convex and compact for the Prohorov metric, by applying the minimax theorem (see Strasser (1985)), we have

$$B(M_{s,p,q}(C), \varepsilon) = \sup_{\pi \in M_{s,p,q}(C)} B(\pi),$$

where $B(\pi)$ is the Bayes risk of π. Then, our goal is to construct $\mathcal{M}_{s,p,q}(C)$ a subclass of $M_{s,p,q}(C)$ which elements have a simpler structure and such that

$$\sup_{\pi \in M_{s,p,q}(C)} B(\pi) = \sup_{\tilde{\pi} \in \mathcal{M}_{s,p,q}(C)} B(\tilde{\pi}).$$

For π in $M_{s,p,q}(C)$, we construct $\tilde{\pi}$ as the following independent product of marginal distributions :

$$\tilde{\pi} = \prod_j \prod_{k \in \mathcal{I}_j} \tilde{\pi}_j,$$

where $\tilde{\pi}_j$ denotes the average of the univariate marginal distributions of π at the level j :

$$\tilde{\pi}_j = \frac{1}{|\mathcal{I}_j|} \sum_{k \in \mathcal{I}_j} \pi_{jk},$$

where $\mathcal{I}_j$ is defined in (3.4), and we set

$$\mathcal{M}_{s,p,q}(C) = \{\tilde{\pi} : \ \pi \in M_{s,p,q}(C)\}.$$

By using standard arguments (see, for instance, Johnstone (1994)), we have

$$B(M_{s,p,q}(C), \varepsilon) = \sup_{\tilde{\pi} \in \mathcal{M}_{s,p,q}(C)} B(\tilde{\pi}).$$

What is more, it is relevant to note that under a prior of $\mathcal{M}_{s,p,q}(C)$, the Bayes estimator of β_{jk} depends only on y_{jk}.

Finally, we apply a rescaling to the model : if F is a probability measure, we note $s_\varepsilon F$ the probability measure defined by $s_\varepsilon F(\mathcal{A}) = F(\varepsilon^{-1}\mathcal{A})$. We note $x_{j1} = \varepsilon^{-1} y_{j1}$ and F_{j1} (or rather F_j if there is no risk of confusion) the distribution of $u_{j1} = \varepsilon^{-1}\beta_{j1}$. We get :

$$B(M_{s,p,q}(C), \varepsilon) = \varepsilon^{p'} \sup_{s_\varepsilon F \in \mathcal{M}_{s,p,q}(C)} \inf_{(d_j)} \sum_j 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} |\mathcal{I}_j| \ \mathbb{E}_{F_j} \ \mathbb{E}_{u_{j1}} |d_j(x_{j1}) - u_{j1}|^{p'}.$$

3.4 Results, discussions and simulations

The main tool for exhibiting the results of this chapter is the evaluation of the minimax risk for each weak Besov ball $\mathcal{WB}_{s,p,q}(C)$ and for $\mathcal{B}_{s',p',p'}$-norms as loss functions. So, in the following, we suppose that we are given a Besov loss $\mathcal{B}_{s',p',p'}$ with $1 \leq p' < \infty$ and $0 \leq s' < \infty$, and we set

$$\mathcal{R} = \left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) > p'\left(s'+\frac{1}{2}\right)\right\} \cup \{(s,p,q): \quad p' \leq p\},$$
$$\mathcal{C} = \left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) = p'\left(s'+\frac{1}{2}\right)\right\}.$$

Respectively, these zones correspond to the regular case and the critical case introduced by Donoho, Johnstone, Kerkyacharian and Picard (1996). In the critical case, we need a minimal hypothesis on the regularity of f to control the size of the β_{jk}'s at high levels. That is the reason why we suppose from now on until the end that in addition, in the critical case, f lies in $\mathcal{B}_{\eta,p',\infty}(C)$ (with $\eta > s'$ but $\eta - s'$ eventually very small). So, let us set

$$\Theta = \mathcal{WB}_{s,p,q}(C) \quad \text{on } \mathcal{R},$$

$$\Theta = \mathcal{WB}_{s,p,q}(C) \cap \mathcal{B}_{\eta,p',\infty}(C) \quad \text{on } \mathcal{C},$$

and the minimax risk we consider is from now on :

$$R_\varepsilon = \inf_{\hat{\beta}} \sup_{\beta\in\Theta} \mathbb{E}_\beta \sum_j 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^{p'}. \tag{3.11}$$

3.4.1 Notations and technical tools

In this section, we precise some notations that will be useful in the following. For this purpose, we suppose that we are given $0 < s, p, q, C < \infty$.

- For any x, δ_x denotes Dirac measure at the point x.
- If x is a real number, $\lfloor x \rfloor$ denotes the greatest integer smaller or equal to x.
  If h_1 and h_2 denote two positive functions of ε,
- $h_1(\varepsilon) \overset{\varepsilon\to 0}{\sim} h_2(\varepsilon)$ means that
$$\lim_{\varepsilon\to 0} \frac{h_1(\varepsilon)}{h_2(\varepsilon)} = 1.$$
- $h_1(\varepsilon) \simeq h_2(\varepsilon)$ means that

$$\forall\, \varepsilon > 0, \quad h_2(\varepsilon) < h_1(\varepsilon) \leq 2\, h_2(\varepsilon).$$

- $h_1(\varepsilon) \approx h_2(\varepsilon)$ means that there exist positive constants A and B, depending only on s, p, q, s', p' such that
$$\forall\, \varepsilon > 0, \quad A\, h_1(\varepsilon) \le h_2(\varepsilon) \le B\, h_1(\varepsilon).$$

Let us fix κ a real number belonging to $\left]1, \frac{2p'(s'+\frac{1}{2})}{2p'(s'+\frac{1}{2})-1}\right[$. We set for all $\varepsilon > 0$, $(j_*, j^*) \in \mathbb{N}^2$ such that
$$2^{j_*} \simeq \left(\frac{C}{\varepsilon}\right)^{\frac{1}{s+\frac{1}{2}}}, \quad 2^{j^*} \simeq \left(\frac{C}{\varepsilon}\right)^{\frac{\kappa}{s+\frac{1}{2}}}.$$

Now, we define two integers j_1 and j_2, and a sequence of non negative real numbers $(\alpha_j)_j$ depending on the zone.

- If (s, p, q) lies in $\mathcal{C}$ and if $q > p$, we set $j_1 = j_*$, $j_2 = j^*$ and for all j,
$$\alpha_j = \begin{cases} V_\varepsilon 2^{-j(s+\frac{1}{2})} & j < j_1 \\ V_\varepsilon 2^{-j(s+\frac{1}{2})} j & j \in \{j_1, \dots, j_2\} \\ 0 & \text{otherwise,} \end{cases}$$
where V_ε is a real number independent of j such that
$$\sum_j 2^{jq(s+\frac{1}{2})}\, \alpha_j^q = \left(\frac{C}{\varepsilon}\right)^q. \tag{3.12}$$
The real number V_ε that is also used in the other two cases, verifies
$$V_\varepsilon \overset{\varepsilon\to 0}{\sim} (q+1)^{\frac{1}{q}} (\kappa^{q+1} - 1)^{-\frac{1}{q}} \left(\frac{C}{\varepsilon}\right) j_*^{-1-\frac{1}{q}}.$$

- If (s, p, q) lies in $\mathcal{C}$ and if $q \le p$, we set $j_1 = j_2 = j^*$ and for all j,
$$\alpha_j = \begin{cases} V_\varepsilon 2^{-j(s+\frac{1}{2})} & j < j_* \\ 0 & j \ge j_* \text{ and } j \ne j_1, \end{cases}$$
and α_{j_1} is such that (3.12) is checked.

- If (s, p, q) lies in $\mathcal{R}$, we set $j_1 = j_2 = j_*$ and for all j,
$$\alpha_j = \begin{cases} V_\varepsilon 2^{-j(s+\frac{1}{2})} & j < j_* \\ 0 & j \ge j_* \text{ and } j \ne j_1, \end{cases}$$
and α_{j_1} is such that (3.12) is checked.

To each sequence $(\alpha_j)_j$, we associate two sequences $(c_j)_j$ and $(\mu_j)_j$ such that $\phi(\mu_j + c_j) = \left(\frac{\alpha_j}{\mu_j}\right)^p \phi(c_j)$, where ϕ denotes the standard Gaussian density function. Furthermore, if $(\alpha_j)_j$ tends to 0, we require in addition that $(c_j)_j$ and $(\mu_j)_j$ tend to $+\infty$ with $c_j = o(\mu_j)$. This yields that

$$\mu_j \overset{\alpha_j \to 0}{\sim} (-2\log \alpha_j^p)^{\frac{1}{2}}.$$

Sequences defined by similar procedures have been used to build least favorable priors for l_p balls (see Donoho and Johnstone (1994b)) or for weak l_p balls (see Johnstone (1994)).

3.4.2 Main results and comments

By using notations defined in section 3.4.1, we have the following main result :

Theorem 3.1. *Let us fix $0 < s, p, q, C < \infty$, such that $s > s'$.*

- *If $\vartheta = \frac{s-s'}{s+\frac{1}{2}}$, and*

$$\Psi(C,\varepsilon) = \begin{cases} C^{p'(1-\vartheta)}\varepsilon^{p'\vartheta} & \text{on } \mathcal{R} \\ C^{p'(1-\vartheta)}\varepsilon^{p'\vartheta}\left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})_+} & \text{on } \mathcal{C}, \end{cases}$$

  we have

$$C_1 \le \liminf_{\varepsilon\to 0} R_\varepsilon \Psi(C,\varepsilon)^{-1} \le \limsup_{\varepsilon\to 0} R_\varepsilon \Psi(C,\varepsilon)^{-1} \le C_2,$$

  where C_1 and C_2 are positive constants depending only on s, p, q, s', p'. Besides, on $\mathcal{C}$, C_2 also depends on η.

- *If we set π_ε as the distribution of a sequence of independent variables $(\beta_{jk})_{j,k}$ such that the distribution of β_{jk} is symmetric about 0 and*

$$|\beta_{jk}| = \begin{cases} \varepsilon\alpha_j & \text{if } j < j_* \\ \varepsilon\min(\alpha_j X_{jk}, \mu_j) & \text{otherwise,} \end{cases}$$

  where X_{jk} is a Pareto(p)-variable, then π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$.

- *The following thresholding rule*

$$\hat{f}_\varepsilon^* = \sum_j \sum_k sign(y_{jk})\left(|y_{jk}| - \varepsilon\lambda_j\right)_+ \psi_{jk},$$

  where $\lambda_j = 0$ if $j < j_*$ and $\lambda_j = (-2\log \alpha_j^p)^{\frac{1}{2}}$ otherwise, attains the minimax rate up to constants.

The proof of Theorem 1 is given in section 5. More precisely, in section 3.5.1, to get the upper bound of the minimax risk R_ε, we exploit the method exhibited by Johnstone (1994) and Donoho and Johnstone (1994b) : the use of the asymptotically least favorable priors for the Bayes threshold risk. This method allows to exhibit minimax thresholding rules. And as expected, the Bayes threshold λ_j verifies the same kind of relationship as the one checked by the thresholds of the minimax thresholding rules exhibited by Johnstone (1994) or by Donoho and Johnstone (1994b). Furthermore, we naturally rely on the asymptotically least favorable prior for the Bayes threshold risk to build π_ε. In section 3.5.2, we prove that π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$ but not for $\mathcal{B}_{s,p,q}(C)$. On $\mathcal{R}$, the lower bound of R_ε is provided by the lower bound of the minimax risk for $\mathcal{B}_{s,p,q}(C)$ included into $\mathcal{WB}_{s,p,q}(C)$ (see Donoho, Johnstone, Kerkyacharian and Picard (1996)). On $\mathcal{C}$, we cannot use this argument, since we consider $\Theta = \mathcal{WB}_{s,p,q}(C) \cap \mathcal{B}_{\eta,p',\infty}(C)$ that does not contain $\mathcal{B}_{s,p,q}(C)$. However, on $\mathcal{C}$, by exploiting the results of sections 3.5.2, it is easy to derive the lower bound of R_ε from the Bayes method (see Remark 3.2 in section 3.5.2).
Let us make relevant remarks about the constants. In the regular case, we cannot obtain the optimal values for C_1 and C_2, since we cannot use the constants of the asymptotic results of Theorems 3.2, and 3.3. In the critical case, the constant C_2 comes from the logarithmic term and

$$C_2 = \left(\frac{1}{\eta - s'}\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})_+} \tilde{C}_2(s,p,q,s',p')$$

blows up when η tends to s'. But, if we consider a slight modification of $\mathcal{WB}_{s,p,q}(C)$ by limiting the number of non zero wavelet coefficients, we have the following result :

Proposition 3.2. *On the critical case and if $q \le p$, with*

$$\mathcal{WB}_{s,p,q}(C,n) = \mathcal{WB}_{s,p,q}(C) \cap \{f : \quad \beta_{jk} = 0 \quad \forall\, j > \log_2(n),\ \ \forall\, k \in \mathcal{I}_j\},$$

and if $\varepsilon \equiv \varepsilon(n)$ is such that

$$\varepsilon^p \; n^{p'(s'+\frac{1}{2})} \overset{n\to\infty}{\longrightarrow} +\infty, \tag{3.13}$$

and

$$\varepsilon^{p-\tau} n^{p'(s'+\frac{1}{2})-\frac{1}{2}} \overset{n\to\infty}{\longrightarrow} 0, \tag{3.14}$$

for a positive constant τ, then, if R_ε^n is the minimax risk associated with $\mathcal{WB}_{s,p,q}(C,n)$,

$$R_\varepsilon^n \sim \frac{p'}{p'-p} C^{p'(1-\vartheta)} \varepsilon^{p'\vartheta} \left(-2\log\left[\left(\frac{C}{\varepsilon}\right)^p n^{-p'(s'+\frac{1}{2})}\right]\right)^{\frac{p'\vartheta}{2}}..$$

To prove Proposition 3.2, we mimic the proof of Theorem 3.1. Conditions (3.13) and (3.14) allow to apply Theorems 3.2 and 3.3 and to use the constants of the asymptotic results of these theorems.
Actually, when $p < 2$, $\mathcal{WB}_{\frac{1}{p}-\frac{1}{2},p,p}(C,n)$ can be viewed as a weak l_p ball considered by Johnstone (1994). In this chapter, the L_2-loss was investigated. Of course, the same result was obtained.
The logarithmic case

$$\mathcal{L} = \left\{(s,p,q): \quad p' > p,\ p\left(s+\frac{1}{2}\right) < p'\left(s'+\frac{1}{2}\right)\right\}$$

is very different from the other ones. Donoho, Johnstone, Kerkyacharian and Picard (1996) pointed out that the functions typical of $\mathcal{B}_{s,p,q}(C)$ are very sparse when $(s,p,q) \in \mathcal{L}$. Since $\mathcal{B}_{s,p,q}(C) \subset \mathcal{WB}_{s,p,q}(C)$, we expect the priors produced by our method not to be able to capture the sparsity of most of the functions of $\mathcal{WB}_{s,p,q}(C)$ on $\mathcal{L}$.

3.4.3 Thresholding rules via a Bayesian framework

One goal of this chapter was to exhibit thresholding rules by using a Bayesian approach. Many authors have investigated this problem. We briefly recall the methods exhibited by Abramovich, Sapatinas and Silverman (1998) and by Vidakovic (1998) to construct Bayes thresholds. Abramovich, Sapatinas and Silverman (1998) consider a prior model having the following form :

$$\beta_{jk} \sim \gamma_j \mathcal{N}(0,\tau_j^2) + (1-\gamma_j)\delta_0.$$

The hyperparameters τ_j^2 and γ_j are chosen to ensure that the underlying function f belongs to a given strong Besov space $\mathcal{B}_{s,p,q}$. Then, $\hat{\beta}_{jk}$, the estimator of β_{jk} obtained by using the median of the posterior distribution, is zero if y_{jk} falls into an interval of the form $[-\lambda'_j, \lambda'_j]$. Vidakovic (1998) imposes a symmetric prior on β_{jk} and the marginal model for y_{jk} conditioned to β_{jk} is the double exponential with the density given by

$$f(y_{jk}|\beta_{jk}) = \frac{1}{2}(2\mu)^{\frac{1}{2}} \exp\left(-(2\mu)^{\frac{1}{2}}|y_{jk}-\beta_{jk}|\right).$$

He constructs a procedure that mimics the hard thresholding rule. He estimates β_{jk} by $y_{jk}\mathbb{1}_{\ \eta_{jk}<1}$ where $\eta_{jk} = \mathbb{P}(\beta_{jk}=0 \mid y_{jk})/\mathbb{P}(\beta_{jk} \neq 0 \mid y_{jk})$.
Unfortunately, $\hat{f}_\varepsilon^*$, the estimator exhibited in Theorem 3.1 is not adaptive, since at high levels the threshold λ_j is related to the parameters of π_ε by the relationship :

$$\lambda_j = (-2\log(\alpha_j^p))^{\frac{1}{2}}.$$

However, we have investigated a method to overcome the shortcoming of non adaptivity and to adapt this estimator for the discrete data by using the discrete wavelet transform. This method and other related results will be the subjects of the next chapter.

3.4.4 Comparison of $\mathcal{B}_{s,p,q}(C)$ with $\mathcal{WB}_{s,p,q}(C)$ from the statistical point of view

In section 3.2, we emphasize on the strong relationship between $\mathcal{B}_{s,p,q}(C)$ and $\mathcal{WB}_{s,p,q}(C)$. So, it is natural to compare them from the statistical point of view. The previous theorem and Theorem 1 of Donoho, Johnstone, Kerkyacharian and Picard (1996) enable us to elicit the rates of convergence for $\mathcal{B}_{s,p,q}(C)$ and $\mathcal{WB}_{s,p,q}(C)$ on regular and critical zones, and to conclude that these rates are the same.
Theorem 3.1 shows that π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$ but π_ε is not an asymptotically least favorable prior for $\mathcal{B}_{s,p,q}(C)$. Indeed in section 3.5.2, we prove the following result :

Proposition 3.3. *Under π_ε, $\mathbb{P}(\beta \in \mathcal{B}_{s,p,q}(C)) \overset{\varepsilon\to 0}{\longrightarrow} 0$.*

Let us note that the construction of π_ε uses dense distributions. We note that this is not necessarily the case for the least favorable priors of $\mathcal{B}_{s,p,q}(C)$ exploited by Johnstone (1994). If for Pinsker's case ($p = q = 2$), Johnstone uses Gaussian distributions (so, they are dense), when $p < 2$, the least favorable priors are based on three or two point distributions (for the coarsest levels, the prior distributions are dense, whereas they are sparse for high levels, with a few wavelet coefficients carrying all the energy).

3.4.5 Realizations from asymptotically least favorable priors

To obtain a good representation of the worst functions to be estimated and belonging to weak Besov balls $\mathcal{WB}_{s,p,q}(C)$, we present various realizations of the asymptotically least favorable priors for the regular and the critical cases and for different values of s, p and q. To reach this goal, and as Johnstone (1994) or Abramovich, Sapatinas and Silverman (1998), our approach is based on a periodized form of the discrete wavelet transform (denoted DWT) : if we are given a vector of discrete values of a periodic signal f with unit period

$$\mathbf{f} = \left\{ f(\frac{i}{n}) : \quad 1 \le i \le n = 2^N \right\},$$

we perform the discrete wavelet transform of $\mathbf{f}$: for all $-1 \le j \le N-1, 0 \le k < 2^j$,

$$d_{jk} = (W\mathbf{f})_{jk},$$

where W is the $n \times n$ DWT matrix associated to a mother wavelet ψ. The matrix W is orthogonal, so we reconstruct $\mathbf{f}$ by : $\mathbf{f} = W^T d$. Along this section, we use the DWT matrix associated to Daubechies' least asymmetric wavelet of order 8, $n = 2^{12}$ plotting points and the classical calibration $\varepsilon = n^{-\frac{1}{2}}$. For each realization, we set $C = 1$ and because of its improper nature, we place no prior on d_{-10}. Donoho and Johnstone (1994) noted the approximation

$$n^{\frac{1}{2}} W_{jk}(i) \asymp 2^{\frac{j}{2}} \psi(2^j i/n - k),$$

where $W_{jk}(i)$ is the ith element of the (j,k)th row of W. This approximation leads to the following relationship between discrete wavelet coefficients and ordinary wavelet coefficients :

$$d_{jk} \asymp n^{\frac{1}{2}} \beta_{jk}.$$

So, using Theorem 3.1 (and its notations), to reach our goal, it is natural to impose the following prior on the d_{jk}'s :

$$d_{jk} \sim \begin{cases} F_j & j \geq j_* \\ \frac{1}{2}(\delta_{\alpha_j} + \delta_{-\alpha_j}) & j < j_*, \end{cases}$$

where $F_j = \frac{1}{2}(F_j^+ + F_j^-)$, F_j^+ is the distribution of $\min(\alpha_j X_{jk}, \mu_j)$ and F_j^- is the reflection of F_j^+ about 0. To complete the definition of this prior model, and according to section 3.4.1, we set $j_* \in \mathbb{N}$ such that $2^{j_*} \simeq \left(\frac{C}{\varepsilon}\right)^{\frac{1}{s+\frac{1}{2}}}$, and for the critical case, we set $j^* \in \mathbb{N}$ such that $2^{j^*} \simeq \left(\frac{C}{\varepsilon}\right)^{\frac{\kappa}{s+\frac{1}{2}}}$, where

$$\kappa = \frac{2p'(s' + \frac{1}{2})}{2p'(s' + \frac{1}{2}) - 1} = \frac{2p(s + \frac{1}{2})}{2p(s + \frac{1}{2}) - 1},$$

since on $\mathcal{C}$, $p'(s' + \frac{1}{2}) = p(s + \frac{1}{2})$. This allows to define the integers j_1 and j_2, the sequence $(\alpha_j)_j$ as in section 3.4.1 and we take $\forall\, j \in \{j_1, \ldots, j_2\}$, $\mu_j = \left(-2\log(\alpha_j^p)\right)^{\frac{1}{2}}$. Figure 3.1 shows sample paths we obtain. Of course, realizations for the regular case are more regular than realizations for the critical case. This fact is illustrated by the comparison between (c) and (a). As expected, we note that realizations are more regular when s is great (compare (a) and (f)) or when q is small (compare (a) and (b) or (d) and (e)). Finally, when p decreases, the realizations are less regular and the size of the high peaks increases (compare (a) and (d) or (b) and (e)).

Remark 3.1. *The use of periodic functions on* $[0,1]$*, which leads to exploit the construction of periodic wavelets (see Daubechies (1992)) does not alter the qualitative phenomena we wish to present here. See Johnstone (1994) for further explanations.*

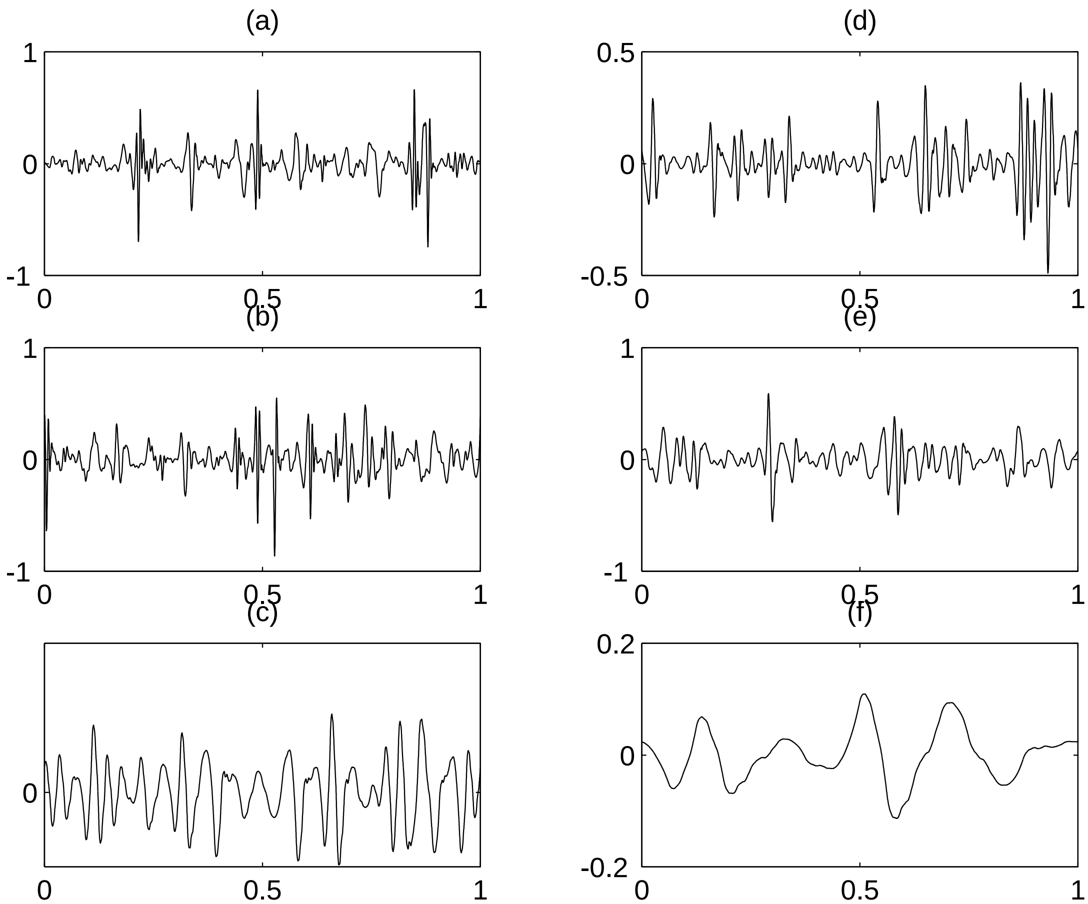

FIG. 3.1 – Realizations with various values of parameters s, p and q, with $C = 1$ and $n = 2^{12} = 4096$ plotting points. The critical case is illustrated by (a), (b), (d), (e) and (f) and the regular case, by (c). (a) : $s = 1$, $p = 1$, $q = 2$; (b) : $s = 1$, $p = 1$, $q = 3$; (c) : $s = 1$, $p = 1$, $q = 2$; (d) : $s = 1$, $p = 2$, $q = 2$; (e) : $s = 1$, $p = 2$, $q = 3$; (f) : $s = 2$, $p = 1$, $q = 2$.

3.5 Proof of Theorem 3.1

In this section, the notations K, K_1, K_2, will keep designating all the positive constants depending only on s, p, q, s', p' we could need. From now on, we suppose for sake of simplicity and without loss of generality, that for all j, $|\mathcal{I}_j| = 2^j$.
For all j_η in $\{0, 1, \ldots, +\infty\}$, we consider

$$\Theta^* = \{\beta : \quad \beta \in \Theta \text{ with } \beta_{jk} = 0, \ \forall \ (j \geq j_\eta, k \in \mathcal{I}_j)\},$$

and

$$R_\varepsilon^* = \inf_{\hat{\beta}} \sup_{\beta \in \Theta^*} \mathbb{E}_\beta \sum_{j < j_\eta} 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^{p'}. \tag{3.15}$$

Using (3.11), (3.15) and the properties of $\mathcal{B}_{\eta,p',\infty}(C)$, we have

$$R_\varepsilon^* \leq R_\varepsilon \leq R_\varepsilon^* + \sup_{\beta \in \mathcal{B}_{\eta,p',\infty}(C)} \sum_{j \geq j_\eta} 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\beta_{jk}|^{p'} \leq R_\varepsilon^* + K\ C^{p'} 2^{p'(s'-\eta)j_\eta}.$$

In the critical case, we choose j_η so that $j_\eta = \lfloor \frac{\vartheta}{\eta - s'} \log_2\left(\frac{C}{\varepsilon}\right) \rfloor$. In the regular case, we set $j_\eta = \infty$ and $R_\varepsilon^* = R_\varepsilon$. So, the first point of Theorem 3.1 will be proved as soon as we prove that

$$C_1 \leq \liminf_{\varepsilon \to 0} R_\varepsilon^* \Psi(C, \varepsilon)^{-1} \leq \limsup_{\varepsilon \to 0} R_\varepsilon^* \Psi(C, \varepsilon)^{-1} \leq C_2.$$

Therefore, we use the following modified minimax Bayes risk :

$$B(M, \varepsilon) = \inf_{\hat{\beta}} \sup_{\pi \in M} \ \mathbb{E}_\pi \ \mathbb{E}_\beta \sum_{j < j_\eta} 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^{p'},$$

where M is the closure of the generalized convex hull of $m_{s,p,q}(C, j_\eta)$ and

$$m_{s,p,q}(C, j_\eta) = \left\{ \pi : \quad \sum_{j < j_\eta} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} [\mathbb{E}_\pi N(j, \lambda)]^{\frac{q}{p}} \leq \left(\frac{C}{\lambda}\right)^q \quad \forall\ \lambda > 0 \right\}.$$

If $\mathcal{M}$ is the subclass of M constructed by the same procedure as for $\mathcal{M}_{s,p,q}(C)$, we have :

$$R_\varepsilon^* \leq B(M, \varepsilon) = \varepsilon^{p'} \sup_{s_\varepsilon F \in \mathcal{M}} \inf_{(d_j)} \sum_{j < j_\eta} 2^{jp'(s'+\frac{1}{2})}\ \mathbb{E}_{F_j}\ \mathbb{E}_{u_{j1}} |d_j(x_{j1}) - u_{j1}|^{p'}. \tag{3.16}$$

3.5.1 Upper bound

Our first goal is to prove that

$$B(M,\varepsilon) \le K\Psi(C,\varepsilon),$$

where K may depend on η on $\mathcal{C}$. We omit the case $q < p$ that will be handled by the case $q = p$ (let us note that $m_{s,p,q}(C, j_\eta) \subset m_{s,p,p}(C, j_\eta)$ if $q < p$). To reach our goal, we use the risk associated with the soft thresholding estimator

$$d_\lambda(x) = \mathrm{sign}(x)\ (|x| - \lambda)_+$$

denoted by

$$r(\lambda,\xi) = \mathbb{E}_\xi\ |d_\lambda(x) - \xi|^{p'},$$

where $x \sim \mathcal{N}(\xi, 1)$. We recall the properties of $\xi \longrightarrow r(\lambda,\xi)$ we use in the following :

Proposition 3.4. *For all $\lambda > 0$, if ϕ denotes the standard Gaussian density function and $\Phi(y) = \int_{-\infty}^{y} \phi(z)dz$ is the standard Gaussian cumulative distribution function,*

1. *$\xi \to r(\lambda,\xi)$ is symmetric about 0.*
2. *$\forall\ \xi > 0$, $\frac{\partial r}{\partial \xi}(\lambda,\xi) = p'\ \xi^{p'-1}\Phi([-\lambda-\xi, \lambda-\xi])$ and $\xi \longrightarrow r(\lambda,\xi)$ is a strictly increasing function.*
3. *$\lim_{\xi\to+\infty} r(\lambda,\xi) = \int_{-\infty}^{+\infty} |z|^{p'}\phi(z+\lambda)dz.$*
4. *$K_1 \exp(\frac{-\lambda^2}{2})\lambda^{-p'-1} \le r(\lambda, 0) \le K_2 \exp(\frac{-\lambda^2}{2})\lambda^{-p'-1}$.*

The proof of this proposition is omitted since it is just an extension of the classical case $p' = 2$ (see Donoho and Johnstone (1994a)).
By (3.16), we have

$$B(M,\varepsilon) \le \varepsilon^{p'} \sup_{s_\varepsilon F \in \mathcal{M}} \inf_{(\lambda_j)} \sum_{j<j_\eta} 2^{jp'(s'+\frac{1}{2})}\mathbb{E}_{F_j} r(\lambda_j, \xi). \tag{3.17}$$

The condition $s_\varepsilon F \in \mathcal{M}$ means that

$$\sum_{j<j_\eta} 2^{jq(s+\frac{1}{2})} \left(\mathbb{E}_{F_j}\left(\mathbb{1}_{\ |\xi|>\lambda\varepsilon^{-1}}\right)\right)^{\frac{q}{p}} \le \left(\frac{C}{\lambda}\right)^q, \quad \forall\ \lambda > 0.$$

So, the prior we investigate is such that

$$\forall -1 \leq j < j_\eta, \quad \forall \lambda > 0, \quad F_j(|\xi| > \lambda \varepsilon^{-1}) \;\leq\; t_j 2^{-jp(s+\frac{1}{2})} \left(\frac{C}{\lambda}\right)^p,$$

with

$$\sum_{j<j_\eta} t_j^{\frac{q}{p}} \leq 1. \tag{3.18}$$

Since $\xi \longrightarrow r(\lambda_j, \xi)$ is symmetric about 0, without loss of generality, we can assume that F_j is supported by $\mathbb{R}_+$ and

$$\forall \lambda > 0, \quad F_j(|\xi| > \lambda\varepsilon^{-1}) \leq t_j 2^{-jp(s+\frac{1}{2})} \left(\tfrac{C}{\lambda}\right)^p$$
$$\Leftrightarrow \; \forall v > 0, \qquad F_j(\xi > v) \leq t_j \; 2^{-jp(s+\frac{1}{2})} \left(\tfrac{C}{\varepsilon}\right)^p v^{-p}.$$

Let us define the following function :

$$\forall v > 0, \quad h(v) = \min\left\{1 \;,\; t_j\, 2^{-jp(s+\frac{1}{2})} \left(\frac{C}{\varepsilon}\right)^p v^{-p}\right\}.$$

Since $\xi \longrightarrow r(\lambda_j, \xi)$ is increasing,

$$\begin{aligned}
\mathbb{E}_{F_j} r(\lambda_j, \xi) &= \int r(\lambda_j, \xi) F_j(d\xi) \\
&= \int F_j(d\xi) \int_0^{r(\lambda_j,\xi)} dt \\
&= \int_0^{+\infty} dt \int \mathbb{1}_{\,]r^{-1}(\lambda_j,t),+\infty[} F_j(d\xi) \\
&= \int_0^{+\infty} F_j\left(\,]r^{-1}(\lambda_j, t), +\infty[\,\right) dt \\
&\leq \int_0^{+\infty} h(r^{-1}(\lambda_j, t)) dt,
\end{aligned}$$

and the equality is obtained if for all $v > 0$, $F_j\left(\,]v, +\infty[\,\right) = h(v)$. This means that F_j is absolutely continuous with respect to the Lebesgue measure and has the following density :

$$f_{\alpha_j}(\xi) = p \mathbb{1}_{\,\xi \,\geq\, \alpha_j}\; \alpha_j^p \; \xi^{-p-1}$$

with $\alpha_j = t_j^{\frac{1}{p}}\; 2^{-j(s+\frac{1}{2})} \left(\frac{C}{\varepsilon}\right) \geq 0$ such that

$$\sum_{j<j_\eta} \alpha_j^q \; 2^{jq(s+\frac{1}{2})} \leq \left(\frac{C}{\varepsilon}\right)^q$$

according to (3.18). And we can conclude by using (3.17) that

$$B(M,\varepsilon) \le \varepsilon^{p'} \sup \left\{ \sum_{j<j_\eta} 2^{jp'(s'+\frac{1}{2})} \inf_{\lambda_j} \mathbb{E}_{f_{\alpha_j}} r(\lambda_j,\xi) : \quad \sum_{j<j_\eta} \alpha_j^q \, 2^{jq(s+\frac{1}{2})} \le \left(\frac{C}{\varepsilon}\right)^q \right\}.$$

So, the following will be devoted to the study of the maximum of the function

$$\tau((\alpha_j)_{-1\le j<j_\eta}) = \sum_{j<j_\eta} 2^{jp'(s'+\frac{1}{2})} \inf_{\lambda_j} \mathbb{E}_{f_{\alpha_j}} r(\lambda_j,\xi)$$

on the set $\mathcal{K} = \left\{ (\alpha_j)_{-1\le j<j_\eta} : \quad \alpha_j \ge 0 \text{ and } \sum_{j<j_\eta} \alpha_j^q 2^{jq(s+\frac{1}{2})} \le \left(\frac{C}{\varepsilon}\right)^q \right\}$.

Let $\tilde{\alpha}$ be the point of $\mathcal{K}$ where τ attains its maximum. In the following, we use j_* in $\mathbb{N}$ (defined in section 3.4.1), so that

$$2^{j_*} \simeq \left(\frac{C}{\varepsilon}\right)^{\frac{1}{s+\frac{1}{2}}}, \tag{3.19}$$

and

$$\tau(\tilde{\alpha}) = S_{\varepsilon,1} + S_{\varepsilon,2},$$

where

$$S_{\varepsilon,1} = \sum_{j=-1}^{j_*-1} 2^{jp'(s'+\frac{1}{2})} \inf_{\lambda_j} \mathbb{E}_{f_{\tilde{\alpha}_j}} r(\lambda_j,\xi),$$

and

$$S_{\varepsilon,2} = \sum_{j=j_*}^{j_\eta-1} 2^{jp'(s'+\frac{1}{2})} \inf_{\lambda_j} \mathbb{E}_{f_{\tilde{\alpha}_j}} r(\lambda_j,\xi).$$

By using the second and the third points of Proposition 3.4 and the value of ϑ given in Theorem 1,

$$\begin{aligned}
S_{\varepsilon,1} &\le \sum_{j<j_*} 2^{jp'(s'+\frac{1}{2})} \inf_{\lambda_j} \int_{-\infty}^{+\infty} |z|^{p'} \phi(z+\lambda_j) dz \int_{\tilde{\alpha}_j}^{+\infty} \xi^{-p-1} \, p \, \tilde{\alpha}_j^p d\xi \\
&\le \int_{-\infty}^{+\infty} |z|^{p'} \phi(z) dz \sum_{j<j_*} 2^{jp'(s'+\frac{1}{2})} \\
&\le K \left(\frac{C}{\varepsilon}\right)^{\frac{p'(s'+\frac{1}{2})}{s+\frac{1}{2}}} \\
&\le K \left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)}.
\end{aligned}$$

To get an upper bound for $S_{\varepsilon,2}$, we need to evaluate the Bayes threshold risk $\mathbb{E}_{f_{\alpha_j}} r(\lambda_j, \xi)$ when α_j tends to 0. We have the following theorem, which generalizes the case $p' = 2$ and $p < 2$ investigated by Johnstone (1994) :

Theorem 3.2. *Let us suppose that we are given $1 \leq p' < \infty$ and $p > 0$. When α tends to 0 and for any threshold $\lambda \geq (-2 \log \alpha^{p \wedge p'})^{\frac{1}{2}}$, we have :*

- *if $p < p'$,*

$$\mathbb{E}_{f_\alpha} r(\lambda, \xi) \sim \frac{p'}{p' - p} \, \alpha^p \, \lambda^{p'-p},$$

- *if $p = p'$,*

$$\mathbb{E}_{f_\alpha} r(\lambda, \xi) \sim p \, \alpha^p \, \log \left(\frac{\lambda}{\alpha} \right),$$

- *if $p > p'$,*

$$\mathbb{E}_{f_\alpha} r(\lambda, \xi) \sim \frac{p}{p - p'} \, \alpha^{p'}.$$

Proof : Using Proposition 3.4, we have

$$\begin{aligned} \mathbb{E}_{f_\alpha} r(\lambda, \xi) &= \int_\alpha^\infty p\alpha^p \xi^{-1-p} r(\lambda, \xi) d\xi \\ &= r(\lambda, \alpha) + p'\alpha^p \int_\alpha^{+\infty} \xi^{p'-p-1} \Phi([-\lambda - \xi, \lambda - \xi]) d\xi. \end{aligned}$$

Let us define

$$I(\lambda, \alpha) = \int_\alpha^{+\infty} \xi^{p'-p-1} \Phi([-\lambda - \xi, \lambda - \xi]) d\xi.$$

The asymptotic values (when α tends to 0 and λ to $+\infty$) of this integral are easily available :

1. If $p < p'$,

$$I(\lambda, \alpha) \sim \frac{1}{p' - p} \lambda^{p'-p}.$$

2. If $p = p'$,

$$I(\lambda, \alpha) \sim \log \left(\frac{\lambda}{\alpha} \right).$$

3. If $p > p'$,

$$I(\lambda, \alpha) \sim \frac{1}{p - p'} \alpha^{p'-p}.$$

Using the second and fourth points of Proposition 3.4, we get

$$r(\lambda,\alpha) = r(\lambda,0) + \int_0^\alpha p'\, \xi^{p'-1}\Phi([-\lambda-\xi,\lambda-\xi])d\xi,$$

and

$$K_1\lambda^{-p'-1}\exp(-\frac{\lambda^2}{2}) + \Phi([-\lambda,\lambda-\alpha])\ \alpha^{p'} \le r(\lambda,\alpha) \le K_2\lambda^{-p'-1}\exp(-\frac{\lambda^2}{2})\ + \alpha^{p'}.$$

So, we just have to take $\lambda \ge (-2\ \log \alpha^{p\wedge p'})^{\frac{1}{2}}$ to obtain :

$$\mathbb{E}_{f_\alpha} r(\lambda,\xi) \sim \alpha^{p'} + p'\alpha^p\ I(\lambda,\alpha),$$

which yields the results of the theorem.

