Habilitation à diriger des recherches

Mathématiques

Vincent RIVOIRARD

Contributions à l'estimation non-paramétrique. Du rôle de la parcimonie en statistique à la calibration théorique et pratique d'estimateurs

Rapporteurs : Sara van de Geer
Pascal Massart
Alexandre Tsybakov

Soutenue le 7 décembre 2009 devant le jury composé de :
Lucien Birgé
Stéphane Mallat
Pascal Massart
Dominique Picard
Jean-Michel Poggi
Alexandre Tsybakov

Au moment de soutenir mon habilitation à diriger des recherches, je tiens en premier lieu à exprimer toute ma gratitude à l'égard de Dominique Picard et Pascal Massart qui me font le plaisir d'être présents aujourd'hui. Dominique a guidé mes premiers pas dans la recherche et me fait partager depuis dix ans son expérience. Toujours disponible, elle m'a témoigné un soutien constant et sa confiance a été d'une importance capitale dans mon travail. Les conseils de Pascal, jamais intrusifs, ont également joué un rôle essentiel depuis mon arrivée à Orsay. Les travaux de ce mémoire doivent beaucoup à ses idées pertinentes et à sa générosité scientifique.

Mes remerciements les plus chaleureux vont ensuite à Sara van de Geer et Sacha Tsybakov pour l'intérêt qu'ils ont manifesté pour mon travail en acceptant de rapporter mon habilitation. Merci également à Sacha pour sa participation à mon jury.

Je suis honoré de la présence de Lucien Birgé, Stéphane Mallat et Jean-Michel Poggi dans mon jury. Lucien a toujours porté une attention bienveillante à mes travaux. Par ses questions et ses remarques, il a été pour moi un interlocuteur des plus stimulants. J'ai également beaucoup appris de mes nombreuses discussions avec Jean-Michel. Il m'a fait partager sa profonde connaissance du monde de la statistique et son rôle dans mon parcours de maître de conférences a été majeur.

Je voudrais exprimer le grand plaisir que j'ai eu à collaborer avec Florent Autin, Karine Bertin, Marie Doumic, Niels Hansen, Marc Hoffmann, Erwan Le Pennec, Jean-Michel Loubes, Thanh Mai Pham Ngoc, Dominique Picard, Judith Rousseau, Gilles Stoltz, Karine Tribouley et Christine Tuleau. Nos nombreux échanges m'ont énormément apporté tant sur le plan scientifique que sur le plan humain. Au delà du fait que j'ai la chance de co-encadrer la thèse de Laure Sansonnet avec elle, Patricia Reynaud-Bouret tient une place particulière parmi mes co-auteurs. Sa très grande culture mathématique, son esprit curieux et son enthousiasme en font une collaboratrice exceptionnelle et elle a profondément marqué ma vision de la statistique.

Bien entendu, je n'oublie pas d'associer à ces remerciements tous les collègues (enseignants, chercheurs, informaticiens et secrétaires) que je côtoie quotidiennement au Laboratoire de Mathématiques d'Orsay et au Département de Mathématiques de l'ENS et qui rendent ces lieux si chaleureux.

Enfin, un très grand merci à ma famille pour son affection et en particulier à Alice pour sa patience et sa compréhension.

Contents

Introduction

Ce mémoire présente les travaux que j'ai effectués tout d'abord au sein du Laboratoire de Probabilités et Modèles Aléatoires des Universités Paris 6 et Paris 7 pendant ma thèse, puis dans l'Equipe Probabilités, Statistique et Modélisation du Département de Mathématiques de l'Université d'Orsay depuis 2003, et enfin également au Département de Mathématiques et Applications de l'Ecole Normale Supérieure à partir de 2007.

Ce manuscrit revient tout d'abord sur l'adage popularisé dans les années 90 :

"Bien estimer, c'est bien approcher."

Pour l'estimation de quantités complexes, la démarche qui consiste à considérer au préalable des objets 'simples' ayant de 'bonnes propriétés d'approximation' est à présent courante en statistique. On peut chercher à aller plus loin et se demander dans quelle mesure la question de l'estimation revient à la question de l'approximation. L'approche maxiset introduite en statistique par Kerkyacharian et Picard, dont le but est de déterminer exactement l'ensemble des fonctions pouvant être estimées par une procédure donnée et avec une précision fixée, constitue un cadre idéal pour fournir des réponses à cette problématique. Plus précisément, elle formalise mathématiquement plusieurs éléments que je décris ci-dessous et dans le premier chapitre de ce manuscrit dans lequel je présente le calcul des maxisets pour trois familles de procédures parmi les plus utilisées :

- les estimateurs à noyau,
- les procédures par seuillage de coefficients d'ondelettes (pour le problème de l'estimation de fonctionnelles quadratiques),
- les procédures par sélection de modèles.

Pour chacune de ces procédures, je décris le cadre, toujours très général, qui permet de déterminer leurs maxisets. Cela revient à dire que, dans ces cas là, la qualité de l'estimation est alors donnée par la qualité de l'approximation. Il faut noter au passage que cela implique l'utilisation d'outils probabilistes puissants qui permettent d'éliminer l'aléa. Ainsi, nous proposons un cadre mathématique qui formalise l'équivalence sous-jacente de l'adage précédent. Mais l'approche maxiset va plus loin puisqu'elle fournit les caractéristiques fonctionnelles qui permettent d'obtenir de bonnes performances d'estimation. En effet, pour résumer sommairement les résultats du premier chapitre, je montre que, de manière générale, lorsqu'une procédure linéaire est considérée, un signal ne peut être bien estimé que s'il est régulier. Ce fait est bien connu et a popularisé l'utilisation des classes fonctionnelles de type Sobolev ou Hölder dans l'approche minimax. Lorsque nous considérons une procédure non-linéaire, plus que la régularité, c'est la parcimonie du signal sous-jacent qui joue le rôle prépondérant, comme je le montre dans le paragraphe 5 du premier chapitre. Enfin, je mentionne que l'approche maxiset permet de donner des définitions très précises de ces notions de 'régularité' et de 'parcimonie' confortant l'intuition que l'on pourrait en avoir.

De l'étude que nous menons dans le premier chapitre, nous déduisons que les procédures d'estimation qui s'appuient sur des dictionnaires de fonctions offrant des décompositions parcimonieuses des signaux à estimer constituent de bons outils. C'est le cas, par exemple, des procédures par seuillage de coefficients d'ondelettes qui présentent en outre l'avantage de s'implémenter

aisément. Mais il faut insister sur le fait que la théorie maxiset, de nature asymptotique, ne permet en aucun cas de calibrer ces procédures ; il faut alors envisager une toute autre approche. Par calibration, j'entends la détermination précise des paramètres (en anglais 'the tuning parameters') de l'estimateur utilisé à des fins pratiques, souvent dans un cadre non-asymptotique. Pour cette problématique, la première approche que j'ai envisagée est de nature bayésienne, ce qui peut sembler paradoxal dans l'optique de calibrer des estimateurs fréquentistes. Le cadre bayésien est pourtant naturel puisqu'il permet :

- d'incorporer l'information disponible concernant le paramètre d'intérêt via l'introduction d'une distribution a priori,
- d'obtenir une construction automatique d'estimateurs grâce au calcul des estimateurs bayésiens associés à la loi a posteriori.

Au vu des conclusions du Chapitre 1, le premier travail consiste donc à construire des lois a priori adaptées à la modélisation de la parcimonie. C'est l'objet de la première partie du Chapitre 2 où nous verrons notamment l'intérêt d'utiliser des distributions fondées sur des densités à queues lourdes. Quand la modélisation est adéquate, l'estimateur bayésien de la médiane a posteriori est de type seuillage où le seuil dépend des hyperparamètres de la distribution a priori. L'étude théorique et pratique des estimateurs bayésiens les plus classiques est menée dans la seconde partie du Chapitre 2.

L'étude de la calibration d'estimateurs est également menée à travers le prisme uniquement fréquentiste dans la dernière partie de ce mémoire. Plus exactement, les seuils calibrés des estimateurs par ondelettes sont fondés sur des inégalités de concentration fines. Pour résumer le propos du Chapitre 4, l'objectif est de combler le fossé qui existe entre un choix théorique du seuil et un choix pratique. Nous verrons dans quelle mesure cet objectif est atteint. Je mentionne simplement le fait que nous parvenons à démontrer, sous certaines conditions, l'existence d'une valeur minimale de la 'constante de seuil', résultat analogue à celui établi par Birgé et Massart en sélection de modèles. L'étude théorique de la calibration est prolongée par une étude par simulations. Ces résultats sont partiellement généralisés dans le cadre du fléau de la dimension où nous ne considérons pas une unique base orthonormée, mais un dictionnaire de fonctions muni de propriétés classiques d'incohérence. Nous considérons pour cela un estimateur construit sous des contraintes de type 'Dantzig'.

Mais avant cela, dans le Chapitre 3, nous revenons sur le problème de l'estimation de densité ou de l'intensité d'un processus de Poisson que nous souhaitons traiter en formulant un minimum d'hypothèses sur le signal à estimer. Nous menons une étude théorique des estimateurs par ondelettes et de type 'Dantzig'. Elle est prolongée par une étude numérique sur des jeux de données simulées et réelles. En particulier, nous axons notre problématique sur la question du support : dans quelle mesure les performances d'une procédure par ondelettes dépendent du support du signal sous-jacent ? Est-il préjudiciable de considérer a priori que le support du signal à estimer est infini (parce qu'inconnu et que l'on ne souhaite pas l'estimer) ? En bref, nous nous demandons dans ce chapitre s'il existe un 'fléau du support' comme il existe un 'fléau de la dimension'. Les réponses à toutes ces questions sont liées une nouvelle fois à la parcimonie. En particulier dans le cadre minimax, les vitesses ne sont pas altérées à la condition d'estimer un signal suffisamment parcimonieux. Nous montrons l'adaptivité de l'estimateur par ondelettes que nous proposons, puisqu'il atteint la vitesse minimax (à un terme logarithmique près) quelle que soit la régularité du signal sous-jacent (ce qui est classique) mais également quelle que soit la taille du support. Du point de vue numérique, la robustesse de notre estimateur à ce fléau du support est également établie.