□

Now, we are able to get the upper bound of $S_{\varepsilon,2}$. Using (3.19), we have

$$\forall\ j_* \le j < j_\eta,\quad \tilde{\alpha}_j^p \le \left(\frac{C}{\varepsilon}\right)^p\ 2^{-jp(s+\frac{1}{2})} \le \left(\frac{C}{\varepsilon}\right)^p\ 2^{-j_*p(s+\frac{1}{2})} < 1.$$

Consequently, taking $\lambda_j = (-2\log(\alpha_j^p))^{\frac{1}{2}}$, we have :

- if $p < p'$,

$$S_{\varepsilon,2} \le K\ \max\left\{\sum_{j=j_*}^{j_\eta-1} 2^{jp'(s'+\frac{1}{2})}\alpha_j^p\left(-\log\alpha_j^p\right)^{\frac{p'-p}{2}}\ :\ \sum_{j=j_*}^{j_\eta-1}\alpha_j^q\ 2^{jq(s+\frac{1}{2})} = \left(\frac{C}{\varepsilon}\right)^q\right\},$$

- if $p = p'$,

$$S_{\varepsilon,2} \le K\ \max\left\{\sum_{j=j_*}^{j_\eta-1} 2^{jp'(s'+\frac{1}{2})}\alpha_j^p\log\left(\frac{1}{\alpha_j}\right) :\ \sum_{j=j_*}^{j_\eta-1}\alpha_j^q\ 2^{jq(s+\frac{1}{2})} = \left(\frac{C}{\varepsilon}\right)^q\right\},$$

- if $p > p'$,

$$S_{\varepsilon,2} \le K\ \max\left\{\sum_{j=j_*}^{j_\eta-1} 2^{jp'(s'+\frac{1}{2})}\alpha_j^{p'}\ :\ \sum_{j=j_*}^{j_\eta-1}\alpha_j^q\ 2^{jq(s+\frac{1}{2})} = \left(\frac{C}{\varepsilon}\right)^q\right\}.$$

To evaluate these maxima, we use the Lagrange mutipliers on the open set

$$U = \{\alpha_j > 0 : \quad \forall\, j_* \leq j < j_\eta\}\,.$$

We can easily show that this method works only for the critical case when $q > p$, and the maximum is attained at the point which coordinates α_j are equivalent to $T2^{-j(s+\frac{1}{2})}j^a$, where $a = \frac{(p'-p)}{2(q-p)}$ and T is a constant depending on ε and η. And, we have :

$$S_{\varepsilon,2} \leq K\left(\frac{C}{\varepsilon}\right)^p j_\eta^{\frac{p'-p}{2}+1-\frac{p}{q}}.$$

For the other cases, the maximum is attained on the boundary of U, which means that at least one coordinate equals zero. By repeating this argument, we finally get that if $(s,p,q) \in \mathcal{R} \cup (\mathcal{C} \cap \{q = p\})$, all the coordinates equal zero except one (α_{j_*} on $\mathcal{R}$, $\alpha_{j_\eta - 1}$ on $\mathcal{C}$ and if $q = p$). It follows :

- on $\mathcal{R}$,

$$\begin{aligned} S_{\varepsilon,2} &\leq K\left(\frac{C}{\varepsilon}\right)^p 2^{j_*(p'(s'+\frac{1}{2})-p(s+\frac{1}{2}))} \\ &\leq K\left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)}, \end{aligned}$$

- on $\mathcal{C}$ and if $q = p$, using $p = p'(1-\vartheta)$,

$$\begin{aligned} S_{\varepsilon,2} &\leq K\left(\frac{C}{\varepsilon}\right)^p 2^{j_\eta(p'(s'+\frac{1}{2})-p(s+\frac{1}{2}))}\left[\log 2^{j_\eta p(s+\frac{1}{2})} - \log\left(\frac{C}{\varepsilon}\right)^p\right]^{\frac{p'-p}{2}} \\ &\leq K(\eta)\left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)}\left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}}, \end{aligned}$$

- on $\mathcal{C}$ and if $q > p$, using $p = p'(1-\vartheta)$,

$$S_{\varepsilon,2} \leq K(\eta)\left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)}\left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+1-\frac{p}{q}}.$$

Finally, we obtain the required inequality :

$$B(M,\varepsilon) \leq K\varepsilon^{p'}\left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)} \qquad \text{on } \mathcal{R},$$

and

$$B(M,\varepsilon) \leq K(\eta)\,\varepsilon^{p'}\left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)}\left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})_+} \quad \text{on } \mathcal{C}.$$

We conclude that the upper bound of Theorem 1 is proved.

□

3.5.2 The prior π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$

Before showing that π_ε is an asymptotically least favorable prior for $\mathcal{WB}_{s,p,q}(C)$, let us provide some explanations about the way we have constructed it. Section 3.5.1 suggests to use a prior based on Pareto(p)-variables and on sequences $(\alpha_j)_j$ of the form given in section 3.4.1. For technical reasons that appear along this section, we introduce the sequence $(\mu_j)_j$ and the integer j_2. For $j < j_*$, section 3.5.1 does not provide any information for the choice of the prior for β_{jk}. But we remember that two point priors are often met in the literature and appear in the definition of least favorable priors for $\mathcal{B}_{s,p,q}(C)$. Since $\mathcal{B}_{s,p,q}(C)$ and $\mathcal{WB}_{s,p,q}(C)$ are close, il seems reasonable to choose two point priors for the coarsest levels ($j < j_*$), independently of the zone, and at level j, that assign $\frac{1}{2}$ mass to $\pm\varepsilon\alpha_j$, with α_j about of the same form as for $j \geq j_*$. Finally, the constant V_ε is chosen to ensure that on $\mathcal{C}$ and when $q > p$, (3.12) is checked. The values of α_{j_1} for the other zones are then obtained by using (3.12). In the following, we shall note F_j, the distribution of $\varepsilon^{-1}\beta_{jk}$ and F_j^+, the distribution of $\varepsilon^{-1}|\beta_{jk}|$.

Asymptotic evaluation of $B(\pi_\varepsilon)$

Our goal is to prove the following proposition :

Proposition 3.5. *We have*

$$B(\pi_\varepsilon) \approx \varepsilon^{p'} \left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)} \qquad on\ \mathcal{R},$$

and

$$B(\pi_\varepsilon) \approx \varepsilon^{p'} \left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)} \left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})_+} \qquad on\ \mathcal{C}.$$

Proof : Let us recall that $\eta > s'$, but $\eta - s'$ is small. So, we can assume that $j_2 \leq j_\eta$. Let us begin by proving :

Lemma 3.1. $B(M,\varepsilon) \geq B(\pi_\varepsilon)$.

Proof : We prove that π_ε belongs to $m_{s,p,q}(C, j_\eta)$. It is easy to show that $\forall\ \lambda > 0$, $\forall\ j < j_\eta$, $\forall\ k \in \mathcal{I}_j$,

$$\mathbb{E}_{\pi_\varepsilon}(\mathbb{1}_{\ |\beta_{jk}|>\lambda}) \leq \alpha_j^p(\lambda\varepsilon^{-1})^{-p}.$$

Consequently, using (3.12), $\forall\ \lambda > 0$,

$$\sum_{j<j_\eta} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left\{\sum_{k\in\mathcal{I}_j} \mathbb{E}_{\pi_\varepsilon}(\mathbb{1}_{\ |\beta_{jk}|>\lambda})\right\}^{\frac{q}{p}} \leq \left(\frac{C}{\lambda}\right)^q,$$

so π_ε lies in $m_{s,p,q}(C, j_\eta)$, which ends the proof of the lemma.

□

On the other hand, we have :

$$\begin{aligned} B(\pi_\varepsilon) &= \inf_{\hat\beta} \sum_{j=-1}^{j_2} 2^{jp'(s'+\frac{1}{2}-\frac{1}{p'})} \sum_k \mathbb{E}_{\pi_\varepsilon}\ \mathbb{E}_{\beta_{jk}} |\hat\beta_{jk} - \beta_{jk}|^{p'} \\ &= \varepsilon^{p'} \sum_{j=-1}^{j_2} 2^{jp'(s'+\frac{1}{2})} \inf_{d_j} \mathbb{E}_{F_j}\ \mathbb{E}_{u_{j1}} |d_j(x_{j1}) - u_{j1}|^{p'} \\ &\geq \varepsilon^{p'} \sum_{j=j_1}^{j_2} 2^{jp'(s'+\frac{1}{2})}\ b(\alpha_j, p'), \end{aligned}$$

where $\forall\ j \in \{j_1, \ldots, j_2\}$, $b(\alpha_j, p')$ is the univariate Bayes risk for F_j, the distribution of $\varepsilon^{-1}\beta_{jk}$, and for the $L_{p'}$-loss :

$$b(\alpha_j, p') = \inf_d\ \mathbb{E}_{F_j} \int |d(x) - \xi|^{p'} \phi(x - \xi) dx.$$

When $p' = 1$, the Bayesian estimator $d(x)$ is easily available since it is the median of the posterior distribution. We have the following result :

Theorem 3.3. *If $p' = 1$, when α tends to 0,*

- *if $p < 1$, $b(\alpha, 1) \sim \frac{1}{1-p}\ \alpha^p(-2\log\alpha^p)^{\frac{1-p}{2}}$,*
- *if $p = 1$, $b(\alpha, 1) \sim \alpha \log\left(\frac{1}{\alpha}\right)$,*
- *if $p > 1$, $b(\alpha, 1) \sim \frac{p}{p-1}\ \alpha$.*

When p' is arbitrary and $p < p'$, when α tends to 0,

$$b(\alpha, p') \sim \frac{p'}{p'-p}\ \alpha^p(-2\log\alpha^p)^{\frac{p'-p}{2}}.$$

This theorem is proved in Appendix.
To get the lower bound when $p \geq p'$, we use the Jensen inequality, which shows that :

$$\begin{aligned} b(\alpha_j, p') &\geq \inf_d \left[\mathbb{E}_{F_j} \int |d(x) - \xi| \phi(x - \xi) dx\right]^{p'} \\ &\geq b(\alpha_j, 1)^{p'}. \end{aligned}$$

Therefore,

- if $p < p'$,

$$B(\pi_\varepsilon) \geq \frac{p'}{p'-p} \ \varepsilon^{p'} \sum_{j=j_1}^{j_2} 2^{jp'(s'+\frac{1}{2})} \ \alpha_j^p \ (-2\log\alpha_j^p)^{\frac{p'-p}{2}} (1+o(\alpha_j)),$$

- if $p \geq p' > 1$, or if $p > p' = 1$,

$$B(\pi_\varepsilon) \geq \left(\frac{p}{p-1}\right)^{p'} \varepsilon^{p'} \sum_{j=j_1}^{j_2} 2^{jp'(s'+\frac{1}{2})} \ \alpha_j^{p'} (1+o(\alpha_j)),$$

- if $p = p' = 1$,

$$B(\pi_\varepsilon) \geq \varepsilon^{p'} \sum_{j=j_1}^{j_2} 2^{jp'(s'+\frac{1}{2})} \ \alpha_j^{p'} \left(\log\frac{1}{\alpha_j}\right)^{p'} (1+o(\alpha_j)).$$

So, we get

$$B(\pi_\varepsilon) \geq K \ \varepsilon^{p'} \left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)} \qquad \text{on } \mathcal{R},$$

$$B(\pi_\varepsilon) \geq K \ \varepsilon^{p'} \left(\frac{C}{\varepsilon}\right)^{p'(1-\vartheta)} \left(\log\left(\frac{C}{\varepsilon}\right)\right)^{\frac{p'\vartheta}{2}+(1-\frac{p}{q})+} \qquad \text{on } \mathcal{C}.$$

Using Lemma 3.1 and the results about the minimax Bayes risk of section 5.1, Proposition 3.5 is proved.

□

The support property

We only deal with the case $p' > p$, $p(s+\frac{1}{2}) = p'(s'+\frac{1}{2})$ and $q > p$. The other cases follow easily from the same arguments. Let $\gamma_\varepsilon = 1 + \left(\log\frac{1}{\varepsilon}\right)^{-1}$ be a real number greater than 1. Along this subsection, we will use that for all $k > p$ and for all j in $\{j_1, \dots, j_2\}$,

$$\mathbb{E}_{\pi_\varepsilon} |\beta_{j1}|^k = \varepsilon^k \left(\frac{k}{k-p} \ \alpha_j^p \ \mu_j^{k-p} - \frac{p}{k-p}\alpha_j^k\right).$$

Our goal is to show the support property :

Proposition 3.6. *Under π_ε, $\mathbb{P}(\beta \in \mathcal{WB}_{s,p,q}(\gamma_\varepsilon C)) \overset{\varepsilon\longrightarrow 0}{\longrightarrow} 1$.*

Proof : We want to prove that under π_ε,

$$\mathbb{P}\left[\sum_{j\leq j_2} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_k \mathbb{1}_{\ |\beta_{jk}|>\lambda}\right)^{\frac{q}{p}} \leq \left(\frac{\gamma_\varepsilon C}{\lambda}\right)^q, \quad \forall\,\lambda > 0\right] \overset{\varepsilon\to 0}{\longrightarrow} 1. \tag{3.21}$$

It will be a consequence of the asymptotic evaluation of $(\mathbb{P}(A_j))_{j_1 \leq j \leq j_2}$, where

$$A_j = \bigcap_{\lambda > 0} \left\{ \frac{1}{2^j} \sum_k \mathbb{1}_{\;|\beta_{jk}|>\lambda} \leq \left(\frac{\varepsilon}{\lambda}\right)^p \gamma_\varepsilon^p \; \alpha_j^p \right\},$$

given by the following lemma :

Lemma 3.2. *Under π_ε, there exists $\tilde{\kappa} > 0$ such that, for ε small enough and for all j in $\{j_1, \ldots, j_2\}$,*

$$\mathbb{P}(A_j^c) \leq 2 \exp \left(-2^{j+1} (\gamma_\varepsilon^p - 1)^2 \left(\frac{\alpha_j}{\mu_j} \right)^{2p} \right) \leq 2 \exp \left(-(\gamma_\varepsilon^p - 1)^2 \left(\frac{C}{\varepsilon} \right)^{\tilde{\kappa}} \right).$$

The second inequality is derived from the fact that for all j in $\{j_1, \ldots, j_2\}$,

$$\begin{aligned} 2^{j+1} \left(\frac{\alpha_j}{\mu_j} \right)^{2p} &\geq K \frac{1}{\mu_{j_2}^{2p}} \left(\frac{C}{\varepsilon} \right)^{2p} 2^{j_2(1-p'(2s'+1))} \, j_2^{-\frac{2p}{q}} \\ &\geq \left(\frac{C}{\varepsilon} \right)^{\tilde{\kappa}}. \end{aligned}$$

For the last inequality, we use that κ belongs to $\left]1, \frac{2p'(s'+\frac{1}{2})}{2p'(s'+\frac{1}{2})-1}\right[$. We have :

$$\begin{aligned} A_j &= \bigcap_{\lambda > 0} \left\{ \frac{1}{2^j} \sum_k \mathbb{1}_{\;|\beta_{jk}|>\lambda} \leq \left(\frac{\varepsilon}{\lambda}\right)^p \gamma_\varepsilon^p \; \alpha_j^p \right\} \\ &= \bigcap_{0<\lambda\leq\varepsilon\mu_j} \left\{ \frac{1}{2^j} \sum_k \mathbb{1}_{\;|\beta_{jk}|>\lambda} \leq \left(\frac{\varepsilon}{\lambda}\right)^p \gamma_\varepsilon^p \; \alpha_j^p \right\} \\ &= \bigcap_{\left(\frac{\alpha_j}{\mu_j}\right)^p \leq v \leq 1} \left\{ \frac{1}{2^j} \sum_k \mathbb{1}_{\;\left(\frac{|\varepsilon^{-1}\beta_{jk}|}{\alpha_j}\right)^{-p} < v} \leq \gamma_\varepsilon^p \; v \right\}. \end{aligned}$$

As the distribution of $|\varepsilon^{-1}\beta_{jk}|$ is F_j^+, the distribution of $\left(\frac{|\varepsilon^{-1}\beta_{jk}|}{\alpha_j}\right)^{-p}$ is :

$$\left(\frac{\alpha_j}{\mu_j} \right)^p \delta_{\left(\frac{\alpha_j}{\mu_j}\right)^p}(\xi) + \mathbb{1}_{\left[\left(\frac{\alpha_j}{\mu_j}\right)^p, 1\right]}(\xi) d\xi.$$

Since we consider the values of v greater than $\left(\frac{\alpha_j}{\mu_j}\right)^p$, the distribution of $\left(\frac{|\varepsilon^{-1}\beta_{jk}|}{\alpha_j}\right)^{-p}$ matches that of U_{jk}, where the U_{jk}'s are independent uniform observations over $[0,1]$. Therefore,

$$\begin{aligned}
\mathbb{P}(A_j^c) &\leq \mathbb{P}\left(\sup_{\left(\frac{\alpha_j}{\mu_j}\right)^p \leq v \leq 1} |F_{2^j}^u(v) - v| > (\gamma_\varepsilon^p - 1)\left(\frac{\alpha_j}{\mu_j}\right)^p\right) \\
&\leq \mathbb{P}\left(\sup_{v\in[0,1]} |F_{2^j}^u(v) - v| > (\gamma_\varepsilon^p - 1)\left(\frac{\alpha_j}{\mu_j}\right)^p\right),
\end{aligned}$$

where $F_{2^j}^u$ is the empirical distribution of the U_{jk}'s for k in $\mathcal{I}_j$. Hence, using the DKW inequality proved by Massart (1990),

$$\mathbb{P}(A_j^c) \leq 2\exp\left(-2^{j+1}(\gamma_\varepsilon^p - 1)^2\left(\frac{\alpha_j}{\mu_j}\right)^{2p}\right),$$

which ends the proof of the lemma.

□

Under π_ε, since

$$\sum_{j\leq j_2} 2^{jq(s+\frac{1}{2})}\,\alpha_j^q = \left(\frac{C}{\varepsilon}\right)^q,$$

$$\mathbb{P}\left[\sum_{j\leq j_2} 2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_k \mathbb{1}_{|\beta_{jk}|>\lambda}\right)^{\frac{q}{p}} \leq \left(\frac{\gamma_\varepsilon C}{\lambda}\right)^q, \quad \forall\,\lambda>0\right] \geq \mathbb{P}\left(A_{-1}\bigcap\left[\bigcap_{j=j_1}^{j_2} A_j\right]\right),$$

where

$$A_{-1} = \bigcap_{\lambda>0}\left\{\sum_{j<j_1} 2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_k \mathbb{1}_{|\beta_{jk}|>\lambda}\right)^{\frac{q}{p}} \leq \gamma_\varepsilon^q\left(\frac{\varepsilon}{\lambda}\right)^q\sum_{j<j_1} 2^{jq(s+\frac{1}{2})}\alpha_j^q\right\}.$$

We have $\mathbb{P}(A_{-1}) = 1$, and

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{j=j_1}^{j_2} A_j\right) &= \prod_{j=j_1}^{j_2}(1 - \mathbb{P}(A_j^c)) \\
&\geq \prod_{j=j_1}^{j_2}\left[1 - 2\exp\left(-(\gamma_\varepsilon^p - 1)^2\left(\frac{C}{\varepsilon}\right)^{\tilde{\kappa}}\right)\right] \\
&\geq \exp\left[-\exp\left(-\left(\frac{C}{\varepsilon}\right)^{\frac{\tilde{\kappa}}{2}}\right)\right],
\end{aligned}$$

for ε small enough, and with $\gamma_\varepsilon = 1 + \log(\frac{1}{\varepsilon})^{-1}$. Therefore,

$$\mathbb{P}\left[\sum_{j\le j_2} 2^{jq(s+\frac{1}{2}-\frac{1}{p})}\left(\sum_k \mathbb{1}_{\ |\beta_{jk}|>\lambda}\right)^{\frac{q}{p}} \le \left(\frac{\gamma_\varepsilon C}{\lambda}\right)^q, \quad \forall\, \lambda>0\right] \ge \exp\left[-\exp\left(-\left(\frac{C}{\varepsilon}\right)^{\frac{\tilde{\kappa}}{2}}\right)\right],$$

and (3.21) is checked. Proposition 3.6 is proved.

□

Remark 3.2. *To get the lower bound of R_ε on $\mathcal{C}$, we apply the Bayes method. We just give here the main arguments since the proof is standard. We denote $\pi_{\varepsilon,\gamma_\varepsilon}$ the asymptotically least favorable prior associated with $\mathcal{WB}_{s,p,q}(C\gamma_\varepsilon^{-1})$. In particular, we have :*

$$\sum_j 2^{jq(s+\frac{1}{2})}\ \alpha_j^q = \left(\frac{C{\gamma_\varepsilon}^{-1}}{\varepsilon}\right)^q.$$

By using some of the previous inequalities, it is easy to prove that $\pi_{\varepsilon,\gamma_\varepsilon}(\Theta^{*c})$ tends to 0 with an exponential rate of convergence. We denote $\nu_{\varepsilon,\gamma_\varepsilon}$ the probability measure defined by

$$\nu_{\varepsilon,\gamma_\varepsilon}(\mathcal{A}) = \pi_{\varepsilon,\gamma_\varepsilon}(\mathcal{A}|\Theta^*) = \frac{1}{\pi_{\varepsilon,\gamma_\varepsilon}(\Theta^*)}\pi_{\varepsilon,\gamma_\varepsilon}(\mathcal{A}\cap\Theta^*),$$

and let $\hat{\beta}^\nu$ and $\hat{\beta}^\pi$ be the Bayesian estimators respectively for $\nu_{\varepsilon,\gamma_\varepsilon}$ and $\pi_{\varepsilon,\gamma_\varepsilon}$. We have

$$B(\nu_{\varepsilon,\gamma_\varepsilon}) \ge \frac{1}{\pi_{\varepsilon,\gamma_\varepsilon}(\Theta^*)}\mathbb{E}_{\pi_{\varepsilon,\gamma_\varepsilon}}\mathbb{E}_\beta\|\hat{\beta}^\pi-\beta\|^{p'}_{\mathcal{B}_{s',p',p'}} - \frac{1}{\pi_{\varepsilon,\gamma_\varepsilon}(\Theta^*)}\mathbb{E}_{\pi_{\varepsilon,\gamma_\varepsilon}}\mathbb{E}_\beta\left(\mathbb{1}_{\ \Theta^{*c}}(\beta)\|\hat{\beta}^\nu-\beta\|^{p'}_{\mathcal{B}_{s',p',p'}}\right).$$

And since each marginal distribution of $\pi_{\varepsilon,\gamma_\varepsilon}$ is compactly supported, we can easily prove that

$$\begin{aligned} B(\nu_{\varepsilon,\gamma_\varepsilon}) &\ge \mathbb{E}_{\pi_{\varepsilon,\gamma_\varepsilon}}\mathbb{E}_\beta\|\hat{\beta}^\pi-\beta\|^{p'}_{\mathcal{B}_{s',p',p'}}(1+o_\varepsilon(1)) \\ &\ge B(\pi_\varepsilon)(1+o_\varepsilon(1)). \end{aligned}$$

Finally, as $\nu_{\varepsilon,\gamma_\varepsilon}$ is supported by Θ^*, $R^*_\varepsilon \ge B(\nu_{\varepsilon,\gamma_\varepsilon})$ and we have $R_\varepsilon \ge B(\pi_\varepsilon)(1+o_\varepsilon(1))$.

Now, we show that π_ε cannot be an asymptotically least favorable prior for $\mathcal{B}_{s,p,q}(C)$, by proving Proposition 3.3. More precisely, we have :

Proposition 3.7. *For all δ_ε in $\left[1, \log\left(\frac{1}{\alpha^p_{j_1}}\right)^{\frac{1}{p}}\right]$, under π_ε, $\mathbb{P}\left(\beta\in\mathcal{B}_{s,p,q}(\delta_\varepsilon C)\right) \overset{\varepsilon\to 0}{\longrightarrow} 0.$*

Proof : Let us fix δ_ε in $\left[1, \log\left(\frac{1}{\alpha^p_{j_1}}\right)^{\frac{1}{p}}\right]$. Proposition 3.7 is a straightforward consequence of the following lemma :

Lemma 3.3. *There exists $\tilde{\kappa} > 0$ such that for ε small enough, for all j in $\{j_1, \dots, j_2\}$, under π_ε,*

$$\mathbb{P}\left(\frac{1}{2^j} \sum_k \frac{|\, \varepsilon^{-1}\beta_{jk}\,|^p}{\alpha_j^p} > (\delta_\varepsilon^q - m_\varepsilon)^{\frac{p}{q}} \right) \geq 1 - K \left(\frac{C}{\varepsilon}\right)^{-\tilde{\kappa}},$$

where

$$m_\varepsilon = \left(\frac{C}{\varepsilon}\right)^{-q} \sum_{j<j_1} 2^{jq(s+\frac{1}{2})} \alpha_j^q = o_\varepsilon(1).$$

Proof of the lemma : We introduce the variables

$$X_{jk} = \frac{\varepsilon^{-1}\beta_{jk}}{\alpha_j}$$

with the following moments :

$$\mathbb{E} \mid X_{jk} \mid^p = p \, \log\left(\frac{\mu_j}{\alpha_j}\right) + 1$$

and

$$\mathbb{E} \mid X_{jk} \mid^{2p} = 2 \, \left(\frac{\mu_j}{\alpha_j}\right)^p - 1.$$

Since $\log\left(\frac{1}{\alpha_{j_1}^p}\right) + \log\left(\mu_{j_1}^p\right) > (\delta_\varepsilon^q - m_\varepsilon)^{\frac{p}{q}}$, for ε small enough,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{2^j} \sum_k \frac{|\, \varepsilon^{-1}\beta_{jk}\,|^p}{\alpha_j^p} > (\delta_\varepsilon^q - m_\varepsilon)^{\frac{p}{q}} \right) &\geq 1 - \mathbb{P}\left(\left| \frac{1}{2^j} \sum_k \mid X_{jk}|^p - \mathbb{E}|X_{j1} \mid^p \right| > 1 \right) \\
&\geq 1 - \mathrm{var}\left(\frac{1}{2^j} \sum_k \mid X_{jk} \mid^p \right) \\
&\geq 1 - K \, \frac{1}{2^j} \, \left(\frac{\mu_j}{\alpha_j}\right)^p \\
&\geq 1 - K \, \left(\frac{C}{\varepsilon}\right)^{-\tilde{\kappa}}.
\end{aligned}$$

For the last inequality, we use the same arguments as for Lemma 3.2.

□

3.6 Appendix : Proof of Theorem 3.3

Let us recall the notations introduced by Johnstone (1994), we use in the following :

Definition 3.3. *For all $\alpha > 0$, we define μ and c, depending on α such that when α tends to 0,*

1. $\mu \longrightarrow +\infty$,
2. $c \longrightarrow +\infty$,
3. $c = o(\mu)$,
4. $\phi(\mu+c) = \left(\frac{\alpha}{\mu}\right)^p \phi(c)$, *where* ϕ *denotes the standard Gaussian density function. These four conditions entail*

$$\mu(\alpha) \overset{\alpha\to 0}{\sim} (-2\log\alpha^p)^{\frac{1}{2}}.$$

We will need the following real numbers : $T = \mu + \frac{c}{2} \longrightarrow \infty$, and $\mu^- = \frac{1}{2T} \longrightarrow 0$. In the following, the notation $o_\alpha(1)$ will keep designating any function that is bounded by a function depending only on α and tending to 0 when α tends to 0. We evaluate

$$b(\alpha, p') = \inf_d \mathbb{E}_{f_\alpha} \int |d(x) - \xi|^{p'} \phi(x-\xi)dx,$$

with

$$f_\alpha^+(\xi)d\xi = p\alpha^p\, \xi^{-1-p} \mathbb{1}_{[\alpha,\mu]}(\xi)d\xi + \left(\frac{\alpha}{\mu}\right)^p \delta_\mu(\xi),$$

and

$$f_\alpha(\xi) = \frac{f_\alpha^+(\xi) + f_\alpha^+(-\xi)}{2}.$$

To get the upper bounds of Theorem 3.3, we notice that, with $d(x) = 0$,

$$b(\alpha, p') \le \mathbb{E}_{f_\alpha} \int |\xi|^{p'} \phi(x-\xi)dx,$$

which allows to obtain the result. The following lemmas will be useful in the following :

Lemma 3.4. *When* α *tends to 0, for all* $\mu^- \le x \le T$,

$$\int_{2\mu^-}^{\mu} \phi(x-\xi)p\alpha^p\xi^{-p-1}d\xi \le \phi(x)\exp(-\frac{1}{2}\mu c)\mu^p T^p = \phi(x)o_\alpha(1).$$

Proof :

$$\begin{aligned}
\int_{2\mu^-}^{\mu} \phi(x-\xi)p\alpha^p\xi^{-p-1}d\xi &\le \phi(x)\int_{2\mu^-}^{\mu} \exp(-\frac{\xi^2}{2} + T\xi)p\alpha^p\xi^{-1-p}d\xi \\
&\le \phi(x)\exp(-\frac{\mu^2}{2} + T\mu)\int_{2\mu^-}^{\mu} p\alpha^p\xi^{-1-p}d\xi \\
&\le \phi(x)\exp(-\frac{\mu^2}{2} + T\mu)\alpha^p T^p \\
&\le \phi(x)\exp(-\frac{1}{2}\mu c)\mu^p T^p.
\end{aligned}$$

As $\exp(-\frac{1}{2}\mu c)\mu^p T^p$ tends to 0, the lemma is proved.

□

Lemma 3.5. *When α tends to 0, for all $\mu^- \leq x \leq T$,*

$$\int_{\alpha}^{2\mu^-} \phi(x-\xi) p\alpha^p \xi^{-p-1} d\xi = \phi(x)(1+o_\alpha(1)).$$

Proof : To prove this lemma, we suppose that the random variable ξ has the density

$$g(z) = p(\alpha^{-p} - (2\mu^-)^{-p})^{-1} 1\!\!1_{[\alpha, 2\mu^-]}(z) z^{-p-1}.$$

We have

$$\int_{\alpha}^{2\mu^-} \phi(x-\xi) p\alpha^p \xi^{-p-1} d\xi \quad \sim \quad \phi(x)\, \mathbb{E}_g \exp\left(x\xi - \frac{\xi^2}{2}\right).$$

As $x \geq \mu^-$, for all ξ in $[\alpha, 2\mu^-]$, $\quad x\xi - \frac{\xi^2}{2} \geq 0$, and

$$\begin{aligned} \left|\exp\left(x\xi - \frac{\xi^2}{2}\right) - 1\right| &\leq \exp\left(2T\mu^- - \frac{(2\mu^-)^2}{2}\right)\left(x\xi - \frac{\xi^2}{2}\right) \\ &\leq \exp(1) x\xi \\ &\leq \exp(1) T\xi. \end{aligned}$$

For all $\varepsilon > 0$, for $\alpha < \varepsilon \exp(-1) T^{-1}$,

$$\begin{aligned} \mathbb{P}_g\left(\left|\exp\left(x\xi - \frac{\xi^2}{2}\right) - 1\right| > \varepsilon\right) &\leq \mathbb{P}_g\left(\xi > \varepsilon \exp(-1) T^{-1}\right) \\ &\leq \alpha^p \varepsilon^{-p} \exp(p) T^p, \end{aligned}$$

which tends to 0 when α tends to 0.
Therefore, as for all ξ in $[\alpha, 2\mu^-]$, $\quad \left|\exp\left(x\xi - \frac{\xi^2}{2}\right) - 1\right| \leq \exp(1)$,

$$\mathbb{E}_g \exp\left(x\xi - \frac{\xi^2}{2}\right) = 1 + o_\alpha(1).$$

The lemma is proved.

□

In the following, we will consider the Bayesian estimators associated with f_α and f_α^+ :

$$d(x) = \arg\inf_m \int_{\mathbb{R}} f_\alpha(\xi)\phi(x-\xi)|\xi - m|^{p'} d\xi,$$

and

$$d^+(x) = \arg\inf_m \int_{\mathbb{R}} f_\alpha^+(\xi)\phi(x-\xi)|\xi - m|^{p'} d\xi.$$

Now, we prove that for all $\mu^- \leq x \leq T$, $d^+(x) \leq 3\mu^-$. For all m in $[3\mu^-, \mu]$, using Lemma 3.5,

$$\begin{aligned}\int_\alpha^{2\mu^-} \phi(x-\xi)p\alpha^p\xi^{-p-1}|\xi - m|^{p'} d\xi &\geq (\mu^-)^{p'} \int_\alpha^{2\mu^-} \phi(x-\xi)p\alpha^p\xi^{-p-1} d\xi \\ &= \phi(x)(\mu^-)^{p'}(1 + o_\alpha(1)),\end{aligned}$$

using Lemma 3.4,

$$\begin{aligned}\int_{2\mu^-}^{\mu} \phi(x-\xi)p\alpha^p\xi^{-p-1}|\xi - m|^{p'} d\xi &\leq \mu^{p'} \int_{2\mu^-}^{\mu} \phi(x-\xi)p\alpha^p\xi^{-p-1} d\xi \\ &\leq \phi(x)\exp(-\frac{1}{2}\mu c)\mu^{p'+p}T^p,\end{aligned}$$

and

$$\begin{aligned}\left(\frac{\alpha}{\mu}\right)^p \int \phi(x-\xi)|\xi - m|^{p'}\delta_\mu(\xi) &\leq \phi(x)\alpha^p\mu^{p'-p}\exp(-\frac{\mu^2}{2} + T\mu) \\ &\leq \phi(x)\mu^{p'}\exp(-\frac{1}{2}\mu c).\end{aligned}$$

As $\exp(-\frac{1}{2}\mu c)\mu^{p'+p}T^p = o((\mu^-)^{p'})$, we have for all m in $[3\mu^-, \mu]$,

$$\begin{aligned}\int_{\mathbb{R}} f_\alpha^+(\xi)\phi(x-\xi)|\xi - m|^{p'} d\xi &\sim \int_\alpha^{2\mu^-} f_\alpha^+(\xi)\phi(x-\xi)|\xi - m|^{p'} d\xi \\ &\geq \int_\alpha^{2\mu^-} f_\alpha^+(\xi)\phi(x-\xi)|\xi - 3\mu^-|^{p'} d\xi \\ &\sim \int_{\mathbb{R}} f_\alpha^+(\xi)\phi(x-\xi)|\xi - 3\mu^-|^{p'} d\xi\end{aligned}$$

and for all x in $[\mu^-, T]$, $\alpha \leq d^+(x) \leq 3\mu^-$. With $f_\alpha^-(\xi) = f_\alpha^+(-\xi)$, for all ξ in $\mathbb{R}$,

$$\begin{aligned}b(\alpha, p') &= \mathbb{E}_{f_\alpha} \int |d(x) - \xi|^{p'}\phi(x-\xi)dx \\ &\geq \frac{1}{2}\inf_m \mathbb{E}_{f_\alpha^+} \int |m - \xi|^{p'}\phi(x-\xi)dx + \frac{1}{2}\inf_m \mathbb{E}_{f_\alpha^-} \int |m - \xi|^{p'}\phi(x-\xi)dx \\ &= \mathbb{E}_{f_\alpha^+} \int |d^+(x) - \xi|^{p'}\phi(x-\xi)dx \\ &\geq \int_{\frac{c}{2}}^{\mu} r(d^+, \xi)p\alpha^p\xi^{-1-p} d\xi + \left(\frac{\alpha}{\mu}\right)^p r(d^+, \mu),\end{aligned}$$

with

$$r(d^+, \xi) = \int |d^+(x) - \xi|^{p'} \phi(x - \xi) dx.$$

For all $\frac{c}{2} \leq \xi \leq \mu$,

$$\begin{aligned} r(d^+, \xi) &\geq \int_{\mu^-}^{T} |d^+(x) - \xi|^{p'} \phi(x - \xi) dx \\ &\geq (\xi - 3\mu^-)^{p'} \int_{\mu^-}^{T} \phi(x - \xi) dx \\ &\geq (\xi - 3\mu^-)^{p'} \int_{\mu^- - \frac{c}{2}}^{\frac{c}{2}} \phi(u) du \\ &\geq (\xi - 3\mu^-)^{p'} s_\alpha, \end{aligned}$$

with s_α that tends to 1 when α tends to 0. Finally, we have when $p < p'$,

$$\begin{aligned} b(\alpha, p') &\geq s_\alpha p \alpha^p \int_{\frac{c}{2}}^{\mu} (\xi - 3\mu^-)^{p'} \xi^{-1-p} d\xi + s_\alpha \left(\frac{\alpha}{\mu}\right)^p (\mu - 3\mu^-)^{p'} \\ &\geq \left(\frac{p}{p' - p} \alpha^p \mu^{p'-p} + \alpha^p \mu^{p'-p}\right) (1 + o_\alpha(1)) \\ &\geq \frac{p'}{p' - p} \alpha^p \mu^{p'-p} (1 + o_\alpha(1)) \\ &\geq \frac{p'}{p' - p} \alpha^p (-2 \log \alpha^p)^{\frac{p'-p}{2}} (1 + o_\alpha(1)). \end{aligned}$$

The second part of Theorem 3.3 is proved.
Now, we suppose that $p' = 1$ and we prove the following lemma :

Lemma 3.6. *For all x such that $|x|$ lies in $[\mu^-, T]$, $\alpha \leq |d(x)| \leq \alpha(1 + o_\alpha(1))$.*

Proof : Without loss of generality, we suppose that $x > 0$. Since $p' = 1$, $d(x)$ is the median of the posterior distribution :

$$\int_{-\infty}^{d(x)} f_\alpha(\xi) \phi(x - \xi) d\xi = \frac{1}{2} \int_{-\infty}^{+\infty} f_\alpha(\xi) \phi(x - \xi) d\xi.$$

Since $x > 0$, we obviously have that $d(x) \geq \alpha$. From Lemma 3.4 and Lemma 3.5, it follows that

$$\int_{\alpha}^{\mu} f_\alpha^+(\xi) \phi(x - \xi) d\xi = \phi(x)(1 + o_\alpha(1)).$$

By using similar arguments as previously, we have

$$\frac{1}{2} \left(\frac{\alpha}{\mu}\right)^p \phi(x + \mu) + \frac{1}{2} \left(\frac{\alpha}{\mu}\right)^p \phi(x - \mu) \leq \phi(x) \exp(-\frac{1}{2}\mu c) = \phi(x) o_\alpha(1).$$

As

$$\begin{aligned}\int_{-\mu}^{-\alpha} f_\alpha^+(-\xi)\phi(x-\xi)d\xi &= \phi(x)\int_\alpha^\mu p\alpha^p\xi^{-1-p}e^{-\frac{\xi^2}{2}-x\xi}d\xi \\ &= \phi(x)\left(e^{-\frac{\alpha^2}{2}-x\alpha}-\left(\frac{\alpha}{\mu}\right)^p e^{-\frac{\mu^2}{2}-x\mu}-\alpha^p\int_\alpha^\mu (x+\xi)\xi^{-p}e^{-\frac{\xi^2}{2}-x\xi}d\xi\right) \\ &= \phi(x)(1+o_\alpha(1)),\end{aligned}$$

we have

$$\int_{-\infty}^{+\infty} f_\alpha(\xi)\phi(x-\xi)d\xi = \phi(x)(1+o_\alpha(1)).$$

Since

$$\begin{aligned}\int_\alpha^{d(x)} f_\alpha(\xi)\phi(x-\xi)d\xi &\geq \int_\alpha^{d(x)} p\alpha^p\xi^{-1-p}d\xi \times \frac{1}{2}\phi(x-\alpha) \\ &\geq \left(1-\left(\frac{\alpha}{d(x)}\right)^p\right)\times\frac{1}{2}\int_{-\infty}^{+\infty} f_\alpha(\xi)\phi(x-\xi)d\xi\times(1+o_\alpha(1)),\end{aligned}$$

we have $d(x) \leq \alpha\ \tau(\alpha)$, where τ is a function that does not depend on x and that tends to 1 when α tends to 0.