J'achève cette introduction en mentionnant que les résultats du Chapitre 3 se prolongent naturellement pour des problèmes plus généraux qui cherchent à étudier l'estimation adaptative d'interactions poissonniennes. Ce thème constitue un sujet de thèse décrit en conclusion de ce mémoire. Il sera traité par Laure Sansonnet et co-encadré avec Patricia Reynaud-Bouret.

CHAPTER 1

Maxiset results

1. The maxiset point of view and arising issues

The maxiset point of view has been introduced by Kerkyacharian and Picard (1993, 2000, 2002) and Cohen, DeVore, Kerkyacharian and Picard (2001). This approach consists in determining the set of all the functions which can be estimated at a specified rate of convergence for a given procedure. More precisely, let us assume we are given a statistical model $\{P_\theta^n,\ \theta \in \Theta\}$, where the P_θ^n's are probability distributions and Θ is the set of parameters. We consider a sequence of estimates $\hat{q}_n$ of a quantity $q(\theta)$, a loss function ρ and a rate of convergence α_n tending to 0. Then, the maxiset associated with the procedure $(\hat{q}_n)_n$, the loss function ρ and the rate α_n is the following set:

$$MS(\hat{q}_n, \rho, \alpha_n) = \left\{ \theta \in \Theta :\ \sup_n \left\{ \alpha_n^{-1} \mathbb{E}_\theta^n \rho(\hat{q}_n, q(\theta)) \right\} < \infty \right\}.$$

Obviously, the larger the maxiset, the better the procedure.

Let us briefly mention the differences with the minimax point of view. To study minimax properties of a procedure, we have to arbitrarily choose a set of functions and look at the worst performances of estimators on this set. The maxiset theory is less pessimistic and instead of a priori fixing a (functional) set such as a Hölder, Sobolev or Besov ball, the problem is handled in a wider context since the parameter set Θ can be very large (Θ can be, for instance, the set of measurable square-integrable functions). The outcome is a set authentically connected to the procedure and the model.

However, still there is a deep parallel between maxiset and minimax theories. For instance, facing a particular situation, the standard procedure to prove that a set $\mathcal{F}$ is the maxiset usually consists (exactly as in minimax theory) in two steps. First, we show that $\mathcal{F} \subset MS(\hat{q}_n, \rho, \alpha_n)$, but this is generally obtained using similar arguments as for proving upper bound inequalities in the minimax setting: we establish that if $\theta \in \mathcal{F}$ then for all n,

$$\mathbb{E}_\theta^n \rho(\hat{q}_n, q(\theta)) \leq C\alpha_n,$$

where C is a constant. Proofs of results stated in this chapter emphasize the gain of the maxiset setting: the second inclusion $MS(\hat{q}_n, \rho, \alpha_n) \subset \mathcal{F}$ is often much simpler than proving lower bounds for minimax rates over complicated spaces. Furthermore, this second step deeply involves the procedure at hand.

Kerkyacharian and Picard and their co-authors have derived following maxiset results giving rise to nice interpretations. It has been established in Kerkyacharian and Picard (1993) that, in a specific context, the maxisets of linear kernel methods are in fact Besov spaces under fairly reasonable conditions on the kernel, whereas the maxisets of thresholding estimates (see Kerkyacharian and Picard (2000) and Cohen, DeVore, Kerkyacharian and Picard (2001)) are specific Lorentz spaces (see below). It has also been observed (see Kerkyacharian and Picard (2002)) that there is a deep connection between oracle inequalities and maxisets, in the sense that verifying an oracle inequality is equivalent to proving that the maxiset of the procedure automatically contains a minimal set associated to the oracle.

However, the maxiset approach raises many issues, some of them were highlighted in the discussion of Kerkyacharian and Picard (2000) by experts of non-parametric statistics.

(1) Can maxiset results be established for the most popular procedures and classical statistical models? In particular, given a statistical model (density estimation, Gaussian white noise model,..) various procedures have now been built and proved to be optimal in the minimax setting. So, the minimax approach cannot decide between them. Can the maxiset theory be a help for this problem?
(2) The outcome of the maxiset approach consists in a functional space, which constitutes a more wealthy answer than a minimax rate. Of course, if $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the maxisets associated with two procedures $\hat{f}_1$ and $\hat{f}_2$ with $\mathcal{F}_1 \subset \mathcal{F}_2$, then $\hat{f}_1$ is naturally said to be outperformed by $\hat{f}_2$ in the maxiset setting. But, is such a comparison always possible and what can we conclude when $\mathcal{F}_1$ and $\mathcal{F}_2$ are not nested?
(3) In previous papers, rates were in fact chosen as a simple function of the tuning parameter (a power of the kernel bandwidth or of the wavelet threshold). Is it really possible to disconnect the rates of convergence and the tuning parameters? In this case, do the maxiset comparisons remain unchanged whatever the rates?
(4) In the minimax approach, an estimator is said to be optimal when it achieves the minimax rate over a large scale of functional spaces. In the oracle approach, optimality is measured within a class of estimators and we aim at mimicking performances of the "oracle estimator" viewed as an ideal. Can the notion of optimality be defined in the maxiset setting?
(5) It was claimed previously that lower bounds for maxiset results are simple to obtain. To what extent, is this statement true in a general way? Is there a general methodology to compute maxisets?
(6) Cohen, Devore, Keryacharian and Picard (2001) shed lights on the role of approximation theory in statistics. Can we go further? Can maxisets help us to mathematically formulate the properties a signal must satisfy to be well estimated?
(7) As said previously, the maxiset theory can be viewed as an alternative to the minimax approach. But, of course, minimax optimality of an estimate can directly be obtained by using maxiset embeddings calculated with minimax rates. Could we apply such a trick to derive new minimax results? More important, minimax optimality of procedures are systematically investigated on functional spaces such as Hölder, Sobolev or Besov spaces designed to capture the regularity of signals. Should the maxiset theory call into question such classical but subjective choices? In this case, what kind of spaces should be considered?

It was a challenging task to provide answers to these questions by investigating further maxiset results. Our purpose in the next sections is to briefly describe the most relevant ones. In Section 2, we first extend results of Kerkyacharian and Picard (1993) by investigating maxisets for kernel rules where the loss function is the sup-norm. In this process, we present a general methodology to compute maxisets. We point out arising difficulties, in particular to characterize maxisets, and the way to overcome them. A very different setting is considered in Section 3 where we investigate the estimation of $\theta(f) = \int f^2$ in the maxiset approach and the wavelet setting. Finally, in Section 4, we consider model selection rules where models are spanned by a dictionary of functions that is not necessarily an orthonormal basis. We emphasize that, in this chapter, the goal is not to present new methodologies to build estimates, but to study existing ones in the maxiset approach. From this study, we shall draw interesting conclusions in terms of sparsity (see Section 5).

2. Maxisets for kernel rules: the methodology to compute maxisets

2.1. General results for kernel rules. This section presents the maxiset results obtained for kernel rules studied in the Gaussian white noise model where the loss function is the sup-norm. Here, we aim at illustrating the general methodology to compute maxisets and the difficulties that arise. The subsequent results can be found in [R8] and constitute extensions of maxiset results for linear estimates established for the $\mathbb{L}_q$-norm ($1 < q < \infty$) and with polynomial rates of

convergence by Kerkyacharian and Picard (1993) in the density estimation setting. We consider the Gaussian white noise model

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}}dW_t, \quad t \in [0,1]^d, \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown function, W is the Brownian sheet in $[0,1]^d$, $\sigma > 0$ is known and $n \in \mathbb{N}^*$, where $\mathbb{N}^*$ is the set of positive integers. We study the estimation of f on $[0,1]^d$ from the observations $\{Y_t, t \in [0,1]^d\}$. For this purpose, we assume that f belongs to $\mathbb{L}_\infty^{\text{per}}(\mathbb{R}^d)$ the set of 1-periodic functions that belong to $\mathbb{L}_\infty(\mathbb{R}^d)$. The quality of an estimator $\hat{f}_n$ is characterized by its risk in sup-norm

$$R_n(\hat{f}_n) = \mathbb{E}\left(\|\hat{f}_n - f\|_\infty^p\right),$$

where $\|g\|_\infty = \text{ess}\sup_{t\in[0,1]^d} |g(t)|$ and $p \geq 1$. In this framework, we set the following definition.