□

Now, to get the lower bound of $b(\alpha, 1)$, we write

$$b(\alpha,1) \geq \int_{m_1(\alpha)}^{m_2(\alpha)} f_\alpha(\xi)\, r(d,\xi)d\xi + \int_{-m_2(\alpha)}^{-m_1(\alpha)} f_\alpha(\xi)\, r(d,\xi)d\xi,$$

with

$$r(d,\xi) = \int |d(x)-\xi|\phi(x-\xi)dx,$$

$$m_1(\alpha) = \alpha, \quad m_2(\alpha) = \log\left(\frac{1}{\alpha}\right)^{-1}, \text{ if } p \geq 1,$$

$$m_1(\alpha) = \frac{c}{2}, \quad m_2(\alpha) = \mu, \text{ if } p < 1.$$

From similar computations as previously, we obtain the required inequalities.

□

Chapitre 4

Bayesian thresholding with priors based on Pareto distributions

In this chapter, we consider wavelet thresholding rules within a Bayesian framework. The prior imposed on the wavelet coefficients is based upon a Pareto distribution. We introduce weak Besov spaces that enable us to measure the sparsity of each estimated signal. At first, we establish a relationship between the parameters of the prior and the parameters of the weak Besov space in which the realizations built from the prior lie. Subsequently, we exhibit a thresholding rule which threshold at each resolution level depends on the prior parameters. It is compared to estimators provided by two well known thresholding procedures : VisuShrink and SureShrink.

4.1 Introduction

Let us suppose we are given noisy data of an unknown function f to be estimated :

$$g_i = f(\frac{i}{n}) + \sigma\varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n. \tag{4.1}$$

We would like to build an efficient procedure that provides a data adaptive estimator of the signal f without having to assume any specific form about this signal. By expanding f on a wavelet basis which atoms are localized in both time and frequency, we expect a parsimonious representation of f : only a small number of the wavelet coefficients are non negligible in which the main part of the information about f is contained. In the following, we consider such *sparse* functions we estimate by using thresholding rules particularly appropriate to this framework. One goal of this chapter is to discuss the choice of the threshold. Many authors have investigated this problem : Donoho and Johnstone (1994) proposed the VisuShrink procedure consisting in choosing the universal threshold $\lambda^u = \sigma\sqrt{2\log(n)}$ for each level. The choice of the threshold for Donoho and Johnstone (1995) is based on the minimization of Stein's unbiased estimate of risk for threshold estimates. Nason (1996) exploited the cross validation approach to choose the threshold. Let us also mention the methods based on the multiple hypotheses testing approach. See for instance Abramovich and Benjamini (1995) or Abramovich, Benjamini, Donoho and Johnstone (2000a) who obtained thresholding rules by adapting the false discovery rate method developed by Benjamini and Hochberg (1995). Abramovich, Sapatinas and Silverman (1998) and Vidakovic (1998) considered thresholding within a Bayesian framework. We adopt this approach and we place a prior model on the wavelet coefficients. But before this, we assume that the function f belongs to a weak Besov space.

The first motivation for the use of weak Besov spaces (denoted $\mathcal{W}^*(r,p)$ in the following) to obtain thresholding rules is provided by their definition : To decide whether f belongs to $\mathcal{W}^*(r,p)$, we introduce at each level j the number of its wavelet coefficients greater than a threshold λ, denoted $N_f(j,\lambda)$. We require a power-law bound $C(f) \times \lambda^{-p}$ on the sum over j of the $N_f(j,\lambda)$ penalized by a weight depending on r (see Definition 4.1). We note that if $p < r$, $\mathcal{W}^*(r,p)$ is very close to $\mathcal{B}_{s,p,p}$ $(s = \frac{r}{2p} - \frac{1}{2})$, a member of the class of the strong Besov spaces $\mathcal{B}_{s,p,q}$ often considered by the statisticians. We have the natural inclusion $\mathcal{B}_{s,p,p} \subset \mathcal{W}^*(r,p)$.

Furthermore, weak Besov spaces appeared in statistics to evaluate the performance of classical estimation procedures. Cohen, DeVore, Kerkyacharian and Picard (2001) and Kerkyacharian and Picard (2000) wondered what is the maximal space over which a procedure attains a prescribed rate of convergence. Kerkyacharian and Picard (2000) roughly proved that if a procedure verifies an oracle inequality then its maxiset contains a weak Besov space. Wavelet thresholding is an example of such a procedure. In section 4.2.2, Proposition 1.1 gives a concrete example of the nature of maxisets associated with wavelet thresholding rules. This result provides the natural relationship between weak Besov spaces and wavelet thresholding rules.

Finally, we note that $\mathcal{W}^*(r,p)$ appears as a generalization of the weak l_p space (denoted wl_p in the following), often considered in approximation theory (see Johnstone (1994), Donoho (1996), Donoho and Johnstone (1996), or Cohen, DeVore and Hochmuth (2000)). Abramovich, Benjamini, Donoho and Johnstone (2000a) used weak l_p spaces to define more precisely the notion of sparsity we mentioned previously. For them, sparsity means that there is a relatively small proportion of relatively large coefficients and they introduce a weak l_p constraint to control this proportion. So, as we shall see in section 4.2.2, weak Besov spaces may appear as natural spaces to capture signals in function of their regularity properties and their sparsity. We shall discuss the roles of the parameters r and p in this framework.

But another way to capture the sparsity of a signal is throughout the use of a Bayesian model.

Most of the authors consider Bayesian models based upon gaussian distributions : For instance, both Clyde, Parmigiani and Vidakovic (1998) and Abramovich, Sapatinas and Silverman (1998) consider a mixture of a normal component and a point mass at zero for the wavelet coefficients. Chipman, Kolaczyk and McCulloch (1997) impose a mixture of two gaussian distributions with different variances for negligible and non negligible wavelet coefficients. Huang and Cressie (2000) assume the underlying signal to be composed of a piecewise-smooth deterministic part plus a zero-mean gaussian part.

As for us, rather than adopting the gaussian point of view, we consider priors based upon Pareto distributions. The reasons of this choice are the following : in chapter 3, we investigated the problem of estimation over weak Besov balls, by using the Bayes method. This approach enables us to exhibit a least favorable prior $\pi^{r,p,C}$ for the wavelet coefficients β_{jk} of a function f lying in the weak Besov ball $\mathcal{W}^*(r,p)(C)$. It takes the following form : The β_{jk}'s are independent and the distribution of each β_{jk} is built from the distribution of $\tilde{\alpha}_j X_{jk}$, where X_{jk} is a Pareto(p) variable and $\tilde{\alpha}_j$ is a level-dependent dilation parameter (see section 4.2.3). The realizations built from $\pi^{r,p,C}$ can be viewed as the worst functions to be estimated and lying in $\mathcal{W}^*(r,p)(C)$. What is more, $\pi^{r,p,C}$ is typical of the ball $\mathcal{W}^*(r,p)(C)$: for instance, we can prove that it cannot be a least favorable prior for the problem of estimation over $\mathcal{B}_{s,p,p}(C)$ ($s = \frac{r}{2p} - \frac{1}{2}$), the strong Besov ball naturally associated with $\mathcal{W}^*(r,p)(C)$. In this chapter, we assume we are given a prior model directly inspired by the least favorable priors $\pi^{r,p,C}$ of the weak Besov balls $\mathcal{W}^*(r,p)(C)$: we suppose that the wavelet coefficients β_{jk} of f are independent, each β_{jk} has a symmetric distribution and $|\beta_{jk}| \sim \min(\alpha_j X_{jk} - \alpha_j, \mu_j)$, where X_{jk} is a Pareto(p) variable, α_j and μ_j are level-dependent parameters precisely defined in section 4.3.1. In section 4.3.2, we establish a relationship between these parameters and the parameters of the weak Besov space in which the function f lies (see Theorem 4.1). We present various realizations that give insight into the meaning of the weak Besov space parameters. In particular, we note that if f is typical of $\mathcal{W}^*(r,p)$, then the regularity of f increases with r. When p is small, f presents very high peaks with a regular behavior between them. When p is great, the peaks are less high and between them, the behavior is less regular.

The rest of the chapter is devoted to the construction of thresholding rules. We consider the model (4.1) and we translate it into the wavelet domain by using the discrete wavelet transform. We assume that the discrete wavelet coefficients of f are provided by the prior model defined in section 4.4.2 and roughly described previously. To estimate each discrete wavelet coefficient, we use the soft thresholding rule. To choose the threshold, we take into account a result proved in chapter 3. In this chapter, a minimax thresholding rule is exhibited. At large resolution levels j, the threshold $\tilde{\lambda}_j$ is proportional to $\sqrt{-2\log(\tilde{\alpha}_j^p)}$, where $\tilde{\alpha}_j$ is the dilation parameter that appears in the definition of the least favorable prior $\pi^{r,p,C}$ associated with $\mathcal{W}^*(r,p)(C)$ (see section 4.2.3). This minimax point of view suggests to use :

$$\lambda_j = \sigma\sqrt{-2\log(\alpha_j^p)}.$$

In section 4.4.2, we give more precise justifications for this choice. We propose a method to estimate the parameters appearing in the definition of λ_j. In section 4.4.3, we measure the performances of the resulting procedure, called ParetoThresh, by using the four test signals : 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'. It is compared to the non Bayesian procedures VisuShrink and SureShrink we have described previously. For this, we use the mean-squared error. Under this criterion, Table 4.3 shows that for the estimation of 'Blocks', 'Bumps' and 'Doppler', ParetoThresh substantially improves VisuShrink and SureShrink. But this is not true any more for 'Heavisine'. Taking into account the properties of 'Heavisine', we explain why, to some extent, this result could have been expected. The chapter is organized as follows : Section 4.2 is devoted to weak Besov spaces. After giving their definition, we recall the results obtained in chapter 3 for the problem of estimation over weak Besov balls. In section 4.3, we define a Bayesian model and we investigate the conditions for the resulting functions to belong to weak Besov spaces. Section 4.4 is devoted to the construction of ParetoThresh, the data adaptive procedure we propose. Finally, in section 4.5, we give the proof of Theorem 4.1.

4.2 Estimation over weak Besov spaces

After an overview of wavelet bases (see section 4.2.1), we introduce in section 4.2.2 weak Besov spaces. Finally, in section 4.2.3, we recall some relevant statistical aspects of weak Besov spaces. From now on, we note $X \sim \mathcal{P}(p)$ $(p > 0)$ to mean that X is a Pareto(p) variable, i.e. X has the density $g(t) = \mathbb{1}_{t\geq 1} p t^{-p-1}$.

4.2.1 Wavelet series representation

In this section, we present some relevant aspects for our survey of the wavelet series representation of a function. For a more complete introduction to wavelets, we refer the reader to Meyer (1992), Daubechies (1992) and Härdle, Kerkyacharian, Picard and Tsybakov (1998). An orthonormal wavelet basis of $L_2(\mathbb{R})$ is generated by translations of a scaling function ϕ and dilations/translations of a wavelet ψ : $\psi_{-1k}(t) = \phi(t-k)$, $\psi_{jk}(t) = 2^{\frac{j}{2}}\psi(2^j t - k)$. With this notation, the wavelet decomposition of a function $f \in L_2(\mathbb{R})$ is

$$f(t) = \sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(t),$$

where the wavelet coefficient β_{jk} is the scalar product of f with ψ_{jk} :

$$\beta_{jk} = \int \psi_{jk}(t) f(t) dt.$$

Actually, for all $m \in \mathbb{N}$, we can build the functions ϕ and ψ to be of 'regularity m' : ϕ and ψ are of class C^m, each of them and their derivatives up to order m have fast decay. Besides, Daubechies (1992) showed that it is possible in addition to require ϕ and ψ to be compactly supported. Recently, the use of wavelets has become very widespread because they provide unconditional bases to various spaces. For instance, wavelet bases are unconditional bases for the class of strong Besov spaces $\mathcal{B}_{s,p,q}$ $(1 \leq p, q \leq \infty,\ 0 < s < \infty)$ (see Meyer (1992)). For a good presentation of strong Besov spaces that model very different forms of spatial inhomogeneity, we refer the reader to Peetre (1976) and DeVore and Lorentz (1993). We just recall that the strong Besov norm of the function f is related to a sequence space norm on the wavelet coefficients of f : Let us assume that the functions ϕ and ψ are of regularity m. If we define

$$\|\beta\|_{b(s,p,q)} = \begin{cases} \left[\sum_{j=-1}^{+\infty} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \|\beta_{j.}\|_p^q \right]^{\frac{1}{q}} & \text{if} \quad 1 \leq q < +\infty \\ \sup_{j \geq -1} 2^{j(s+\frac{1}{2}-\frac{1}{p})} \|\beta_{j.}\|_p & \text{otherwise,} \end{cases}$$

and if $\max(0, \frac{1}{p} - \frac{1}{2}) < s < m$, we have :

$$C_1 \|f\|_{\mathcal{B}_{s,p,q}} \leq \|\beta\|_{b(s,p,q)} \leq C_2 \|f\|_{\mathcal{B}_{s,p,q}},$$

where C_1 and C_2 are constants not depending on f.
Often, for practical reasons, the functions considered in the literature are only defined on a compact set, the interval $[0,1]$ for instance. Cohen, Daubechies and Vial (1993) have described the necessary corrections to adapt wavelets to a bounded interval. As for us, in sections 4.3 and 4.4, we focus on periodic functions f with unit period, and we work with periodic wavelets, we still note ψ_{jk}. This modification, described by Daubechies (1992), implies that the wavelet coefficients are restricted to the indices $\{j \geq -1, k \in \mathcal{I}_j\}$, where

$$\mathcal{I}_j = \{k \in \mathbb{N} : \quad 0 \leq k < 2^j\}. \tag{4.2}$$

4.2.2 Sparsity and definition of weak Besov spaces

Abramovich, Benjamini, Donoho and Johnstone (2000a) introduced the notion of sparsity of an infinite vector $\theta \in \mathbb{R}^{\mathbb{N}}$ through the following approach : the vector θ is said to be sparse if there is a small proportion of relatively large entries. Therefore, they order the components of θ according to their size :

$$|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(n)} \geq \ldots$$

and they control the number of large entries by using a power-law bound on this rearrangement :

$$\sup_n \; n^{\frac{1}{p}} |\theta|_{(n)} < \infty,$$

where $p > 0$. This last condition is equivalent to say that θ belongs to the weak l_p space wlp defined by :

$$wl_p = \left\{ \theta \in \mathbb{R}^{\mathbb{N}} : \quad \sup_{\lambda > 0} \lambda^p \sum_n \mathbb{1}_{\; |\theta_n| > \lambda} < \infty \right\}.$$

As pointed out by DeVore (1989), when $p < 2$, the weak l_p space can be viewed as the collection of all functions on $[0,1]$ that can be approximated in $L^2([0,1])$ at rate N^{-m}, $m = \frac{1}{p} - \frac{1}{2}$.
Now, we define weak Besov spaces as a generalization of weak l_p spaces. Let us consider the following function f expanded in a wavelet series,

$$f(t) = \sum_{j=-1}^{\infty} \sum_k \beta_{jk} \psi_{jk}(t).$$

We define weak Besov spaces as follows :

Definition 4.1. *For all $j \geq -1$ and $\lambda > 0$, we consider $N_f(j,\lambda)$ the number of the wavelet coefficients of f at level j greater than λ :*

$$N_f(j,\lambda) = \sum_k \mathbb{1}_{\; |\beta_{jk}| > \lambda}.$$

If $0 < p, r < \infty$, we say that the function f (or equivalently $\beta = (\beta_{jk})_{j \geq -1, k \in \mathbb{Z}}$) belongs to the weak Besov space $\mathcal{W}^*(r,p)$ if

$$\sup_{\lambda > 0} \lambda^p \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} N_f(j,\lambda) < \infty.$$

To each weak Besov space $\mathcal{W}^*(r,p)$, we associate the balls :

$$\mathcal{W}^*(r,p)(C) = \left\{ f : \quad \sup_{\lambda > 0} \lambda^p \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} \; N_f(j,\lambda) \leq C^p \right\}.$$

The weak Besov space $\mathcal{W}^*(r,p)$ can be viewed as a weighted weak l_p space. The weights penalize the counting of the β_{jk}'s greater than λ for the large scales according to the sign of $r-2$. Therefore, the use of weak Besov spaces may appear as a good device to measure the sparsity of a wavelet expanded signal. Using Markov's inequality, it is easy to prove that for $p<r$, the strong Besov space $\mathcal{B}_{\frac{r}{2p}-\frac{1}{2},p,p}$ is included into $\mathcal{W}^*(r,p)$.
Finally, we recall that Cohen, DeVore, Kerkyacharian and Picard (2001) introduced weak Besov spaces to characterize maxisets for the wavelet thresholding procedure. Among others, they proved the following result :

Proposition 4.1. *Let $1<r<\infty$ and $\alpha\in(0,1)$. We suppose that $f\in L_r([0,1])$. Under the white noise model*

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t\in[0,1],$$

we consider the following thresholding estimator

$$\hat{f}_\varepsilon^T = \sum_{j=-1}^{j_\varepsilon}\sum_k \hat{\beta}_{jk}\mathbb{1}_{|\hat{\beta}_{jk}|>\kappa t_\varepsilon}\psi_{jk},$$

with

- $\hat{\beta}_{jk} = \int \psi_{jk}(t)dY_t$,
- $t_\varepsilon = \varepsilon\sqrt{\log(\varepsilon^{-1})}$
- $2^{-j_\varepsilon} \le \varepsilon^2\log(\varepsilon^{-1}) < 2^{-j_\varepsilon+1}$
- *κ is a constant large enough.*

We have

$$\mathbb{E}\|\hat{f}_\varepsilon^T - f\|_r^r \le K\left(\varepsilon\sqrt{\log(\varepsilon^{-1})}\right)^{\alpha r} \iff f\in\mathcal{B}_{\frac{\alpha}{2},r,\infty}\cap\mathcal{W}^*(r,(1-\alpha)r).$$

4.2.3 Minimax risk and least favorable priors

In chapter 3, we evaluated the minimax risk over weak Besov balls $\mathcal{W}^*(r,p)(C)$ for $\mathcal{B}_{s',p',p'}$ norms by using a Bayesian approach. This enables us to exhibit least favorable priors (noted LFP) which will inspire the prior model chosen in section 4.3.1. Let us recall here the main results we obtain : We restrict our attention to functions f supported by the interval $[0,1]$. They can be written :

$$f(t) = \sum_{j=-1}^{\infty}\sum_{k\in\tilde{\mathcal{I}}_j}\beta_{jk}\psi_{jk}(t),$$

where $\tilde{\mathcal{I}}_j = \{k \in \mathbb{Z} : \beta_{jk} \neq 0\}$. Let us note that with compactly supported wavelets, we have $|\tilde{\mathcal{I}}_j| < \infty$. We introduce two zones we shall denote hereafter respectively as the regular zone and the critical zone :

$$\begin{aligned}\mathcal{R} = &\quad \left\{p' > p,\ \frac{r}{2} > p'\left(s' + \frac{1}{2}\right)\right\} \cup \{p' \leq p\},\\ \mathcal{C} = &\quad \left\{p' > p,\ \frac{r}{2} = p'\left(s' + \frac{1}{2}\right)\right\}.\end{aligned}$$

We consider the white noise model

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t \in [0,1],$$

which means that $\varepsilon > 0$ is known, $f \in L_2([0,1])$ is unknown and for all $\phi \in L_2([0,1])$, $\int_{[0,1]} \phi(t)dY_t = \int_{[0,1]} \phi(t)f(t)dt + \varepsilon \int_{[0,1]} \phi(t)dW_t$ is observable. Among non parametric situations, this statistical model is one of the simplest, at least technically. What is more, it arises as an appropriate large sample limit for more general non parametric models, such as the regression model (4.1) considered previously. (cf. Brown and Low (1996)). These are the reasons why this model is often considered in non parametric situations. We study the asymptotic behavior of the minimax risk

$$R_\varepsilon = \inf_{\hat{f}_\varepsilon} \sup_{f \in \mathcal{W}^*(r,p)(C)} \mathbb{E}_f \|\hat{f}_\varepsilon - f\|^{p'}_{\mathcal{B}_{s',p',p'}},$$

when ε tends to 0. Taking scalar product with ψ_{jk}, the white noise model is translated into the sequence space. We obtain the following sequence of independent variables :

$$\tilde{y}_{jk} = \beta_{jk} + \varepsilon \tilde{z}_{jk}, \quad j \geq -1, k \in \tilde{\mathcal{I}}_j,$$

where $\tilde{z}_{jk} \sim \mathcal{N}(0,1)$. The risk becomes

$$R_\varepsilon = \inf_{\hat{\beta}} \sup_{\beta \in \mathcal{W}^*(r,p)(C)} \mathbb{E}_\beta \|\hat{\beta} - \beta\|^{p'}_{b(s',p',p')}.$$

Under suitable conditions, Theorem 3.1 of chapter 3 proves that on $\mathcal{R}$, the rate of convergence of R_ε is $\varepsilon^{p'\alpha}$, where $\alpha = (s - s')/(s + \frac{1}{2})$ and $s = \frac{r}{2p} - \frac{1}{2}$. On $\mathcal{C}$, the rate of convergence is $\varepsilon^{p'\alpha} \log(\varepsilon^{-1})^{\frac{p'\alpha}{2}}$.

To identify LFP, we first reduce to $M_{r,p}(C)$, the natural set of probability measures associated to $\mathcal{W}^*(r,p)(C)$ and defined as follows :

$$M_{r,p}(C) = \left\{\pi(d\beta) : \quad \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} \mathbb{E}_\pi \sum_{k \in \tilde{\mathcal{I}}_j} \mathbb{1}_{\ |\beta_{jk}| > \lambda} \leq \left(\frac{C}{\lambda}\right)^p, \quad \forall \lambda > 0\right\}. \tag{4.5}$$

For the definition of the LFP, we exhibited two sequences of real numbers $(\tilde{\alpha}_j)_{j\geq -1}$ and $(\tilde{\mu}_j)_{j\geq -1}$ and an integer $\tilde{j}_*$ depending on the zone. We do not recall their exact definitions, which would just add useless technical aspects here, but we briefly recall their properties :

- $(\tilde{\alpha}_j)_{j\geq -1}$ is a sequence of non negative real numbers verifying the condition

$$\sum_{j=-1}^{+\infty} 2^{j\frac{r}{2}}\tilde{\alpha}_j^p = \left(\frac{C}{\varepsilon}\right)^p. \tag{4.6}$$

- $(\tilde{\mu}_j)_{j\geq -1}$ is a sequence of positive real numbers such that $\tilde{\mu}_j \overset{j\to\infty}{\sim} \sqrt{-2\log(\tilde{\alpha}_j^p)}$.
- $\tilde{j}_* \overset{\varepsilon\to 0}{\sim} j(\varepsilon)$, where $j(\varepsilon)$ is the first integer j such that $\tilde{\alpha}_j < 1$.

Now, if we set $\pi_\varepsilon^{r,p,C}$ as the distribution of a sequence of independent variables $(\beta_{jk})_{j\geq -1, k\in\tilde{\mathcal{I}}_j}$ such that

- The distribution of β_{jk} (denoted $\tilde{F}_j$) is symmetric about 0,
- $|\beta_{jk}| = \begin{cases} \varepsilon\min(\tilde{\alpha}_j X_{jk}, \tilde{\mu}_j), \quad \text{where } X_{jk}\sim\mathcal{P}(p), & \text{if } j\geq \tilde{j}_* \\ \varepsilon\tilde{\alpha}_j & \text{otherwise,}\end{cases}$

then, $\pi_\varepsilon^{r,p,C}$, belonging to $M_{r,p}(C)$, is a LFP for the problem of estimation over $\mathcal{W}^*(r,p)(C)$. Indeed, if for each prior π of $M_{r,p}(C)$, we define its Bayes risk denoted $B(\pi)$ by

$$B(\pi) = \inf_{\hat{\beta}} \mathbb{E}_\pi \mathbb{E}_\beta \|\hat{\beta}-\beta\|^{p'}_{b(s',p',p')},$$

then, up to constants, the supremum of B over $M_{r,p}(C)$ is attained for $\pi_\varepsilon^{r,p,C}$ and the asymptotic values of $B(\pi_\varepsilon^{r,p,C})$ are the same as the asymptotic values of R_ε : there exist three constants K_1, K_2, K_3 only depending on r, p, C, s' and p', such that

$$K_1\ B(\pi_\varepsilon^{r,p,C}) \leq R_\varepsilon \leq K_2\ B(\pi_\varepsilon^{r,p,C}), \tag{4.7}$$

and

$$\sup_{\pi\in M_{r,p}(C)} B(\pi) \leq K_3\, B(\pi_\varepsilon^{r,p,C}). \tag{4.8}$$

What is more, asymptotically, the support of $\pi_\varepsilon^{r,p,C}$ is 'almost' included into $\mathcal{W}^*(r,p)(C)$. It means that there exists $(\gamma_\varepsilon)_{\varepsilon>0}$ larger than 1, tending to 1 when ε tends to 0, such that

$$\pi_\varepsilon^{r,p,C}\left(\beta\in\mathcal{W}^*(r,p)(\gamma_\varepsilon C)\right) \overset{\varepsilon\to 0}{\longrightarrow} 1. \tag{4.9}$$

It is interesting to note that the realizations built from the LFP provide a good representation of the worst functions of $\mathcal{W}^*(r,p)(C)$ to be estimated.
Finally, we have the following result : The thresholding rule defined by

$$\hat{f}_\varepsilon = \sum_{j=-1}^{\infty} \sum_{k \in \tilde{\mathcal{I}}_j} \operatorname{sign}(\tilde{y}_{jk})(|\tilde{y}_{jk}| - \tilde{\lambda}_j)_+ \psi_{jk},$$

with

$$\tilde{\lambda}_j = \begin{cases} \varepsilon\sqrt{-2\log(\tilde{\alpha}_j^p)} & \text{if } j \geq \tilde{j}_*, \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

attains the minimax rate of convergence up to constants. We shall inspire from this minimax rule to build constructive estimators in section 4.4.2.

4.3 Construction of functions typical of weak Besov spaces

From now on, following section 4.2.1, we consider a periodic signal

$$f(t) = \sum_{j=-1}^{\infty} \sum_{k \in \mathcal{I}_j} \beta_{jk} \psi_{jk}(t), \tag{4.11}$$

where $\mathcal{I}_j$ is given in (4.2). As pointed out by Johnstone (1994), this computational simplification affects only a fixed number of wavelet coefficients at each level j. We place a prior model on the wavelet coefficients of f to capture its sparsity. But section 4.2.2 pointed out that the sparsity of a signal can be revealed by the weak Besov space $\mathcal{W}^*(r,p)$ in which the signal lies. So, through Theorem 4.1, we connect these two approaches of sparsity by giving a relationship between the parameters of the prior model and $\mathcal{W}^*(r,p)$. This enables us to build functions typical of $\mathcal{W}^*(r,p)$.

4.3.1 The prior model

To fix a prior model, we exploit the LFP $\pi_\varepsilon^{r,p,C}$ defined in section 4.2.3 and that is naturally connected to the weak Besov ball $\mathcal{W}^*(r,p)(C)$: we suppose that the β_{jk}'s are independent and for $j \geq 0$ and $k \in \mathcal{I}_j$, the distribution of each β_{jk} is $F_j^{\alpha_j,\mu_j,p}$, where

- $F_j^{\alpha_j,\mu_j,p} = \frac{1}{2}(F_j^+ + F_j^-)$,
- F_j^- is the reflection of F_j^+ about 0,
- F_j^+ is the distribution of $\min(\alpha_j X_j - \alpha_j, \mu_j)$, where $X_j \sim \mathcal{P}(p)$,

- α_j and μ_j are positive real numbers.

Because of its improper nature, we place no prior on the scaling coefficient β_{-10}. The distribution $F_j^{\alpha_j,\mu_j,p}$ is a slight modification of $\tilde{F}_j$ that appeared in the definition of $\pi_\varepsilon^{r,p,C}$. Indeed, to avoid any discontinuity in the definition of the support of β_{jk}, we translate the variable $\alpha_j\mathcal{P}(p)$ by α_j. This slight modification enables us to capture very small values of β_{jk}. We suppose that the parameter α_j has the form

$$\alpha_j = C2^{-j\delta}$$

where C and δ are positive constants. Whereas the parameter $\tilde{\mu}_j$ verified the relation $\tilde{\mu}_j \overset{j\to\infty}{\sim} \sqrt{-2\log(\tilde{\alpha}_j^p)}$ in the definition of least favorable priors, here we set

$$\mu_j = \sqrt{\max\left(M^2, -2\log(\alpha_j^p)\right)},$$

where M is a positive constant eventually very large. The value of M can be chosen as the a priori maximal size of the wavelet coefficients of the function we want to estimate. We note that $\forall\, \lambda > 0$,

$$\mathbb{P}(|\beta_{jk}| > \lambda) = \begin{cases} \left(\frac{\alpha_j}{\lambda+\alpha_j}\right)^p & \text{if } \lambda < \mu_j \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, even if the support of β_{jk} is bounded at each level j, we expect that this prior model, coming from a heavy-tailed distribution may capture a great number of large coefficients. Under a good choice of the parameters δ and p, we can obtain very inhomogeneous functions. We investigate in the following section the type of inhomogeneity this prior model enables us to obtain.

4.3.2 The main result and simulations

In this section, we assume that the prior model defined in the previous section is placed on the wavelet coefficients. Through the following theorem, we show that under a good choice of the parameters of the prior model, we can generate functions that are typical of the weak Besov space $\mathcal{W}^*(r,p)$:

Theorem 4.1. *We consider the function f given in (4.11). Let $0 < p < \infty$, $0 < r < \infty$. Given three positive real numbers δ, C and M, we define for all $j \geq 0$, $\alpha_j = C2^{-j\delta}$, and $\mu_j = \sqrt{\max\left(M^2, -2\log(\alpha_j^p)\right)}$. Let us assume that the wavelet coefficients β_{jk} of f are independent* *and for $j \geq 0$ and $k \in \mathcal{I}_j$, β_{jk} has the distribution $F_j^{\alpha_j,\mu_j,p}$ given in section 4.3.1 . For all fixed value of β_{-10},*

$$f \in \mathcal{W}^*(r,p)\, a.s. \quad \Longleftrightarrow \quad \frac{r}{2} < \delta p.$$

The proof of this theorem is given in section 4.5.

Remark 4.1. *The condition $\frac{r}{2} < \delta p$ is equivalent to say that*

$$\sum_{j=-1}^{\infty} 2^{j\frac{r}{2}} \alpha_j^p < \infty.$$

This can be connected to the condition (4.6) verified by the sequences $(\tilde{\alpha}_j)_{j \geq -1}$ used to define the LFP for the weak Besov balls $\mathcal{W}^*(r,p)(C)$.

Theorem 4.1 gives us a help to have a good understanding of weak Besov spaces. Figure 4.1 presents various typical realizations with different values for the parameters δ and p. Since the values of C and M do not play a role in the shape of the realizations we get, we set $C = 0.1$ and $M = 2$ for each realization. We used Daubechies's least asymmetric wavelet of order 8.

Naturally, we note that when p is fixed, the realizations are more regular when δ is great (compare (b) and (f) or (c) and (e) or (a) and (d)). The same conclusion is true when δ is fixed and p is great (compare (a), (b) and (e) or (c) and (d)). In fact, as expected, the realizations are smoother when the product δp is great. It is interesting to wonder what happens when the product δp is fixed, and when we take different values for δ and p. Comparing (d), (e) and (f) or (b) and (c), we notice that when p is small, the realizations show very high peaks with a regular behavior between the peaks. When p is great, the peaks are less high, and between the peaks the behavior is less homogenous. To sum up, we can say that when p decreases, the number of negligible coefficients increases, but the few remaining coefficients may be very large. In chapter 3, we drew the same conclusions by using a different approach that exploits the LFP associated with the weak Besov balls $\mathcal{W}^*(r,p)(C)$. Note that these two approaches complement one another :

- The first one produces typical functions of weak Besov spaces but we do not control the radius of the weak Besov ball that contains a function f provided by our prior model. This radius may be very great, which may pollute our perception of the regularity of the function f.
- The second one controls the radius of the weak Besov balls, but each LFP is typical of a set of probability measures. For instance, in section 4.2.3, we chose $M_{r,p}(C)$ defined by (4.5).

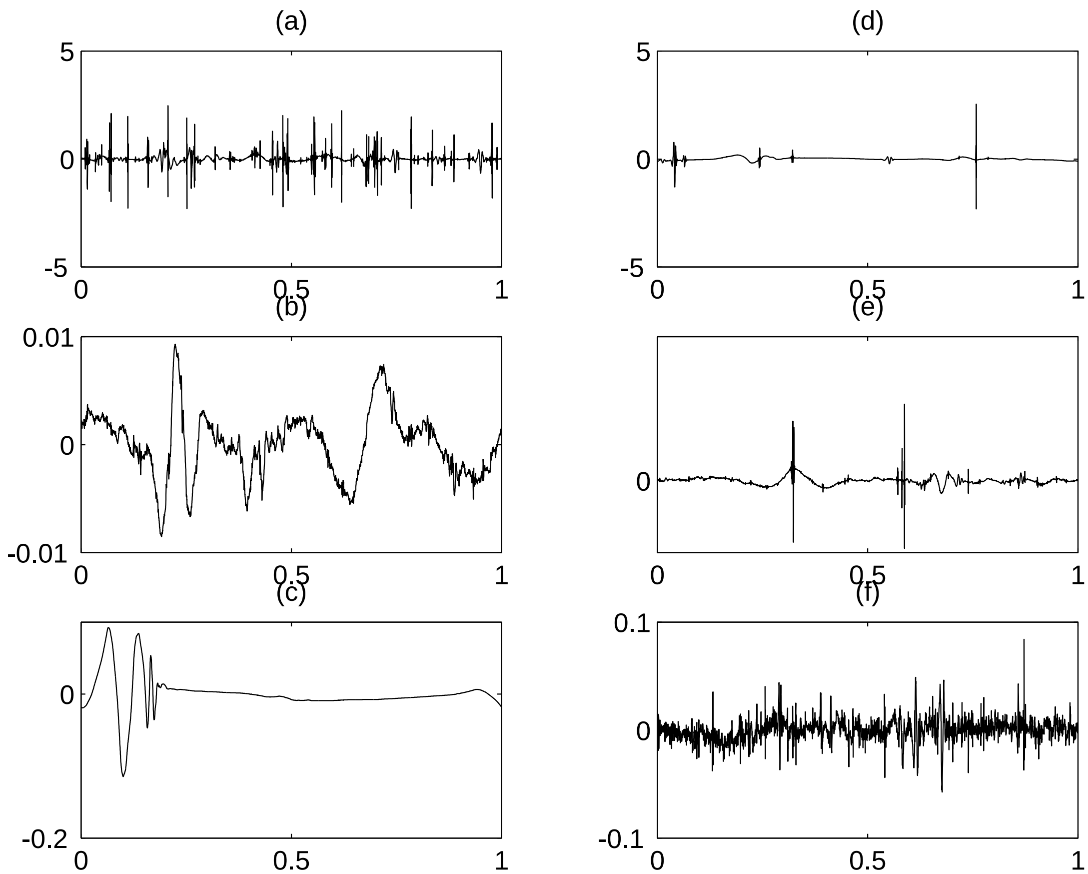

FIG. 4.1 – Realizations with various values of δ and p ; $\beta_{-10} = 0$; n=4096 plotting points ; (a) : $p = 0.5$, $\delta = 1$. (b) : $p = 2$, $\delta = 1$. (c) : $p = 1$, $\delta = 2$. (d) : $p = 0.5$, $\delta = 2$. (e) : $p = 1$, $\delta = 1$. (f) : $p = 2$, $\delta = 0.5$.

To some extent, this choice was judicious since we obtained the properties (4.7), (4.8) and (4.9), but we could have made another choice.

4.4 Thresholding rules

The rest of this chapter is devoted to exhibiting a constructive method to estimate a noisy function f. We shall exploit the results of the previous sections. But before this, let us precise our statistician model.

4.4.1 Model and discrete wavelet transform

Let us consider the standard regression problem :

$$g_i = f(\frac{i}{n}) + \sigma\varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad 1 \leq i \leq n, \tag{4.12}$$

where $n = 2^N, N \in \mathbb{N}$. We introduce the discrete wavelet transform (denoted DWT) of the vector $f^0 = (f(\frac{i}{n}), \quad 1 \leq i \leq n)^T$:

$$d := \mathcal{W} f^0.$$

The DWT matrix $\mathcal{W}$ is orthogonal. Therefore, we can reconstruct f^0 by the relation

$$f^0 = \mathcal{W}^T d.$$

These transformations performed by Mallat's fast algorithm require only $O(n)$ operations (see Mallat (1998)). The DWT provides n discrete wavelet coefficients d_{jk}, $-1 \leq j \leq N-1, k \in \mathcal{I}_j$. They are related to the wavelet coefficients β_{jk} of f by the simple relation

$$d_{jk} \approx \beta_{jk} \times \sqrt{n}. \tag{4.13}$$

Using the DWT, the regression model (4.12) is reduced to the following one :

$$y_{jk} = d_{jk} + \sigma z_{jk}, \quad -1 \leq j \leq N-1, \quad k \in \mathcal{I}_j,$$

where

$$y := (y_{jk})_{j,k} = \mathcal{W} g$$

and

$$z := (z_{jk})_{j,k} = \mathcal{W} \varepsilon.$$

Since $\mathcal{W}$ is orthogonal, z is a vector of independent $\mathcal{N}(0,1)$ variables. Now, instead of estimating f, we estimate the d_{jk}'s.

We suppose in the following that σ is known. Nevertheless, it could robustly be estimated by the median absolute deviation of the $(d_{N-1,k})_{k\in\mathcal{I}_{N-1}}$ divided by 0.6745 (see Donoho and Johnstone (1994)).

4.4.2 Choice of the threshold

In the following, as explained in Introduction, the discrete wavelet coefficients are estimated by using thresholding rules associated to level-dependent thresholds $(\lambda_j)_j$. Following Donoho and Johnstone (1994), many procedures are based on the hard and soft thresholding rules respectively defined by :

$$\eta^{HT}(y,\lambda) = y\mathbb{1}_{|y|>\lambda},$$
$$\eta^{ST}(y,\lambda) = \mathrm{sign}(y)\left(|y|-\lambda\right)_+.$$

This chapter considers soft thresholding although hard thresholding is a possible alternative. However, the soft thresholding rule is smoother ($y \longrightarrow \eta^{ST}(y,\lambda)$ is continuous) and it makes it possible to build minimax estimators over weak Besov spaces, as recalled in section 4.2.3. The choice of the threshold is then crucial. If λ_j is too small (respectively too large) then the estimator tends to overfit (respectively underfit) the data. Let us describe two non Bayesian procedures that have minimax properties : Donoho and Johnstone (1994) proposed their VisuShrink procedure with $\lambda_j = \lambda^u := \sigma\sqrt{2\log(n)}$. If λ^u often underfits the data, it 'guarantees' a noise-free reconstruction since

$$\mathbb{P}(\max_{j,k}|z_{jk}| > \lambda^u) \overset{N\to+\infty}{\longrightarrow} 0.$$

Unlike VisuShrink that seems too universal, SureShrink provides level-dependent thresholds. They are obtained by minimizing Stein's unbiased estimate of risk for threshold estimates, provided we have the following sparsity condition :

$$2^{-j}\sum_{k\in\mathcal{I}_j}(y_{jk}^2-1) > j^{\frac{3}{2}}2^{-\frac{j}{2}}.$$

Otherwise, we choose the threshold $\lambda_j = \sigma\sqrt{2\log(2^j)}$.

Under a Bayes model, a natural approach to build estimators could be to use the mean of the posterior distribution which is the Bayes rule under the squared error loss. But the posterior mean does not involve in general thresholding rules. That is the reason why Abramovich, Sapatinas and Silverman (1998) focus on the posterior median within the following framework : They consider a prior model having the following form :

$$d_{jk} \sim \gamma_j\mathcal{N}(0,\tau_j^2) + (1-\gamma_j)\delta_0.$$

The hyperparameters τ_j^2 and γ_j are chosen to ensure that the underlying function f belongs to a given strong Besov space $\mathcal{B}_{s,p,q}$. Then, $\hat{d}_{jk}$, the estimator of d_{jk} obtained by using the median of the posterior distribution, is zero if y_{jk} falls into an interval of the form $[-\lambda'_j; \lambda'_j]$. Vidakovic (1998) imposes a symmetric prior on d_{jk} and the marginal model for y_{jk} conditioned to d_{jk} is the double exponential with the density given by

$$f(y_{jk}|d_{jk}) = \frac{1}{2}(2\mu)^{\frac{1}{2}} \exp\left(-(2\mu)^{\frac{1}{2}}|y_{jk} - d_{jk}|\right).$$

He constructs a procedure that mimics the hard thresholding rule. He estimates d_{jk} by $y_{jk}\mathbb{1}_{\eta_{jk}<1}$ where $\eta_{jk} = \mathbb{P}(d_{jk} = 0 \mid y_{jk})/\mathbb{P}(d_{jk} \neq 0 \mid y_{jk})$.

As for us, we place the following prior on the discrete wavelet coefficients : We suppose that the d_{jk}'s are independent and for all $j \geq 0$, $k \in \mathcal{I}_j$, $\sigma^{-1}d_{jk} \sim F_j^{\alpha_j,\mu_j,p}$, where $F_j^{\alpha_j,\mu_j,p}$ is given in section 4.3.1. The parameters α_j and μ_j are given by $\alpha_j = C2^{-j\delta}$ and $\mu_j = \sqrt{\max\left(M^2, -2\log(\alpha_j^p)\right)}$, where δ, C and M are positive constants. To estimate d_{jk}, we propose

$$d^*_{jk} = \eta^{ST}(y_{jk}, \lambda_j),$$

where λ_j has the following form

$$\lambda_j = \begin{cases} \sigma\sqrt{-2\log(\alpha_j^p)} & \text{if } j \geq j_*, \\ 0 & \text{otherwise,} \end{cases} \tag{4.14}$$

as suggested by (4.10). Indeed, since our prior model is very close to the LFP over weak Besov balls $\mathcal{W}^*(r,p)(C)$, using (4.14) to define λ_j where j_* is the first integer j such that $\alpha_j < 1$ seems judicious. Then, the threshold λ_j can be rewritten as follows :

$$\lambda_j = \sigma\sqrt{\max\left(0, -2\log(\alpha_j^p)\right)}. \tag{4.15}$$

To apply this procedure, it is necessary to specify the values of C and δ that define α_j and the value of p. If we know the weak Besov space in which the function to be estimated lies and if an efficient method is provided to estimate δ or p, by using Theorem 4.1, it is easy to estimate the other parameter. However, we shall ignore this strategy and the choice of the value of p will be made in section 4.4.3. For the estimation of (C, δ), we set

$$\hat{N}_j(\lambda^u) = \frac{1}{2^j}\sum_{k\in\mathcal{I}_j} \mathbb{1}_{|y_{jk}|>\lambda^u},$$

where λ^u is the universal threshold defined by $\lambda^u = \sigma\sqrt{2\log(n)}$, and we set

$$\hat{\alpha}_j = \sigma^{-1}\lambda^u\hat{N}_j(\lambda^u)^{\frac{1}{p}}(1 - \hat{N}_j(\lambda^u)^{\frac{1}{p}})^{-1}.$$

We estimate C and δ by using the linear regression :

$$(\hat{C}, \hat{\delta}) = \arg\min_{C,\delta} \sum_{j\in\mathcal{S}} (\log(\hat{\alpha}_j) - \log(C) + j\delta\log(2))^2, \tag{4.16}$$

where

$$\mathcal{S} = \{j \in \{1, \dots, N-1\} : \quad \hat{\alpha}_j \in (0, +\infty)\} .$$

But $(\hat{C}, \hat{\delta})$ are well defined only if $\mathrm{card}(\mathcal{S}) \geq 2$. So, when $\mathrm{card}(\mathcal{S}) \geq 2$, we set

$$\hat{\lambda}_j = \sigma\sqrt{\max\left(0, -2p\log(\hat{C}2^{-j\hat{\delta}})\right)}. \tag{4.17}$$

If $\mathrm{card}(\mathcal{S}) \leq 1$, we set

$$\hat{\lambda}_j = \begin{cases} 0 & \text{if } \hat{\alpha}_j = +\infty, \\ \sigma\sqrt{\max(0, -2p\log(\hat{\alpha}_j))} & \text{for } j \in \mathcal{S}, \\ \lambda^u & \text{if } \hat{\alpha}_j = 0. \end{cases} \tag{4.18}$$

Before going further, let us give a precise justification for this procedure : We notice that for all $\lambda < \sigma\mu_j$,

$$\mathbb{P}(|d_{jk}| > \lambda) = \left(\frac{\alpha_j}{\alpha_j + \sigma^{-1}\lambda}\right)^p .$$

But, using extended Glivenko-Cantelli's Theorem,

$$\sup_{\lambda>0} \left| \frac{1}{2^j} \sum_{k\in\mathcal{I}_j} \mathbb{1}_{|d_{jk}|>\lambda} - \mathbb{P}(|d_{jk}| > \lambda) \right| \stackrel{j\to\infty}{\longrightarrow} 0 \text{ a.s.}$$

Therefore, for all $\lambda > 0$, $\left(\frac{\alpha_j}{\alpha_j+\sigma^{-1}\lambda}\right)^p$ is well approximated by

$$N_j(\lambda) = \frac{1}{2^j} \sum_{k\in\mathcal{I}_j} \mathbb{1}_{|d_{jk}|>\lambda}.$$

We choose $\lambda = \lambda^u$, and we estimate $N_j(\lambda^u)$ by $\hat{N}_j(\lambda^u)$. Using the four test functions ('Blocks', 'Bumps', 'Heavisine', and 'Doppler'), Table 4.1 compares for $j \in \{0, \dots, 9\}$ the values of $N_j(\lambda^u)$ and the average over 100 replications of the values of $\hat{N}_j(\lambda^u)$. It shows that our approximation is acceptable.

So,

$$\left(\frac{\alpha_j}{\alpha_j + \sigma^{-1}\lambda^u}\right)^p \approx \hat{N}_j(\lambda^u)$$

and

$$\alpha_j = C2^{-j\delta} \approx \hat{\alpha}_j = \sigma^{-1}\lambda^u\hat{N}_j(\lambda^u)^{\frac{1}{p}}(1 - \hat{N}_j(\lambda^u)^{\frac{1}{p}})^{-1}.$$

Level	Blocks		Bumps		Heavisine		Doppler	
j	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$
j=0	1	1	1	0.76	1	1	1	0.99
j=1	2	2	1	1	2	2	2	2
j=2	4	4	3	3	4	4	3	3
j=3	6	6.02	4	4.67	6	5.44	7	6.62
j=4	9	9.50	10	9.71	0	0.29	6	6.06
j=5	8	9.07	10	11.03	0	0.06	7	6.61
j=6	6	6.45	16	16.16	0	0.11	7	7.15
j=7	5	4.87	16	14.74	0	0.02	5	4.17
j=8	4	3.34	7	7.80	0	0.05	0	0.39
j=9	0	0.60	2	2.79	0	0.13	0	0.15

TAB. 4.1 – Comparison of the values of $N_j = N_j(\lambda^u)$ and $\hat{N}_j = \hat{N}_j(\lambda^u)$ for 'Blocks', 'Bumps', 'Heavisine' and 'Doppler' ; n=1024 ; rsnr=3 ($\sigma = 7/3$).

This provides a justification for (4.16). The pair of equations (4.17) and (4.18) are naturally justified by (4.15).
Now, we set

$$\hat{d}_{jk} = \eta^{ST}(y_{jk}, \hat{\lambda}_j),$$

for all $j \geq 0$, $k \in \mathcal{I}_j$, and $\hat{d}_{-10} = y_{-10}$. Finally, to estimate the signal, we use the reconstruction formula and an estimator of f^0 is provided by :

$$\hat{f}^0 = \mathcal{W}^T \hat{d}.$$

The performances of this Bayesian thresholding procedure denoted from now on as ParetoThresh, are analyzed in the next section. With $p = 1.3$, Table 4.2 gives the average over 100 replications of the values of the level-dependent threshold $\hat{\lambda}_j$ associated with the four test functions.
Donoho and Johnstone (1994) noted that for the coarsest levels the coefficients should not be shrunk to 0. Huang and Cressie (2000) and Abramovich and Benjamini (1995) showed that the choice of these levels is essential for VisuShrink and SureShrink. Let us note that ParetoThresh automatically provides the levels where the coefficients are not shrunk, and this, with a data adaptive method.

Level j	Blocks	Bumps	Heavisine	Doppler
j=0	0	0	0	0
j=1	0	0	0	0
j=2	0	0	0	0
j=3	0	0	0	0
j=4	0	0	6.91	0
j=5	0	0	8.27	0
j=6	1.52	0.10	9.10	2.16
j=7	4.07	2.68	9.75	4.20
j=8	5.53	3.95	10.30	5.51
j=9	6.68	4.89	10.80	6.57

TAB. 4.2 – Values of $\hat{\lambda}_j$ ($p = 1.3$) associated with 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'; n=1024; rsnr=3 ($\sigma = 7/3$); $\lambda^u = 8.69$.

4.4.3 Examples and discussion

In this section, we apply our ParetoThresh procedure to one-dimensional signal processing. We use the four test functions : 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'. These functions have been chosen by Donoho and Johnstone (1994) to represent a large variety of inhomogeneous signals. More precisely, our procedure deals with the 1024 equally spaced values on $[0, 1]$ of these signals. In the subsequent applications of ParetoThresh, we take $p = 1.3$ for every function, which provides quite good results. However, we shall discuss below the effect of varying p. We compare our procedure to VisuShrink and SureShrink, described in section 4.4.2 for which we do not threshold the five coarsest levels. Daubechies's least asymmetric wavelet of order 8 is used for all the methods. The performance of each procedure is measured by using the mean-squared error associated to an estimator $\hat{f}$:

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(\frac{i}{n}) - f(\frac{i}{n})\right)^2.$$

Table 4.3 shows the average mean-squared error (denoted AMSE) using 100 replications for VisuShrink, SureShrink and ParetoThresh with different values for the root signal to noise ratio (RSNR).

Figures 4.2, 4.3, 4.4 and 4.5 show the reconstructions we obtain for these three methods when the RSNR is equal to 3.

RSNR	Signal	VisuShrink	SureShrink	ParetoThresh ($p = 1.3$)
RSNR=3	Blocks	3.3143	1.7850	*1.4559*
	Bumps	5.6100	2.0378	*1.8025*
	Heavisine	0.3136	*0.3042*	0.3059
	Doppler	2.1588	1.0911	*0.9221*
RSNR=5	Blocks	1.8624	0.7645	*0.6928*
	Bumps	2.7345	0.8523	*0.8128*
	Heavisine	0.1946	*0.1816*	0.1851
	Doppler	1.0358	0.4378	*0.4310*
RSNR=8	Blocks	0.9745	0.3449	*0.3248*
	Bumps	1.3139	*0.3032*	0.3714
	Heavisine	0.1312	0.1028	*0.0818*
	Doppler	0.5374	0.2434	*0.2140*

TAB. 4.3 – AMSEs for VisuShrink, SureShrink and ParetoThresh ($p = 1.3$) with various test functions and various values of the RSNR.

Table 4.3 shows that ParetoThresh has generally smaller mean-squared error over the three test functions 'Blocks', 'Bumps', and 'Doppler', than SureShrink second and VisuShrink third in the rankings. For 'Heavisine', we recall that the estimation of the level-dependent thresholds for ParetoThresh is based upon a small number of data passing the threshold λ^u (see Table 4.1). Furthermore, to some extent, by its regularity properties, 'Heavisine' can not be viewed as belonging to the class of 'the worst functions to be estimated'. Consequently, the prior model we adopt does not seem to be a good model for this signal. If ParetoThresh behaves well under the AMSE approach, we note that high-frequency artefacts appear, whereas VisuShrink provides the best method for removing the noise. SureShrink lies in between. But these artefacts may partially disappear if we take small values of p as illustrated by Figure 4.6. This effect may be expected taking into account the conclusions we have drawn from the realizations of section 4.3.2. We remark that this improvement has a cost : the AMSE increases. When p is greater than 1.3, the AMSEs are worse and the artefacts are more numerous. Finally, let us mention that a possible alternative is to use the hard thresholding rule with $(\hat{\lambda}_j)_j$. However, the resulting constructions are less regular. Another alternative is to use a Bayes rule. Let us note that it is

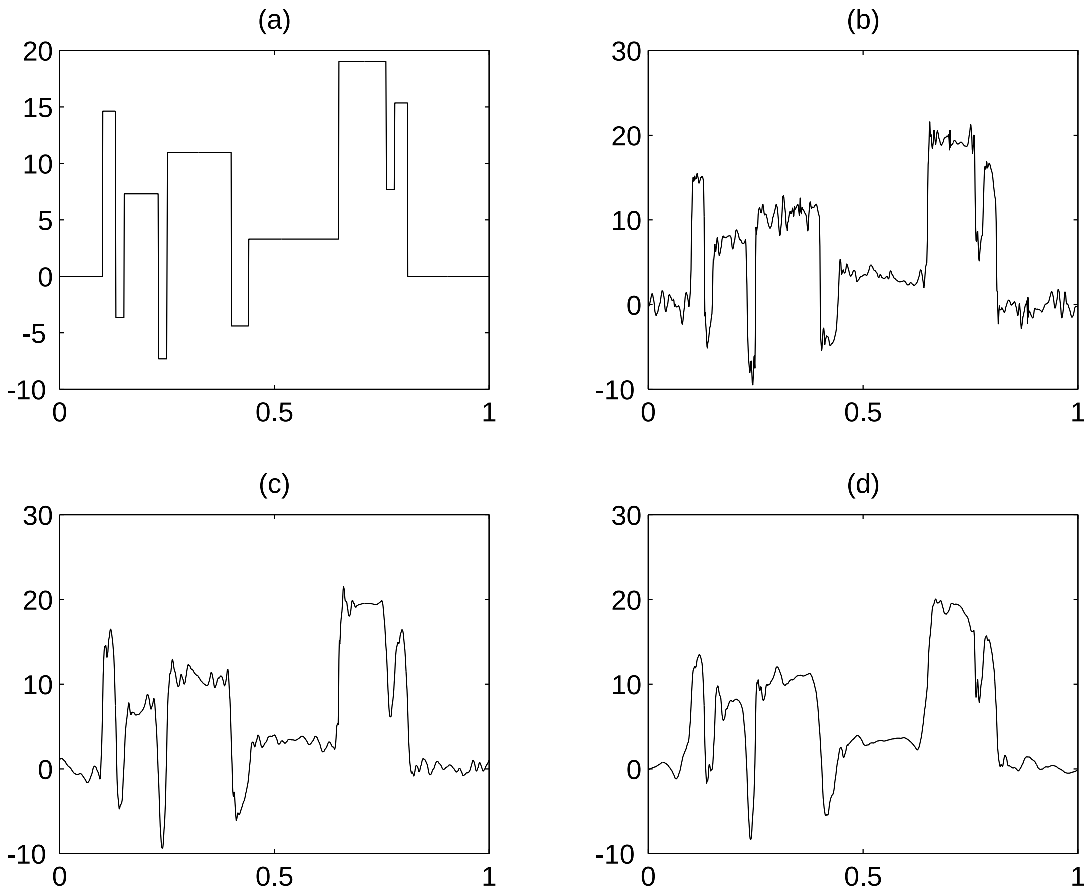

FIG. 4.2 – Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink ; (a) : 'Blocks' (b) : ParetoThresh ($p = 1.3$) (c) : SureShrink (d) : VisuShrink

easy to implement the hard and soft thresholding rules. This is not necessarily the case for a Bayes rule since it results from the minimization of the Bayes risk, even if under a good choice of the loss function we can exhibit the explicit form of the Bayes rule (for instance, the posterior mean, or the posterior median). We can add that neither the use of the hard thresholding rule with $(\hat{\lambda}_j)_j$, nor the use of a Bayes rule (the posterior mean or the posterior median) provides better results as far as the AMSE is concerned.

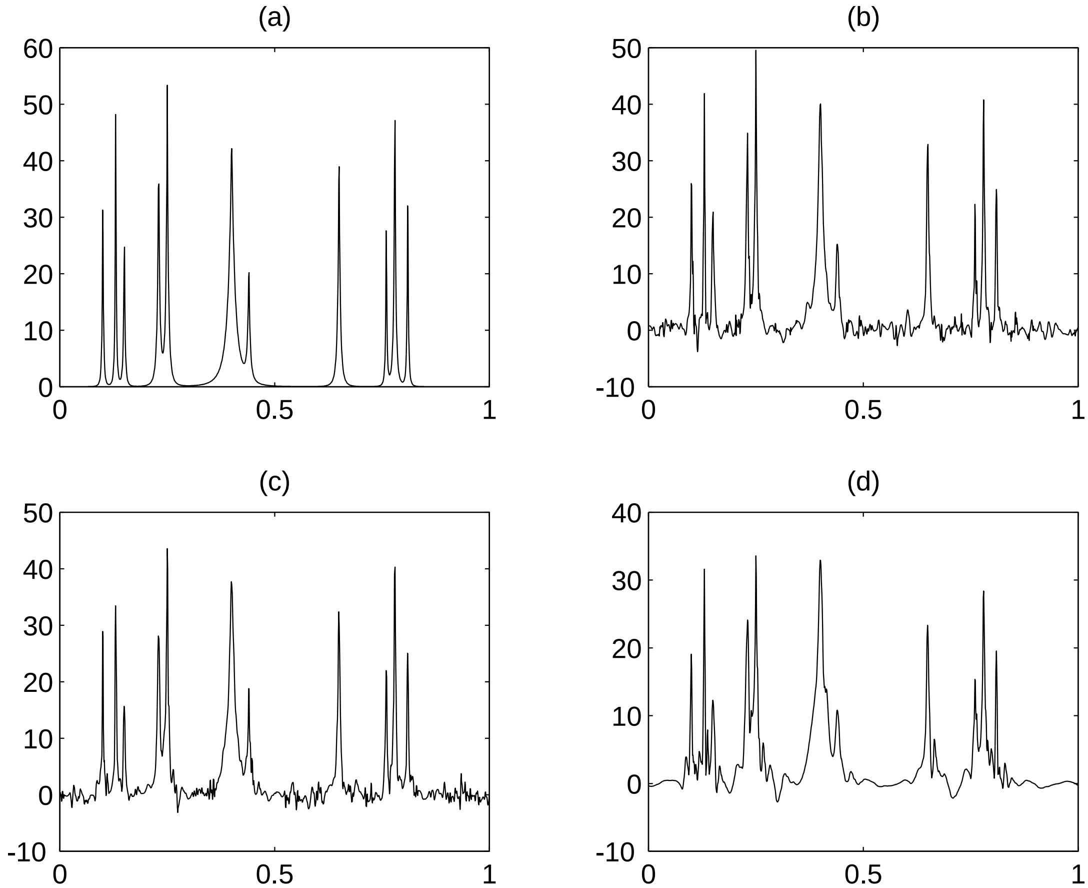

FIG. 4.3 – Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink ; (a) : 'Bumps' (b) : ParetoThresh ($p = 1.3$) (c) : SureShrink (d) : VisuShrink

4.5 Appendix : Proof of Theorem 4.1

Proof of necessity : Let us assume that $f \in \mathcal{W}^*(r,p)$a.s. For all $\lambda \in]0, \mu_0[$, we consider,

$$\forall\, n \in \mathbb{N}, \quad U_n(\lambda) = \frac{1}{\sqrt{c_n}} \left| \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{\,|\beta_{jk}|>\lambda} - \mathbb{P}(|\beta_{jk}| > \lambda) \right) \right|,$$

where

$$c_n = \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(r-2)}.$$

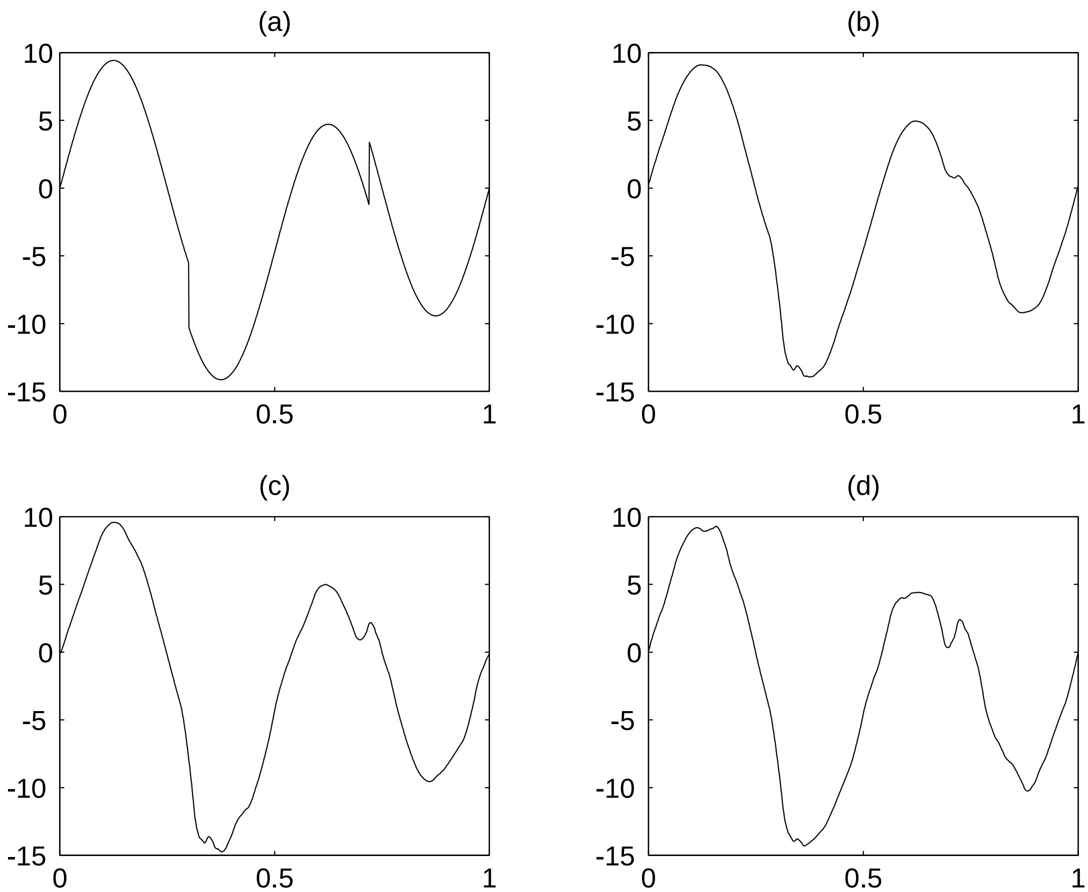

FIG. 4.4 – Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink ; (a) : 'Heavisine' (b) : ParetoThresh ($p = 1.3$) (c) : SureShrink (d) : VisuShrink

Since $\lambda < \mu_0$,

$$U_n(\lambda) = \frac{1}{\sqrt{c_n}} \left| \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{X^*_{jk} \leq \lambda} - F^*_j(\lambda) \right) \right|,$$

where $X^*_{jk} = \alpha_j X_{jk} - \alpha_j$, F^*_j is its continuous distribution function and $X_{jk} \sim \mathcal{P}(p)$. As in Shorack and Wellner (1986) (p 117), we set

$$\overline{F}_n = \frac{1}{c_n} \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(r-2)} F^*_j,$$

$$T_{jk} = \overline{F}_n(X^*_{jk}),$$

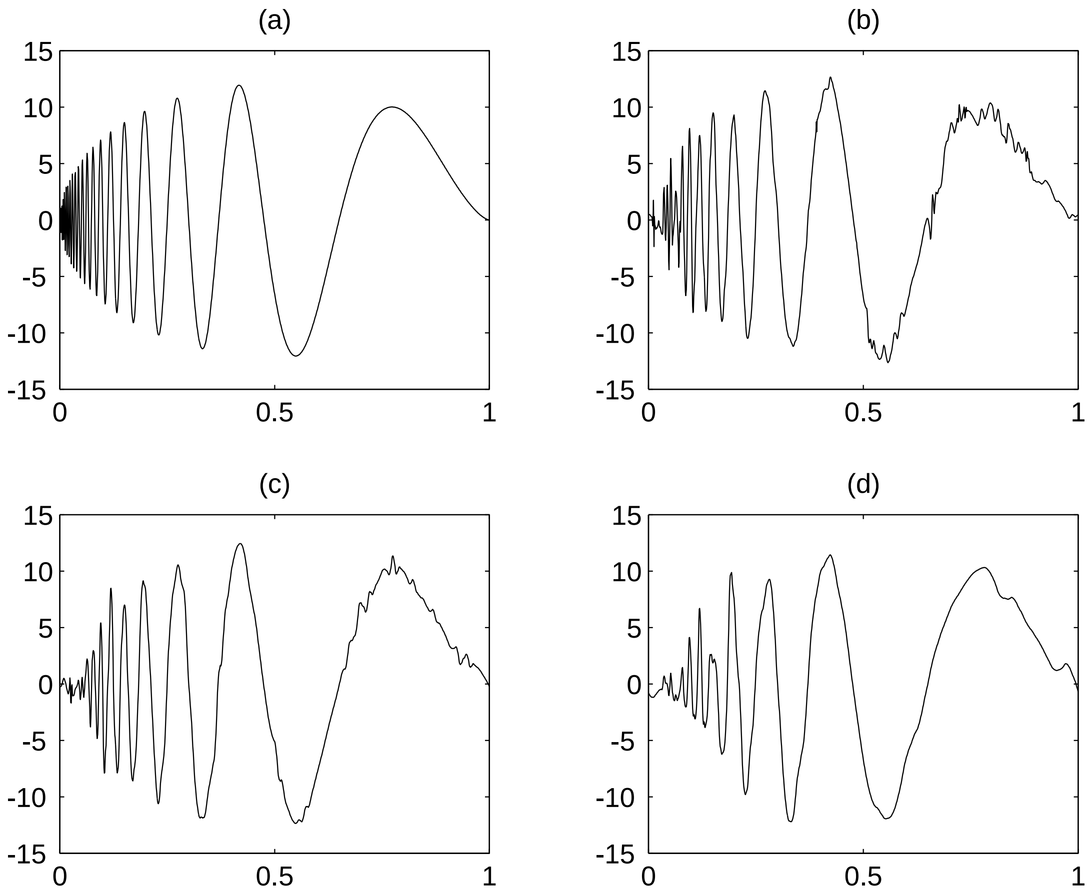

FIG. 4.5 – Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink ; (a) : 'Doppler' (b) : ParetoThresh ($p = 1.3$) (c) : SureShrink (d) : VisuShrink

and we consider the weighted empirical process of the T_{jk}'s,

$$Z_n(\lambda) = \frac{1}{\sqrt{c_n}} \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{\ T_{jk} \leq \lambda} - G_j(\lambda) \right),$$

where $G_j = F_j^* \circ \overline{F}_n^{-1}$ is the distribution function of T_{jk}. So $U_n(\lambda)$ can be written

$$U_n(\lambda) = \left| Z_n(\overline{F}_n(\lambda)) \right|.$$

Now, we suppose that

$$r - 1 - 2\delta p < 0, \tag{4.19}$$

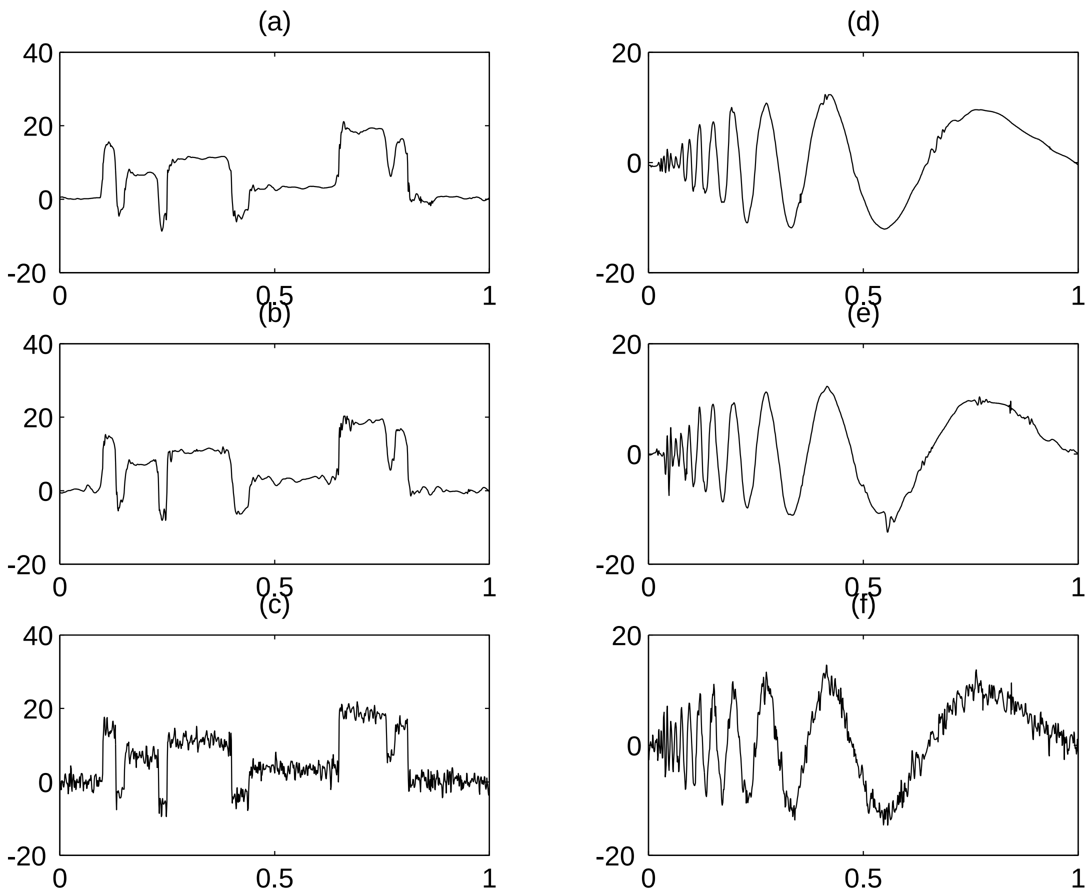

FIG. 4.6 – Various reconstructions of 'Blocks' and 'Doppler' using ParetoThresh with different values of p ; RSNR=3 ; (a) : $p = 0.5$, AMSE=1.7940. (b) : $p = 1$, AMSE=1.5776. (c) : $p = 2$, AMSE=1.5806. (d) : $p = 0.5$, AMSE=1.0644. (e) : $p = 1$, AMSE=0.9502. (f) : $p = 2$, AMSE=1.0865.

and we consider δ' a real number smaller than δ such that $r-1-2\delta' p<0$. We consider the sequence $(\lambda_n)_{n\in\mathbb{N}}$ defined by

$$\forall\, n\in\mathbb{N}, \quad \lambda_n=\min(C\,2^{-\delta' n},\frac{\mu_0}{2}).$$

We have $\forall\, n\in\mathbb{N}$,

$$\begin{aligned}\mathbb{P}\left(\lambda_n^p\sqrt{c_n}U_n(\lambda_n)\geq 1\right) &= \mathbb{P}\left(\left|Z_n(\overline{F}_n(\lambda_n))\right|\geq \lambda_n^{-p}c_n^{-\frac{1}{2}}\right)\\ &\leq \mathbb{P}\left(w_n(1)\geq \lambda_n^{-p}c_n^{-\frac{1}{2}}\right),\end{aligned}$$

where w_n is the modulus of continuity of Z_n. So, using Shorack and Wellner (1986) (p 119),

$$\mathbb{P}\left(\lambda_n^p\sqrt{c_n}U_n(\lambda_n)\geq 1\right)\leq K\left(\lambda_n^{-p}c_n^{-\frac{1}{2}}\right)^{-4},$$

where K is a universal constant. Since $r-1-2\delta' p<0$,

$$\sum_n \mathbb{P}\left(\lambda_n^p\sqrt{c_n}U_n(\lambda_n)\geq 1\right)<\infty,$$

and

$$\sup_n\left|\lambda_n^p\sqrt{c_n}U_n(\lambda_n)\right|<\infty \text{a.s.},$$

which means that

$$\sup_n\left|\lambda_n^p\sum_{j=0}^{n}\sum_{k\in\mathcal{I}_j}2^{j(\frac{r}{2}-1)}\left(\mathbb{1}_{\ |\beta_{jk}|>\lambda_n}-\mathbb{P}(|\beta_{jk}|>\lambda_n)\right)\right|<\infty \text{a.s.}$$

But, using the definition of $\mathcal{W}^*(r,p)$,

$$\begin{aligned}\sup_n\lambda_n^p\sum_{j=0}^{n}\sum_{k\in\mathcal{I}_j}2^{j(\frac{r}{2}-1)}\mathbb{1}_{\ |\beta_{jk}|>\lambda_n} &\leq \sup_{\lambda>0}\lambda^p\sum_{j=0}^{\infty}\sum_{k\in\mathcal{I}_j}2^{j(\frac{r}{2}-1)}\mathbb{1}_{\ |\beta_{jk}|>\lambda}\\ &< \infty\text{a.s.}\end{aligned}$$

Therefore,

$$\begin{aligned} &\sup_n\lambda_n^p\sum_{j=0}^{n}\sum_{k\in\mathcal{I}_j}2^{j(\frac{r}{2}-1)}\mathbb{P}(|\beta_{jk}|>\lambda_n)<\infty\\ \Rightarrow\ &\sup_n\lambda_n^p\sum_{j=0}^{n}2^{j\frac{r}{2}}\left(\frac{\alpha_j}{\lambda_n+\alpha_j}\right)^p<\infty\\ \Rightarrow\ &\sup_n\sum_{j=\lfloor\frac{n\delta'}{\delta}\rfloor+1}^{n}2^{j\frac{r}{2}}\alpha_j^p\left(1+\alpha_j\lambda_n^{-1}\right)^{-p}<\infty.\end{aligned}$$

But for all $j \geq \frac{n\delta'}{\delta}$, $\alpha_j \lambda_n^{-1} \leq 1$. Consequently, the last inequality implies that

$$\sup_n \sum_{j=\lfloor \frac{n\delta'}{\delta} \rfloor + 1}^{n} 2^{j\frac{r}{2}} \alpha_j^p < \infty,$$

which means that $\frac{r}{2} < \delta p$.
Using the embedding $\mathcal{W}^*(r,p) \subset \mathcal{W}^*(r',p)$, when $r' < r$, we can omit the hypothesis (4.19) and assert that if f is in $\mathcal{W}^*(r,p)$ almost everywhere then $\frac{r}{2} < \delta p$.

□

Proof of sufficiency : Here K denotes a constant, eventually depending on r, p, δ, C and that may be different at each line. We suppose that $\frac{r}{2} < \delta p$ and we consider for all $\lambda > 0$,

$$Y_\lambda = \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}| > \lambda}.$$

We are going to prove that

$$\sup_{\lambda > 0} Y_\lambda < \infty \text{a.s.} \tag{4.20}$$

First, we note that using the three-series theorem, the condition $\frac{r}{2} < \delta p$ implies $Y_\lambda < \infty$a.s. for all $\lambda > 0$.
Let us begin by studying Y_λ when λ tends to 0. For all $n \in \mathbb{N}$, we set $Y^n = Y_{\lambda_n}$, where $\lambda_n = 2^{-n}$. We consider $j_0(n)$ the greatest integer j such that $\alpha_j \geq \lambda_n$.

$$\begin{aligned}
\mathbb{E}(Y^n) &= \lambda_n^p \sum_{j=0}^{\infty} 2^{j\frac{r}{2}} \left(\frac{\alpha_j}{\lambda_n + \alpha_j} \right)^p \\
&\leq \lambda_n^p \sum_{j=0}^{j_0-1} 2^{j\frac{r}{2}} + \sum_{j=j_0}^{\infty} 2^{j\frac{r}{2}} \alpha_j^p \\
&\leq K \, \lambda_n^p \, 2^{j_0\frac{r}{2}} \\
&\leq K \, \lambda_n^{p-\frac{r}{2\delta}}.
\end{aligned}$$

Therefore, $\sum_n \mathbb{E}(Y^n) < \infty$, which implies $\mathbb{P}(Y^n > 1 \quad \text{i.o.}) = 0$ and

$$\sup_n Y^n < \infty \text{a.s.}$$

Now, let $0 < \lambda \leq 1$ a fixed real number. There exists an integer n such that $\lambda_{n+1} < \lambda \leq \lambda_n$.

$$\begin{aligned}
Y_\lambda &= \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda} \\
&\leq \lambda_n^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda_{n+1}} \\
&\leq \left(\frac{\lambda_n}{\lambda_{n+1}}\right)^p Y^{n+1} \\
&\leq 2^p \sup_n Y^n.
\end{aligned}$$

Therefore

$$\sup_{0<\lambda\leq 1} Y_\lambda < \infty \text{a.s.}$$

Finally, we study Y_λ when λ is large. Let n be a fixed integer.

$$\mathbb{E}(Y_n) = n^p \sum_{j=0}^{\infty} 2^{j\frac{r}{2}} \, \mathbb{P}(|\beta_{jk}| > n).$$

But

$$\mathbb{P}(|\beta_{jk}| > n) = \begin{cases} \left(\frac{\alpha_j}{n+\alpha_j}\right)^p & \text{if } n < \mu_j, \\ 0 & \text{otherwise.} \end{cases}$$

If $m_n = \left[\frac{n^2+2p\log(C)}{2p\delta\log(2)}\right] + 1$, for $n \geq M$,

$$\begin{aligned}
\mathbb{E}(Y_n) &= n^p \sum_{j=m_n}^{\infty} 2^{j\frac{r}{2}} \left(\frac{\alpha_j}{n+\alpha_j}\right)^p \\
&\leq \sum_{j=m_n}^{\infty} 2^{j\frac{r}{2}} \alpha_j^p \\
&\leq K 2^{m_n(\frac{r}{2}-\delta p)}.
\end{aligned}$$

As previously, $\sum_n \mathbb{E}(Y_n) < \infty$, and

$$\sup_n Y_n < \infty \text{a.s.}$$

Now, let $\lambda \geq 1$ a fixed real number. There exists $n \in \mathbb{N}^*$ such that $n \leq \lambda < n+1$, and

$$\begin{aligned} Y_\lambda &= \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda} \\ &\leq (n+1)^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>n} \\ &\leq \left(\frac{n+1}{n}\right)^p Y_n \\ &\leq 2^p \sup_n Y_n. \end{aligned}$$

Therefore,

$$\sup_{\lambda \geq 1} Y_\lambda < \infty \text{a.s.}$$

and (4.20) is true. Theorem 4.1 is proved.

□

Chapitre 5

Maxisets for linear procedures

In this chapter, we study maxisets for linear procedures in the framework of the heteroscedastic white noise model. This allows to exhibit the nature of the spaces naturally connected to these procedures and to compare the performance of linear and non linear estimates by comparing their respective maxisets.