DEFINITION 1.1. *Let $1 \leq p < \infty$, $\psi = (\psi_n)_n$ a decreasing sequence of positive real numbers and let $\hat{f} = (\hat{f}_n)_n$ be an estimation procedure. The maxiset of $\hat{f}$ associated with the rate ψ and the $\|.\|_\infty^p$-loss is:*

$$MS(\hat{f}, \psi, p) = \left\{ f \in \mathbb{L}_\infty^{per}(\mathbb{R}^d) : \quad \sup_n \left[\psi_n^{-p}\mathbb{E}\left(\|\hat{f}_n - f\|_\infty^p\right)\right] < \infty \right\}.$$

We focus on kernel rules, denoted $\tilde{f}_{K,h} = \left(\tilde{f}_{K,h_n}\right)_n$ in the sequel, where

$$\tilde{f}_{K,h_n}(t) = \frac{1}{h_n^d}\int_{\mathbb{R}^d} K\left(\frac{t-u}{h_n}\right) dY_u, \quad t \in [0,1]^d, \tag{1.2}$$

$K : \mathbb{R}^d \to \mathbb{R}$ is a compactly supported function satisfying $\|K\|_2^2 = \int_{\mathbb{R}^d} K^2(u)du < \infty$ and $h = (h_n)_n$ is a sequence of bandwidth parameters that tends to 0. Note that some boundary problems arise to define $\tilde{f}_{K,h_n}(t)$ for t close to 0 or 1 but they can easily be overcome by periodizing observations (see Section 2.1 of [R8] for more details). In particular, we have:

$$\mathbb{E}\left[\tilde{f}_{K,h_n}(t)\right] = K_{h_n} * f(t),$$

where for any $t \in \mathbb{R}^d$, $K_{h_n}(t) = \frac{1}{h_n^d}K\left(\frac{t}{h_n}\right)$. We first state that if $MS\left(\tilde{f}_{K,h}, \psi, p\right)$ is not empty, then $h = (h_n)_n$ cannot go to 0 too quickly and an approximation property of functions belonging to the maxiset can be derived.

THEOREM 1.1. *Let $\psi = (\psi_n)_n$ be a positive sequence such that $\lim_{n\to\infty} n\psi_n^2 = +\infty$. Let us assume that there exists a function f satisfying for all $n \in \mathbb{N}$*

$$\mathbb{E}\|\tilde{f}_{K,h_n} - f\|_\infty^p \leq \psi_n^p.$$

Then, for all $0 < \varepsilon < 1$, there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$h_n \geq (1-\varepsilon)\left(\frac{\log(nc_0\psi_n^2)}{dnc_0\psi_n^2}\right)^{1/d}, \tag{1.3}$$

where

$$c_0 = \frac{2}{d\sigma^2\|K\|_2^2}. \tag{1.4}$$

Furthermore,

$$\sup_n \left[\psi_n^{-1}\|K_{h_n} * f - f\|_\infty\right] \leq 1. \tag{1.5}$$

Observe that the result is true for any arbitrary rate $(\psi_n)_n$ as soon as this rate is slower than the parametric rate. Theorem 1.1 is proved in [R8] but, to shed lights on the maxiset methodology, let us mention the main tools that allow to prove this result. In the minimax setting for any estimator f_n^* and $p \geq 1$, we use the classic decomposition of the risk in bias and variance terms:

$$\mathbb{E}\|f_n^* - f\|_\infty^p \leq 2^{p-1}\left(\|\mathbb{E}f_n^* - f\|_\infty^p + \mathbb{E}\|f_n^* - \mathbb{E}f_n^*\|_\infty^p\right).$$

In the maxiset setting, and in particular to prove Theorem 1.1, we use the following converse result that shows that controlling the risk allows to control the bias and the variance terms.

LEMMA 1.1. *For any estimator f_n^*, we have:*

$$\|\mathbb{E}f_n^* - f\|_\infty^p \leq \mathbb{E}\|f_n^* - f\|_\infty^p,$$
$$\mathbb{E}\|f_n^* - \mathbb{E}f_n^*\|_\infty^p \leq 2^p\mathbb{E}\|f_n^* - f\|_\infty^p.$$

The proof of Theorem 1.1 also relies on the following proposition concerning the variance term that actually provides the lower bound for the bandwidth parameter.

PROPOSITION 1.1. *For any $\delta > 0$, there exists $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$,*

$$\mathbb{E}\|\tilde{f}_{K,h_n} - \mathbb{E}\tilde{f}_{K,h_n}\|_\infty^p \geq (1-\delta)\left(\frac{2d\sigma^2\|K\|_2^2|\log(h_n)|}{nh_n^d}\right)^{p/2}.$$

By using Theorem 1.1, we can now deduce an optimal choice for the bandwidth parameter h_n. Indeed, the smaller h_n, the smaller the bias term $\|K_{h_n} * f - f\|_\infty$. Therefore, considering both (1.3) and (1.5) leads to the optimal choice:

$$h_n = \left(\frac{\log(nc_0\psi_n^2)}{dnc_0\psi_n^2}\right)^{1/d} \text{ with } c_0 = \frac{2}{d\sigma^2\|K\|_2^2}. \tag{1.6}$$

Now, we can state the maxiset result associated with the kernel estimate with this choice of bandwidth parameter. As explained in Section 1, it is established in two steps. The first one uses similar arguments as for proving upper bound inequalities in the minimax setting. The second one is simply based on Theorem 1.1. In the sequel, we consider the following additional condition on the kernel K. For all $t \in \mathbb{R}^d$ such that $\|t\| \leq 1$, $\int_{\mathbb{R}^d}(K(t+u) - K(u))^2du \leq C\|t\|^{2v}$, where $\|\cdot\|$ is a norm of $\mathbb{R}^d$, C is a positive constant and $v \in (0,1]$.

THEOREM 1.2. *Let $\psi = (\psi_n)_n$ be a positive sequence such that $\lim_{n\to\infty} n\psi_n^2 = +\infty$. If h_n satisfies (1.6),*

$$MS\left(\tilde{f}_{K,h},\psi,p\right) = \left\{f \in \mathbb{L}_\infty^{per}(\mathbb{R}^d) : \quad \sup_n\left[\psi_n^{-1}\|K_{h_n} * f - f\|_\infty\right] < \infty\right\}. \tag{1.7}$$

Note that $MS\left(\tilde{f}_{K,h},\psi,p\right)$ does not depend on the parameter p. But this maxiset depends on the kernel K and on the rate ψ_n. The next step consists in characterizing this space in terms of classical functional spaces. Unfortunately, this cannot be done without further assumptions on the rate. This issue can be addressed if we consider the classic rate $\psi = \psi(\beta,d) = (\psi_n(\beta,d))_n$ with

$$\psi_n(\beta,d) := C\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}, \tag{1.8}$$

where $\beta > 0$ and C is a positive constant. In the sequel, we restrain our attention on the following class of kernel functions.

DEFINITION 1.2. *For $N \in \mathbb{N}^*$, $\mathcal{K}(N)$ is the set of the functions $K : \mathbb{R}^d \to \mathbb{R}$ that satisfies conditions stated before and*

- $\int_{\mathbb{R}^d} K(u)du = 1$,

- *for any* $(\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$*, such that* $\sum_{i=1}^d \alpha_i \leq N$*, we have*

$$\int_{\mathbb{R}^d} \left| \frac{\partial^{\alpha_1}}{\partial t_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial t_d^{\alpha_d}} K(t) \right| dt < \infty,$$

- *for all polynomial* P *of degree less than* N *such that* $P(0) = 0$*,*

$$\int_{\mathbb{R}^d} P(u)K(u)du = 0.$$

The sets $\mathcal{K}(N)$ contain kernels commonly used in estimation (see Section 2.3.2 of [R8]). We are now ready to state the main maxiset result of this section. We respectively denote $\mathcal{B}^{\beta}_{\infty,\infty}$ and $\Sigma(\beta)$ the Besov and Hölder spaces of parameter β. Remember that the Besov spaces can be introduced through different points of view. Due to the setting of the maxiset approach considered here, Section 2.3.1 of [R8] introduce them by using the point of view of the approximation theory. We also recall that when $\beta \notin \mathbb{N}^*$ the Hölder space $\Sigma(\beta)$ and the Besov space $\mathcal{B}^{\beta}_{\infty,\infty}$ are identical (see for instance (Meyer (1990) p. 52–53). This is not true when β is an integer and $\Sigma(\beta)$ is strictly included in $\mathcal{B}^{\beta}_{\infty,\infty}$. For $\beta > 0$, we denote

$$\lceil \beta \rceil = \min \{ l \in \mathbb{N} : \quad l > \beta \}.$$

THEOREM 1.3. *Consider the procedure* $\tilde{f}_{K,h}$ *with* $K \in \mathcal{K}(\lceil \beta \rceil)$ *and* h *given by (1.6).*

(1) *If* β *is not an integer*

$$MS(\tilde{f}_{K,h}, \psi(\beta, d), p) = \mathcal{B}^{\beta}_{\infty,\infty},$$

(2) *if* β *is an integer*

$$\Sigma(\beta) \subset MS(\tilde{f}_{K,h}, \psi(\beta, d), p) \subset \mathcal{B}^{\beta}_{\infty,\infty}.$$

This result establishes that the set of functions that can be estimated at the classic rate $\psi(\beta, d)$ is exactly the functions that belong to $\mathcal{B}^{\beta}_{\infty,\infty}$ when β is not an integer. When β is an integer, there is a slight ambiguity resulting from the strict inclusion of $\Sigma(\beta)$ in $\mathcal{B}^{\beta}_{\infty,\infty}$. It was already known that β-Hölder functions can be estimated at the rate $\psi(\beta, d)$, but our maxiset results prove that these functions are the only ones. We mention that simulations are performed to study practical performances of the kernel estimate associated with the bandwidth parameter proposed previously. We refer the reader to Section 3 of [R8] for more details.