5.1 Introduction

One of the most traditional methods to measure the performance of an estimation procedure consists in investigating its minimax rates of convergence over large function classes. However, this minimax approach has undoubtedly two drawbacks : the choice for the function classes is quite subjective and exhibiting an estimator well adapted to the worst functions of this class seems too pessimistic for practical purposes. Among others, these are the reasons why Cohen, DeVore, Kerkyacharian and Picard (2001) or Kerkyacharian and Picard (2000, 2002) focused on an alternative to the minimax setting : the maxiset approach which consists in investigating the maximal space (called maxiset) where an estimation procedure achieves a given rate of convergence. For instance, these authors applied this theory for two well known efficient procedures : wavelet thresholding and local bandwidth selection. Under some conditions, since the maxiset for the first one is included into the maxiset for the second one, they could conclude that local bandwidth selection is at least as good as the thresholding procedure. For the statistical model considered in this chapter (see below), in chapter 6, we established that the Bayesian procedures for the mean and the median associated with a very general Bayesian model, achieve the same performance as the thresholding one from the maxiset approach. We can note that this approach is less pessimistic and, above all, it exhibits function spaces directly connected to the estimation

procedure.
Actually, if the maxiset point of view has recently been conceptualized in statistics from the similar notion already existing in approximation theory, it was underlying in Kerkyacharian and Picard (1993)'s paper that deals with the problem of density estimation with linear estimates. If we are given $X_1, \ldots, X_n$, n random variables having the common density f to be estimated, Kerkyacharian and Picard proved that when $2 \leq p < \infty$, then the minimax rate of convergence $n^{-\tau/(2\tau+1)}$ for the L_p-loss function on the balls of the Besov space $\mathcal{B}_{\tau,p,\infty}$ for $0 < \tau < \infty$ is achieved by linear estimators of the form

$$\hat{E}_{j_n}(x) = \frac{1}{n}\sum_{i=1}^{n} E_{j_n}(x, X_i) = \frac{1}{n}\sum_{i=1}^{n} 2^{j_n} E(2^{j_n}x, 2^{j_n}X_i),$$

where E is a kernel associated with a projection on the space V_0 of a multiresolution analysis and $2^{j_n} = n^{1/(2\tau+1)}$. But they investigated a "converse" result and proved that if a functional set V has the minimax rate $n^{-\tau/(2\tau+1)}$ for the L_p-estimation ($2 \leq p < \infty$) of an unknown density if one restricts to linear estimates, then V is included in a ball of the Besov space $\mathcal{B}_{\tau,p,\infty}$ and there is no advantage in restricting to V. Furthermore, this maxiset-type result underlines the strong link between Besov spaces $\mathcal{B}_{\tau,p,\infty}$ and linear estimates, for the model of density estimation.
In this chapter, and as Kerkyacharian and Picard (2000) for thresholding procedures or as in chapter 6 for Bayesian procedures, we consider the following heteroscedastic white noise model :

$$x_k = \theta_k + \varepsilon\sigma_k\xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k \in \mathbb{N}^*, \tag{5.1}$$

where $\theta = (\theta_k)_{k\in\mathbb{N}^*}$ is a sequence to be estimated and $\varepsilon > 0$. This heteroscedastic white noise model has extensively been used, for instance, as a sequence space framework for statistical inverse problems where $(\sigma_k^{-2})_{k\in\mathbb{N}^*}$ are the eigenvalues sequence of a known operator (see the references cited in section 5.2.1). In the wavelet context, Johnstone and Silverman (1997) proved that the heteroscedastic model appears as an appropriate large sample limit of the classical non parametric regression problem with correlated errors. So, along this chapter, we assume that $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$ is a known sequence of positive real numbers.
We can add that for inverse problems considered in the framework of the heteroscedastic white noise model, many authors consider linear estimates of the form

$$\hat{\theta}^{\lambda}(x) = (\lambda_k x_k)_{k\in\mathbb{N}^*}, \tag{5.2}$$

where $(\lambda_k)_{k\in\mathbb{N}^*}$, that can be interpreted as a smoothing parameter, is a sequence of non random weights that may depend on the parameters of the problem. Consequently, under the heteroscedastic white noise model, it seems relevant to determine maxisets for this linear procedure. It will be the main goal of this chapter. Another goal will be to answer the following questions :

- are the maxisets exhibited of the same nature as for the model of density estimation ?
- can we compare the performances of linear and non linear procedures by comparing their respective maxisets ?

The results we obtain are the following. We consider the model (5.1) and the risk for linear estimates is evaluated for a weighted l_p-norm ($1 \leq p < \infty$), that can be the l_p-norm, a Sobolev norm or a Besov norm. The maxisets are exhibited under very general assumptions on the λ_k's that are verified by commonly used weights (projection, Pinsker or Tikhonov-Phillips weights). Furthermore, if we consider the l_p-risk, a rate of convergence of the form $(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(2s+1)}$, and if $\sum_k \sigma_k^p < \infty$ or if σ has a polynomial growth or decay, the maxisets for our linear procedures are strictly included into the maxisets for the thresholding one. It means that linear estimators are outperformed by non linear ones. It is worthwhile to note that, unlike the linear estimators, the thresholding procedure involved in this chapter is adaptive. Then, we study maxisets associated with the problem of the L_p-estimation ($1 < p < \infty$) with projection weights, of functions f decomposed on a basis $(\psi_k)_{k\in\mathbb{N}^*}$: $f = \sum_{k\in\mathbb{N}^*} \theta_k\psi_k$. For this purpose, we exploit two important properties the ψ_k's must verify and that have been isolated by Kerkyacharian and Picard (2000). More precisely, we focus on unconditional bases of L_p-spaces that also verify a superconcentration inequality (see section 5.3.2). One more time, we prove that from the maxiset point of view, if $\forall\, k \in \mathbb{N}^*$, $\sigma_k = 1$, linear estimates are outperformed by thresholding ones. Let us note that in a different context (see Donoho, Johnstone, Kerkyacharian and Picard (1996b)), the minimax approach allowed to draw the same conclusion.
Finally, this chapter precisely describes the nature of the function spaces naturally linked to linear procedures under the model (5.1). Indeed, whatever the choice of the risk, the maxisets obtained are of the same type. And for the L_p-risk, under some conditions on the basis, the maxisets are exactly Besov spaces $\mathcal{B}_{s,p,\infty}$ as for the model of density estimation.
The chapter is organized as follows. Section 5.2 is devoted to the description of the statistical model and the function spaces involved in this chapter and in section 5.3, we describe the maxisets obtained for linear estimates.

5.2 Model and function spaces

5.2.1 Heteroscedastic white noise model

In the following, we assume that we are given the model (5.1) where $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$ is a known sequence of positive real numbers. This heteroscedastic white noise model appears as a generalization of the classical white noise model (for which, we have $\forall\, k \in \mathbb{N}^*$, $\sigma_k = 1$) often considered

by statisticians. We use the heteroscedastic white noise model when we have to estimate the solution of a linear operator equation $g = Af$, with noisy observations of g. Most of the time, to deal with such a problem, we exploit the singular value decomposition of A and the sequence $(\sigma_k^{-2})_{k\in\mathbb{N}^*}$ is then the eigenvalues sequence of the operator A^*A, with A^* the adjoint of A. Some well-posed inverse problems with noise can be reduced to (5.1) with $\sigma_k \overset{k\to\infty}{\longrightarrow} 0$. The condition $\sigma_k \overset{k\to\infty}{\longrightarrow} +\infty$ characterizes ill-posed problems. For instance, the sequences $(\sigma_k)_{k\geq 1}$ associated with operators such as integration considered by Ruymgaart (1993), the Radon transform (see Cavalier and Tsybakov (2000)), convolution for the case studied by Cavalier and Tsybakov (2000) or operators for some elliptic differential equations (see Mair and Ruymgaart (1996)) have a polynomial growth. But, the σ_k's may grow exponentially. See for instance Pereverzev and Schock (1999) who considered the problem of satellite geodesy or the inverse problems associated with partial differential equations such as the heat equation (see Mair and Ruymgaart (1996)).
In the wavelet context, Johnstone and Silverman (1997) showed that the heteroscedastic white noise model can also be used to represent direct observations with correlated structure. More precisely, let us assume that we are given the following non parametric regression model :

$$Y_i = f(\frac{i}{n}) + e_i, \quad 1 \leq i \leq n, \tag{5.3}$$

where the e_i's are drawn from a stationary Gaussian process. By studying the autocorrelation function of the errors, Johnstone and Silverman (1997) showed that under a good choice of the noise level, the model (5.1) appears as a good approximation of the model (5.3). To illustrate this statement, let us assume that $\lim_{k\to+\infty} k^\alpha \mathrm{cov}(e_i, e_{i+k}) = K < \infty$, with $0 < \alpha < 1$. In this case, the noise level at the resolution level j is proportional to $2^{-j(1-\alpha)/2}$.

5.2.2 Function spaces

As explained in Introduction, the final outcome of the maxiset approach is a function space where an estimation procedure achieves a given rate of convergence. So, let us introduce the following function spaces that will be useful throughout this chapter.

Definition 5.1. *For all $1 \leq p < \infty$ and $0 < \eta < \infty$, and for any non negative measure μ on $\mathbb{N}^*$, we set :*

$$B^\eta_{p,\infty}(\mu) = \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^{p\eta} \sum_{k\geq\lambda} \mu_k |\theta_k|^p < \infty \right\}.$$

When D is an interval of $\mathbb{R}^d$, $d \geq 1$, and $\mathcal{B} = \{\psi_k, k \in \mathbb{N}^*\}$ is an unconditional basis of $L_p(D)$,

$1 < p < \infty$, *we set*

$$B^{\eta}_{p,\infty}(\mathcal{B}) = \left\{ f = \sum_k \theta_k \psi_k : \quad \sup_{\lambda > 0} \lambda^{\eta} \| \sum_{k \geq \lambda} \theta_k \psi_k \|_{L_p} < \infty \right\}.$$

Under certain conditions on D and on the basis $\mathcal{B}$, the space $B^{\eta}_{p,\infty}(\mathcal{B})$ is a Besov space (see section 5.3.2). We shall see that the spaces $B^{\eta}_{p,\infty}(\mu)$ and $B^{\eta}_{p,\infty}(\mathcal{B})$, that are of the same nature, are strongly connected to linear procedures. It illustrates the following statement underlined as an advantage in Introduction : unlike the minimax setting, the maxiset approach exhibits function spaces naturally linked to the chosen estimation procedure. To be more convincing, let us provide here the maxisets associated with the thresholding procedures, and which nature is very different from the spaces introduced previously. We still consider the model (5.1) and for all $\varepsilon > 0$, we assume that we are given a real number $\Lambda_\varepsilon > 0$ only depending on ε and tending to $+\infty$ when ε tends to 0, and we set :

$$\hat{\theta}^t_k(x_k) = \begin{cases} x_k \mathbb{1}_{|x_k| \geq \kappa_t t_{k,\varepsilon}} & \text{if } k < \Lambda_\varepsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where κ_t is a constant and $t_{k,\varepsilon} = \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}$ is the universal threshold. The maxiset for this procedure has been exhibited by Kerkyacharian and Picard (2000) who obtained the following result :

Theorem 5.1. *Let* $1 \leq p < \infty$ *be a fixed real number and* $0 < r < \infty$. *We suppose that*

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \Lambda_\varepsilon = \left(\varepsilon \sqrt{\log(1/\varepsilon)}\right)^{-r},$$

where ε_0 *is such that* $\varepsilon_0 \sqrt{\log(1/\varepsilon_0)} = 1$, *and there exists a positive constant* C_t, *such that* $\forall\, 0 < \varepsilon \leq \varepsilon_0$,

$$\varepsilon^{\frac{\kappa_t^2}{16}} \log(1/\varepsilon)^{-\frac{1}{4} - \frac{p}{2}} \sum_{k < \Lambda_\varepsilon} \sigma_k^p \leq C_t.$$

Let ν *be a fixed real number such that* $0 < \nu < \infty$.
Then, if $\kappa_t \geq \sqrt{2p}$, *there exists a positive constant* C *such that*

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \left(\mathbb{E}\|\hat{\theta}^t - \theta\|^p_{l_p}\right)^{\frac{1}{p}} \leq C \left(\varepsilon \sqrt{\log(1/\varepsilon)}\right)^{2\nu/(2\nu+1)}$$

$$\iff \theta \in wl_{p,p/(2\nu+1)}(\sigma) \cap B^{\frac{2\nu}{r(2\nu+1)}}_{p,\infty}(\mu),$$

where for all $0 < q < p$,

$$\begin{aligned} wl_{p,q}(\sigma) &= \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^q \sum_k \mathbb{1}_{\,|\theta_k|>\lambda\sigma_k} \sigma_k^p < \infty \right\}, \\ &= \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^{q-p} \sum_k \mathbb{1}_{\,|\theta_k|\le\lambda\sigma_k} |\theta_k|^p < \infty \right\}, \end{aligned}$$

and μ *is the counting measure on* $\mathbb{N}^*$.

Notations : In the sequel, we shall use the following notations : if f_1 and f_2 are two positive functions of a variable x, we shall note $f_1 \approx f_2$, if and only if there exist two positive constants C_1 and C_2 independent of x such that $f_1(x)/f_2(x) \in [C_1, C_2]$.
For all $m \in \mathbb{R}_+^*$, we shall note $[m]$ the integer such that $m - 1 \le [m] < m$.

5.3 Maxisets and linear estimates

5.3.1 Our main result

Along this section, we study linear estimators of the form given in (5.2) which risk is evaluated under weighted l_p-norms $(1 \le p < \infty)$: for any non negative measure μ on $\mathbb{N}^*$, we set :

$$R_{p,\mu}(\hat{\theta}^\lambda) = \left(\mathbb{E}\|\hat{\theta}^\lambda - \theta\|^p_{l_p(\mu)}\right)^{\frac{1}{p}} = \left(\sum_{k\in\mathbb{N}^*} \mu_k \mathbb{E}|\lambda_k x_k - \theta_k|^p\right)^{\frac{1}{p}}.$$

Of course, if μ is the counting measure on $\mathbb{N}^*$, then the risk is studied under the l_p-norm. If $\forall\, k \in \mathbb{N}^*$, $\mu_k = k^{2w}$ $(w > 0)$, then $l_2(\mu)$ can be identified with W_w a Sobolev scale in $L_2(0,1)$:

$$W_w = \left\{ f = \sum_{k\in\mathbb{N}^*} \theta_k \phi_k : \quad \sum_{k\in\mathbb{N}^*} k^{2w}\theta_k^2 < \infty \right\},$$

where $(\phi_k)_{k\in\mathbb{N}^*}$ is an orthonormal basis of $L_2(0,1)$. Let us recall that when $w \in \mathbb{N}$, and $(\phi_k)_{k\in\mathbb{N}^*}$ is the Fourier basis,

$$\begin{aligned} W_w = \quad & \Big\{ f \in L_2(0,1) : f^{(w-1)} \text{ is absolutely continuous }, f^{(w)} \in L_2(0,1), \\ & f^{(2j)}(0) = f^{(2j)}(1) = 0, \quad \forall\, j \in \mathbb{N} \cap [0, (w-1)/2] \Big\}. \end{aligned}$$

We can generalize this example by considering measures μ associated with isotropic Besov spaces $\mathcal{B}_{z,p,p}(\mathbb{R}^d)$, $d \ge 1$. Indeed, let us assume that we are given a pair of scaling function and wavelet

ϕ and ψ and let us consider the following wavelet-tensor product basis denoted

$$\left\{\phi_{\mathbf{l},0}, \psi_{j,\mathbf{l},H} : \quad j \in \mathbb{N}, \mathbf{l} = (l_1, \dots, l_d) \in \mathbb{Z}^d, H \in \mathcal{H}_d\right\},$$

where $\mathcal{H}_d$ is the set of all the non void subsets of $\{1, \dots, d\}$ and

$$\phi_{\mathbf{l},0}(x_1, \dots, x_d) = \prod_{i=1}^{d} \phi(x_i - l_i),$$

$$\psi_{j,\mathbf{l},H}(x_1, \dots, x_d) = \prod_{i \in H} 2^{\frac{j}{2}} \psi(2^j x_i - l_i) \prod_{i \notin H} 2^{\frac{j}{2}} \phi(2^j x_i - l_i).$$

Then, under standard properties of regularity and moment vanishing of the wavelet (see Meyer (1992)) if

$$f = \sum_{\mathbf{l} \in \mathbb{Z}^d} \alpha_{\mathbf{l}} \phi_{\mathbf{l},0} + \sum_{j \in \mathbb{N}} \sum_{\mathbf{l} \in \mathbb{Z}^d} \sum_{H \in \mathcal{H}_d} \beta_{j,\mathbf{l},H} \psi_{j,\mathbf{l},H},$$

and $\forall\, 0 \le z < \infty$, $\forall\, 1 \le q \le \infty$,

$$f \in \mathcal{B}_{z,p,q}(\mathbb{R}^d) \iff \|\alpha\|_{l_p} + \left(\sum_{j \in \mathbb{N}} 2^{jq(z + d(\frac{1}{2} - \frac{1}{p}))} \|\beta_{j,\cdot,\cdot}\|_{l_p}^q \right)^{1/q} < \infty,$$

with the obvious modifications for $q = \infty$. So, with $k = (j, \mathbf{l}, H)$ and $\mu_k = 2^{jp(z + d(\frac{1}{2} - \frac{1}{p}))}$, $l_p(\mu)$ can be identified with $\mathcal{B}_{z,p,p}(\mathbb{R}^d)$. The maxisets obtained in this framework are the following :

Theorem 5.2. *Let $1 \le p < \infty$ and $0 < s < \infty$ be two real numbers. For any $m > 0$, we suppose we are given $(\lambda_k(m))_{k \ge 1}$, a non increasing sequence of weights lying in $[0, 1]$ such that*

- (A_1) *: there exists $\tilde{C} < 1$ such that $\forall\, m > 0$, $\forall\, k \ge m$, $\lambda_k(m) \le \tilde{C}$,*
- (A_2) *: there exists $C_s \in \mathbb{R}$ such that, with $\lambda_0 = 1$, $\forall\, m > 1$,*

$$\sum_{1 \le k < m} (\lambda_{k-1}(m) - \lambda_k(m)) (1 - \lambda_k(m))^{p-1} \left(\frac{k}{m} \right)^{-ps} \le C_s.$$

Then, for all $\varepsilon > 0$, we suppose we are given $m_\varepsilon > 0$ such that

1. *there exists ε such that $m_\varepsilon \le 1$,*
2. *$\varepsilon \to m_\varepsilon$ is continuous,*
3. *$\lim_{\varepsilon \to 0} m_\varepsilon = +\infty$.*

We consider the model (5.1) and the estimator $\hat{\theta}^{\lambda(m_\varepsilon)} = (\lambda_k(m_\varepsilon) \times x_k)_{k\in\mathbb{N}^*}$. *We suppose that there exists a positive constant* L *such that*

$$\forall\, \varepsilon > 0, \quad (\varepsilon m_\varepsilon^s)^p \sum_{k\in\mathbb{N}^*} \mu_k \sigma_k^p (\lambda_k(m_\varepsilon))^p \leq L. \tag{5.6}$$

Then, there exists a positive constant C *such that*

$$\forall\, \varepsilon > 0, \quad R_{p,\mu}(\hat{\theta}^{\lambda(m_\varepsilon)}) \leq C\, m_\varepsilon^{-s} \tag{5.7}$$
$$\iff \theta \in B^s_{p,\infty}(\mu).$$

This theorem is proved at the end of section 5.3.1. To shed light on this result, we consider the l_p-norm (we note μ the counting measure on $\mathbb{N}^*$) and $\forall\, k \in \mathbb{N}^*$, $\sigma_k = k^b$, $b > -1/p$. To check (5.6), we take $m_\varepsilon = \varepsilon^{-p/(ps+pb+1)}$, and with $\lambda_k(m) = \mathbb{1}_{\,k<m}$, we obtain :

$$\exists\, C > 0,\ \text{st}\ \forall\, \varepsilon > 0, \quad R_{p,\mu}(\hat{\theta}^{\lambda(m_\varepsilon)}) \leq C\varepsilon^{\frac{ps}{ps+pb+1}} \iff \theta \in B^s_{p,\infty}(\mu).$$

In particular, when $p = 2$, $\varepsilon^{2s/(2s+2b+1)}$ is the minimax rate achieved on the balls of the Sobolev space W_s (see Cavalier and Tsybakov (2000)) but also on the balls of $B^s_{2,\infty}(\mu)$ that contains W_s. Let us note that $B^s_{2,\infty}(\mu)$ can be identified with the Besov space $\mathcal{B}_{s,2,\infty}([0,1])$, and when $b = 0$ ($\forall\, k \in \mathbb{N}^*$, $\sigma_k = 1$), we obtain the classical rate $\varepsilon^{2s/(2s+1)}$.
It is worthwhile to apply this theorem with weights often met in the literature of inverse problems. For this purpose, we assume we are given for all $m > 0$, positive real numbers denoted T_m, such that $m \to T_m$ is not decreasing, continuous and $\lim_{m\to+\infty} T_m = +\infty$. For instance, let us focus on the weights considered by Cavalier, Golubev, Picard and Tsybakov (2000). That is to say, we consider

1. the projection weights : $\lambda_k^{(1)}(m) = \mathbb{1}_{\,k<T_m}, \quad m > 0,$
2. the Tikhonov-Phillips weights : $\lambda_k^{(2)}(m) = \frac{1}{1+(k/T_m)^\alpha}, \quad m > 0, \quad \alpha > 0,$
3. the Pinsker weights : $\lambda_k^{(3)}(m) = (1 - (k/T_m)^\alpha)_+, \quad m > 0, \quad \alpha > 0$, where $x_+ = \max(x; 0)$.

Projection weights are quite natural in the linear context and are extensively used when inverse problems are considered. The Tikhonov-Phillips weights appeared in the definition of a smoothing spline estimator under a non parametric regression model and with the penalty functional $\int_0^1 (f^{(N)}(u))^2 du$ (see Wahba (1990)). In this case, $\alpha = 2N$. The linear estimators associated with the weights $(\lambda_k^{(3)}(m)_{k\geq 1}$ have been exhibited by Pinsker (1980) who showed that they achieve the minimax rates of convergence on Sobolev classes under the white noise model. By straightforward computations (given in Appendix), we can easily prove the following result :

Proposition 5.1. *The weights $\lambda_k^{(i)}(m)$ verify assumptions (A_1) and (A_2) of Theorem 5.2 if and only if $\alpha > s$, if $i \in \{2,3\}$, and there exist two positive constants β_1 and β_2 such that $\forall\, m \geq 1$,*

$$\begin{aligned} T_m \leq [m+1], \ \textit{and}\ m \leq \beta_2 T_m, \qquad & \textit{if} \quad i = 1, \\ \beta_1 T_m \leq [m+1], \ \textit{and}\ m \leq \beta_2 T_m, \qquad & \textit{if} \quad i \in \{2,3\}. \end{aligned}$$

Generally, the previous weights appeared when $(\sigma_k)_{k\geq 1}$ grows not faster than a power of k, which is supposed by Cavalier, Golubev, Picard and Tsybakov (2000). So, it seems natural to define weights depending on the σ_k's and having the same form as the previous ones. Consequently, we define :

4. $\lambda_k^{(4)}(m) = 1_{\sigma_k < T_m}, \quad m > 0,$
5. $\lambda_k^{(5)}(m) = \frac{1}{1+(\sigma_k/T_m)^\alpha}, \quad m > 0, \quad \alpha > 0,$
6. $\lambda_k^{(6)}(m) = (1 - (\sigma_k/T_m)^\alpha)_+, \quad m > 0, \quad \alpha > 0.$

These weights, or closed forms of them, have been used by Ruymgaart (1993), Mair and Ruymgaart (1996) or Cavalier and Tsybakov (2000). If the weights $(\lambda_k^{(i)})_{k\geq 1}$ for $i \in \{1,2,3\}$ are well adapted to a polynomial growth of $(\sigma_k)_{k\geq 1}$, the following result (proved by straightforward computations given in Appendix) shows how to choose the parameters of $(\lambda_k^{(i)})_{k\geq 1}$ for $i \in \{4,5,6\}$ when $(\sigma_k)_{k\geq 1}$ has at least an exponential growth.

Proposition 5.2. *We assume that there exists $b > 1$ such that*

$$\forall\, k \in \mathbb{N}^*, \quad \sigma_k \geq b\sigma_{k-1}.$$

We define :

$$\forall\, x > 0, \quad \sigma^{-1}(x) = \begin{cases} k & \textit{if } \sigma_k = x, \\ k+1 & \textit{if } \sigma_k < x < \sigma_{k+1}. \end{cases}$$

The weights $\lambda_k^{(i)}(m)$ verify assumptions (A_1) and (A_2) of Theorem 5.2 if and only if there exist two positive constants β_1 and β_2 such that $\forall\, m \geq 1$,

$$\begin{aligned} T_m \leq \sigma_{[m+1]}, \ \textit{and}\ m \leq \beta_2\, \sigma^{-1}(T_m), \qquad & \textit{if} \quad i = 4, \\ \beta_1 T_m \leq \sigma_{[m+1]}, \ \textit{and}\ m \leq \beta_2\, \sigma^{-1}(T_m), \qquad & \textit{if} \quad i \in \{5,6\}. \end{aligned}$$

Now, let us compare linear estimates with adaptive thresholding estimates from the maxiset point of view for the l_p-norm and with the rate of convergence $(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(2s+1)}$, $(0 < s < \infty)$. Most of the time, the thresholding procedure of Theorem 5.1 is applied with $r = 2$. Nevertheless, we shall exploit the results of Theorem 5.1 with $r \geq 2$ and $\kappa_t \geq \sqrt{2p}$ and those of Theorem 5.2 with the projection weights $\lambda_k(m) = \mathbb{1}_{\,k<m}$ and we take $m_\varepsilon = (\varepsilon\sqrt{\log(1/\varepsilon)})^{-2/(2s+1)}$. We still note μ the counting measure on $\mathbb{N}^*$.

If $\sum_k \sigma_k^p < \infty$.

This assumption yields that condition (5.6) of Theorem 5.2 is fulfilled. $B_{p,\infty}^{\frac{2s}{r(2s+1)}}(\mu) \subset l_p(\mu) \subset wl_{p,p/(2s+1)}(\sigma)$. Consequently, $B^s_{p,\infty}(\mu)$ is strictly included into $B_{p,\infty}^{\frac{2s}{r(2s+1)}}(\mu) \cap wl_{p,p/(2s+1)}(\sigma) = B_{p,\infty}^{\frac{2s}{r(2s+1)}}(\mu)$, which means that linear estimators are outperformed by non linear ones.

Polynomial growth/decay of $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$**.**

We suppose that there exist $c_1 > 0$, $c_2 > 0$ and $b \in \mathbb{R}$ such that, $\forall\, k \in \mathbb{N}^*$,

$$c_1 k^b \leq \sigma_k \leq c_2 k^b.$$

This case includes the homoscedastic case ($\sigma_k = 1$, $\forall\, k \in \mathbb{N}^*$).
Condition (5.6) of Theorem 5.2 is fulfilled if and only if $2(bp+1) \leq p$ and we take $\kappa_t \geq \sqrt{\max(2p; 16r(pb+1))}$.

Lemma 5.1. *Since $r \geq 2$ and $2(bp+1) \leq p$, $B^s_{p,\infty}(\mu)$ is strictly included into $B_{p,\infty}^{\frac{2s}{r(2s+1)}}(\mu) \cap wl_{p,p/(2s+1)}(\sigma)$.*

Proof : Let us assume that $bp > -1$. Then, $\forall\, \theta = (\theta_k)_{k\in\mathbb{N}^*}$, $\forall\, 0 < \lambda \leq 1$, $\forall\, N_\lambda \geq 1$,

$$\begin{aligned}
\lambda^{-\frac{2ps}{2s+1}} \sum_k \mathbb{1}_{\,|\theta_k| \leq \sigma_k \lambda} |\theta_k|^p &\leq \lambda^{-\frac{2ps}{2s+1}} \sum_{k \geq N_\lambda} |\theta_k|^p + \lambda^{-\frac{2ps}{2s+1}} \sum_{k < N_\lambda} \mathbb{1}_{\,|\theta_k| \leq \sigma_k \lambda} |\theta_k|^p \\
&\leq \lambda^{-\frac{2ps}{2s+1}} \sum_{k \geq N_\lambda} |\theta_k|^p + c_2^p \lambda^{\frac{p}{2s+1}} \sum_{k < N_\lambda} k^{pb} \qquad (5.8) \\
&\leq \lambda^{-\frac{2ps}{2s+1}} \sum_{k \geq N_\lambda} |\theta_k|^p + K \lambda^{\frac{p}{2s+1}} N_\lambda^{pb+1},
\end{aligned}$$

where K is a positive constant. With $N_\lambda = \lambda^{-2/(2s+1)}$, and if $\theta \in B^s_{p,\infty}(\mu)$, since $2(bp+1) \leq p$, we have :

$$\sup_{\lambda \leq 1} \lambda^{-\frac{2ps}{2s+1}} \sum_k \mathbb{1}_{\,|\theta_k| \leq \sigma_k \lambda} |\theta_k|^p < \infty.$$

Furthermore, $\forall\, \theta = (\theta_k)_{k\in\mathbb{N}^*}$,

$$\sup_{\lambda \geq 1} \lambda^{-\frac{2ps}{2s+1}} \sum_k \mathbb{1}_{\,|\theta_k| \leq \sigma_k \lambda} |\theta_k|^p \leq \sum_k |\theta_k|^p.$$

So, $B^s_{p,\infty}(\mu) \subset wl_{p,p/(2s+1)}(\sigma)$. Now, since $\frac{p}{2} - pb \geq 1$, there exists $u > 0$, such that

$$\frac{1}{2} \leq \frac{(pb + u^{-1})(2s+1)}{p} < \frac{(2s+1)}{2},$$

so we set

$$v = \frac{(upb+1)(2s+1)}{p} \in \left[\frac{u}{2}, \frac{u(2s+1)}{2}\right),$$

and

$$\forall\, l \in \mathbb{N}^*, \quad \theta_l = \begin{cases} \sigma_l \times \frac{1}{k^v} & \text{if } l = [k^u + 1], \\ 0 & \text{otherwise.} \end{cases}$$

For any $0 < \lambda \leq 1$,

$$\begin{aligned} \lambda^{p/(2s+1)} \sum_l \mathbb{1}_{\ |\theta_l| > \sigma_l \lambda} \sigma_l^p &\approx \lambda^{p/(2s+1)} \sum_{k \leq \lambda^{-1/v}} k^{upb} \\ &\leq K \lambda^{p/(2s+1) - (upb+1)/v} \\ &\leq K, \end{aligned}$$

where K is a positive constant. This implies that $\theta \in wl_{p,p/(2s+1)}(\sigma)$. For any $\lambda \geq 1$,

$$\begin{aligned} \sum_{l \geq \lambda} |\theta_l|^p &\approx \sum_{k \geq \lambda^{1/u}} k^{upb-pv} \\ &\approx \sum_{k \geq \lambda^{1/u}} k^{vp/(2s+1)-1-pv} \\ &\approx \sum_{k \geq \lambda^{1/u}} k^{-2vps/(2s+1)-1} \\ &\approx \lambda^{-ps \times \frac{2v}{u(2s+1)}}. \end{aligned}$$

Since $2v < u(2s+1)$, then $\theta \notin B^s_{p,\infty}(\mu)$. But, since $r \geq 2$ and $2v \geq u$, we have $\theta \in B^{\frac{2s}{r(2s+1)}}_{p,\infty}(\mu)$.
Consequently, $B^s_{p,\infty}(\mu)$ is strictly included into $B^{\frac{2s}{r(2s+1)}}_{p,\infty}(\mu) \cap wl_{p,p/(2s+1)}(\sigma)$.
If $bp = -1$, then we reach the same conclusion by using (5.8).
Let us note that the case $bp < -1$ is handled by the case $\sum_k \sigma_k^p < \infty$.

□

The result of Lemma 5.1 means that linear estimators are outperformed by non linear ones. Let us note that the assumption $r \geq 2$ is essential to claim this statement. Indeed, if $r < 2$ and for small values of s, $B^{\frac{2s}{r(2s+1)}}_{p,\infty}(\mu)$ is strictly included into $B^s_{p,\infty}(\mu)$. Finally, if σ has an exponential growth, then we could apply neither the thresholding procedure, nor the linear one with $m_\varepsilon = (\varepsilon\sqrt{\log(1/\varepsilon)})^{-2/(2s+1)}$. But with a logarithmic rate of convergence and by adapting the results of Theorem 5.1, we expect from the maxiset point of view, thresholding procedures to be preferable to linear ones. Let us also note that these last results remain valid for Pinsker and Tikhonov-Phillips weights for an appropriate choice of α.

Proof of Theorem 5.2 : We have :

$$\begin{aligned}\sum_k \mu_k (1-\lambda_k)^p |\theta_k|^p &\leq 2^{p-1}\left[\sum_k \mu_k \mathbb{E}|\lambda_k\ x_k - \theta_k|^p + \sum_k \mu_k \mathbb{E}|\lambda_k (x_k-\theta_k)|^p\right] \\ &\leq 2^{p-1}\left[\mathbb{E}\|\hat{\theta}^{\lambda} - \theta\|^p_{l_p(\mu)} + \varepsilon^p\ \mathbb{E}|\xi_1|^p \sum_k \mu_k \lambda_k^p \sigma_k^p\right].\end{aligned}$$

Therefore, under (5.6), if (5.7) is true, for $\varepsilon > 0$,

$$(1-\tilde{C})^p \sum_{k\geq m_\varepsilon} \mu_k |\theta_k|^p \leq \sum_k \mu_k (1-\lambda_k)^p |\theta_k|^p \leq K m_\varepsilon^{-ps},$$

where K is a positive constant. Using the hypotheses on m_ε, this implies that $\theta \in B^s_{p,\infty}(\mu)$. Conversely, let us assume that $\theta \in B^s_{p,\infty}(\mu)$, then, we have that there exists a positive constant C, such that

$$\forall\ k \geq 1, \quad s_k = \sum_{l\geq k} \mu_l |\theta_l|^p \leq C\ k^{-ps}.$$

We have, for $\varepsilon > 0$:

$$\begin{aligned}\sum_{k<m_\varepsilon} \mu_k (1-\lambda_k)^p |\theta_k|^p &= \sum_{k=1}^{[m_\varepsilon]} (1-\lambda_k)^p s_k - \sum_{k=1}^{[m_\varepsilon]} (1-\lambda_k)^p s_{k+1} \\ &\leq \sum_{k=1}^{[m_\varepsilon]} \left((1-\lambda_k)^p - (1-\lambda_{k-1})^p\right) s_k \\ &\leq pC \sum_{k=1}^{[m_\varepsilon]} (\lambda_{k-1}-\lambda_k)(1-\lambda_k)^{p-1} k^{-ps} \\ &\leq pCC_s m_\varepsilon^{-ps}.\end{aligned}$$

Then, for $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}\|\hat{\theta}^{\lambda(m_\varepsilon)} - \theta\|^p_{l_p(\mu)} &\leq 2^{p-1}\left[\sum_k \mu_k (1-\lambda_k)^p |\theta_k|^p + \sum_k \mu_k \mathbb{E}|\lambda_k (x_k-\theta_k)|^p\right] \\ &\leq 2^{p-1}\left[\sum_{k<m_\varepsilon} \mu_k (1-\lambda_k)^p |\theta_k|^p + \sum_{k\geq m_\varepsilon} \mu_k |\theta_k|^p\right] \\ &\quad +2^{p-1}\varepsilon^p \mathbb{E}|\xi_1|^p \sum_k \mu_k \lambda_k^p \sigma_k^p \\ &\leq K m_\varepsilon^{-ps},\end{aligned}$$

where K is a positive constant. Theorem 5.2 is proved.

□

5.3.2 Maxisets for the L_p-risk and with projection weights

In this section, we estimate functions of $L_p(D)$, where $D = [0,1]^d$ or $D = \mathbb{R}^d$. For this purpose, we exploit a wavelet basis of $L_2(D)$ noted $\mathcal{B} = \{\psi_k, k \in \mathbb{N}^*\}$. More precisely, we assume that $(\psi_k)_{k\in\mathbb{N}^*}$ is a wavelet-tensor product basis constructed on compactly supported wavelets. So, if $1 < p < \infty$, Meyer (1992) proved that $\mathcal{B}$ is an unconditional basis of $L_p(D)$, which means that

- for any $f \in L_p(D)$, there exists a unique sequence $\theta = (\theta_k)_{k\in\mathbb{N}^*}$ such that

$$f = \sum_k \theta_k \psi_k, \tag{5.9}$$

- there exists an absolute constant K such that if $\forall\, k \in \mathbb{N}^*$, $|\theta_k| \le |\theta'_k|$, then

$$\left\| \sum_k \theta_k \psi_k \right\|_{L_p} \le K \left\| \sum_k \theta'_k \psi_k \right\|_{L_p}.$$

The restriction $p \in (1,\infty)$ is due to the fact that in general, there is no unconditional bases if $p \notin (1,\infty)$. Furthermore, we assume that $\{\sigma_k\psi_k,\ k \in \mathbb{N}^*\}$ verifies the following superconcentration inequality : for any $0 < r < \infty$, there exists a constant $C(p,r)$ such that for all $F \subset \mathbb{N}^*$,

$$\left\| \left[\sum_{k\in F} |\sigma_k \psi_k|^r \right]^{\frac{1}{r}} \right\|_{L_p} \le C(p,r) \left\| \sup_{k\in F} |\sigma_k\psi_k| \right\|_{L_p}. \tag{5.10}$$

Remark 5.1. *Kerkyacharian and Picard (2000) noted that if the σ_k's depend only on the resolution levels and if at level j, the noise level is proportional to 2^{bj}, with $b > -\frac{d}{2}$, then $\{\sigma_k\psi_k,\ k \in \mathbb{N}^*\}$ verifies the previous superconcentration inequality.*

Under these hypotheses, we can exhibit the maxiset associated with the linear procedure with projection weights for the L_p-risk. We have :

Theorem 5.3. *Let $0 < s < \infty$ be a fixed real number. For all $\varepsilon > 0$, we suppose we are given $m_\varepsilon > 0$ such that*

1. *there exists ε such that $m_\varepsilon \le 1$,*

2. *$\varepsilon \to m_\varepsilon$ is continuous,*

3. $\lim_{\varepsilon\to 0} m_\varepsilon = +\infty$.

To estimate a function of the form given in (5.9) under the model (5.1), we suppose that there exists a positive constant $\tilde{L}$ *such that*

$$\forall\, \varepsilon > 0, \quad (\varepsilon m_\varepsilon^s)^p \sum_{k<m_\varepsilon} \int |\psi_k|^p \sigma_k^p \le \tilde{L}. \tag{5.11}$$

Then, if $\lambda_k(m) = \mathbb{1}_{\,k<m}$*, there exists a positive constant* $\tilde{C}$ *such that*

$$\forall\, \varepsilon > 0, \quad m_\varepsilon^s \left[\mathbb{E}\|\sum_k (\lambda_k(m_\varepsilon)x_k - \theta_k)\psi_k\|_{L_p}^p\right]^{1/p} \le \tilde{C} \tag{5.12}$$
$$\Longleftrightarrow\; f = \sum_k \theta_k\psi_k \in B_{p,\infty}^s(\mathcal{B}).$$

Remark 5.2. *For the proof of Theorem 5.3, we exploit the particular form of the projection weights that can take only two values : 0 or 1, and the superconcentration inequality checked by* $\{\sigma_k\psi_k,\ k\in\mathbb{N}^*\}$*. To obtain a similar result for other weights* $\lambda_k(m_\varepsilon)$*, it is enough to check that,* $\forall\,\varepsilon>0$, $\{\lambda_k(m_\varepsilon)\sigma_k\psi_k,\ k\in\mathbb{N}^*\}$ *verifies a superconcentration inequality.*

To shed light on this result, we assume until the end of this section that $\forall\, k\in\mathbb{N}^*$, $\sigma_k = 1$, and $\mathcal{B}$ is a wavelet-tensor basis on $[0,1]^d$, denoted $\mathcal{B} = \{\psi_{j\mathbf{l}}:\quad j\in\mathbb{N}, \mathbf{l}\in A_j\}$, where $\forall\, j\in\mathbb{N}$, A_j is a set with cardinality proportional to 2^{jd}. Then, under standard properties of regularity and moment vanishing of the wavelet (see Meyer (1992)), $B_{p,\infty}^{\eta/d}(\mathcal{B})$ can be identified with the Besov space

$$\mathcal{B}_{\eta,p,\infty}([0,1]^d) = \left\{ f = \sum_{j\in\mathbb{N}}\sum_{\mathbf{l}\in A_j} \beta_{j\mathbf{l}}\psi_{j\mathbf{l}} : \quad \sup_{J\ge 0} 2^{J\eta}\|\sum_{j\ge J}\sum_{\mathbf{l}\in A_j}\beta_{j\mathbf{l}}\psi_{j\mathbf{l}}\|_{L_p} < \infty \right\},$$

and this shows that whatever the choice of the model (density estimation or heteroscedastic white noise model), Besov spaces $\mathcal{B}_{\eta,p,\infty}$ are strongly connected to linear procedures. Furthermore, using $\int|\psi_{jk}|^p \approx 2^{jd(\frac{p}{2}-1)}$, the upper bound obtained for $\mathcal{B}_{s,p,\infty}([0,1]^d)$ and with the L_p-risk is $\varepsilon^{2s/(d+2s)}$. What is more, and as in section 5.3.1, using Theorems 5.1, 5.2 and 6.2 of Kerkyacharian and Picard (2000), we show that linear estimators are outperformed by non linear ones for the rate $(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(d+2s)}$. With the classical calibration $\varepsilon\approx n^{-1/2}$, this rate is, up to a logarithmic term, the minimax rate that appears for the model of density estimation (see Donoho, Johnstone, Kerkyacharian and Picard (1996b)) or for the model of non parametric regression (see Donoho, Johnstone, Kerkyacharian and Picard (1996a)). Let us note that for the model (5.1), Kerkyacharian and Picard (2000) exhibited an adaptive thresholding procedure that achieves the rate

$(\varepsilon\sqrt{\log(1/\varepsilon)})^{2s/(d+2s)}$ on the balls of $\mathcal{B}_{s,p,\infty}([0,1]^d)$. Finally, let us give the proof of Theorem 5.3 :

Proof of Theorem 5.3 : We recall that $\mathcal{B}$ is an unconditional basis of $L_p(D)$ if and only if there exists $M>0$ such that for any set $F\subset\mathbb{N}^*$, and any choice of the coefficients c_k's

$$M^{-1}\left\|\sum_{k\in F}c_k\psi_k\right\|_{L_p}\le\left\|\left(\sum_{k\in F}|c_k\psi_k|^2\right)^{\frac{1}{2}}\right\|_{L_p}\le M\left\|\sum_{k\in F}c_k\psi_k\right\|_{L_p}. \tag{5.13}$$

Furthermore, since $\{\sigma_k\psi_k,\ k\in\mathbb{N}^*\}$ verifies a superconcentration inequality, there exist two positive constants c_p and C_p such that for any $F\subset\mathbb{N}^*$, we have :

$$c_p\int\sum_{k\in F}|\sigma_k\psi_k|^p\le\int\left(\sum_{k\in F}|\sigma_k\psi_k|^2\right)^{\frac{p}{2}}\le C_p\int\sum_{k\in F}|\sigma_k\psi_k|^p. \tag{5.14}$$

In the following, the notation K will keep designating a constant independent of ε that may be different at each line.
If $\varepsilon>0$, let us set

$$A_\varepsilon=\mathbb{E}\|\sum_{k<m_\varepsilon}(x_k-\theta_k)\psi_k\|^p_{L_p}.$$

Using (5.13),

$$\begin{aligned}A_\varepsilon &= \varepsilon^p\mathbb{E}\|\sum_{k<m_\varepsilon}\sigma_k\xi_k\psi_k\|^p_{L_p}\\ &\le K\varepsilon^p\mathbb{E}\int\left(\sum_{k<m_\varepsilon}\sigma_k^2\xi_k^2\psi_k^2\right)^{\frac{p}{2}}.\end{aligned}$$

If $p\le 2$, since $l_p\subset l_2$,

$$A_\varepsilon \le K\varepsilon^p\sum_{k<m_\varepsilon}\sigma_k^p\int|\psi_k|^p.$$

If $p\ge 2$, using the generalized Minkowski inequality and (5.14),

$$\begin{aligned}A_\varepsilon &\le K\varepsilon^p\int\left(\sum_{k<m_\varepsilon}\sigma_k^2\psi_k^2(\mathbb{E}|\xi_k|^p)^{\frac{2}{p}}\right)^{\frac{p}{2}}\\ &\le K\varepsilon^p\sum_{k<m_\varepsilon}\sigma_k^p\int|\psi_k|^p.\end{aligned}$$

Now, using (5.11), we have :

$$\begin{aligned}\|\sum_{k\geq m_\varepsilon}\theta_k\psi_k\|_{L_p}^p &= \|\sum_k(1-\lambda_k)\theta_k\psi_k\|_{L_p}^p\\ &\leq 2^{p-1}(A_\varepsilon+\mathbb{E}\|\sum_k(\lambda_k x_k-\theta_k)\psi_k\|_{L_p}^p)\\ &\leq K m_\varepsilon^{-ps},\end{aligned}$$

if (5.12) is true, which implies that $f\in B^s_{p,\infty}(\mathcal{B})$.
Conversely, if we assume that $f\in B^s_{p,\infty}(\mathcal{B})$,

$$\begin{aligned}\mathbb{E}\|\sum_k(\lambda_k x_k-\theta_k)\psi_k\|_{L_p}^p &\leq 2^{p-1}(A_\varepsilon+\|\sum_{k\geq m_\varepsilon}\theta_k\psi_k\|_{L_p}^p)\\ &\leq K m_\varepsilon^{-ps},\end{aligned}$$

by using (5.11), which ends the proof of the theorem.

□

5.4 Appendix : Proof of Propositions 5.1 and 5.2

In this section, we give a brief sketch of the proof of Propositions 5.1 and 5.2, by investigating the conditions the six weights defined in section 5.3.1 must check to fulfill assumptions (A_1) and (A_2) of Theorem 5.2. We shall note $\forall\ 0<u<\infty$,

$$\forall\ m>1,\quad U_m=\sum_{1\leq k<m}\left(\lambda_{k-1}(m)-\lambda_k(m)\right)\left(1-\lambda_k(m)\right)^{p-1}\left(\frac{k}{m}\right)^{-pu},$$

and we suppose we are given $1\leq p<\infty$. Let us also note that (A_1) is fulfilled if and only if there exists $\tilde{C}>0$ such that $\forall\ m>0$, $\lambda_{[m+1]}(m)\leq\tilde{C}<1$.

1. Projection weights : $\lambda_k(m)=1_{k<T_m}$.
Assumption (A_1) is fulfilled if and only if

$$\forall\ m>0,\quad T_m\leq[m+1].$$

We have, $\forall\ m>1$, $\forall\ 1\leq k\leq[m]$:

$$\left(\lambda_{k-1}(m)-\lambda_k(m)\right)\left(1-\lambda_k(m)\right)^{p-1}\neq 0\iff k-1<T_m\leq k.$$

So, $(U_m)_{m>1}$ is bounded if and only if there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \leq \beta_2 T_m.$$

2. Tikhonov-Phillips weights : $\lambda_k(m) = \frac{1}{1+(k/T_m)^\alpha}$.
Assumption (A_1) is fulfilled if and only if there exists $\beta_1 > 0$, such that

$$\forall\, m > 0, \quad \beta_1 T_m \leq [m+1].$$

Furthermore, we have, $\forall\, m > 1$,

$$\begin{aligned} U_m &= \sum_{k=1}^{[m]} \frac{T_m^\alpha (k^\alpha - (k-1)^\alpha)}{(k^\alpha + T_m^\alpha)((k-1)^\alpha + T_m^\alpha)} \left(\frac{k^\alpha}{k^\alpha + T_m^\alpha}\right)^{p-1} \left(\frac{k}{m}\right)^{-pu} \\ &\approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[m]} k^{p(\alpha-u)-1} \left(1 + \left(\frac{k}{T_m}\right)^\alpha\right)^{-p-1}. \end{aligned}$$

If $[m] \leq T_m$,we have :

$$U_m \approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[m]} k^{p(\alpha-u)-1}.$$

Using $\beta_1 T_m \leq [m+1]$, $(U_m)_{m>1}$ is bounded if and only if $\alpha > u$.
If $[m] > T_m$,we have :

$$\begin{aligned} U_m &\approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[T_m]} k^{p(\alpha-u)-1} + \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=[T_m]+1}^{[m]} k^{p(\alpha-u)-1} \left(\frac{k}{T_m}\right)^{-\alpha(p+1)} \\ &\approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[T_m]} k^{p(\alpha-u)-1} + \left(\frac{m}{T_m}\right)^{pu}. \end{aligned}$$

Using $\beta_1 T_m \leq [m+1]$, $(U_m)_{m>1}$ is bounded if and only if $\alpha > u$ and there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \leq \beta_2 T_m.$$

3. Pinsker weights : $\lambda_k(m) = \left(1 - \left(\frac{k}{T_m}\right)^\alpha\right)_+$.
Assumption (A_1) is fulfilled if and only if there exists $\beta_1 > 0$, such that

$$\forall\, m > 0, \quad \beta_1 T_m \leq [m+1].$$

For any $m > 1$, if $[m] \leq T_m$, we have :

$$\begin{aligned} U_m &\approx \sum_{k=1}^{[m]} \left(\left(\frac{k}{T_m}\right)^{\alpha} - \left(\frac{k-1}{T_m}\right)^{\alpha} \right) \left(\frac{k}{T_m}\right)^{\alpha(p-1)} \left(\frac{k}{m}\right)^{-pu} \\ &\approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[m]} k^{p(\alpha-u)-1}. \end{aligned}$$

Using $\beta_1 T_m \leq [m+1]$, $(U_m)_{m>1}$ is bounded if and only if $\alpha > u$.
If $[m] > T_m$,we have :

$$\begin{aligned} U_m &= \sum_{k=1}^{[T_m]} \left(\left(\frac{k}{T_m}\right)^{\alpha} - \left(\frac{k-1}{T_m}\right)^{\alpha} \right) \left(\frac{k}{T_m}\right)^{\alpha(p-1)} \left(\frac{k}{m}\right)^{-pu} \\ &\qquad + \left(1 - \left(\frac{[T_m]}{T_m}\right)^{\alpha}\right) \left(\frac{[T_m]+1}{m}\right)^{-pu} \\ &\approx \frac{m^{pu}}{T_m^{p\alpha}} \sum_{k=1}^{[T_m]} k^{p(\alpha-u)-1} + \left(1 - \left(\frac{[T_m]}{T_m}\right)^{\alpha}\right) \left(\frac{[T_m]+1}{m}\right)^{-pu}. \end{aligned}$$

Using $\beta_1 T_m \leq [m+1]$, $(U_m)_{m>1}$ is bounded if and only if $\alpha > u$ and there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \leq \beta_2 T_m.$$

So, using the hypotheses on $m \to T_m$, assumptions (A_1) and (A_2) of Theorem 5.2 are checked if and only if there exist two positive constants β_1 and β_2, such that

$$\forall\, m \geq 1, \quad \beta_2^{-1} m \leq T_m \leq \beta_1^{-1}[m+1],$$

with $\beta_1 \geq 1$ for projection weights and $\alpha > s$ for Tikhonov-Phillips and Pinsker weights.

In the following, we shall assume that there exists $b > 1$ such that

$$\forall\, k \in \mathbb{N}^* \quad \sigma_k \geq b\sigma_{k-1}.$$

We shall note indifferently $\sigma(k) = \sigma_k$. We shall use the following function :

$$\forall\, x \in \mathbb{R}_+, \quad \sigma^{-1}(x) = \begin{cases} k & \text{if } \sigma_k = x, \\ k+1 & \text{if } \sigma_k < x < \sigma_{k+1}. \end{cases}$$

This implies that

$$\forall\, x \in \mathbb{R}_+, \quad \sigma(\sigma^{-1}(x) - 1) < x \leq \sigma(\sigma^{-1}(x)).$$

We can observe that $\forall\, m > 0$, if $\sigma_{[m]} \le T_m$ then $[m] \le \sigma^{-1}(T_m)$. Indeed, if $[m] > \sigma^{-1}(T_m)$ then, $\sigma_{[m]} > \sigma(\sigma^{-1}(T_m)) \ge T_m$ and we obtain a contradiction.

4. Projection weights : $\lambda_k(m) = 1_{\sigma_k < T_m}$**.**
Assumption (A_1) is fulfilled if and only if

$$\forall\, m > 0, \quad T_m \le \sigma_{[m+1]}.$$

Furthermore, we have, $\forall\, m > 1$, $\forall\, 1 \le k \le [m]$:

$$(\lambda_{k-1}(m) - \lambda_k(m))\,(1 - \lambda_k(m))^{p-1} \ne 0 \iff k = \sigma^{-1}(T_m).$$

So, $(U_m)_{m>1}$ is bounded if and only if there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \le \beta_2 \sigma^{-1}(T_m).$$

5. Tikhonov-regularization weights : $\lambda_k(m) = \frac{1}{1+(\sigma_k/T_m)^\alpha}$**.**
Assumption (A_1) is fulfilled if and only if there exists $\beta_1 > 0$, such that

$$\forall\, m > 0, \quad \beta_1 T_m \le \sigma_{[m+1]}.$$

Furthermore, we have, $\forall\, m > 1$,

$$\begin{aligned} U_m &= \sum_{k=1}^{[m]} \frac{T_m^\alpha(\sigma_k^\alpha - \sigma_{k-1}^\alpha)}{(\sigma_k^\alpha + T_m^\alpha)(\sigma_{k-1}^\alpha + T_m^\alpha)} \left(\frac{\sigma_k^\alpha}{\sigma_k^\alpha + T_m^\alpha}\right)^{p-1} \left(\frac{k}{m}\right)^{-pu} \\ &\approx \sum_{k=1}^{[m]} \frac{T_m^\alpha \sigma_k^{p\alpha}}{(\sigma_k^\alpha + T_m^\alpha)^p(\sigma_{k-1}^\alpha + T_m^\alpha)} \left(\frac{k}{m}\right)^{-pu} \end{aligned}$$

If $\sigma_{[m]} \le T_m$, then $U_m \approx \left(\frac{\sigma_{[m]}}{T_m}\right)^{p\alpha} \le 1$.
If $\sigma_{[m]} > T_m$, then $[m] \ge \sigma^{-1}(T_m)$, and

$$\begin{aligned} U_m &\approx \left(\frac{m}{\sigma^{-1}(T_m)}\right)^{pu} + \sum_{k=\sigma^{-1}(T_m)+1}^{[m]} \left(\frac{T_m}{\sigma_{k-1}}\right)^\alpha \left(\frac{k}{m}\right)^{-pu} \\ &\approx \left(\frac{m}{\sigma^{-1}(T_m)}\right)^{pu}. \end{aligned}$$

So, $(U_m)_{m>1}$ is bounded if and only there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \le \beta_2 \sigma^{-1}(T_m).$$

6. Pinsker weights : $\lambda_k(m) = \left(1 - \left(\frac{\sigma_k}{T_m}\right)^\alpha\right)_+$**.**
Assumption (A_1) is fulfilled if and only if there exists $\beta_1 > 0$, such that

$$\forall\, m > 0, \quad \beta_1 T_m \leq \sigma_{[m+1]}.$$

For any $m > 1$, if $\sigma_{[m]} \leq T_m$, then

$$\begin{aligned} U_m &= \sum_{k=1}^{[m]} \left(\left(\frac{\sigma_k}{T_m}\right)^\alpha - \left(\frac{\sigma_{k-1}}{T_m}\right)^\alpha\right) \left(\frac{\sigma_k}{T_m}\right)^{\alpha(p-1)} \left(\frac{k}{m}\right)^{-pu} \\ &\approx \sum_{k=1}^{[m]} \left(\frac{\sigma_k}{T_m}\right)^{p\alpha} \left(\frac{k}{m}\right)^{-pu} \\ &\approx \left(\frac{\sigma_{[m]}}{T_m}\right)^{p\alpha} \leq 1. \end{aligned}$$

If $\sigma_{[m]} > T_m$, then $[m] \geq \sigma^{-1}(T_m)$, and

$$\begin{aligned} U_m &= \sum_{k=1}^{\sigma^{-1}(T_m)-1} \left(\left(\frac{\sigma_k}{T_m}\right)^\alpha - \left(\frac{\sigma_{k-1}}{T_m}\right)^\alpha\right) \left(\frac{\sigma_k}{T_m}\right)^{\alpha(p-1)} \left(\frac{k}{m}\right)^{-pu} \\ &\qquad + \left(1 - \left(\frac{\sigma(\sigma^{-1}(T_m) - 1)}{T_m}\right)^\alpha\right) \left(\frac{\sigma^{-1}(T_m)}{m}\right)^{-pu} \\ &\approx \left(\frac{m}{\sigma^{-1}(T_m)}\right)^{pu}. \end{aligned}$$

So, $(U_m)_{m>1}$ is bounded if and only if there exists $\beta_2 > 0$, such that

$$\forall\, m > 1, \quad m \leq \beta_2 \sigma^{-1}(T_m).$$

So, using the hypotheses on $m \to T_m$, assumptions (A_1) and (A_2) of Theorem 5.2 are checked if and only if there exist two positive constants β_1 and β_2, such that $\forall\, m \geq 1$,

$$T_m \leq \beta_1^{-1} \sigma_{[m+1]},$$

$$m \leq \beta_2 \sigma^{-1}(T_m).$$

with $\beta_1 \geq 1$ for projection weights.

□

Chapitre 6

Bayesian modelization of sparse sequences and maxisets for Bayes rules

In this chapter, our aim is to estimate sparse sequences in the framework of the heteroscedastic white noise model. To modelize sparsity, we consider a Bayesian model composed of a mixture of a heavy-tailed density and a point mass at zero. To evaluate the performance of the Bayes rules (the median or the mean of the posterior distribution), we exploit an alternative to the minimax setting developed in particular by Kerkyacharian and Picard : we evaluate the maxisets for each of these estimators. Using this approach, we compare the performance of Bayesian procedures with thresholding ones. Furthermore, the maxisets obtained can be viewed as weighted versions of weak l_q spaces that naturally modelize sparsity. This remark leads us to investigate the following problem : how can we choose the prior parameters to build typical realizations of weak l_q spaces ?

6.1 Introduction

In this chapter, we consider the heteroscedastic white noise model where we are given observations of an unknown sequence $\theta = (\theta_k)_{k\in\mathbb{N}^*}$, subject to noise :

$$x_k = \theta_k + \varepsilon\sigma_k\xi_k, \quad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1), \quad k \in \mathbb{N}^*. \tag{6.1}$$

This heteroscedastic white noise model has extensively been used, for instance, as a sequence space framework for statistical inverse problems where $(\sigma_k^{-2})_{k\in\mathbb{N}^*}$ are the eigenvalues of a known operator (see the references cited in section 6.2.1). In the wavelet context, Johnstone and Silverman (1997) and Johnstone (1999) proved that the heteroscedastic model appears as an appropriate large sample limit of the classical non parametric regression problem with correlated errors. So, along this chapter, we assume that $\sigma = (\sigma_k)_{k\in\mathbb{N}*}$ is a known sequence of positive real numbers.

In this chapter, we suppose that the sequence θ to be estimated is sparse. It means that only a small proportion of the θ_k's are non negligible. When the θ_k's are wavelet coefficients, this assumption is natural since the underlying property of wavelets is that a function can be well approximated by a function with a relatively small proportion of nonzero wavelet coefficients.

The first goal of this chapter is to discuss the modelization of sparsity throughout a Bayesian approach. Most of the authors consider Bayes models based on normal distributions. For instance, Johnstone and Silverman (1998) following Abramovich, Sapatinas and Silverman (1998) and Clyde, Parmigiani and Vidakovic (1998) consider a mixture of a normal component and a point mass at zero for the wavelet coefficients. Chipman, Kolaczyk and McCulloch (1997) impose a mixture of two Gaussian distributions with different variances for negligible and non negligible wavelet coefficients. But Johnstone and Silverman (2002a, 2002b) did not use Gaussian distributions and showed the advantages from the minimax point of view in considering heavy-tailed distributions. In this chapter, the θ_k's are not necessarily wavelet coefficients but θ is assumed to be sparse and we consider the following Bayesian model : the θ_k's are supposed to be independent and

$$\theta_k \sim w_{k,\varepsilon}\gamma_{k,\varepsilon}(\theta_k) + (1 - w_{k,\varepsilon})\delta_0(\theta_k), \tag{6.2}$$

with $w_{k,\varepsilon} \in (0, 1)$. The nonzero part of the prior, $\gamma_{k,\varepsilon}$, is assumed to be the dilation of a fixed symmetric, positive and unimodal density γ. We shall see below that for the theoretical approach of this chapter, it is preferable to consider heavy-tailed densities γ.

To estimate the θ_k's, we consider the Bayes rules associated with our Bayesian model. Most of the time, statisticians use the posterior mean, but Abramovich, Sapatinas and Silverman (1998) considered the posterior median. For their Bayesian model that has the same form as (6.2), and unlike the posterior mean, the posterior median is a true thresholding rule. In the wavelet context, they showed several simulated examples for which this approach improves most of the traditional methods. If from the practical point of view, the median seems preferable to the mean, what happens under a theoretical approach ? From the minimax point of view, Johnstone and Silverman (2002a, 2002b) showed that the posterior median of their Bayes model achieves optimal rates of convergence under Besov body constraints and for l_q-losses, with $0 < q \leq 2$. If the posterior mean is used, optimal rates are also achieved but only if $1 \leq q \leq 2$. This provides some theoretical justification for preferring the posterior median over the posterior mean.

But to compare estimators from the theoretical point of view, we can adopt an alternative to the minimax approach : the maxiset point of view. This approach consists in investigating the

maximal set (called maxiset) where an estimation procedure achieves a specific rate of convergence. Not only does it provide a functional set directly connected to the procedure, but it is less pessimistic than the minimax setting. Cohen, DeVore, Kerkyacharian and Picard (2000) and Kerkyacharian and Picard (2000, 2002) applied this theory for two well known efficient procedures : wavelet thresholding and local bandwidth selection. Under some conditions, since the maxiset for the first one is included into the maxiset for the second one, they could conclude that local bandwidth selection is at least as good as the thresholding procedure. For the model (6.1), in chapter 5, we establish that linear estimators are outperformed by thresholding ones. So, it seems relevant to investigate the maxisets for Bayesian procedures associated with the median or the mean of the posterior distribution. It will enable us to compare the performance of these estimators with more traditional procedures by comparing their respective maxisets. It will be the second goal of this chapter.
Now, let us describe our results. First, we establish asymptotic properties of the median and the mean of the posterior distribution of θ_k conditioned to x_k, noted respectively $\hat{\theta}_k^{b_1}(x_k)$ and $\hat{\theta}_k^{b_2}(x_k)$. For this, we assume that $w_{k,\varepsilon} = w_\varepsilon$ depends only on ε. Moreover, we assume that the tails of γ are exponential or heavier. We prove that $\hat{\theta}_k^{b_1}(x_k)$ is a thresholding rule. The posterior mean does not have this thresholding property, but is a shrinkage rule (see Propositions 6.2, 6.3 and 6.4). Then, the maxisets for these procedures are easily available if for each of them, we assume in addition that θ_k is estimated by 0, if k is large enough. We note that the hypotheses on the tails of γ are essential to get maxisets as large as possible and the maximal set of sequences where the risk of our Bayes rules achieves a given rate of convergence is the intersection of two spaces denoted $B^\eta_{p,\infty}$ and $wl_{p,q}(\sigma)$ (see section 6.2.2). The first one is necessary to impose minimal regularity on the θ_k's for the large values of k. The spaces $wl_{p,q}(\sigma)$ can be viewed as weighted versions of weak l_q spaces. These spaces appeared in approximation theory and coding and have been studied by DeVore (1989), Johnstone (1994), Donoho and Johnstone (1996) or Cohen, DeVore and Hochmuth (2000). As noted by DeVore (1989), if $q < 2$, weak l_q spaces can be viewed as collections of functions on $[0,1]$ that can be approximated in $L_2([0,1])$ at rate N^{-s}, $s = 1/q - 1/2$. In statistics, these spaces appeared in the maxiset framework and for instance when Kerkyacharian and Picard (2000) considered the thresholding rules associated with the universal threshold $\lambda_{k,\varepsilon} = \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}$. Actually, the main result of this chapter is the following. As far as the maxiset point of view is concerned, and roughly under the same conditions, both Bayesian procedures achieve exactly the same performance as the thresholding one. We can then claim that each of the Bayesian procedures outperforms the linear algorithm. Thus, our Bayesian model and $wl_{p,q}(\sigma)$ spaces are connected throughout the maxiset approach. It is worthwhile to notice that the prior model has been constructed to modelize the sparsity of

the sequences to be estimated. But as recalled in section 6.2.2, $wl_{p,q}(\sigma)$ spaces are natural spaces to measure the sparsity of a sequence by controlling the proportion of non negligible θ_k's. In section 6.4, we give another illustration of the strong link between our Bayesian model and $wl_{p,q}(\sigma)$ spaces. Indeed, Theorem 6.7 provides some conditions over the prior parameters to ensure that a sequence coming from the Bayes model belongs to $wl_{p,q}(\sigma)$ almost surely. In particular, we exhibit the condition $\sup_{\lambda>0} \lambda^q \int_\lambda^{+\infty} \gamma(x)dx < \infty$, which means that the tails of γ cannot be heavier than those of a Pareto(q)-variable. Consequently, as Theorem 3.1 of chapter 3, this result illustrates the strong connections between Pareto(q)-distributions and $wl_{p,q}(\sigma)$ spaces.
The chapter is organized as follows. After defining the statistical model, Section 6.2 is devoted to the sequence spaces $B^\eta_{p,\infty}$ and $wl_{p,q}(\sigma)$ and their roles in the maxiset setting associated with the thresholding procedure. In section 6.3, we evaluate the maxisets obtained for our Bayesian procedures under the l_p-risk. Furthermore, from these results, we determine the maxisets obtained for Bayesian procedures associated with the L_p-risk, with $1 < p < \infty$. (see section 6.3.5). Section 6.4 investigates the relationships between the Bayesian model and $wl_{p,q}(\sigma)$ spaces. Finally, in section 6.5, we prove the results concerning the asymptotic properties of our Bayes rules.

6.2 Model and maxiset point of view

6.2.1 Heteroscedastic white noise model

In the following, we assume that we are given the model (6.1) where $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$ is a known sequence of positive real numbers. This heteroscedastic white noise model appears as a generalization of the classical white noise model (for which, we have $\forall\ k \in \mathbb{N}^*$, $\sigma_k = 1$) often considered by statisticians. We use the heteroscedastic white noise model when we have to estimate the solution of a linear operator equation $g = Af$, with noisy observations of g. For instance, the operator A can be integration, convolution, or the Radon transform. Often, to deal with such a problem, we exploit the singular value decomposition of A and the sequence $(\sigma_k^{-2})_{k\in\mathbb{N}^*}$ is then the eigenvalues sequence of the operator A^*A, where A^* is the adjoint of A. There is a considerable literature about statistical inverse problems. Let us cite Korostelev and Tsybakov (1993), Donoho (1995), Mair and Ruymgaart (1996), Golubev and Khasminskii (1999), Goldenshluger and Pereverzev (2000), Cavalier and Tsybakov (2000), Cavalier, Golubev, Picard and Tsybakov (2000).
In the wavelet context, Johnstone and Silverman (1997) and Johnstone (1999) showed that the heteroscedastic white noise model can also be used to represent direct observations with correlated structure. More precisely, let us assume that we are given the following non parametric

regression model :

$$Y_i = f(t_i) + e_i, \quad 1 \leq i \leq n, \tag{6.3}$$

where the e_i's are drawn from a stationary process. By studying the autocorrelation function of the errors, Johnstone and Silverman (1997) and Johnstone (1999) showed that under a good choice of the noise level at each resolution level, the model (6.1) appears as a good approximation of the model (6.3). To illustrate this statement, let us assume that $\lim_{k \to +\infty} k^{\alpha} \mathrm{cov}(e_t, e_{t+k}) = K < \infty$, with $0 < \alpha < 1$. In this case, the noise level at the resolution level j is proportional to $2^{-j(1-\alpha)/2}$.

Along this chapter, we consider estimators of the form $\hat{\theta}(x) = (\hat{\theta}_k(x_k))_{k \in \mathbb{N}^*}$, which risk is studied under l_p-norms with $1 \leq p < \infty$:

$$R_p(\hat{\theta}) = \left(\mathbb{E} \|\hat{\theta} - \theta\|_{l_p}^p \right)^{\frac{1}{p}} = \left(\sum_{k \in \mathbb{N}^*} \mathbb{E} |\hat{\theta}_k(x_k) - \theta_k|^p \right)^{\frac{1}{p}}..$$

6.2.2 Sequence spaces and sparsity

As explained in Introduction, the maxiset point of view has two starting points : a procedure of estimation and a rate of convergence. Its final outcome is the largest set of functions where the risk for the given procedure attains the given rate. Let us introduce the following sequence spaces often met in the maxiset approach, as illustrated by Theorem 6.1, and sections 6.3.3 and 6.3.4.

Definition 6.1. *For all $1 \leq p < \infty$ and $0 < \eta < \infty$, we set :*

$$B^{\eta}_{p,\infty} = \left\{ \theta = (\theta_k)_{k \in \mathbb{N}^*} : \quad \sup_{\lambda > 0} \lambda^{p\eta} \sum_{k \geq \lambda} |\theta_k|^p < \infty \right\},$$

and if q is a real number such that $0 < q < p$, we set

$$wl_{p,q}(\sigma) = \left\{ \theta = (\theta_k)_{k \in \mathbb{N}^*} : \quad \sup_{\lambda > 0} \lambda^{q} \sum_{k} \mathbb{1}_{\,|\theta_k| > \lambda \sigma_k} \sigma_k^p < \infty \right\}.$$

We can easily prove the following result (see Kerkyacharian and Picard (2000)) :

Proposition 6.1. *For any $1 \leq p < \infty$, and $0 < q < p$,*

$$wl_{p,q}(\sigma) = \left\{ \theta = (\theta_k)_{k \in \mathbb{N}^*} : \quad \sup_{\lambda > 0} \lambda^{q-p} \sum_{k} \mathbb{1}_{\,|\theta_k| \leq \lambda \sigma_k} |\theta_k|^p < \infty \right\}.$$

In the maxiset approach, we use the spaces $B^{\eta}_{p,\infty}$ to control the θ_k's for the large values of k. As for the spaces $wl_{p,q}(\sigma)$, they can be viewed as weighted wl_q spaces considered by Johnstone (1994), Donoho (1996), Donoho and Johnstone (1996) or Abramovich, Benjamini, Donoho and Johnstone (2000a). These weak versions of l_q spaces are defined as follows :

$$wl_q = \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^q \sum_k \mathbb{1}_{\,|\theta_k|>\lambda} < \infty \right\}.$$

But, we can note that if we order the components of a sequence θ according to their size :

$$|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(n)} \cdots,$$

then

$$\theta \in wl_q \iff \sup_n n^{\frac{1}{q}} |\theta|_{(n)} < \infty.$$

To control the proportion of large θ_k's, and as suggested by Abramovich, Benjamini, Donoho and Johnstone (2000a), it is natural to use the previous lower bound. Then, as wl_q spaces, the weighted weak l_q spaces $wl_{p,q}(\sigma)$ appear as natural spaces to measure the sparsity of a signal. Theorem 6.7 of section 6.4 illustrates this last remark. In the following section, we recall the maxisets obtained for thresholding rules in the framework of the heteroscedastic white noise model.

6.2.3 Thresholding rules

In this section, under the model (6.1), we estimate each θ_k by using thresholding rules. An estimator $\hat{\theta}^t_k(x_k)$ of θ_k will be called a thresholding rule if there exists $t_k > 0$ such that $x \longrightarrow \hat{\theta}^t_k(x)$ is antisymmetric, increasing on $\mathbb{R}$ and

$$\hat{\theta}^t_k(x_k) = 0 \iff |x_k| \leq t_k.$$

More precisely, we focus on thresholding rules associated with the universal threshold $\lambda_{k,\varepsilon} = \sigma_k \varepsilon \sqrt{\log(1/\varepsilon)}$ (see Donoho and Johnstone (1994)) : for all $\varepsilon > 0$, we assume that we are given a real number $\Lambda_\varepsilon > 0$ only depending on ε and tending to $+\infty$ when ε tends to 0, and we set :

$$\hat{\theta}^t_k(x_k) = \begin{cases} x_k \mathbb{1}_{\,|x_k|\geq\kappa_*\lambda_{k,\varepsilon}} & \text{if } k < \Lambda_\varepsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where κ_* is a constant. Kerkyacharian and Picard (2000) have studied the maxisets for this procedure. They obtained the following result for $\hat{\theta}^t = (\hat{\theta}^t_k)_{k\in\mathbb{N}^*}$:

Theorem 6.1. *Let $1 \leq p < \infty$ be a fixed real number and $0 < r < \infty$. We suppose that*

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad \Lambda_\varepsilon = \left(\varepsilon \sqrt{\log(1/\varepsilon)}\right)^{-r},$$

where ε_0 is such that $\varepsilon_0 \sqrt{\log(1/\varepsilon_0)} = 1$, and there exists a positive constant T, such that $\forall\, 0 < \varepsilon \leq \varepsilon_0$,

$$\varepsilon^{\frac{\kappa_*^2}{16}} \log(1/\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \leq T.$$

Let q be a fixed positive real number such that $q < p$.
Then, if $\kappa_* \geq \sqrt{2p}$, there exists a positive constant C such that

$$\forall\, 0 < \varepsilon \leq \varepsilon_0, \quad R_p(\hat{\theta}^t) \leq C \left(\varepsilon \sqrt{\log(1/\varepsilon)}\right)^{(1-q/p)}$$

$$\iff \theta \in wl_{p,q}(\sigma) \cap B_{p,\infty}^{\frac{1}{r}(1-q/p)}.$$

In the next section, we consider a Bayesian model and we evaluate the maxisets respectively for the median and the mean of the posterior distribution.

6.3 Bayesian procedures

6.3.1 The prior model

The sequence θ to be estimated is supposed to be sparse. With this in mind, we wish to estimate each θ_k by using Bayes rules and we consider the following Bayesian model : we suppose that we are given a fixed unimodal density γ, assumed to be positive on $\mathbb{R}$, symmetric about 0 and such that there exist two positive constants M and M_1 such that

$$\sup_{\theta \geq M_1} \left| \frac{d}{d\theta} \log \gamma(\theta) \right| = M < \infty. \qquad (H_1)$$

Then, we assume that the θ_k's are independent and $\forall\, k \in \mathbb{N}^*$,

$$\theta_k \sim w_{k,\varepsilon} \gamma_{k,\varepsilon}(\theta_k) + (1 - w_{k,\varepsilon}) \delta_0(\theta_k), \qquad (M_1)$$

where $\forall\, \theta \in \mathbb{R}$,

$$\gamma_{k,\varepsilon}(\theta) = s_{k,\varepsilon} \gamma(s_{k,\varepsilon} \theta), \quad s_{k,\varepsilon} = (\varepsilon \sigma_k)^{-1},$$

and $w_{k,\varepsilon}$ is a real number lying in $(0, 1)$. The hypothesis (H_1) will be useful to determine maxisets for Bayes rules. It means that the tails of γ have to be exponential or heavier. Indeed, under (H_1), we have :

$$\forall\, u \geq M_1, \quad \gamma(u) \geq \gamma(M_1) \exp(-M(u - M_1)).$$

In the minimax approach, the priors of Johnstone and Silverman (2002a, 2002b) also verified (H_1), which is not surprising, since the maxiset approach is not very far from the minimax setting.
Johnstone and Silverman (2002a, 2002b) proved that if γ is a normal density, whatever the value of $w_{k,\varepsilon}$, the posterior median satisfies

$$|\hat{\theta}^{b_1}(x_k)| \leq (1-\alpha)|x_k|,$$

for $\alpha > 0$ and the same inequality holds for the posterior mean. It yields for $\theta_k > 0$,

$$\mathbb{E}|\hat{\theta}_k^{b_1}(x_k) - \theta_k|^p \geq \frac{1}{2}\alpha^p \theta_k^p.$$

Moreover, in the case where γ has tails equivalent to $\exp(-C|t|^\lambda)$ for some $\lambda \in (1,2)$, Johnstone and Silverman showed that for large θ_k,

$$|\hat{\theta}_k^{b_1}(x_k) - \theta_k| \geq C|\theta_k|^{\lambda-1}.$$

Thus (H_1) cannot be essentially relaxed without obtaining smaller maxisets.
Actually, till the end of section 6.3, we shall assume that $w_{k,\varepsilon} = w_\varepsilon$ depends only on ε. In the following, we shall note ϕ the density of ξ_1 and ϕ_k the density of $\varepsilon\sigma_k\xi_k$. Throughout this chapter, we shall use :

$$\pi_\varepsilon = (1-w_\varepsilon)w_\varepsilon^{-1},$$

and we shall assume that

1. $\varepsilon \longrightarrow \pi_\varepsilon$ is continuous,
2. $\inf_{\varepsilon>0} \pi_\varepsilon > 1$,
3. $\pi_1 = \exp(1)$,
4. $\pi_\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} +\infty$,
5. $\varepsilon\sqrt{\log \pi_\varepsilon} \overset{\varepsilon\to 0}{\longrightarrow} 0$.

6.3.2 Bayes rules

The posterior distribution of θ_k conditioned to x_k is

$$\gamma_{k,\varepsilon}^{\phi}(\theta_k|x_k) = \frac{\phi_k(x_k-\theta_k)[w_\varepsilon \gamma_{k,\varepsilon}(\theta_k) + (1-w_\varepsilon)\delta_0(\theta_k)]}{w_\varepsilon \int_{-\infty}^{+\infty} \phi_k(x_k-\theta)\gamma_{k,\varepsilon}(\theta)d\theta + (1-w_\varepsilon)\phi_k(x_k)}. \tag{6.4}$$

In this chapter, we consider the Bayes rules associated with the l_1-loss (the median of the posterior distribution) and with the l_2-loss (the mean of the posterior distribution). For all $\varepsilon > 0$, we assume that we are given a real number $\Lambda_\varepsilon > 1$ depending only on ε and tending to $+\infty$ when ε tends to 0. We estimate each θ_k by $\hat{\theta}_k^{b_1}(x_k)$ or by $\hat{\theta}_k^{b_2}(x_k)$ defined by the following procedure.

- If $k < \Lambda_\varepsilon$, $\hat{\theta}_k^{b_1}(x_k)$ (respectively $\hat{\theta}_k^{b_2}(x_k)$) is the median (respectively the mean) of the posterior distribution of θ_k given x_k.
- If $k \geq \Lambda_\varepsilon$, $\hat{\theta}_k^{b_1}(x_k) = \hat{\theta}_k^{b_2}(x_k) = 0$.

Let us recall the properties these Bayes rules verify. For further details, see Johnstone and Silverman (2002a, 2002b).