2.2. Maxisets for the Lepski procedure. In the last years, in the minimax approach, the question of adaptation (that consists in building optimal procedures that do not require the knowledge of the regularity of the underlying signal) has been extensively considered. Many answers have been proposed to handle this issue, as, for instance the method proposed by Lepski (see Lepski 1992)). The last result of this section concerns the maxisets associated with the Lepski method applied to kernel rules. Let $B = \{\beta_1, \ldots, \beta_L\}$ a finite subset such that $\beta_i < \beta_j$ if $i < j$ and the β_i's are non-integer. For each $\beta \in B$, we denote $\hat{f}_\beta = \tilde{f}_{K,h}$ with $K \in \mathcal{K}(\lceil \beta_L \rceil)$, h_n given by (1.6) and ψ defined in (1.8). We set

$$\hat{\beta} = \max \left\{ u \in B : \quad \|\hat{f}_{n,u} - \hat{f}_{n,\gamma}\|_\infty \leq \eta_n(\gamma),\ \forall\, \gamma \leq u \right\},$$

with

$$\eta_n(\gamma) = C_1 \psi_n(\gamma, d),$$

and C_1 is a constant assumed to be large enough. Denote this procedure $\hat{f} = (\hat{f}_{n,\hat{\beta}})_n$. The Lepski procedure is based on the fact that while $\gamma \leq \beta \leq \delta$ and f is of regularity δ, the bias of $\hat{f}_{n,\beta} - \hat{f}_{n,\gamma}$ is bounded from above by a term of order $\psi_n(\gamma, d)$. We have the following theorem.

THEOREM 1.4. *Let* $\beta \in B$*. We have*

$$MS(\hat{f}, \psi(\beta, d), p) = \mathcal{B}^{\beta}_{\infty,\infty}.$$

This result proves that the adaptive kernel procedure $\hat{f}$ achieves the same performance as $\hat{f}_\beta$ from the maxiset point of view. To prove Theorem 1.4, we first use arguments of Bertin (2005) to derive the inclusion $\mathcal{B}^\beta_{\infty,\infty} \subset MS(\hat{f}, \psi(\beta,d), p)$. The inclusion $MS(\hat{f}, \psi(\beta,d), p) \subset \mathcal{B}^\beta_{\infty,\infty}$ is expected since we guess that the maxiset performances of $\hat{f}$ cannot be stronger than those of $\hat{f}_\beta$. Technical details of this proof are given in Section 4 of [R8].

3. The maxiset point of view for estimating integrated quadratic functionals

3.1. Presentation of the problem. Our aim, in this section, is to investigate the estimation of $\theta(f) = \int f^2$, where f is the functional parameter of the unidimensional white noise model:

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t \in [0,1]. \tag{1.9}$$

We illustrate three original and significant facts related to the maxiset approach:

(1) Maxisets can be established for very general convergence rates.
(2) Two procedures are not always ordered in this approach and maxiset comparisons can differ when convergence rates vary.
(3) Maxisets allow to derive minimax properties of estimates and can help the statistician to choose between optimal minimax procedures.

We consider a $\mathbb{L}_2$-orthonormal wavelet basis $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$ (the indice $j=-1$ corresponds to the father wavelets and the indices $j \geq 0$ to the mother wavelets. Wavelets are assumed to be compactly supported). We translate the original functional model (1.9) into the sequence space model:

$$y_{jk} = \beta_{jk} + \varepsilon\, z_{jk},\ j \in \{-1\} \cup \mathbb{N}, \quad k \in \mathbb{Z},$$

where (z_{jk}) is a sequence of i.i.d. standard Gaussian variables, (y_{jk}) is the sequence of observed variables, and (β_{jk}) are the wavelet coefficients of f: $f = \sum_{j=-1}^{+\infty} \sum_k \beta_{jk}\psi_{jk}$. Since the wavelet basis is orthonormal, the parameter to be estimated is $\theta = \theta(f) = \sum_{j=-1}^{\infty} \sum_k \beta_{jk}^2$.

This problem has been intensively studied in the minimax theory and is now completely solved (see in particular Ibragimov and Khas'minskii (1980), Bickel and Ritov (1988), Donoho and Nussbaum (1990), Fan (1991), Efromovich and Low (1996), Tribouley (2000), Laurent and Massart (2000) and Cai and Low (2005)). Generally, f is assumed to belong to the Besov space $\mathcal{B}^\alpha_{p,\infty}$ for $\alpha > 0$, $p \geq 1$. One gets different rates according to the regularity α of the function f. If f is regular enough, it is possible to estimate $\theta(f)$ with the parametric rate. Otherwise, the (non-parametric) rates depend on α and on p when $p < 2$. Moreover, as in the problem of estimating the entire function f, two forms of rates have been pointed out when f is *dense* ($p \geq 2$) or f is *sparse* ($p < 2$). Procedures have been proposed to achieve the minimax rate in each case. Under some conditions, in the case where $p \geq 2$, quadratic methods or global thresholding methods are shown to be minimax or adaptive minimax (see Tribouley (2000)); in the case $p < 2$, Cai and Low (2005) have proved that a local thresholding method is minimax. For further details, see Section 4 of [R6].

We consider the maxiset setting:

DEFINITION 1.3. *Let $R > 0$ and let $\rho_\varepsilon > 0$ be the target rate. If $\hat{\theta}$ denotes an estimator of θ, the maxiset of $\hat{\theta}$ of radius R for the rate ρ_ε is denoted $MS(\hat{\theta}, \rho_\varepsilon)(R)$, and is defined by*

$$MS(\hat{\theta}, \rho_\varepsilon)(R) = \left\{ f \in \mathbb{L}_2([0,1]) : \quad \sup_\varepsilon \rho_\varepsilon^{-1} \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \leq R^2 \right\}.$$

We use the following convention:

CONVENTION 1.1. *We write*

$$MS(\hat{\theta}, \rho_\varepsilon) =: \mathcal{A}$$

to mean that

$$\forall R,\ \exists R',\ MS(\hat{\theta}, \rho_\varepsilon)(R) \subset \mathcal{A}(R')$$

and

$$\forall R',\ \exists R,\ \mathcal{A}(R') \subset MS(\hat{\theta}, \rho_\varepsilon)(R),$$

where $R, R' > 0$ are the radii of balls of $MS(\hat{\theta}, \rho_\varepsilon)$ and $\mathcal{A}$ respectively.

We consider the procedures briefly mentioned previously. For local thresholding, we refer to Cai and Low (2005), and for global thresholding to Tribouley (2000). Let j_0, j_1 be levels (chosen later) such that $j_0 \leq j_1$. We consider estimates

$$\hat{\theta} = \sum_{j=-1}^{j_1-1} \sum_k \hat{\theta}_{jk},$$

where for $j = -1, \ldots, j_0 - 1$, and all k, $\hat{\theta}_{jk} = y_{jk}^2 - \varepsilon^2$.

(1) If $j_0 = j_1$, $\hat{\theta}$ is the *classical quadratic estimator*. In this case, we note $\hat{\theta} = \hat{\theta}^Q$.
(2) If $j_0 < j_1$, we consider
- either the *soft local thresholding procedure*: for all $j \in \{j_0, \ldots, j_1 - 1\}$ and all k,

$$\hat{\theta}_{jk} = \hat{\theta}_{jk}^L = \left(y_{jk}^2 - \mu\varepsilon^2\right) 1_{|y_{jk}| > \varepsilon\sqrt{\tau}} - \varepsilon^2 \mathbb{E}\left(z_{jk}^2 - \mu\right) 1_{|z_{jk}| > \sqrt{\tau}},$$

with $\mu = \tau = \kappa j$ for a constant κ large enough (we note $\hat{\theta} = \hat{\theta}^L$);
- or the *soft global thresholding procedure*: for all $j \in \{j_0, \ldots, j_1 - 1\}$ and all k,

$$\hat{\theta}_{jk} = \hat{\theta}_{jk}^G = 2^{-j} \sum_k \left(y_{jk}^2 - \lambda\varepsilon^2\right) 1_{\sum_k \left(y_{jk}^2 - \varepsilon^2\right) > \varepsilon^2 \sqrt{2^j \tau}},$$

with $\tau = \kappa j$ for a constant κ large enough and $\lambda = 1 + 2^{-j/2}\sqrt{\tau}$ (we note $\hat{\theta} = \hat{\theta}^G$).

3.2. Maxiset results for general rates. This section shows that maxiset results can be established for very general rates. To define them, we consider a continuous function $u : [0, 1] \longrightarrow \mathbb{R}_+$ such that

$$\exists\, \delta > 0, \exists\, M > 0, \forall\, x \in [0, 1], \forall\, y \in [x, 1], \quad u(y) y^{\delta - 2} \leq M u(x) x^{\delta - 2}. \tag{1.10}$$

We consider quadratic, local and global procedures.

THEOREM 1.5. *Let γ and γ' two constants such that $0 < \gamma < \gamma'$. We assume that (1.10) is satisfied for $\delta \geq \max(\gamma, 1)$.*

- *If $2^{j_0} = 2^{j_1} = \varepsilon^{-2\gamma}$, then*

$$MS(\hat{\theta}^Q, u^2(\varepsilon)) =: \mathcal{B}_{2,\gamma,\infty}(u),$$

where

$$\mathcal{B}_{2,\gamma,\infty}(u)(R) := \left\{ f : \quad \sup_{\lambda > 0} u(\lambda)^{-1} \sum_{j \geq -2\gamma \log_2(\lambda)} \sum_k \beta_{jk}^2 \leq R^2 \right\}.$$

- *If $2^{j_0} = \varepsilon^{-2\gamma}$ and $2^{j_1} = \varepsilon^{-2\gamma'}$, then*

$$MS(\hat{\theta}^L, u^2(\varepsilon)) =: W_\gamma^L(u) \cap \mathcal{B}_{2,\gamma',\infty}(u),$$

where

$$W_\gamma^L(u)(R) := \left\{ f : \quad \sup_{\lambda > 0} u(\lambda)^{-1} \sum_{j \geq -2\gamma \log_2(\lambda)} \sum_k \beta_{jk}^2 1_{|\beta_{jk}| \leq \lambda\sqrt{j}} \leq R^2 \right\}.$$

- *If* $2^{j_0} = \varepsilon^{-2\gamma}$ *and* $2^{j_1} = \varepsilon^{-2\gamma'}$*, then*

$$MS(\hat{\theta}^G, u^2(\varepsilon)) :=: W_\gamma^G(u) \cap \mathcal{B}_{2,\gamma',\infty}(u),$$

where

$$W_\gamma^G(u)(R) := \left\{ f : \quad \sup_{\lambda>0} u(\lambda)^{-1} \sum_{j \geq -2\gamma \log_2(\lambda)} \sum_k \beta_{jk}^2 1_{\sum_k \beta_{jk}^2 \leq \lambda^2 2^{j/2}\sqrt{j}} \leq R^2 \right\}.$$

Note that the larger γ, the larger the maxiset. When polynomial rates are studied, i.e. $u(\varepsilon)$ is of the form $u(\varepsilon) = \varepsilon^r$ for $r > 0$, then

$$\mathcal{B}_{2,\gamma,\infty}(u)(R) = \left\{ f : \quad \sup_{\lambda>0} \lambda^{-r} \sum_{j \geq -2\gamma \log_2(\lambda)} \sum_k \beta_{jk}^2 \leq R^2 \right\}.$$

If the multiresolution analysis associated with the wavelet basis is regular enough, $\mathcal{B}_{2,\gamma,\infty}(u)(R)$ corresponds to the classic Besov ball $\mathcal{B}_{2,\infty}^{r/(4\gamma)}(R)$.

For the balls $W_\gamma^L(u)(R)$, we focus on the number of the wavelet coefficients that are smaller than a prescribed threshold. For the balls $W_\gamma^G(u)(R)$, we do the same job, but level by level and for a function of the wavelet coefficients. These spaces are close to the weak Besov spaces that have already been introduced in the maxiset context in statistics (see Cohen, DeVore, Kerkyacharian, and Picard (2001), Kerkyacharian, and Picard (2000, 2002), [R1], or [R5]) and in approximation theory (see Cohen, DeVore, and Hochmuth (2000) for instance). The main difference lies in the level j's we consider: we do not care about wavelet coefficients when $j < -2\gamma \log_2(\lambda)$, and this difference is crucial for the following maxiset comparisons.

Maxisets corresponding to the polynomial rates $u^2(\varepsilon) = \varepsilon^{2r}$ for $r > 0$ are studied in details in Sections 5.1 and 5.2 of [R6]. But Theorem 1.5 is especially interesting when the target rate is

$$u^2(\varepsilon) = \varepsilon^{2r} |\log(\varepsilon)|^{2r'}$$

for $r' \geq 0$ because they appear in the minimax adaptive framework (note that (1.10) is satisfied with $\delta \geq \max(\gamma, 1)$ if $M = 1$, $\delta = 2 - r$ for $0 < r \leq 1$ and $\gamma \leq 2 - r$). It is of particular interest to compare thresholding and quadratic procedures when the rate is of this form. These comparisons are based on fine studies of the spaces $\mathcal{B}_{2,\gamma,\infty}(u)$, $W_\gamma^L(u)$ and $W_\gamma^G(u)$ established in Section 5.4 of [R6]. They lead to the following result showing the large variety of maxiset results. In particular, we note that the power of the logarithmic term plays a key role. For the sake of brevity and clarity, thresholding procedures are applied with $j_1 = \infty$ but following results remain true provided γ' is large enough (see Section 5.2 and Remark 3 of Section 7 in [R6]).

THEOREM 1.6. *When* $\gamma < 2 - r$*, the thresholding procedures outperform the quadratic one and local and global thresholding are not comparable since*

$$MS(\hat{\theta}^Q, u^2(\varepsilon)) \subsetneq MS(\hat{\theta}^L, u^2(\varepsilon)) \quad \textit{and} \quad MS(\hat{\theta}^Q, u^2(\varepsilon)) \subsetneq MS(\hat{\theta}^G, u^2(\varepsilon)),$$

and

$$MS(\hat{\theta}^L, u^2(\varepsilon)) \not\subset MS(\hat{\theta}^G, u^2(\varepsilon)) \quad \textit{and} \quad MS(\hat{\theta}^G, u^2(\varepsilon)) \not\subset MS(\hat{\theta}^L, u^2(\varepsilon)).$$

Then, let us assume that $\gamma = 2 - r$*.*

- *If* $r' > 1/2$*, the previous conclusions remain valid.*
- *If* $r' < 1/2$*, the quadratic procedure and the global thresholding one achieve the same performance and the local thresholding procedure outperforms the other ones since*

$$MS(\hat{\theta}^Q, u^2(\varepsilon)) :=: MS(\hat{\theta}^G, u^2(\varepsilon)) \subsetneq MS(\hat{\theta}^L, u^2(\varepsilon)).$$

The maxiset comparison for the case $\gamma = 2 - r$ and $r' = 1/2$ remains an open question.

For polynomial rates $u^2(\varepsilon) = \varepsilon^{2r}$, and with an optimal choice of parameters (i.e. $\gamma = 2 - r$), we establish that the local thresholding procedure is always the best in the sense that it achieves the given target rate on the largest set of functions. This conclusion remains true if $u^2(\varepsilon)$ has the classical form

$$u^2(\varepsilon) = (|\log \varepsilon|^{1/4}\varepsilon)^{16\alpha/(1+4\alpha)}$$

for $\alpha > 0$. Non-comparability of local and global thresholding when $\gamma < 2 - r$ could appear as an illustration of a drawback of the maxiset setting where the order is not total. However, we can draw interesting conclusions from these maxiset results in the lights of counter-examples of Section 7 of [R6]. Indeed, we point out what are the functions that belong to the maxiset of one procedure and not to the maxiset of the other one, according to their sparsity. And as a conclusion, roughly speaking, local thresholding is convenient when estimating sparse functions, global thresholding for dense ones. Further conclusions are drawn from Theorem 1.6 in Section 6.2 of [R6].

3.3. Maxiset for minimax. In this section, we use previous maxiset results to simply and automatically establish minimax properties of $\hat{\theta}^Q$, $\hat{\theta}^L$ and $\hat{\theta}^G$. Indeed, to prove that a procedure is minimax on $\mathcal{F}$, we can point out the minimax rate ρ_ε associated with $\mathcal{F}$. Then we compute the maxiset of the procedure for the rate ρ_ε by using theorems of the previous section and prove that $\mathcal{F}$ is included in the maxiset. Of course, some of the minimax results established below were already known (see Section 4 of [R6]), but Theorems 1.7 and 1.8 generalize these known minimax results. We recall that the minimax rate on $\mathcal{B}^\alpha_{p,\infty}$ is ε^2 if $p \geq 2$, $\alpha \geq 1/4$, or $p < 2$, $\alpha > 1/(2p)$. It is also the adaptive minimax rate. When $p \geq 2$, $\alpha < 1/4$, the minimax rate is $\varepsilon^{16\alpha/(1+4\alpha)}$, but the adaptive minimax rate is $(|\log \varepsilon|^{1/4}\varepsilon)^{16\alpha/(1+4\alpha)}$. We have the following result.

THEOREM 1.7. *The quadratic procedure built with $\gamma = 1$ is minimax on $\mathcal{B}^\alpha_{p,\infty}$ if $p \geq 2,\ \alpha \geq 1/4$ or $p < 2,\ \alpha \geq 1/p - 1/4$. The quadratic procedure built with $\gamma = 2/(1 + 4\alpha)$ is minimax on $\mathcal{B}^\alpha_{p,\infty}$ if $p \geq 2,\ \alpha \leq 1/4$. The same conclusions are true for the global soft thresholding procedure built respectively with $\gamma = 1$ and $\gamma = 2/(1 + 4\alpha)$.*

The local soft thresholding procedure built with $\gamma = 1$ is minimax on $\mathcal{B}^\alpha_{p,\infty}$ if $p \geq 2,\ \alpha \geq 1/4$ or $p < 2,\ \alpha > 1/(2p)$. The local soft thresholding procedure built with $\gamma = 2/(1 + 4\alpha)$ is minimax on $\mathcal{B}^\alpha_{p,\infty}$ if $p \geq 2,\ \alpha \leq 1/4$.

Finally, the last result concerns adaptation results of thresholding procedures applied with $j_1 = \infty$ and $\gamma = 1$.

THEOREM 1.8. *The adaptive soft local procedure is not adaptive minimax on $\mathcal{B}^\alpha_{p,\infty}$ for $p \geq 2$, $\alpha < 1/4$. The adaptive soft global procedure is adaptive minimax on $\mathcal{B}^\alpha_{p,\infty}$ for $p \geq 2$, $\alpha > 0$.*

4. Maxisets for model selection

We have now a general idea of maxiset results for non-linear rules. More precisely, at this stage, maxisets have essentially been identified for thresholding rules and signals decomposed on a wavelet basis (or on a unconditional well localized basis, see Kerkyacharian and Picard (2000)). Our goal now is to determine maxisets for model selection rules, which constitutes a first generalization of previous results (since thresholding can be viewed as penalized rules for specific model collections as recalled in Section 4.3). The second direction for extending results obtained so far is to consider models spanned by a dictionary of functions that is not necessarily an orthonormal basis. The results for penalized estimators published in [R9] are described in the sequel. They enhance, in particular, the role of approximation theory in statistics.