Proposition 6.2. *For all $k \in \mathbb{N}^*$, since γ is symmetric, positive and unimodal, we have for $i \in \{1, 2\}$,*

- *for any (x_k^1, x_k^2), such that $x_k^1 \leq x_k^2$, $\hat{\theta}_k^{b_i}(x_k^1) \leq \hat{\theta}_k^{b_i}(x_k^2)$,*
- *for any x_k, $\hat{\theta}_k^{b_i}(-x_k) = -\hat{\theta}_k^{b_i}(x_k)$,*
- *for any $x_k \geq 0$, $0 \leq \hat{\theta}_k^{b_i}(x_k) \leq x_k$,*
- *for any (x_k, θ_k), $|\hat{\theta}_k^{b_i}(x_k) - \theta_k| \leq \max(|\theta_k|; |x_k - \theta_k|)$.*

6.3.3 Maxisets for the posterior median

In this section, we assume that (H_1) is true. Before evaluating the maxiset for $\hat{\theta}^{b_1} = (\hat{\theta}_k^{b_1})_{k \in \mathbb{N}^*}$, we need to study the asymptotic behavior of $\hat{\theta}_k^{b_1}(x_k)$. We have the following result :

Proposition 6.3. *For all $k < \Lambda_\varepsilon$, $\hat{\theta}_k^{b_1}(x_k)$ is a thresholding rule and there exists $t = t(\pi_\varepsilon)$ such that*

$$\hat{\theta}_k^{b_1}(x_k) = 0 \iff s_{k,\varepsilon}|x_k| \leq t(\pi_\varepsilon),$$

where the threshold $t(\pi_\varepsilon)$ verifies for π_ε large enough, $t(\pi_\varepsilon) \geq \sqrt{2\log(\pi_\varepsilon)}$ and

$$\lim_{\pi_\varepsilon \to +\infty} \frac{t(\pi_\varepsilon)}{\sqrt{2\log(\pi_\varepsilon)}} = 1.$$

Furthermore, there exists a positive constant C such that

$$\limsup_{\pi_\varepsilon \to +\infty} |s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_1}(x_k)| \mathbb{1}_{|s_{k,\varepsilon}x_k| > 2t(\pi_\varepsilon)} \leq C.$$

The proof of this proposition is given in Appendix. Now, since $\forall\, k < \Lambda_\varepsilon$, $\hat{\theta}_k^{b_1}$ is a thresholding rule, it is easy to evaluate the maxiset for this procedure. On one hand, we have :

Theorem 6.2. *Let $0 < r < \infty$ and $1 \leq p < \infty$ be two fixed real numbers. We suppose that $\forall\, \varepsilon > 0$,*

$$\Lambda_\varepsilon = (\varepsilon\sqrt{\log \pi_\varepsilon})^{-r},$$

and there exist two positive constant T_1 and T_2, such that $\forall\, \varepsilon > 0$,

$$\varepsilon^{-p}\ \pi_\varepsilon^{-1}(\log \pi_\varepsilon)^{-\frac{1}{2}-\frac{p}{2}} \leq T_1,$$
$$\pi_\varepsilon^{-\frac{1}{8}}(\log \pi_\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \leq T_2.$$

Let q be a fixed positive real number such that $q < p$.
If $\theta \in B_{p,\infty}^{\frac{1}{r}(1-q/p)} \cap wl_{p,q}(\sigma)$, then there exists a positive constant C such that

$$\forall\, \varepsilon > 0, \quad R_p(\hat{\theta}^{b_1}) \leq C(\varepsilon\sqrt{\log \pi_\varepsilon})^{1-q/p}.$$

The proof of this theorem that uses Proposition 6.3 is omitted since it is inspired by the proof of Theorem 3.1 of Kerkyacharian and Picard (2000) and is very similar to the proof of Theorem 6.4 of section 6.3.4.
On the other hand, we have :

Theorem 6.3. *Let $0 < r < \infty$ and $1 \leq p < \infty$ be two fixed real numbers. We suppose that $\forall\, \varepsilon > 0$,*

$$\Lambda_\varepsilon = (\varepsilon\sqrt{\log \pi_\varepsilon})^{-r}.$$

Let q be a fixed positive real number such that $q < p$.
If there exists a positive constant C such that

$$\forall\, \varepsilon > 0, \quad R_p(\hat{\theta}^{b_1}) \leq C(\varepsilon\sqrt{\log \pi_\varepsilon})^{1-q/p},$$

then $\theta \in B_{p,\infty}^{\frac{1}{r}(1-q/p)} \cap wl_{p,q}(\sigma)$.

Proof : Since $\pi_1 = \exp(1)$, we have $\Lambda_1 = 1$ and $\|\theta\|_{l_p}^p \leq C$.
For any $\varepsilon > 0$,

$$\sum_{k\geq\Lambda_\varepsilon} |\theta_k|^p \leq \mathbb{E}\|\hat{\theta}^{b_1} - \theta\|_{l_p(\mu)}^p \leq C\Lambda_\varepsilon^{\frac{-(p-q)}{r}},$$

and since $\varepsilon \to \Lambda_\varepsilon$ is continuous with $\lim_{\varepsilon\to 0} \Lambda_\varepsilon = +\infty$, we have :

$$\sup_{\lambda>0} \lambda^{\frac{p}{r}(1-\frac{q}{p})} \sum_{k\geq\lambda} |\theta_k|^p \leq C,$$

and $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})}$.
In the following, we shall use $t(\pi_\varepsilon)$ (noted t if there is no risk of confusion) introduced in Proposition 6.3 and the following inequality : for any $m > 0$, and with $Z \sim \mathcal{N}(0,1)$, for ε small enough,

$$\begin{aligned} \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \geq \frac{t}{m}\right) &\leq \mathbb{P}\left(|Z| \geq \frac{1}{m}\sqrt{2\log \pi_\varepsilon)}\right) \\ &\leq K(\log \pi_\varepsilon)^{-1/2}\ \pi_\varepsilon^{-1/m^2}, \end{aligned} \tag{6.7}$$

where K depends only on m. So, for $k < \Lambda_\varepsilon$, and for ε small enough,

$$\begin{aligned} |\theta_k|^p \mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} &= |\theta_k|^p \mathbb{E}(\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} \mathbb{1}_{\ |s_{k,\varepsilon}x_k| \geq t}) + \mathbb{E}(|\theta_k|^p \mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} \mathbb{1}_{\ |s_{k,\varepsilon}x_k| < t}) \\ &\leq |\theta_k|^p \mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} \mathbb{P}(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \geq \frac{t}{2}) + \mathbb{E}|\hat{\theta}_k^{b_1}(x_k) - \theta_k|^p \\ &\leq \frac{1}{2}|\theta_k|^p \mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} + \mathbb{E}|\hat{\theta}_k^{b_1}(x_k) - \theta_k|^p. \end{aligned}$$

Therefore, for ε small enough,

$$\begin{aligned} \sum_k |\theta_k|^p \mathbb{1}_{\ |s_{k,\varepsilon}\theta_k| \leq \frac{t}{2}} &\leq 2\mathbb{E}\|\hat{\theta}^{b_1} - \theta\|^p_{l_p(\mu)} \\ &\leq 2C(\varepsilon\sqrt{\log \pi_\varepsilon})^{p-q}. \end{aligned}$$

Since $\varepsilon \to \pi_\varepsilon$ is continuous, it implies that there exists $\lambda_0 > 0$ such that

$$\sup_{\lambda < \lambda_0} \lambda^{q-p} \sum_k |\theta_k|^p \mathbb{1}_{\ |\theta_k| \leq \sigma_k \lambda} < \infty.$$

Since $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})}$,

$$\sup_{\lambda \geq \lambda_0} \lambda^{q-p} \sum_k |\theta_k|^p \mathbb{1}_{\ |\theta_k| \leq \sigma_k \lambda} \leq \lambda_0^{q-p} \sum_k |\theta_k|^p < \infty.$$

We have proved that $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$.

□

We can conclude that the spaces $B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$ appear as maximal spaces where this Bayesian procedure attains specific rates of convergence.

6.3.4 Maxisets for the posterior mean

In this section, we still consider the model (M_1) and we suppose that (H_1) is true. Then, we have the following results for $\hat{\theta}^{b_2} = (\hat{\theta}_k^{b_2})_{k\in\mathbb{N}^*}$:

Proposition 6.4. *Let $k < \Lambda_\varepsilon$ be fixed. There exists two functions ε_1 and ε_2 bounded on $[1, +\infty)$, such that $\varepsilon_1(x) \overset{x\to\infty}{\longrightarrow} 0$, and*

$$\hat{\theta}_k^{b_2}(x_k) = x_k \times \frac{1 + \varepsilon_1(s_{k,\varepsilon}x_k)}{1 + \pi_\varepsilon \phi(s_{k,\varepsilon}x_k)\gamma(s_{k,\varepsilon}x_k)^{-1}\varepsilon_2(s_{k,\varepsilon}x_k)}. \tag{6.8}$$

If $t = t(\pi_\varepsilon)$ is the threshold introduced in Proposition 6.3, there exists a positive constant C such that

$$\limsup_{\pi_\varepsilon\to+\infty} |s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_2}(x_k)|\mathbb{1}_{\ |s_{k,\varepsilon}x_k|>2t(\pi_\varepsilon)} \leq C.$$

Proposition 6.4 is proved in Appendix.

Remark 6.1. *For all $k < \Lambda_\varepsilon$, unlike $\hat{\theta}_k^{b_1}$, $\hat{\theta}_k^{b_2}$ is not a thresholding rule since $\hat{\theta}_k^{b_2}(x_k) \neq 0$ if $x_k \neq 0$ and we can easily prove that under (H_1),*

$$\lim_{s_{k,\varepsilon}x_k\to 0} \left(x_k \times \frac{\int_{-\infty}^{+\infty} u^2 \exp(-\frac{1}{2}u^2)\gamma(u)du}{\int_{-\infty}^{+\infty} \exp(-\frac{1}{2}u^2)\gamma(u)du + \pi_\varepsilon} \right)^{-1} \times \hat{\theta}_k^{b_2}(x_k) = 1.$$

Remark 6.2. *By using the results of Proposition 6.4, we have for π_ε large enough, $|\hat{\theta}_k^{b_2}(s_{k,\varepsilon}^{-1} \times \frac{t}{2})| \leq \pi_\varepsilon^{-\frac{1}{2}} s_{k,\varepsilon}^{-1}\sqrt{\log \pi_\varepsilon}$.*

As in section 6.3.3, we prove the following results :

Theorem 6.4. *Let $0 < r < \infty$ and $1 \leq p < \infty$ be two fixed real numbers. We suppose that $\forall\, \varepsilon > 0$,*

$$\Lambda_\varepsilon = (\varepsilon\sqrt{\log \pi_\varepsilon})^{-r},$$

and there exist two positive constants T_1 and T_2, such that $\forall\, \varepsilon > 0$,

$$\varepsilon^{-p}\ \pi_\varepsilon^{-\frac{1}{4}}(\log \pi_\varepsilon)^{-\frac{1}{2}-\frac{p}{2}} \leq T_1,$$

$$\pi_\varepsilon^{-\frac{1}{32}}(\log \pi_\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \leq T_2. \tag{6.10}$$

Let q be a fixed positive real number such that $q < p$.
If $\theta \in B_{p,\infty}^{\frac{1}{r}(1-q/p)} \cap wl_{p,q}(\sigma)$, then there exists a positive constant C such that

$$\forall\, \varepsilon > 0, \quad R_p(\hat{\theta}^{b_2}) \leq C(\varepsilon\sqrt{\log \pi_\varepsilon})^{1-q/p}.$$

Proof : In the following, the notation K will keep designating a constant independent of ε that may be different at each line. Let $t = t(\pi_\varepsilon)$ be the threshold introduced in Proposition 6.3. For all $k < \Lambda_\varepsilon$,

$$\mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p = A + B + C,$$

with

$$A = \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}},$$
$$B = \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t},$$

and

$$C = \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k| > 2t}.$$

For the first term, we have, using Remark 6.2 and (6.7),

$$\begin{aligned} A &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}} \\ &\le 2^{p-1}\mathbb{E}|\hat{\theta}_k^{b_2}(x_k)|^p \mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}} + 2^{p-1}|\theta_k|^p \mathbb{E}\mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}} \\ &\le K s_{k,\varepsilon}^{-p} \pi_\varepsilon^{-\frac{p}{2}} (\log \pi_\varepsilon)^{\frac{p}{2}} + 2^{p-1}|\theta_k|^p \, \mathbb{E}[\mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| > t}] \\ &\quad + 2^{p-1}|\theta_k|^p \, \mathbb{E}[\mathbb{1}_{|s_{k,\varepsilon}x_k| \le \frac{t}{2}} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| \le t}] \\ &\le K\sigma_k^p \varepsilon^p \pi_\varepsilon^{-\frac{p}{2}} (\log \pi_\varepsilon)^{\frac{p}{2}} + 2^{p-1}|\theta_k|^p \, \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge \frac{t}{2}\right) + 2^{p-1}|\theta_k|^p \, \mathbb{1}_{|s_{k,\varepsilon}\theta_k| \le t} \\ &\le K\sigma_k^p \varepsilon^p \pi_\varepsilon^{-\frac{p}{2}} (\log \pi_\varepsilon)^{\frac{p}{2}} + K|\theta_k|^p \pi_\varepsilon^{-\frac{1}{4}} (\log \pi_\varepsilon)^{-\frac{1}{2}} + 2^{p-1}|\theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}\theta_k| \le t}. \end{aligned}$$

The second term can be split into $B = B_1 + B_2 + B_3$, with

$$\begin{aligned} B_1 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| > 3t}, \\ B_2 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{\frac{t}{4} < |s_{k,\varepsilon}\theta_k| \le 3t}, \\ B_3 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| \le \frac{t}{4}}. \end{aligned}$$

Using Proposition 6.2 and (6.7),

$$\begin{aligned} B_1 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| > 3t} \\ &\le \mathbb{E}|x_k - \theta_k|^p \mathbb{1}_{|x_k - \theta_k| \ge |\theta_k|} \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| > 3t} \\ &\qquad + |\theta_k|^p \mathbb{E}\mathbb{1}_{|x_k - \theta_k| < |\theta_k|} \mathbb{1}_{\frac{t}{2} < |s_{k,\varepsilon}x_k| \le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k| > 3t} \\ &\le \left(\mathbb{E}|x_k - \theta_k|^{2p}\right)^{\frac{1}{2}} \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge t\right)^{\frac{1}{2}} + |\theta_k|^p \, \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge t\right) \\ &\le K\varepsilon^p \sigma_k^p \pi_\varepsilon^{-\frac{1}{2}} (\log \pi_\varepsilon)^{-\frac{1}{4}} + K|\theta_k|^p \pi_\varepsilon^{-1} (\log \pi_\varepsilon)^{-\frac{1}{2}}. \end{aligned}$$

$$
\begin{aligned}
B_2 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2}<|s_{k,\varepsilon}x_k|\le 2t} \mathbb{1}_{\frac{t}{4}<|s_{k,\varepsilon}\theta_k|\le 3t} \\
&\le \mathbb{E}|x_k - \theta_k|^p \mathbb{1}_{|x_k-\theta_k|\ge|\theta_k|} \mathbb{1}_{\frac{t}{4}<|s_{k,\varepsilon}\theta_k|} + |\theta_k|^p\, \mathbb{E}\mathbb{1}_{|x_k-\theta_k|<|\theta_k|} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le 3t} \\
&\le K\varepsilon^p\sigma_k^p\, \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge \frac{t}{4}\right)^{\frac{1}{2}} + |\theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le 3t} \\
&\le K\varepsilon^p\sigma_k^p \pi_\varepsilon^{-\frac{1}{32}} (\log \pi_\varepsilon)^{-\frac{1}{4}} + |\theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le 3t}.
\end{aligned}
$$

$$
\begin{aligned}
B_3 &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{\frac{t}{2}<|s_{k,\varepsilon}x_k|\le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} \\
&\le \mathbb{E}|x_k - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|\ge\frac{t}{4}} \\
&\le \left(\mathbb{E}|x_k - \theta_k|^{2p}\right)^{\frac{1}{2}} \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge \frac{t}{4}\right)^{\frac{1}{2}} \\
&\le K\varepsilon^p\sigma_k^p \pi_\varepsilon^{-\frac{1}{32}} (\log \pi_\varepsilon)^{-\frac{1}{4}}.
\end{aligned}
$$

For the last term, we denote $\tau(k,\varepsilon) = s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_2}(x_k)$ and recall that ξ_k is the noise. Using Proposition 6.4 and (6.7), we have

$$
\begin{aligned}
C &= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \\
&= \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le t} + \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t} \\
&\le s_{k,\varepsilon}^{-p}\mathbb{E}\left(|\xi_k - \tau(k,\varepsilon)|^{2p} \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t}\right)^{\frac{1}{2}} \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge t\right)^{\frac{1}{2}} \\
&\qquad + s_{k,\varepsilon}^{-p}\mathbb{E}|\xi_k - \tau(k,\varepsilon)|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t} \\
&\le K s_{k,\varepsilon}^{-p}\mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge t\right)^{\frac{1}{2}} + K s_{k,\varepsilon}^{-p} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t} \\
&\le K\varepsilon^p\sigma_k^p \left(\pi_\varepsilon^{-\frac{1}{2}} (\log \pi_\varepsilon)^{-\frac{1}{4}} + \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t}\right).
\end{aligned}
$$

Finally, for ε small enough, and $\forall\, k < \Lambda_\varepsilon$,

$$
\begin{aligned}
\mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p \le\; & K[\varepsilon^p\sigma_k^p \pi_\varepsilon^{-\frac{1}{32}} (\log \pi_\varepsilon)^{-\frac{1}{4}} + \varepsilon^p\sigma_k^p \mathbb{1}_{|\theta_k|>\varepsilon\sigma_k\sqrt{2\log\pi_\varepsilon}} \\
& + |\theta_k|^p \mathbb{1}_{|\theta_k|\le 4\varepsilon\sigma_k\sqrt{2\log\pi_\varepsilon}} + |\theta_k|^p \pi_\varepsilon^{-\frac{1}{4}} (\log \pi_\varepsilon)^{-\frac{1}{2}}].
\end{aligned}
$$

We conclude by observing that

$$
\begin{aligned}
\mathbb{E}\|\hat{\theta}^{b_2} - \theta\|_{l_p(\mu)}^p &= \sum_{k<\Lambda_\varepsilon} \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p + \sum_{k\ge\Lambda_\varepsilon} |\theta_k|^p \\
&\le K(\varepsilon\sqrt{\log \pi_\varepsilon})^{p-q},
\end{aligned}
$$

since $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$ and $\Lambda_\varepsilon = (\varepsilon\sqrt{\log \pi_\varepsilon})^{-r}$.

□

As in section 6.3.3, we have a converse result, but unlike Theorem 6.3, we need to control the size of the σ_k's.

Theorem 6.5. *Let $0 < r < \infty$ and $1 \le p < \infty$ be two fixed real numbers. We suppose that $\forall\, \varepsilon > 0$,*

$$\Lambda_\varepsilon = (\varepsilon\sqrt{\log \pi_\varepsilon})^{-r}.$$

Let q be a fixed positive real number such that $q < p$.
If there exists a positive constant C such that

$$\forall\, \varepsilon > 0, \quad R_p(\hat\theta^{b_2}) \le C(\varepsilon\sqrt{\log \pi_\varepsilon})^{1-q/p},$$

then $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$ as soon as there exists a positive constant T such that

$$\pi_\varepsilon^{-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \le T. \tag{6.11}$$

Proof : To prove that $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})}$, we refer the reader to the proof of Theorem 6.3. Then, we want to show that

$$\sup_{\lambda>0} \lambda^{q-p} \sum_k |\theta_k|^p \mathbb{1}_{\,|\theta_k|\le\lambda\sigma_k} < \infty.$$

For this, we still use the threshold $t(\pi_\varepsilon)$ of Proposition 6.3. Using (6.7), for all $k < \Lambda_\varepsilon$, and for ε small enough,

$$\begin{aligned}
|\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} &= |\theta_k|^p \mathbb{E}(\mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} \mathbb{1}_{\,|s_{k,\varepsilon}x_k|\ge\frac{t}{2}}) + |\theta_k|^p \mathbb{E}(\mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} \mathbb{1}_{\,|s_{k,\varepsilon}x_k|<\frac{t}{2}}) \\
&\le |\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}}\, \mathbb{P}\left(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| \ge \frac{t}{4}\right) \\
&\quad + |\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} \mathbb{P}\left(|s_{k,\varepsilon}x_k| < \frac{t}{2}\right) \\
&\le \frac{1}{2}|\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} + |\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}} \mathbb{P}\left(|s_{k,\varepsilon}x_k| < \frac{t}{2}\right).
\end{aligned}$$

By using Remark 6.2, we have for ε small enough,

$$|\hat\theta_k^{b_2}(s_{k,\varepsilon}^{-1}\frac{t}{2})| \le \pi_\varepsilon^{-\frac{1}{2}} s_{k,\varepsilon}^{-1}\sqrt{\log \pi_\varepsilon}.$$

Therefore,

$$\begin{aligned}
\sum_k |\theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le \frac{t}{4}} &\le 2\sum_k |\theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le \frac{t}{4}} \mathbb{P}\left(|s_{k,\varepsilon}x_k| < \frac{t}{2}\right) \\
&\le 2^p \sum_k \mathbb{E}[|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p + |\hat{\theta}_k^{b_2}(x_k)|^p]\mathbb{1}_{|s_{k,\varepsilon}x_k|<\frac{t}{2}} \\
&\le 2^p \sum_k \mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p + 2^p \sum_{k<\Lambda_\varepsilon} \mathbb{E}|\hat{\theta}_k^{b_2}(x_k)|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|<\frac{t}{2}} \\
&\le 2^p\, \mathbb{E}\|\hat{\theta}^{b_2} - \theta\|^p_{l_p(\mu)} + 2^p \sum_{k<\Lambda_\varepsilon} s_{k,\varepsilon}^{-p}\pi_\varepsilon^{-\frac{p}{2}}(\log \pi_\varepsilon)^{\frac{p}{2}} \\
&\le 2^p\, C(\varepsilon\sqrt{\log \pi_\varepsilon})^{p-q} + 2^p\varepsilon^p\pi_\varepsilon^{-\frac{p}{2}}\log(\pi_\varepsilon)^{\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p.
\end{aligned}$$

Consequently, under condition (6.11), for ε small enough,

$$\sum_k |\theta_k|^p \mathbb{1}_{|\theta_k|\le\frac{1}{4}\sigma_k\varepsilon\sqrt{\log\pi_\varepsilon}} \le 2^p(C+T)(\varepsilon\sqrt{\log \pi_\varepsilon})^{p-q}.$$

Using the same arguments as for the proof of Theorem 6.3, the last inequality implies that $\theta \in wl_{p,q}(\sigma)$.

□

One more time, we can conclude that the spaces $B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$ appear as maximal spaces where this Bayesian procedure attains specific rates of convergence.

Let us end this section by comparing the different procedures involved in this chapter. We recall that unlike $\hat{\theta}_k^{b_1}$, $\hat{\theta}_k^{b_2}$ has not the advantage of being a thresholding rule ($k < \Lambda_\varepsilon$) and this explains the differences between the hypotheses that are necessary to determine the respective maxisets associated with our Bayesian procedures. For instance, this explains why, unlike $\hat{\theta}^{b_1}$, if $\hat{\theta}^{b_2}$ achieves the given rate of convergence, we need a condition on the σ_k's to prove that $\theta \in B_{p,\infty}^{\frac{1}{r}(1-\frac{q}{p})} \cap wl_{p,q}(\sigma)$. But, we can note that condition (6.11) is less restrictive than condition (6.10) since $p/2 \ge 1/2 > 1/32$. Furthermore, to obtain the upper bound for $R_p(\hat{\theta}^{b_2})$, we use a decomposition in eleven terms for $\mathbb{E}|\hat{\theta}_k^{b_2}(x_k) - \theta_k|^p$. Since it is a thresholding rule, the corresponding decomposition for $\hat{\theta}_k^{b_1}$ is simpler. This explains why assumptions of Theorem 6.4 are a bit more restrictive than assumptions of Theorem 6.2. Actually, to obtain assumptions of Theorem 6.4, we just have to replace π_ε with $\pi_\varepsilon^{1/4}$ in the assumptions of Theorem 6.2. But, since we consider a rate of the form $\varepsilon\sqrt{\log\pi_\varepsilon}$ without focusing on the optimal constant, the Bayesian procedures achieve exactly the same performance from the maxiset point of view. When π_ε is a power of ε, then, by using Theorem 6.1, we can compare our Bayesian procedures with the thresholding one.

We can conclude that each of them achieves the same performance as the thresholding one. Finally, since linear estimates are outperformed by thresholding ones (see chapter 5), they are also outperformed by our Bayesian procedures.

6.3.5 L_p-risk for Bayesian procedures $(1 < p < \infty)$

In this section, we estimate functions of $L_p(D)$, where $D = [0,1]^d$ or $D = \mathbb{R}^d$. For this purpose, we exploit a wavelet basis of $L_2(D)$ noted $\mathcal{B} = \{\psi_k, k \in \mathbb{N}^*\}$. More precisely, we assume that $(\psi_k)_{k\in\mathbb{N}^*}$ is the wavelet-tensor product constructed on compactly supported wavelets (see Meyer (1992)). So if $1 < p < \infty$, Meyer (1992) proved that $\mathcal{B}$ is an unconditional basis of $L_p(D)$, which means that

- for any $f \in L_p(D)$, there exists a unique sequence θ such that $f = \sum_k \theta_k \psi_k$,
- there exists an absolute constant K such that if $\forall\, k \in \mathbb{N}^*$, $|\theta_k| \leq |\theta'_k|$, then

$$\left\| \sum_k \theta_k \psi_k \right\|_{L_p} \leq K \left\| \sum_k \theta'_k \psi_k \right\|_{L_p}.$$

Furthermore, we assume that $\{\sigma_k \psi_k,\ k \in \mathbb{N}^*\}$ verifies the following superconcentration inequality : for any $0 < r_1 < \infty$, there exists a constant $C(p, r_1)$ such that for all $F \subset \mathbb{N}^*$,

$$\left\| \left[\sum_{k\in F} |\sigma_k \psi_k|^{r_1} \right]^{\frac{1}{r_1}} \right\|_{L_p} \leq C(p, r_1) \left\| \sup_{k\in F} |\sigma_k \psi_k| \right\|_{L_p}.$$

Remark 6.3. *Kerkyacharian and Picard (2001) proved that if the σ_k's depend only on the resolution level and if at level j, the noise level is proportional to 2^{bj}, with $b > -\frac{d}{2}$, then $\{\sigma_k \psi_k,\ k \in \mathbb{N}^*\}$ verifies the superconcentration inequality.*

Under these hypotheses, we can exhibit the maxiset associated with either of our Bayesian procedure for the L_p-risk $(1 < p < \infty)$. For any $\eta > 0$, $0 < q < p$, we introduce :

$$B^{\eta}_{p,\infty}(\mathcal{B}) = \left\{ f = \sum_k \theta_k \psi_k : \quad \sup_{\lambda>0} \lambda^{\eta} \| \sum_{k\geq\lambda} \theta_k \psi_k \|_{L_p} < \infty \right\},$$

$$wl_{p,q}(\sigma)(\mathcal{B}) = \left\{ f = \sum_k \theta_k \psi_k : \quad \sup_{\lambda>0} \lambda^{q} \sum_k \mathbb{1}_{\,|\theta_k|>\lambda\sigma_k} \sigma_k^p \|\psi_k\|^p_{L_p} < \infty \right\}.$$

Under (H_1), we have :

Theorem 6.6. *Let $0 < r < \infty$ be a fixed real number. We suppose that*

$$\forall\, \varepsilon > 0, \quad \Lambda_\varepsilon = (\varepsilon \sqrt{\log \pi_\varepsilon})^{-r},$$

and there exist two positive constants T_1, and T_2 such that for any $\varepsilon > 0$,

$$\varepsilon^{-p} \left[\pi_\varepsilon^{-1} (\log \pi_\varepsilon)^{-\frac{1}{2}} \right]^{\frac{1}{2}\min(p;2)} (\log \pi_\varepsilon)^{-\frac{p}{2}} \leq T_1,$$

and

$$\pi_\varepsilon^{-\frac{1}{8}} (\log \pi_\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \|\psi_k\|_{L_p}^p \leq T_2.$$

Let q be a fixed positive real number such that $q < p$.
Then, under the model (6.1), there exists a positive constant C such that

$$\forall\, \varepsilon > 0, \quad \left(\varepsilon \sqrt{\log \pi_\varepsilon} \right)^{-(1-q/p)} \left[\mathbb{E} \| \sum_k (\hat{\theta}_k^{b_1}(x_k) - \theta_k) \psi_k \|_{L_p}^p \right]^{1/p} \leq C$$
$$\iff f = \sum_k \theta_k \psi_k \in wl_{p,q}(\sigma)(\mathcal{B}) \cap B_{p,\infty}^{\frac{1}{r}(1-q/p)}(\mathcal{B}).$$

The analogous result for the procedure associated with the posterior mean is obtained if we assume that for any $\varepsilon > 0$,

$$\varepsilon^{-p} \left[\pi_\varepsilon^{-\frac{1}{4}} (\log \pi_\varepsilon)^{-\frac{1}{2}} \right]^{\frac{1}{2}\min(p;2)} (\log \pi_\varepsilon)^{-\frac{p}{2}} \leq T_3,$$

and

$$\pi_\varepsilon^{-\frac{1}{32}} (\log \pi_\varepsilon)^{-\frac{1}{4}-\frac{p}{2}} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \|\psi_k\|_{L_p}^p \leq T_4,$$

where T_3 and T_4 are two positive constants.

Remark 6.4. *The restriction $1 < p < \infty$ is due to the fact that there is no unconditional basis if $p \notin (1, \infty)$.*

Proof : One more time, we only give the proof for the procedure associated with the mean. We recall that $\mathcal{B}$ is an unconditional basis of $L_p(D)$ if and only if there exists $M > 0$ such that for any set $F \subset \mathbb{N}^*$, and any choice of the coefficients c_k's

$$M^{-1} \left\| \sum_{k\in F} c_k \psi_k \right\|_{L_p} \leq \left\| \left(\sum_{k\in F} |c_k \psi_k|^2 \right)^{\frac{1}{2}} \right\|_{L_p} \leq M \left\| \sum_{k\in F} c_k \psi_k \right\|_{L_p}. \tag{6.12}$$

Furthermore, since $\{\sigma_k\psi_k,\ k\in\mathbb{N}^*\}$ verifies a superconcentration inequality, there exist two positive constants c_p and C_p such that for any $F\subset\mathbb{N}^*$, we have :

$$c_p\int\sum_{k\in F}|\sigma_k\psi_k|^p\le\int\left(\sum_{k\in F}|\sigma_k\psi_k|^2\right)^{\frac{p}{2}}\le C_p\int\sum_{k\in F}|\sigma_k\psi_k|^p. \tag{6.13}$$

In the following, the notation K will keep designating a constant independent of ε that may be different at each line. We use $t=t(\pi_\varepsilon)$ the threshold introduced in Proposition 6.3 and the results of Proposition 6.2. Let us assume that $f\in wl_{p,q}(\sigma)(\mathcal{B})\cap B_{p,\infty}^{\frac{1}{r}(1-q/p)}(\mathcal{B})$. Then, $\mathbb{E}\|\sum_k(\hat\theta_k^{b_2}-\theta_k)\psi_k\|_{L_p}^p$ is bounded by $K\sum_{i=1}^{11}A_i$, with the A_i's defined as follows.

$$A_1=\|\sum_{k\ge\Lambda_\varepsilon}\theta_k\psi_k\|_{L_p}^p\le K\left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{(p-q)},$$

since $f\in B_{p,\infty}^{\frac{1}{r}(1-q/p)}(\mathcal{B})$.

$$\begin{aligned}A_2&=\mathbb{E}\|\sum_{k<\Lambda_\varepsilon}\theta_k\psi_k\mathbb{1}_{|s_{k,\varepsilon}x_k|\le\frac{t}{2}}\mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t}\|_{L_p}^p\\&\le K\mathbb{E}\int(\sum_{k<\Lambda_\varepsilon}\theta_k^2\psi_k^2\mathbb{1}_{|s_{k,\varepsilon}x_k-s_{k,\varepsilon}\theta_k|\ge\frac{t}{2}})^{p/2},\end{aligned}$$

by using (6.12).

If $p\le 2$, by using (6.7), the Jensen inequality and (6.12),

$$\begin{aligned}A_2&\le K\int(\sum_{k<\Lambda_\varepsilon}\theta_k^2\psi_k^2\mathbb{P}(|s_{k,\varepsilon}x_k-s_{k,\varepsilon}\theta_k|\ge t/2))^{p/2}\\&\le K(\pi_\varepsilon^{-1/4}(\log\pi_\varepsilon)^{-1/2})^{p/2}\|\sum_{k<\Lambda_\varepsilon}\theta_k\psi_k\|_{L_p}^p.\end{aligned}$$

If $p\ge 2$, by using (6.7), the generalized Minkowski inequality and (6.12)

$$\begin{aligned}A_2&\le K\int(\sum_{k<\Lambda_\varepsilon}\theta_k^2\psi_k^2\mathbb{P}(|s_{k,\varepsilon}x_k-s_{k,\varepsilon}\theta_k|\ge t/2)^{2/p})^{p/2}\\&\le K\pi_\varepsilon^{-1/4}(\log\pi_\varepsilon)^{-1/2}\|\sum_{k<\Lambda_\varepsilon}\theta_k\psi_k\|_{L_p}^p.\end{aligned}$$

$$\begin{aligned}A_3&=\mathbb{E}\|\sum_{k<\Lambda_\varepsilon}\theta_k\psi_k\mathbb{1}_{|s_{k,\varepsilon}x_k|\le\frac{t}{2}}\mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le t}\|_{L_p}^p\\&\le K\int(\sum_{k<\Lambda_\varepsilon}\theta_k^2\psi_k^2\mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le t})^{p/2},\end{aligned}$$

by using (6.12).

$$\begin{aligned} A_4 &= \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} \hat{\theta}_k^{b_2}\psi_k \mathbb{1}_{\,|s_{k,\varepsilon}x_k|\leq \frac{t}{2}} \|_{L_p}^p \\ &\leq K\mathbb{E}\int (\sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2})^2 \mathbb{1}_{\,|s_{k,\varepsilon}x_k|\leq \frac{t}{2}} \psi_k^2)^{p/2} \\ &\leq K\varepsilon^p \pi_\varepsilon^{-p/2}(\log \pi_\varepsilon)^{p/2} \int (\sum_{k<\Lambda_\varepsilon} \sigma_k^2\psi_k^2)^{p/2} \\ &\leq K\varepsilon^p \pi_\varepsilon^{-p/2}(\log \pi_\varepsilon)^{p/2} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \|\psi_k\|_{L_p}^p, \end{aligned}$$

by using (6.12), Remark 6.2 and (6.13). Using (6.12), we have :

$$\begin{aligned} A_5 &= \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{\,\frac{t}{2}<|s_{k,\varepsilon}x_k|\leq 2t} \mathbb{1}_{\,\frac{t}{4}<|s_{k,\varepsilon}\theta_k|\leq 3t} \mathbb{1}_{\,|x_k-\theta_k|\geq|\theta_k|} \|_{L_p}^p \\ &\leq K\mathbb{E}\int (\sum_{k<\Lambda_\varepsilon} (x_k-\theta_k)^2\psi_k^2 \mathbb{1}_{\,|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|\geq \frac{t}{4}})^{p/2}. \end{aligned}$$

If $p \leq 2$,

$$\begin{aligned} A_5 &\leq K \int \sum_{k<\Lambda_\varepsilon} |\psi_k|^p \mathbb{E}(|x_k-\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|\geq \frac{t}{4}}) \\ &\leq K\varepsilon^p \pi_\varepsilon^{-\frac{1}{32}}(\log \pi_\varepsilon)^{-\frac{1}{4}} \int \sum_{k<\Lambda_\varepsilon} |\psi_k|^p \sigma_k^p, \end{aligned}$$

by using (6.7) and the Cauchy Schwartz inequality.
If $p \geq 2$, by using (6.7), the generalized Minkowski inequality and (6.13),

$$\begin{aligned} A_5 &\leq K \int (\sum_{k<\Lambda_\varepsilon} \psi_k^2 (\mathbb{E}|x_k-\theta_k|^p \mathbb{1}_{\,|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|\geq \frac{t}{4}})^{2/p})^{p/2} \\ &\leq K\varepsilon^p \pi_\varepsilon^{-\frac{1}{32}}(\log \pi_\varepsilon)^{-\frac{1}{4}} \sum_{k<\Lambda_\varepsilon} \int |\psi_k|^p \sigma_k^p. \end{aligned}$$

$$\begin{aligned} A_6 &= \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{\,\frac{t}{2}<|s_{k,\varepsilon}x_k|\leq 2t} \mathbb{1}_{\,\frac{t}{4}<|s_{k,\varepsilon}\theta_k|\leq 3t} \mathbb{1}_{\,|x_k-\theta_k|<|\theta_k|} \|_{L_p}^p \\ &\leq K \int (\sum_{k<\Lambda_\varepsilon} \theta_k^2\psi_k^2 \mathbb{1}_{\,|s_{k,\varepsilon}\theta_k|\leq 3t})^{p/2}, \end{aligned}$$

by using (6.12). By using (6.12), we have :

$$A_7 = \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{\frac{t}{2}<|s_{k,\varepsilon}x_k|\le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>3t} \mathbb{1}_{|x_k-\theta_k|\ge|\theta_k|}\|_{L_p}^p \le KA_5,$$

$$A_8 = \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{\frac{t}{2}<|s_{k,\varepsilon}x_k|\le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>3t} \mathbb{1}_{|x_k-\theta_k|<|\theta_k|}\|_{L_p}^p \le KA_2,$$

$$A_9 = \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{\frac{t}{2}<|s_{k,\varepsilon}x_k|\le 2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le\frac{t}{4}}\|_{L_p}^p \le KA_5.$$

$$\begin{aligned} A_{10} &= \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|\le t}\|_{L_p}^p \\ &\le K\mathbb{E}\int (\sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)^2\psi_k^2 \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|>t})^{p/2}, \end{aligned}$$

by using (6.12).
If $p \le 2$, by using (6.7), the Cauchy Schwartz inequality and Proposition 6.4,

$$\begin{aligned} A_{10} &\le K\int \sum_{k<\Lambda_\varepsilon} |\psi_k|^p \mathbb{E}|\hat{\theta}_k^{b_2} - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|>t} \\ &\le K\int \sum_{k<\Lambda_\varepsilon} |\psi_k|^p \varepsilon^p \sigma_k^p \mathbb{P}(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| > t)^{1/2} \\ &\le K\varepsilon^p \pi_\varepsilon^{-1/2} (\log \pi_\varepsilon)^{-1/4} \sum_{k<\Lambda_\varepsilon} \sigma_k^p \int |\psi_k|^p. \end{aligned}$$

If $p \ge 2$, by using the generalized Minkowski inequality, the Cauchy Schwartz inequality, (6.7), (6.13) and Proposition 6.4,

$$\begin{aligned} A_{10} &\le K\int (\sum_{k<\Lambda_\varepsilon} \psi_k^2 (\mathbb{E}|\hat{\theta}_k^{b_2} - \theta_k|^p \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k|>t})^{2/p})^{p/2} \\ &\le K\int (\sum_{k<\Lambda_\varepsilon} \psi_k^2 \sigma_k^2 \varepsilon^2 \mathbb{P}(|s_{k,\varepsilon}x_k - s_{k,\varepsilon}\theta_k| > t)^{1/p})^{p/2} \\ &\le K\varepsilon^p \pi_\varepsilon^{-1/2} (\log \pi_\varepsilon)^{-1/4} \sum_{k<\Lambda_\varepsilon} \int |\psi_k|^p \sigma_k^p. \end{aligned}$$

$$A_{11} = \mathbb{E}\| \sum_{k<\Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k)\psi_k \mathbb{1}_{|s_{k,\varepsilon}x_k|>2t} \mathbb{1}_{|s_{k,\varepsilon}\theta_k|>t}\|_{L_p}^p$$

$$\leq \quad K\mathbb{E}\int(\sum_{k<\Lambda_\varepsilon}(\hat{\theta}_k^{b_2}-\theta_k)^2\psi_k^2 \mathbb{1}_{\ |s_{k,\varepsilon}x_k|>2t}\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t})^{p/2}.$$

If $p \leq 2$, by using Proposition 6.4,

$$\begin{aligned} A_{11} &\leq K\int\sum_{k<\Lambda_\varepsilon}|\psi_k|^p\mathbb{E}|\hat{\theta}_k^{b_2}-\theta_k|^p\mathbb{1}_{\ |s_{k,\varepsilon}x_k|>2t}\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t} \\ &\leq K\varepsilon^p\sum_{k<\Lambda_\varepsilon}\int|\psi_k|^p\sigma_k^p\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t}. \end{aligned}$$

If $p \geq 2$, by using (6.13) and Proposition 6.4,

$$\begin{aligned} A_{11} &\leq K\int(\sum_{k<\Lambda_\varepsilon}\psi_k^2\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t}(\mathbb{E}|\hat{\theta}_k^{b_2}-\theta_k|^p\mathbb{1}_{\ |s_{k,\varepsilon}x_k|>2t})^{2/p})^{p/2} \\ &\leq K\varepsilon^p\int(\sum_{k<\Lambda_\varepsilon}\psi_k^2\sigma_k^2\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t})^{p/2} \\ &\leq K\varepsilon^p\sum_{k<\Lambda_\varepsilon}\int|\psi_k|^p\sigma_k^p\mathbb{1}_{\ |s_{k,\varepsilon}\theta_k|>t}. \end{aligned}$$

Finally, we conclude that

$$\limsup_{\varepsilon\to 0}\left\{\left(\varepsilon\sqrt{\log\pi_\varepsilon}\right)^{-(1-q/p)}\left[\mathbb{E}\|\sum_k(\hat{\theta}_k^{b_1}-\theta_k)\psi_k\|_{L_p}^p\right]^{1/p}\right\}<\infty,$$

by using the following result proved by Kerkyacharian and Picard (2000) : if $f=\sum_k\theta_k\psi_k\in wl_{p,q}(\sigma)(\mathcal{B})$, then $\forall\ \lambda>0$,

$$\int(\sum_k\theta_k^2\psi_k^2\mathbb{1}_{\ |\theta_k|\leq\sigma_k\lambda})^{p/2}\leq K\lambda^{p-q}.$$

Now, let us assume that

$$\limsup_{\varepsilon\to 0}\left\{\left(\varepsilon\sqrt{\log\pi_\varepsilon}\right)^{-(1-q/p)}\left[\mathbb{E}\|\sum_k(\hat{\theta}_k^{b_1}-\theta_k)\psi_k\|_{L_p}^p\right]^{1/p}\right\}<\infty.$$

To bound the following term, we just use (6.12) :

$$\begin{aligned} A_{12} &= \| \sum_{k \geq \Lambda_\varepsilon} \theta_k \psi_k \|_{L_p}^p \\ &= \mathbb{E} \| \sum_{k \geq \Lambda_\varepsilon} (\hat{\theta}_k^{b_2} - \theta_k) \psi_k \|_{L_p}^p \\ &\leq K \mathbb{E} \int (\sum_k (\hat{\theta}_k^{b_2} - \theta_k)^2 \psi_k^2)^{p/2} \\ &\leq K \mathbb{E} \| \sum_k (\hat{\theta}_k^{b_2} - \theta_k) \psi_k \|_{L_p}^p \\ &\leq K \left(\varepsilon \sqrt{\log (1/\varepsilon)} \right)^{p-q}, \end{aligned}$$

which proves that $f = \sum_k \theta_k \psi_k \in B_{p,\infty}^{\frac{1}{r}(1-q/p)}(\mathcal{B})$. Now, using (6.12),

$$\| \sum_k \theta_k \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} \theta_k| \leq \frac{t}{4}} \|_{L_p}^p \quad \leq \quad K(A_{12} + A_{13}),$$

with

$$\begin{aligned} A_{13} &= \| \sum_{k < \Lambda_\varepsilon} \theta_k \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} \theta_k| \leq \frac{t}{4}} \|_{L_p}^p \\ &\leq \mathbb{E} \| \sum_{k < \Lambda_\varepsilon} \theta_k \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} \theta_k| \leq \frac{t}{4}} \mathbb{1}_{\ |s_{k,\varepsilon} x_k| > \frac{t}{2}} \|_{L_p}^p + A_{14} \\ &\leq \frac{1}{2} \| \sum_{k < \Lambda_\varepsilon} \theta_k \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} \theta_k| \leq \frac{t}{4}} \|_{L_p}^p + A_{14} \\ &\leq \frac{1}{2} A_{13} + A_{14} \\ &\leq 2 A_{14}, \end{aligned}$$

for ε small enough (see the upper bound of A_2), and with

$$A_{14} = \mathbb{E} \| \sum_{k < \Lambda_\varepsilon} \theta_k \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} \theta_k| \leq \frac{t}{4}} \mathbb{1}_{\ |s_{k,\varepsilon} x_k| \leq \frac{t}{2}} \|_{L_p}^p.$$

But, using (6.12),

$$\begin{aligned} A_{14} &\leq K \mathbb{E} \| \sum_k (\hat{\theta}_k^{b_2} - \theta_k) \psi_k \|_{L_p}^p + K \mathbb{E} \| \sum_{k < \Lambda_\varepsilon} \hat{\theta}_k^{b_2} \psi_k \mathbb{1}_{\ |s_{k,\varepsilon} x_k| \leq \frac{t}{2}} \|_{L_p}^p \\ &\leq K \mathbb{E} \| \sum_k (\hat{\theta}_k^{b_2} - \theta_k) \psi_k \|_{L_p}^p + K A_4 \\ &\leq K \left(\varepsilon \sqrt{\log (1/\varepsilon)} \right)^{(p-q)}, \end{aligned}$$

which implies that

$$\|\sum_k \theta_k \psi_k \mathbb{1}_{|s_{k,\varepsilon}\theta_k| \leq \frac{t}{4}}\|_{L_p}^p \leq K \left(\varepsilon\sqrt{\log(1/\varepsilon)}\right)^{(p-q)}.$$

Now, we use Lemma 5.1 of Kerkyacharian and Picard (2000), which ends the proof of the Theorem.