4.1. General results. We still consider the classical Gaussian white noise model

$$dY_{n,t} = s(t)dt + \frac{1}{\sqrt{n}}dW_t, \quad t \in \mathcal{D},$$

where $\mathcal{D} \subset \mathbb{R}$, s is the unknown function, W is the Brownian motion in $\mathbb{R}$. The model selection methodology consists in constructing an estimator by minimizing an empirical contrast γ_n over a given set, called a model. In non-parametric estimation, performances of estimators are usually measured by using the quadratic norm, which gives rise to the following empirical quadratic contrast

$$\gamma_n(u) = -2Y_n(u) + \|u\|^2 = -2\int_{\mathcal{D}} u(t)dY_{n,t} + \|u\|^2$$

for any function u, where $\|\cdot\|$ denotes the norm associated to $\mathbb{L}_2(\mathcal{D})$. We assume that we are given a dictionary of functions of $\mathbb{L}_2(\mathcal{D})$, denoted by $\Phi = (\varphi_i)_{i\in\mathcal{I}}$ where $\mathcal{I}$ is a countable set and we consider $\mathcal{M}_n$, a collection of models spanned by some functions of Φ. For any $m \in \mathcal{M}_n$, we denote by $\mathcal{I}_m$ the subset of $\mathcal{I}$ such that

$$m = \text{span}\{\varphi_i : \ i \in \mathcal{I}_m\}$$

and $D_m \leq |\mathcal{I}_m|$ the dimension of m. Let $\hat{s}_m$ be the function that minimizes the empirical quadratic criterion $\gamma_n(u)$ with respect to $u \in m$. Now, the issue is the selection of the best model $\hat{m}$ from the data which gives rise to the model selection estimator $\hat{s}_{\hat{m}}$. For this purpose, a penalization rule is considered, which aims at selecting an estimator, close enough to the data, but still lying in a small space to avoid overfitting issues. Following the classical model selection literature, we only use penalties proportional to the dimension D_m of m:

$$\text{pen}_n(m) = \frac{\lambda_n}{n} D_m, \tag{1.11}$$

with λ_n to be specified. The model $\hat{m}$ is selected using the penalized criterion

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \left\{\gamma_n(\hat{s}_m) + \text{pen}_n(m)\right\}. \tag{1.12}$$

The asymptotic behavior of model selection estimators has been studied by many authors. We refer to Massart (2007) for general references. Then, our goal is to determine:

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha)(R) = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{ \rho_{n,\alpha}^{-2} \mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2\right]\right\} \leq R^2 \right\},$$

where

$$\rho_{n,\alpha} = \left(\frac{\lambda_n}{n}\right)^{\frac{\alpha}{1+2\alpha}} \tag{1.13}$$

for any $\alpha > 0$. To describe maxisets, we introduce the key deterministic quantity

$$Q(s,n) = \inf_{m\in\mathcal{M}_n} \left\{ \|s_m - s\|^2 + \frac{\lambda_n}{n} D_m \right\}, \tag{1.14}$$

where s_m stands for the best approximation (in the $\mathbb{L}_2$ sense) of the function s by a function of m. In other words s_m is the orthogonal projection of s onto m. Let us state the following result.

THEOREM 1.9. *Let $0 < \alpha_0 < \infty$ be fixed. Let us assume that the sequence of model collections satisfies for any n*

$$\mathcal{M}_n \subset \mathcal{M}_{n+1}, \tag{1.15}$$

and that the sequence of positive numbers $(\lambda_n)_n$ is non-decreasing and satisfies

$$\lim_{n\to+\infty} n^{-1}\lambda_n = 0, \tag{1.16}$$

and there exist $n_0 \in \mathbb{N}^*$ and two constants $0 < \delta \leq \frac{1}{2}$ and $0 < p < 1$ such that for $n \geq n_0$,

$$\lambda_{2n} \leq 2(1-\delta)\lambda_n, \tag{1.17}$$

$$\sum_{m\in\mathcal{M}_n} e^{-\frac{(\sqrt{\lambda_n}-1)^2 D_m}{2}} \leq p \tag{1.18}$$

and

$$\lambda_{n_0} \geq \Upsilon(\delta, p, \alpha_0), \tag{1.19}$$

where $\Upsilon(\delta, p, \alpha_0)$ is a positive constant only depending on α_0, p and δ. Then, the penalized rule $\hat{s}_{\hat{m}}$ is such that for any $\alpha \in (0, \alpha_0]$, for any $R > 0$, there exists $R' > 0$ such that for $s \in \mathbb{L}_2(\mathcal{D})$,

$$\sup_n \left\{\rho_{n,\alpha}^{-2} \mathbb{E}\left[\|\hat{s}_{\hat{m}} - s\|^2\right]\right\} \leq R^2 \Rightarrow \sup_n \left\{\rho_{n,\alpha}^{-2} Q(s,n)\right\} \leq (R')^2.$$

Assumptions are very mild. In particular, we make no assumption on Φ. Note that (1.15) does not imply a strong structure on the model collection for a given n. In particular, this does not imply that the models are nested. Assumption (1.16) is necessary to deal with rates converging to 0. The classical cases $\lambda_n = \lambda_0$ or $\lambda_n = \lambda_0 \log(n)$ satisfy (1.16), (1.17) and (1.19) with λ_0 large enough. It is worth noticing that Assumption (1.18) is very close to Assumption (4.5) of Theorem 4.2 of Massart (2007) and allows to control the number of models with the same dimension. The assumption $\alpha \in (0, \alpha_0]$ can be relaxed for particular model collections, which is highlighted in Proposition 2 of [R9]. Finally, Assumption (1.15) can be removed for some special choice of model collection $\mathcal{M}_n$ at the price of a slight overpenalization as it is shown in Proposition 1 of [R9].

Combining Theorem 1.9 and the oracle type inequality given by Theorem 4.2 of Massart (2007), we obtain a first characterization of the maxiset of the model selection procedure $\hat{s}_{\hat{m}}$ (we use the Convention 1.1). In particular, since the maxiset only depends on $Q(s,n)$ that is the main term of Massart's oracle inequality, the next result emphasizes the connections between oracle and maxiset approaches.

COROLLARY 1.1. *Let $\alpha_0 < \infty$ be fixed. Assume that Assumptions (1.15), (1.16), (1.17) (1.19) are satisfied. If there exist two constants $\kappa > 1$ and $0 < p < 1$ such that for any n,*

$$\sum_{m \in \mathcal{M}_n} e^{-\frac{(\sqrt{\kappa^{-1}\lambda_n} - 1)^2 D_m}{2}} \leq p \tag{1.20}$$

then for any $\alpha \in (0, \alpha_0]$,

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{\rho_{n,\alpha}^{-2} Q(s,n)\right\} < \infty \right\}.$$

The maxiset of $\hat{s}_{\hat{m}}$ is characterized by a deterministic approximation property of s with respect to the models $\mathcal{M}_n$. It can be related to some classical approximation properties of s in terms of approximation rates if the functions of Φ are orthonormal.

4.2. The case of orthonormal bases. From now on, $\Phi = \{\varphi_i\}_{i \in \mathcal{I}}$ is assumed to be an orthonormal basis (for the $\mathbb{L}_2$ scalar product). We also assume that the model collections $\mathcal{M}_n$ are constructed through restrictions of a single model collection $\mathcal{M}$. Namely, given a collection of models $\mathcal{M}$ we introduce a sequence $\mathcal{J}_n$ of increasing subsets of the indices set $\mathcal{I}$ and we define the intermediate collection $\mathcal{M}'_n$ as

$$\mathcal{M}'_n = \{m' = \text{span}\{\varphi_i : \ i \in \mathcal{I}_m \cap \mathcal{J}_n\} : \ m \in \mathcal{M}\}. \tag{1.21}$$

The model collections $\mathcal{M}'_n$ do not necessarily satisfy the embedding condition (1.15). Thus, we define

$$\mathcal{M}_n = \bigcup_{k \leq n} \mathcal{M}'_k$$

so $\mathcal{M}_n \subset \mathcal{M}_{n+1}$. The assumptions on Φ and on the model collections allow to give an explicit characterization of the maxisets. We denote $\widetilde{\mathcal{M}} = \cup_n \mathcal{M}_n = \cup_n \mathcal{M}'_n$. Remark that without any further assumption $\widetilde{\mathcal{M}}$ can be a larger model collection than $\mathcal{M}$. Now, let us denote by $V = (V_n)_n$ the sequence of approximation spaces defined by

$$V_n = \text{span}\{\varphi_i : \ i \in \mathcal{J}_n\}$$

and consider the corresponding approximation space

$$\mathcal{L}_V^\alpha = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_n \left\{ \rho_{n,\alpha}^{-1} \|P_{V_n} s - s\| \right\} < \infty \right\},$$

where $P_{V_n} s$ is the projection of s onto V_n. Define also another kind of approximation sets:

$$\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha = \left\{ s \in \mathbb{L}_2(\mathcal{D}) : \quad \sup_{M>0} \left\{ M^\alpha \inf_{\{m \in \widetilde{\mathcal{M}} :\, D_m \le M\}} \|s_m - s\| \right\} < \infty \right\}.$$

The corresponding balls of radius $R > 0$ are defined, as usual, by replacing ∞ by R in the previous definitions. We have the following result.

THEOREM 1.10. *Let $\alpha_0 < \infty$ be fixed. Assume that (1.16), (1.17), (1.19) and (1.20) are satisfied. Then, the penalized rule $\hat{s}_{\hat{m}}$ satisfies the following result: for any $\alpha \in (0, \alpha_0]$,*

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha \cap \mathcal{L}_V^\alpha.$$

The result pointed out in Theorem 1.10 links the performance of the estimator to an approximation property for the estimated function. This approximation property is decomposed into a linear approximation measured by $\mathcal{L}_V^\alpha$ and a non-linear approximation measured by $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha$. The linear condition is due to the use of the reduced model collection $\mathcal{M}_n$ instead of $\mathcal{M}$, which is often necessary to ensure either the calculability of the estimator or Condition (1.20). It plays the role of a minimum regularity property that is easily satisfied.

Observe that if we have one model collection, that is for any k and k', $\mathcal{M}_k = \mathcal{M}_{k'} = \mathcal{M}$, $\mathcal{J}_n = \mathcal{I}$ for any n and thus $\widetilde{\mathcal{M}} = \mathcal{M}$. Then

$$\mathcal{L}_V^\alpha = \text{span}\,\{\varphi_i : \ i \in \mathcal{I}\}$$

and Theorem 1.10 gives

$$MS(\hat{s}_{\hat{m}}, \rho_\alpha) :=: \mathcal{A}_{\mathcal{M}}^\alpha.$$

The spaces $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha$ and $\mathcal{L}_V^\alpha$ highly depend on the models and the approximation space. At first glance, the best choice seems to be $V_n = \mathbb{L}_2(\mathcal{D})$ and

$$\mathcal{M} = \{m : \ \mathcal{I}_m \subset \mathcal{I}\}$$

since the infimum in the definition of $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha$ becomes smaller when the collection is enriched. There is however a price to pay when enlarging the model collection: the penalty has to be larger to satisfy (1.20), which deteriorates the convergence rate. A second issue comes from the tractability of the minimization (1.12) itself which will further limit the size of the model collection. Note that $\mathcal{M}_n$, $\mathcal{L}_V^\alpha$ and $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha$ can be defined in a similar fashion for any arbitrary dictionary Φ. However, one can only obtain the inclusion $MS(\hat{s}_{\hat{m}}, \rho_\alpha) \subset \mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha \cap \mathcal{L}_V^\alpha$ in the general case.

4.3. A brief illustration. In Section 3 of [R9], we exemplify previous maxiset results for different model selection estimators built with wavelet methods by identifying precisely the spaces $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha$ and $\mathcal{L}_V^\alpha$. Let us give a brief meaningful illustration that shed lights on maxisets for popular non-linear procedures. For this purpose, we denote $(\psi_{jk})_{j \ge -1, k}$ a wavelet basis on $\mathcal{D} = [0,1]$ (see Paragraph 3.1) and we consider the model collections

$$\mathcal{M}_n = \mathcal{M}'_n = \{m = \text{span}\{\psi_{jk} : \ (j,k) \in \mathcal{I}_m\} : \ \mathcal{I}_m \in \mathcal{P}_{j_0(n)}\}$$

where $\mathcal{P}_{j_0(n)}$ is the set of all subsets of indices (j,k) such that $-1 \le j \le j_0(n)$, where

$$2^{j_0(n)} \le n\lambda_n^{-1} < 2^{j_0(n)+1}.$$

The classical logarithmic penalty

$$\text{pen}_n(m) = \frac{\lambda_0 \log(n) D_m}{n},$$

which corresponds to $\lambda_n = \lambda_0 \log(n)$, is sufficient to ensure Condition (1.20) as soon as λ_0 is a constant large enough. If the multiresolution analysis associated with the basis is regular enough, $\mathcal{L}_V^\alpha$ is a Besov space:

$$\mathcal{L}_V^\alpha = \mathcal{B}_{2,\infty}^{\frac{\alpha}{1+2\alpha}}.$$

The identification of the non-linear approximation space gives: $\mathcal{A}_{\widetilde{\mathcal{M}}}^\alpha = \mathcal{W}_{\frac{2}{1+2\alpha}}$ with for any $p \in]0,2[$,

$$\mathcal{W}_p := \left\{ s = \sum_{j\geq -1} \sum_k \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \quad \sup_{u>0} u^p \sum_{j=-1}^{+\infty} \sum_k 1_{|\beta_{jk}|>u} < \infty \right\}. \tag{1.22}$$

This result is not new, since for this model collection the corresponding estimate is a thresholding rule:

$$\hat{s}_{\hat{m}} = \sum_{j=-1}^{j_0(n)} \sum_k \hat{\beta}_{jk} \mathbf{1}_{|\hat{\beta}_{jk}|>\sqrt{\frac{\lambda_n}{n}}} \psi_{jk}, \quad \text{with} \quad \hat{\beta}_{jk} = \int_0^1 \psi_{jk}(t) dY_{n,t}.$$

In this setting, Theorem 1.10 corresponds thus to the maxiset result established by Kerkyacharian and Picard (2000). We only focus on this procedure that can be easily presented and give a good overview of maxisets for model selection based on orthonormal bases. But I mention that more intricate model selection procedures based on wavelets are studied in Section 3 of [R9]. Comparisons between the performances of these estimators are provided and discussed.

5. The role of approximation theory in statistics

Previous maxiset results allow to clarify the links between statistics and approximation theory. We noticed that two types of spaces emerge according to the nature of the procedure. Indeed, as shown by Sections 2 and 3 and Kerkyacharian and Picard (1993), maxisets for linear procedures are characterized by the approximation properties of orthogonal projections on linear subspaces specified by the setting (the statistical model, the loss function,...). In addition, under mild conditions, these maxisets are characterized in terms of Besov balls. For non-linear rules, maxisets are no longer Besov spaces, but appear as weak versions of Besov spaces (see Embedding (2.2) in the next section) so are denoted *weak Besov spaces* in the sequel. They belong to the class of Lorentz spaces (see Lorentz (1950, 1966) or DeVore and Lorentz (1993)).

DEFINITION 1.4. *Let Ω a space equipped with a measure μ. For any $0 < p < \infty$, the Lorentz space $\mathcal{L}_{p,\infty}(\mu)$ is the set of μ-measurable functions f: $\Omega \longrightarrow \mathbb{R}$ such that*

$$\|f\|^p_{\mathcal{L}_{p,\infty}(\mu)} := \sup_{\lambda>0} \lambda^p \mu(|f| > \lambda) < \infty.$$

If $\Omega = \mathbb{N}^*$ (or can be identified with $\mathbb{N}^*$), we shall note $w\ell_p(\mu) = \mathcal{L}_{p,\infty}(\mu)$ and $w\ell_p = w\ell_p(\mu)$ if μ is the counting measure:

$$w\ell_p = \left\{ \theta = (\theta_n)_n : \quad \sup_{\lambda>0} \lambda^p \sum_n \mathbb{1}_{|\theta_n|>\lambda} < \infty \right\}.$$

Thus, we can write

$$\mathcal{W}_p = \left\{ s = \sum_{j\geq -1} \sum_k \beta_{jk}\psi_{jk} \in \mathbb{L}_2 : \quad (\beta_{jk})_{jk} \in w\ell_p \right\}.$$

We notice that $\mathcal{W}_p$ seems to strongly depend on the wavelet basis. But, under some conditions (in particular, the multiresolution analysis associated with the wavelet basis has to be regular enough), $\mathcal{W}_p$ can be viewed as an interpolation space between $\mathbb{L}_2$ and a suitable Besov space, which proves that the dependency on the wavelet basis can be relaxed. For more details, we refer the reader to Cohen, DeVore and Hochmuth (2000). See also DeVore (1989), DeVore and Lorentz (1993),

DeVore, Konyagin and Temlyakov (1998), Temlyakov (1999) or Cohen (2000) for more results about Lorentz space in approximation theory.

The sequence spaces $w\ell_p$ are naturally connected to sparsity. Indeed, let us give a sequence $\theta = (\theta_n)_n$ and its non-increasing rearrangement:

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

Then,

$$\theta \in w\ell_p \iff \sup_{n\in\mathbb{N}^*} n^{\frac{1}{p}}|\theta|_{(n)} < \infty. \tag{1.23}$$

We deduce

$$\ell_p \subset w\ell_p \subset \ell_{p+\delta}, \quad \delta > 0.$$

Thus, the $w\ell_p$-space can be viewed as a weak version of the classical ℓ_p-space. Furthermore, Condition (1.23) gives a polynomial control of the decreasing rate of the sequence $(|\theta|_{(n)})_n$ and the smaller p, the sparser the signal. So, the spaces $\mathcal{W}_p$ constitute an ideal class to measure the sparsity of a wavelet decomposed signal. Furthermore, replacing $w\ell_p$ with $w\ell_p(\mu)$ with a suitable choice for μ, regularity can also be measured. Such interpretations allow to give a clear mathematical formulation of the statement that claims that performances of non-linear statistical procedures depend on regularity and sparsity properties of the estimated signal. We mention that pioneering works about the role of Lorentz spaces in statistics are due to Donoho (1993), Johnstone (1994) and Donoho and Johnstone (1996).

We end this section by giving embeddings between Besov and weak Besov spaces. We have:

$$\mathcal{B}^{\alpha}_{2,\infty} \subsetneq \mathcal{W}_{\frac{2}{1+2\alpha}} \cap \mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}.$$

This result establishes a maxiset comparison between linear and non-linear procedures, and as expected, non-linear procedures outperform linear ones. See [R1] for more details and for extensions of this result in a very general setting.

6. Conclusions

Previous sections provide answers to most of the maxiset issues raised in Section 1, although, of course, we could pursue our investigations to obtain some more results. However, the problem of optimality in the maxiset setting seems to be unresolved. Nevertheless, I mention some partial results obtained about this question. Indeed, as in the oracle approach, the notion of optimality can be defined within a class of procedures. For this purpose, in Autin (2006, 2008), three large classes of shrinkage procedures were introduced, respectively denoted *the class of limited, elitist and hereditary procedures*. For each class, an ideal maxiset is derived and is used as a benchmark. In this setting, optimal procedures are proposed by Autin and suggestions are given to build estimation procedures with large maxisets.