□

One more time, when π_ε is a power of ε (i.e. for the rate $\varepsilon\sqrt{\log(1/\varepsilon)}$), and by using Theorems 5.1 and 5.2 of Kerkyacharian and Picard (2000), we can conclude that each of the Bayesian procedures achieves the same performance as the thresholding one for the L_p-risk.

6.4 Relationships between (M_1) and $wl_{p,q}(\sigma)$ spaces

In this chapter, our aim is to estimate sparse sequences, and we modelize sparsity throughout a Bayes approach, and more precisely by using the model (M_1). We noticed that under this model, the maxisets for the previous Bayes rules are defined by using $wl_{p,q}(\sigma)$ spaces. To some extent, this result is not surprising, since section 6.2.2 recalled that these spaces are weighted versions of weak l_p spaces that naturally modelize sparsity. Then, the model (M_1) and $wl_{p,q}(\sigma)$ spaces are connected via a maxiset approach. So, it is natural to wonder whether we can establish other more natural connections between our Bayesian approach to modelize sparsity and $wl_{p,q}(\sigma)$ spaces. The following result gives a positive answer.

Theorem 6.7. *Let us suppose that we are given $1 \leq p < \infty$ and $0 < q < p$. We still consider the model (M_1) with $\varepsilon = 1$. We note $w_k = w_{k,1}$ and $\forall\ \lambda \geq 0$, $\tilde{F}(\lambda) = 2\int_\lambda^{+\infty}\gamma(x)dx$. If there exists a constant C such that*

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda} \leq C^q \quad a.s.,$$

then

$$\sup_{\lambda>0} \lambda^q \tilde{F}(\lambda) \sum_k w_k \sigma_k^p \leq C^q.$$

Conversely, if there exists a constant C such that

$$\sup_{\lambda>0} \lambda^q \tilde{F}(\lambda) \sum_k w_k \sigma_k^p \leq C^q, \tag{6.14}$$

then

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda} < \infty \quad a.s.$$

Proof : If

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{\ |\theta_k|>\sigma_k\lambda} \leq C^q \quad a.s.,$$

we have :

$$\begin{aligned}
\sup_{\lambda>0} \lambda^q \tilde{F}(\lambda) \sum_k w_k \sigma_k^p &= \sup_{\lambda>0} \mathbb{E}\left(\lambda^q \sum_k \sigma_k^p \mathbb{1}_{\ |\theta_k|>\sigma_k\lambda}\right) \\
&\leq \mathbb{E}\left(\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{\ |\theta_k|>\sigma_k\lambda}\right) \\
&\leq C^q.
\end{aligned}$$

Conversely, let us suppose that (6.14) is true. To establish the required inequality, we exploit Theorem 0.3. of Marcus and Zinn (1984). Let $(r_k)_{k\geq 1}$ be a Rademacher sequence, independent of $(\theta_k)_{k\geq 1}$ and S_n the partial sum of the symmetrization of random variables $\left(\sigma_k^{p-q}|\theta_k|^q\right)_{k\geq 1}$:

$$S_n = \sum_{k=1}^n Y_k,$$

where

$$Y_k = r_k\ \sigma_k^{p-q}|\theta_k|^q.$$

We have

$$\begin{aligned}
\mathbb{E}(Y_k \mathbb{1}_{\ |Y_k|\leq 1}) &= 0, \\
\mathbb{P}(|Y_k| > 1) &= \mathbb{P}\left(\left|\frac{\theta_k}{\sigma_k}\right|^q > \sigma_k^{-p}\right) \\
&= w_k \tilde{F}(\sigma_k^{-p/q}) \\
&\leq C^q \left(\sum_k w_k \sigma_k^p\right)^{-1} w_k \sigma_k^p,
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}(Y_k \mathbb{1}_{\ |Y_k|\leq 1}) &= \mathbb{E}(Y_k^2 \mathbb{1}_{\ |Y_k|\leq 1}) \\
&= \sigma_k^{2p} \mathbb{E}\left[(\sigma_k^{-1}|\theta_k|)^{2q} \mathbb{1}_{\ \sigma_k^{-1}|\theta_k|\leq \sigma_k^{-p/q}}\right] \\
&= 2w_k \sigma_k^{2p} \int_0^{\sigma_k^{-p/q}} x^{2q}\gamma(x)dx \\
&\leq 2w_k \sigma_k^{2p} \int_0^{\sigma_k^{-p/q}} qx^{2q-1}\tilde{F}(x)dx \\
&\leq 2C^q \left(\sum_k w_k \sigma_k^p\right)^{-1} w_k \sigma_k^{2p} \int_0^{\sigma_k^{-p/q}} qx^{q-1}dx
\end{aligned}$$

$$\leq \quad 2C^q \left(\sum_k w_k \sigma_k^p \right)^{-1} \quad w_k \sigma_k^p.$$

Using the three series theorem, when n tends to $+\infty$, S_n converges with probability 1. Therefore, if ε is a fixed positive real number, and for any increasing sequence of positive real numbers $(b_n)_n$ with $\lim_{n\to+\infty} b_n = +\infty$,

$$\limsup_{n\to+\infty} \frac{1}{b_n} |S_n| \leq \varepsilon \quad \text{a.s.}$$

Obviously,

$$\limsup_{n\to+\infty} \sup_{1\leq j\leq n} \sup_{\lambda>0} \frac{1}{b_n} \lambda^q \Big| \sum_{k=1}^{j} \sigma_k^p \mathbb{E}(\mathbb{1}_{|\theta_k|>\sigma_k\lambda}) \Big| \leq \varepsilon \quad \text{a.s.}$$

Then, we can apply Theorem 0.3. of Marcus and Zinn (1984), which shows that

$$\limsup_{n\to+\infty} \frac{1}{b_n} \sup_{\lambda>0} \lambda^q \left| \sum_{k=1}^{n} \sigma_k^p \left(\mathbb{1}_{|\theta_k|>\sigma_k\lambda} - \mathbb{P}(|\theta_k| > \sigma_k\lambda) \right) \right| \quad \leq \quad 1160\ \varepsilon.$$

Since

$$\limsup_{n\to+\infty} \frac{1}{b_n} \sup_{\lambda>0} \lambda^q \sum_{k=1}^{n} \sigma_k^p \mathbb{P}(|\theta_k| > \sigma_k\lambda) \quad = \quad 0,$$

it yields

$$\limsup_{n\to+\infty} \frac{1}{b_n} \sup_{\lambda>0} \lambda^q \sum_{k=1}^{n} \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda} \quad \leq \quad 1160\ \varepsilon.$$

With $\varepsilon \to 0$, we have proved that for any increasing sequence of positive real numbers $(b_n)_n$ with $\lim_{n\to+\infty} b_n = +\infty$,

$$\lim_{n\to+\infty} \frac{1}{b_n} \sup_{\lambda>0} \lambda^q \sum_{k=1}^{n} \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda} \quad = \quad 0 \quad \text{a.s.}$$

If the random sequence $A_n = \sup_{\lambda>0} \lambda^q \sum_{k=1}^{n} \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda}$ were not bounded, we could construct an increasing function Φ, with $\lim_{n\to+\infty} \Phi(n) = +\infty$, such that $A_{\Phi(n)} > n$. By considering an increasing sequence $(b_n)_n$, with $b_{\Phi(n)} = n$, we obtain a contradiction. So, there exists a finite random variable Y, such that

$$\forall\, n \in \mathbb{N}^*, \quad A_n \leq Y, \quad \text{a.s.}$$

It implies that

$$\sup_{\lambda>0} \lambda^q \sum_k \sigma_k^p \mathbb{1}_{|\theta_k|>\sigma_k\lambda} < \infty \quad \text{a.s.}$$

The result is proved.

□

So, to ensure that a sequence coming from the Bayesian model (M_1) belongs to $wl_{p,q}(\sigma)$ almost surely, we should not consider densities γ having tails heavier than those of Pareto(q)-distributions. In the wavelet framework, with special values for the σ_k's, in chapter 3 we have already noted the strong connections between Pareto(q)-distributions and $wl_{p,q}(\sigma)$ spaces since least favorable priors for these spaces were built from Pareto(q) distributions.

6.5 Appendix : Proof of Propositions 6.3 and 6.4

In this section, the notation $O_x(1)$ will keep designating any function of x that is bounded when $x \to +\infty$. The notation $o_{\pi_\varepsilon}(1)$ will keep designating any function that is bounded by a function depending only on π_ε and that tends to 0 when π_ε tends to $+\infty$. We shall exploit the following lemma :

Lemma 6.1. *For any $x > 0$ and $0 < \tau < x$, let us define :*

$$K_\tau(x) = \int_\tau^x \exp(-\frac{1}{2}v^2)\gamma(x-v)dv,$$

and

$$I(x) = \int_0^{+\infty} \exp(-\frac{1}{2}v^2 + xv)\gamma(v)dv.$$

Under (H_1), there exist four positive constants M_2, M_3, C_1 and C_2 such that

$$M_2 \int_\tau^x \exp(-\frac{1}{2}v^2 - Mv)dv \leq K_\tau(x)\gamma(x)^{-1} \leq M_3 \int_\tau^{+\infty} \exp(-\frac{1}{2}v^2 + Mv)dv,$$

and

$$C_1 \leq \liminf_{x\to+\infty} I(x)\gamma(x)^{-1}\phi(x) \leq \limsup_{x\to+\infty} I(x)\gamma(x)^{-1}\phi(x) \leq C_2 < \infty.$$

Proof of Lemma 6.1 : Under (H_1), and since γ is positive, symmetric and unimodal, it is easy to show that there exist two constants M_2 and M_3 such that,

$$\forall\ (a,b) \in \mathbb{R}^2, \quad M_2 \exp(-M(|a-b|) \leq \gamma(a)\gamma(b)^{-1} \leq M_3 \exp(M|a-b|).$$

We immediately get the first inequality. Now, let us define $\forall\ x > 0$,

$$J(x) = \int_0^{+\infty} \exp(-\frac{1}{2}v^2)\gamma(x+v)dv.$$

As previously,

$$M_2\int_0^{+\infty}\exp(-\frac{1}{2}v^2-Mv)dv\leq J(x)\gamma(x)^{-1}\leq M_3\int_0^{+\infty}\exp(-\frac{1}{2}v^2+Mv)dv.$$

By simple computations, we have :

$$I(x)=\exp(\frac{1}{2}x^2)(J(x)+K_0(x)),$$

which implies the result.

□

Now, let us give the proof of Propositions 6.3 and 6.4.

Proof of Proposition 6.3 : Without loss of generality, we can assume that $x_k>0$. Then, using Proposition 6.2,

$$\begin{aligned}
& \hat{\theta}_k^{b_1}(x_k)=0 \\
\Longleftrightarrow\quad & \mathbb{P}(\theta_k>0|x_k)<\frac{1}{2} \\
\Longleftrightarrow\quad & 2w_\varepsilon\int_0^{+\infty}s_{k,\varepsilon}\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta< \\
& w_\varepsilon\int_{-\infty}^{+\infty}s_{k,\varepsilon}\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta+(1-w_\varepsilon)s_{k,\varepsilon}\phi(s_{k,\varepsilon}x_k) \\
\Longleftrightarrow\quad & \int_0^{+\infty}s_{k,\varepsilon}\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta \\
& \quad -\int_{-\infty}^{0}s_{k,\varepsilon}\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta<\pi_\varepsilon s_{k,\varepsilon}\phi(s_{k,\varepsilon}x_k).
\end{aligned}$$

Then $\hat{\theta}_k^{b_1}(x_k)=0 \iff s_{k,\varepsilon}x_k\leq t$, with t such that

$$\int_0^{+\infty}\exp(-\frac{1}{2}u^2)\gamma(u)(\exp(tu)-\exp(-tu))du=\pi_\varepsilon.$$

We have that t is a function depending only on π_ε and as $\pi_\varepsilon\longrightarrow+\infty$,

$$\int_0^{+\infty}\exp(-\frac{1}{2}u^2)\gamma(u)\exp(tu)du=\pi_\varepsilon(1+o_{\pi_\varepsilon}(1)).$$

Using Lemma 6.1, we have, for π_ε large enough,

$$\sqrt{2\log(\pi_\varepsilon)}\leq t(\pi_\varepsilon)\leq\sqrt{2\log(\pi_\varepsilon)}(1+o_{\pi_\varepsilon}(1)),$$

and the first point of Proposition 6.3 is proved.
For the second point, we assume that $s_{k,\varepsilon}x_k > 2t$, which implies that $\hat{\theta}_k^{b_1}(x_k) > 0$. Using (6.4), we have :

$$\begin{aligned}
\mathbb{P}(\theta_k \leq \hat{\theta}_k^{b_1}(x_k)|x_k) = \frac{1}{2} \quad &\Longleftrightarrow \quad 2w_\varepsilon \int_{-\infty}^{\hat{\theta}_k^{b_1}} \phi_k(x_k-\theta)\gamma_{k,\varepsilon}(\theta)d\theta + (1-w_\varepsilon)\phi_k(x_k) \\
&\qquad = w_\varepsilon \int_{-\infty}^{+\infty} \phi_k(x_k-\theta)\gamma_{k,\varepsilon}(\theta)d\theta \\
&\Longleftrightarrow \quad 2\int_{-\infty}^{s_{k,\varepsilon}\hat{\theta}_k^{b_1}} \exp(s_{k,\varepsilon}x_k u - \frac{1}{2}u^2)\gamma(u)du + \pi_\varepsilon \\
&\qquad = \int_{-\infty}^{+\infty} \exp(s_{k,\varepsilon}x_k u - \frac{1}{2}u^2)\gamma(u)du.
\end{aligned}$$

Using Lemma 6.1, since $s_{k,\varepsilon}x_k \geq 2t$, we prove easily that

$$\pi_\varepsilon \times I(s_{k,\varepsilon}x_k)^{-1} = o_{\pi_\varepsilon}(1).$$

Therefore,

$$2\int_{-\infty}^{s_{k,\varepsilon}\hat{\theta}_k^{b_1}} \exp(s_{k,\varepsilon}x_k u - \frac{1}{2}u^2)\gamma(u)du \quad = \quad I(s_{k,\varepsilon}x_k)(1+o_{\pi_\varepsilon}(1)).$$

By using again Lemma 6.1, it implies that there exists a positive constant V such that for π_ε large enough,

$$\int_{-\infty}^{s_{k,\varepsilon}\hat{\theta}_k^{b_1}-s_{k,\varepsilon}x_k} \exp(-\frac{1}{2}v^2)\gamma(v+s_{k,\varepsilon}x_k)dv \times \gamma(s_{k,\varepsilon}x_k)^{-1} \geq V$$

$$\Longleftrightarrow K_{s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_1}}(s_{k,\varepsilon}x_k)\gamma(s_{k,\varepsilon}x_k)^{-1} \geq V,$$

with the notations of Lemma 6.1. Finally, there exists a positive constant C such that

$$\limsup_{\pi_\varepsilon \to +\infty} |s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_1}(x_k)|\mathbb{1}_{|s_{k,\varepsilon}x_k|>2t(\pi_\varepsilon)} \leq C.$$

□

Proof of Proposition 6.4 : Without loss of generality, we can assume that $x_k > 0$. We have :

$$\begin{aligned}
\hat{\theta}_k^{b_2}(x_k) &= \int_{-\infty}^{+\infty} \theta \gamma_{k,\varepsilon}^{\phi}(\theta|x_k)d\theta \\
&= \frac{\int_{-\infty}^{+\infty} s_{k,\varepsilon}\theta\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta}{\int_{-\infty}^{+\infty} s_{k,\varepsilon}\phi(s_{k,\varepsilon}(x_k-\theta))s_{k,\varepsilon}\gamma(s_{k,\varepsilon}\theta)d\theta + \pi_\varepsilon s_{k,\varepsilon}\phi(s_{k,\varepsilon}x_k)} \\
&= \frac{1}{s_{k,\varepsilon}} \times \frac{\int_{-\infty}^{+\infty} u\exp(-\frac{1}{2}u^2 + s_{k,\varepsilon}x_k u)\gamma(u)du}{\int_{-\infty}^{+\infty} \exp(-\frac{1}{2}u^2 + s_{k,\varepsilon}x_k u)\gamma(u)du + \pi_\varepsilon}.
\end{aligned}$$

Let us set

$$I_1(x) = \int_{-\infty}^{+\infty} u\exp(-\frac{1}{2}u^2 + xu)\gamma(u)du,$$

and

$$I_2(x) = \int_{-\infty}^{+\infty} \exp(-\frac{1}{2}u^2 + xu)\gamma(u)du,$$

which implies that

$$\hat{\theta}_k^{b_2}(x_k) = \frac{1}{s_{k,\varepsilon}} \times \frac{I_1(s_{k,\varepsilon}x_k)}{I_2(s_{k,\varepsilon}x_k) + \pi_\varepsilon}. \tag{6.15}$$

On one hand, when $x \to +\infty$, using Lemma 6.1,

$$C_1 \leq \liminf_{x\to+\infty} I_2(x)\gamma(x)^{-1}\phi(x) \leq \limsup_{x\to+\infty} I_2(x)\gamma(x)^{-1}\phi(x) \leq C_2. \tag{6.16}$$

On the other hand, Lemma 6.1 yields,

$$\begin{aligned}
\exp(-\frac{1}{2}x^2)I_1(x) &= \int_{-\infty}^{+\infty} u\exp(-\frac{1}{2}(x-u)^2)\gamma(u)du \\
&= \int_{-\infty}^{+\infty} (v+x)\exp(-\frac{1}{2}v^2)\gamma(v+x)dv \\
&= x\exp(-\frac{1}{2}x^2)I_2(x) + \int_{-\infty}^{+\infty} v\exp(-\frac{1}{2}v^2)\gamma(v+x)dv.
\end{aligned}$$

But, it is easy to prove that

$$\int_0^{+\infty} v\exp(-\frac{1}{2}v^2)\gamma(v+x)dv = \gamma(x)O_x(1),$$

$$\int_{-x}^{0} v\exp(-\frac{1}{2}v^2)\gamma(v+x)dv = \gamma(x)O_x(1),$$

and obviously,

$$\lim_{x\to\infty} \gamma(x)^{-1}\int_{-\infty}^{-x} v\exp(-\frac{1}{2}v^2)\gamma(v+x)dv = 0.$$

Therefore,

$$\exp(-\frac{1}{2}x^2)I_1(x) = x\exp(-\frac{1}{2}x^2)I_2(x) + \gamma(x)O_x(1). \tag{6.17}$$

Using (6.16), we obtain equation (6.8). Now, let us prove the second point of Proposition 6.4. Let us suppose that $s_{k,\varepsilon}x_k \geq 2t(\pi_\varepsilon)$. Using (6.15) and (6.17),

$$\begin{aligned} 0 \leq s_{k,\varepsilon}x_k - s_{k,\varepsilon}\hat{\theta}_k^{b_2}(x_k) &= s_{k,\varepsilon}x_k - \frac{I_1(s_{k,\varepsilon}x_k)}{I_2(s_{k,\varepsilon}x_k) + \pi_\varepsilon} \\ &= \frac{\pi_\varepsilon s_{k,\varepsilon}x_k + I_2(s_{k,\varepsilon}x_k)T(s_{k,\varepsilon}x_k)}{I_2(s_{k,\varepsilon}x_k) + \pi_\varepsilon}, \end{aligned}$$

where T is a bounded function. But we suppose that $s_{k,\varepsilon}x_k \geq 2t(\pi_\varepsilon)$. So, (6.16) implies that for π_ε large enough

$$I_2(s_{k,\varepsilon}x_k) \geq C_1\gamma(s_{k,\varepsilon}x_k)\phi(s_{k,\varepsilon}x_k)^{-1}.$$

Therefore, by using again $s_{k,\varepsilon}x_k \geq 2t(\pi_\varepsilon)$,

$$\limsup_{\pi_\varepsilon \to \infty} \frac{\pi_\varepsilon s_{k,\varepsilon}x_k}{I_2(s_{k,\varepsilon}x_k)} < \infty,$$

which ends the proof of the proposition.

□

Bibliographie

Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and Statistics*, pages 5–14. Springer, New York.

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2000a). Adapting to unknown sprsity by controlling the false discovery rate. *Technical Report.*

Abramovich, F., Besbeas, P., and Sapatinas, T. (2002). Empirical bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis*, 39 :435–451.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4) :725–749.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (2000b). Stochastic expansions in an overcomplete wavelet dictionary. *Probab. Theory Related Fields*, 117(1) :133–144.

Bakushinskii, A. B. (1969). on the construction of regularizing algorithms under random noise. *Soviet Math. Doklady*, 189 :231–233.

Belitser, E. N. and Levit, B. Y. (1995). On minimax filtering over ellipsoids. *Math. Methods Statist.*, 4(3) :259–273.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300.

Bergh, J. and Löfström, J. (1976). *Interpolation spaces. An introduction.* Springer-Verlag, Berlin. Grundlehren der Mathematischen Wissenschaften, No. 223.

Bickel, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, 9(6) :1301–1309.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2) :181–237.

Birgé, L. (1985). Nonasymptotic minimax risk for Hellinger balls. *Probab. Math. Statist.*, 5(1) :21–29.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités : risque minimax. *Z. Wahrsch. Verw. Gebiete*, 47(2) :119–137.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6) :2384–2398.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(3) :627–641.

Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean. *Ann. Statist.*, 9(4) :870–878.

Cavalier, L. (1998). Asymptotically efficient estimation in a problem related to tomography. *Math. Methods Statist.*, 7(4) :445–456 (1999).

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2000). Oracle inequalities for inverse problems. *Preprint Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII.*

Cavalier, L. and Tsybakov, A. B. (2000). Sharp adaptation for inverse problems with random noise. *Preprint Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII.*

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1) :33–61 (electronic).

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92 :1413–1421.

Clyde, M. and George, E. I. (1998). Robust empirical bayes estimation in wavelets. *Technical Report.*

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(4) :681–698.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2) :391–401.

Cohen, A. (2000). Wavelet methods in numerical analysis. In *Handbook of numerical analysis, Vol. VII*, pages 417–711. North-Holland, Amsterdam.

Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1) :54–81.

Cohen, A., DeVore, R. A., and Hochmuth, R. (2000). Restricted nonlinear approximation. *Constr. Approx.*, 16(1) :85–113.

Cohen, A., DeVore, R. A., Kerkyacharian, G., and Picard, D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2) :167–191.

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.*, 46(4) :886–902.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41(7) :909–996.

Daubechies, I. (1992). *Ten lectures on wavelets.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Davis, G., Mallat, S., and Zhang, Z. (1994). Adaptive time-frequency approximations with matching pursuits. In *Wavelets : theory, algorithms, and applications (Taormina, 1993)*, pages 271–293. Academic Press, San Diego, CA.

DeVore, R. A. (1989). Degree of nonlinear approximation. In *Approximation theory VI, Vol. I (College Station, TX, 1989)*, pages 175–201. Academic Press, Boston, MA.

DeVore, R. A. (1998). Nonlinear approximation. In *Acta numerica, 1998*, pages 51–150. Cambridge Univ. Press, Cambridge.

DeVore, R. A., Konyagin, S. V., and Temlyakov, V. N. (1998). Hyperbolic wavelet approximation. *Constr. Approx.*, 14(1) :1–26.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation.* Springer-Verlag, Berlin.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.*, 1(1) :100–115.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2) :101–126.

Donoho, D. L. (1996). Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Anal.*, 3(4) :388–392.

Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455.

Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, 99(2) :277–303.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224.

Donoho, D. L. and Johnstone, I. M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli*, 2(1) :39–62.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3) :879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369. With discussion and a reply by the authors.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2) :508–539.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1997). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam*, pages 183–218. Springer, New York.

Farrell, R. H. (1967). On the lack of a uniformly consistent sequence of estimators of a density function in certain cases. *Ann. Math. Statist.*, 38 :471–474.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4) :731–747.

Goldenshluger, A. and Pereverzev, S. V. (2000). Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory Related Fields*, 118(2) :169–186.

Golubev, G. K. and Khasminskii, R. Z. (1999). A statistical approach to some inverse problems for partial differential equations. *Problemy Peredachi Informatsii*, 35(2) :51–66.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*. Springer-Verlag, New York.

Huang, H.-C. and Cressie, N. (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, 42(3) :262–276.

Ibragimov, I. A. and Khasminskii, R. Z. (1981). *Statistical estimation*. Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

Jansen, M., Malfait, M., and Bultheel, A. (1997). Generalized cross validation for wavelet thresholding. *Signal Processing*, 56 :33–44.

Johnstone, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical decision theory and related topics, V (West Lafayette, IN, 1992)*, pages 303–326. Springer, New York.

Johnstone, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems : adaptivity results. *Statist. Sinica*, 9(1) :51–83.

Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.*, 18(1) :251–280.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B*, 59(2) :319–351.

Johnstone, I. M. and Silverman, B. W. (1998). Empirical bayes approaches to mixture problems and wavelet regression. *Technical Report*.

Johnstone, I. M. and Silverman, B. W. (2002a). Empirical bayes selection of wavelet thresholds. *Technical Report*.

Johnstone, I. M. and Silverman, B. W. (2002b). Risk bounds for empirical bayes estimates of sparse sequences. *Technical Report*.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett.*, 13(1) :15–24.

Kerkyacharian, G. and Picard, D. (1993). Density estimation by kernel and wavelets methods : optimality of Besov spaces. *Statist. Probab. Lett.*, 18(4) :327–336.

Kerkyacharian, G. and Picard, D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2) :283–344. With comments, and a rejoinder by the authors.

Kerkyacharian, G. and Picard, D. (2001). Entropy, universal coding, approximation and bases properties. *Preprint Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII.*

Kerkyacharian, G. and Picard, D. (2002). Minimax or maxisets ? *Bernoulli*, 8(2) :219–253.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41 :495–502.

Koo, J.-Y. (1993). Optimal rates of convergence for nonparametric statistical inverse problems. *Ann. Statist.*, 21(2) :590–599.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction.* Springer-Verlag, New York.

Kress, R. (1989). *Linear integral equations.* Springer-Verlag, Berlin.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B*, 40(2) :113–146. With discussion.

Lepskii, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4) :645–659.

Lindenstrauss, J. and Tzafriri, L. (1977). *Classical Banach spaces. I.* Springer-Verlag, Berlin. Sequence spaces, Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 92.

Lorentz, G. G. (1950). Some new functional spaces. *Ann. of Math. (2)*, 51 :37–55.

Lorentz, G. G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.*, 72 :903–937.

Mair, B. A. and Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5) :1424–1444.

Mallat, S. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbf{r})$. *Trans. Amer. Math. Soc.*, 315(1) :69–87.

Mallat, S. (1998). *A wavelet tour of signal processing.* Academic Press Inc., San Diego, CA.

Marcus, M. B. and Zinn, J. (1984). The bounded law of the iterated logarithm for the weighted empirical distribution process in the non-i.i.d. case. *Ann. Probab.*, 12(2) :335–360.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3) :1269–1283.

Meyer, Y. (1992). *Wavelets and operators.* Cambridge University Press, Cambridge. Translated from the 1990 French original by D. H. Salinger.

Müller, P. and Vidakovic, B. (1995). Wavelet shrinkage with affine bayes rules with applications. *Technical Report.*

Müller, P. and Vidakovic, B., editors (1999). *Bayesian inference in wavelet-based models.* Springer-Verlag, New York.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, 58(2) :463–479.

Nemirovskii, A. S. (1986). Nonparametric estimation of smooth regression functions. *J. Comput. Syst. Sci.*, 23(6) :1–11.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6) :2399–2430.

Ogden, T. and Parzen, E. (1996a). Change-point approach to data analytic wavelet thresholding. *Statistics and Computing*, 6 :93–99.

Ogden, T. and Parzen, E. (1996b). Data dependent wavelet thresholding in non-parametric regression with change-point applications. *Computational Statistics and Data Analysis*, 22 :53–70.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4) :502–527. With comments and a rejoinder by the author.

Peetre, J. (1976). *New thoughts on Besov spaces.* Mathematics Department, Duke University, Durham, N.C. Duke University Mathematics Series, No. 1.

Pereverzev, S. and Schock, E. (1999). Error estimates for band-limited spherical regularization wavelets in an inverse problem of satellite geodesy. *Inverse Problems*, 15(4) :881–890.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. (russian). *Problems Inform. Transmission*, 16(2) :52–68.

Ruymgaart, F. H. (1992). Regularized operator inversion as an approach to minimax rate density estimation. *J. Math. Systems Estim. Control*, 2(3) :363–380.

Ruymgaart, F. H. (1993). A unified approach to inversion problems in statistics. *Math. Methods Statist.*, 2(2) :130–146.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics.* John Wiley & Sons Inc., New York.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B*, 47(1) :1–52. With discussion.

Steinberg, D. M. (1990). A Bayesian approach to flexible modeling of multivariable response functions. *J. Multivariate Anal.*, 34(2) :157–172.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4) :1040–1053.

Strasser, H. (1985). *Mathematical theory of statistics.* Walter de Gruyter & Co., Berlin. Statistical experiments and asymptotic decision theory.

Sudakov, V. N. and Khalfin, L. A. (1964). Statistical approach to ill-posed problems in mathematical physics. *Soviet Math. Doklady*, 157 :1094–1096.

Temlyakov, V. N. (1999). Greedy algorithms and m-term approximation with regard to redundant dictionaries. *J. Approx. Theory*, 98(1) :117–145.

Tsybakov, A. B. (2000). Adaptive estimation for inverse problems : a logarithmic effect in l_2. *Preprint Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII.*

van Rooij, A. C. M. and Ruymgaart, F. H. (1991). Regularized deconvolution on the circle and the sphere. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, pages 679–690. Kluwer Acad. Publ., Dordrecht.

van Rooij, A. C. M. and Ruymgaart, F. H. (1996). Asymptotic minimax rates for abstract linear estimators. *J. Statist. Plann. Inference*, 53(3) :389–402.

Vannucci, M. and Corradi, F. (1999). Modeling dependence in the wavelet domain. In *Bayesian inference in wavelet-based models*, pages 173–186. Springer, New York.

Vapnik, V. (1982). *Estimation of dependences based on empirical data.* Springer-Verlag, New York. Translated from the Russian by Samuel Kotz.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.*, 93(441) :173–179.

Vidakovic, B. and Ruggeri, F. (2001). BAMS method : theory and simulations. *Sankhyā Ser. B*, 63(2) :234–249. Special issue on wavelets.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4) :651–667.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40(3) :364–372.

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.*, 9(1) :146–156.

Wahba, G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B*, 20 :334–343.

Young, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika*, 64(2) :309–317.

Zhang, C.-H. (2002). General empirical bayes wavelet methods. *Technical Report.*

Résumé :

Dans le cadre d'une analyse par ondelettes, nous nous intéressons à l'étude statistique d'une classe particulière d'espaces de Lorentz : les espaces de Besov faibles qui apparaissent naturellement dans le contexte de la théorie maxiset. Avec des hypothèses de type "bruit blanc gaussien", nous montrons, grâce à des techniques bayésiennes, que les vitesses minimax des espaces de Besov forts ou faibles sont les mêmes. Les distributions les plus défavorables que nous exhibons pour chaque espace de Besov faible sont construites à partir des lois de Pareto et diffèrent en cela de celles des espaces de Besov forts. Grâce aux simulations de ces distributions, nous construisons des représentations visuelles des "ennemis typiques". Enfin, nous exploitons ces distributions pour bâtir une procédure d'estimation minimax, de type "seuillage" appelée ParetoThresh, que nous étudions d'un point de vue pratique. Dans un deuxième temps, nous nous plaçons sous le modèle hétéroscédastique de bruit blanc gaussien et sous l'approche maxiset, nous établissons la sous-optimalité des estimateurs linéaires par rapport aux procédures adaptatives de type "seuillage". Puis, nous nous interrogeons sur la meilleure façon de modéliser le caractère "sparse" d'une suite à travers une approche bayésienne. À cet effet, nous étudions les maxisets des estimateurs bayésiens classiques - médiane, moyenne - associés à une modélisation construite sur des densités à queues lourdes. Les espaces maximaux pour ces estimateurs sont des espaces de Lorentz, et coïncident avec ceux associés aux estimateurs de type "seuillage". Nous prolongeons de manière naturelle ce résultat en obtenant une condition nécessaire et suffisante sur les paramètres du modèle pour que la loi a priori se concentre presque sûrement sur un espace de Lorentz précis.

Mots-clés : Estimation adaptative, Ondelettes, Théorie minimax, Théorie maxiset, Vitesse de convergence, Sparsité, Modélisation bayésienne, Modèle de bruit blanc gaussien, Modèle hétéreoscédastique, Problème statistique inverse, Estimateur bayésien, Estimateur de type "seuillage", Estimateur linéaire, Espace de Lorentz, Espace de Besov faible, Espace de Besov fort, Distributions les plus défavorables, Loi de Pareto.

Abstract :

In the framework of a wavelet analysis, we study the statistical meaning of a special class of Lorentz spaces : the weak Besov spaces are naturally appearing in the maxiset theory. With 'Gaussian white noise' type assumptions, we show, using Bayesian tools, that the minimax rates associated with strong or weak Besov spaces are the same. We exhibit the least favorable priors for each weak Besov space. They are associated with Pareto distributions and highly differ from the priors of strong Besov spaces that are Gaussian. Using simulations of these distributions, we build visual representations of the 'typical enemies'. Finally, we exploit these distributions to build a minimax thresholding estimation procedure, called ParetoThresh, that we study from a practical point of view. Subsequently, we consider the heteroscedastic white noise model and under the maxiset approach, we prove that linear estimators are outperformed by adaptive thresholding ones. Finally, we investigate the best way to modelize the sparsity of a sequence throughout a Bayesian approach. For this purpose, we study the maxisets of the classical estimators - median, mean - associated with a model built on heavy tailed densities. The maximal spaces for these rules are Lorentz spaces, and coincide with maxisets associated with thresholding estimators. This result is reinforced by a necessary and sufficient condition on the parameters of the model in order to make sure that the prior is almost surely concentrated on a precise Lorentz space.

Key words : Adaptive estimation, Wavelets, Minimax theory, Maxiset theory, Rate of convergence, Sparsity, Bayesian model, Gaussian white noise model, Heteroscedastic model, Statistical inverse problem, Bayes rule, Thresholding rule, Linear estimator, Lorentz space, Weak Besov space, Strong Besov space, Least favorable priors, Pareto distribution.