The maxiset theory has shown that weak Besov spaces, that naturally classify signals in function of their sparsity, play a capital role in statistics. In the minimax perspective, it is then natural to investigate optimal estimators on the class of weak Besov spaces. This will be the goal of the next chapter: we first derive minimax rates for this class and then we build optimal estimation procedures. This problem is handled by using the Bayesian approach and by keeping in mind the notion of sparsity that has been proved to play a capital role.

CHAPTER 2

Sparsity in the Bayesian setting

1. Introduction

The previous chapter has shown, via the maxiset theory, the key role of sparsity in statistics. In this chapter, we consider the Bayes setting that constitutes a natural framework to express sparsity as well. Indeed, most of the time, any *a priori* information about the signal can be easily incorporated by using a convenient prior model and hierarchical modelling (see Robert (2006)). Furthermore, once elicitation of hyperparameters is performed, natural estimates of the signal are obtained by using standard Bayes rules such as the mean or the median of the posterior distribution. In particular, for specific prior models (see below), we obtain shrinkage and sometimes thresholding rules, which is, in the wavelet setting for instance, of particular interest. As illustrated in Section 4.3, the Bayesian approach gives in general estimates with quite satisfying numerical performances. Finally, we note that when the prior distribution is selected in the natural conjugate family of the noise distribution, computations can be easily performed.

So, in our framework, the major issue is to investigate a convenient prior model to capture the sparsity of the underlying signal, which has already been widely investigated (see the references given in Section 3). For this purpose, we establish a link between deterministic and Bayes approaches of sparsity. Since Chapter 1 reveals the key role of Lorentz spaces to model sparsity of non-random signals, we aim at building typical realizations of such spaces. Section 3 considers a prior model on the coefficients θ_k's of a given signal of the form

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

where $w_k \in [0, 1]$ and γ_k is a density. Then, we investigate a necessary and sufficient condition on these parameters to obtain a signal belonging, almost surely, to a prescribed Lorentz space.

Before this, as explained in Section 6 of Chapter 1, we study weak Besov spaces from the statistical point of view. This study is strongly connected to the Bayesian modelling of sparsity since it points out Bayes models naturally associated with weak Besov spaces. More precisely, the starting point is the following. We wish to get a representation of the 'typical enemies' for classical non-linear procedures. Since maxisets for these procedures are characterized by weak Besov spaces, it is natural to look for these signals in weak Besov balls. Given a weak Besov ball, denoted $\mathcal{WB}^{\alpha}_{p,p}(C)$ in the sequel, we consider least favorable priors associated with $\mathcal{WB}^{\alpha}_{p,p}(C)$. As explained in Section 2, such priors have a Bayes risk that is asymptotically equal to the minimax risk and their support belongs asymptotically to $\mathcal{WB}^{\alpha}_{p,p}(C)$. Realizations of these distributions provide good representations of the worst functions of $\mathcal{WB}^{\alpha}_{p,p}(C)$ to be estimated. So, in the next section, for each weak Besov ball, we lead the study of the minimax risk and least favorable priors.

Next sections show the importance of heavy-tailed prior models in our framework. More precisely, the study of least favorable priors in Section 2 and Theorem 2.2 in Section 3 reveal that Pareto distributions play a key role in our context.

Finally, we study various Bayesian thresholding procedures. In the process, we revisit the problem of choosing the threshold. This calibration issue has been widely investigated. See for instance Donoho and Johnstone (1994, 1995), Nason (1996), Abramovich and Benjamini (1995) or Abramovich, Benjamini, Donoho and Johnstone (2006). In the Bayesian wavelet setting, we can mention Abramovich, Sapatinas and Silverman (1998), Vidakovic (1998) and Johnstone and Silverman (2004, 2005).

In the sequel, we still use notations of Section 3 of Chapter 1 and denote $(\psi_{jk})_{j\geq -1,k}$ a wavelet basis with standard properties of smoothness and moment vanishing allowing the sequential characterization of Besov spaces based on wavelet coefficients.

2. Minimax study of weak Besov spaces and least favorable priors

Our first issue is to point out minimax rates of convergence for weak Besov balls in the framework of the classical white noise model:

$$dY_t = f(t)dt + \varepsilon dW_t, \quad t \in [0,1]. \tag{2.1}$$

By using the wavelet basis $(\psi_{jk})_{j\geq -1,k}$, the model (2.1) is reduced to a sequence space model. We obtain the following sequence of independent variables:

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

where the β_{jk}'s denote the unknown wavelet coefficients of f. We extend the definition of weak Besov spaces introduced in Chapter 1 as follows.

DEFINITION 2.1. *Let $0 < \alpha, p < \infty$. We say that*

$$f = \sum_{j=-1}^{+\infty} \sum_{k} \beta_{jk}\psi_{jk}$$

belongs to the weak Besov space of parameters α and p, denoted $\mathcal{WB}^{\alpha}_{p,p}$, if

$$\sup_{\lambda>0} \lambda^p \sum_{j=-1}^{+\infty} 2^{jp(\alpha+\frac{1}{2}-\frac{1}{p})} \sum_{k} 1_{|\beta_{jk}|>\lambda} < \infty.$$

With each weak Besov space $\mathcal{WB}^{\alpha}_{p,p}$, we associate the balls:

$$\mathcal{WB}^{\alpha}_{p,p}(C) = \left\{ f : \quad \sup_{\lambda>0} \lambda^p \sum_{j=-1}^{+\infty} 2^{jp(\alpha+\frac{1}{2}-\frac{1}{p})} \sum_{k} 1_{|\beta_{jk}|>\lambda} \leq C^p \right\}.$$

If $p < 2$, we obviously have: $\mathcal{WB}^{\alpha}_{p,p} = \mathcal{W}_p$ with $\alpha = \frac{1}{p} - \frac{1}{2}$. More generally, if we consider the measure μ such that

$$\mu(j,k) = 2^{jp(\alpha+\frac{1}{2}-\frac{1}{p})},$$

we have $f \in \mathcal{WB}^{\alpha}_{p,p} \iff (\beta_{jk})_{jk} \in w\ell_p(\mu)$. The Markov inequality shows that for any α and any p,

$$\mathcal{B}^{\alpha}_{p,p}(C) \subset \mathcal{WB}^{\alpha}_{p,p}(C), \tag{2.2}$$

which justifies *a posteriori* the terminology "weak Besov space" used in Chapter 1. Straightforward computations show that actually this embedding is strict. We mention that the definition of weak Besov spaces can be further extended, so that each Besov ball $\mathcal{B}^{\alpha}_{p,q}(C)$ is associated with a weak Besov ball by using a similar trick. For the sake of clarity, I do not present this technical extension here, but I refer the reader to [R4] where subsequent results are established in a more general setting. In the sequel, we make a slight abuse of notation by using sometimes the coefficients of a function f instead of f itself. In particular, we sometimes write $\beta \in \mathcal{WB}^{\alpha}_{p,p}(C)$ where $\beta = (\beta_{jk})_{jk}$.

We now consider the minimax risk associated with $\mathcal{WB}^{\alpha}_{p,p}(C)$. The loss function is the $\mathcal{B}^{\alpha'}_{p',p'}$-loss for $1 \leq p' < \infty$ and $\alpha' \geq 0$ fixed until the end of this section. In particular, the value $\alpha' = 0$ provides a conjecture of the minimax rates for the $\mathbb{L}_{p'}$-loss. More precisely, we consider the following setting. We introduce two distinct zones respectively denoted $\mathcal{R}$ and $\mathcal{C}$ and named in the sequel the regular and the critical zones. For $1 \leq p < \infty$ and $0 < \alpha < \infty$, we say that

$$(\alpha,p) \in \mathcal{R} \iff \left\{ p' > p \text{ and } p\left(\alpha+\frac{1}{2}\right) > p'\left(\alpha'+\frac{1}{2}\right) \right\} \text{ or } p' \leq p$$

$$(\alpha,p)\in\mathcal{C} \quad\Longleftrightarrow\quad p'>p \text{ and } p\left(\alpha+\frac{1}{2}\right)=p'\left(\alpha'+\frac{1}{2}\right).$$

The logarithmic zone that corresponds to $p'>p$ and $p\left(\alpha+\frac{1}{2}\right)<p'\left(\alpha'+\frac{1}{2}\right)$ is very different from the other ones and is not considered in this section. In the critical case, to evaluate the minimax risk, we need a minimal assumption on the regularity of f to control the size of the β_{jk}'s at high levels. That is the reason why we suppose that in addition, in the critical case, f lies in $\mathcal{B}^{\eta}_{p',\infty}(C)$ (with $\eta>\alpha'$ but $\eta-\alpha'$ eventually very small). So, we set

$$\Theta=\begin{cases} \mathcal{WB}^{\alpha}_{p,p}(C) & \text{on } \mathcal{R},\\ \mathcal{WB}^{\alpha}_{p,p}(C)\cap\mathcal{B}^{\eta}_{p',\infty}(C) & \text{on } \mathcal{C},\end{cases}$$

and the minimax risk we consider is from now on

$$R_\varepsilon=\inf_{\hat\beta}\sup_{\beta\in\Theta}\mathbb{E}_\beta\|\hat\beta-\beta\|^{p'}_{\mathcal{B}^{\alpha'}_{p',p'}}, \tag{2.3}$$

where the infimum is taken over all estimators. In [R4], we establish the following theorem:

THEOREM 2.1. *We set* $r=\frac{\alpha-\alpha'}{\alpha+\frac{1}{2}}$, *and*

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

We have

$$C_1\le\liminf_{\varepsilon\to0}R_\varepsilon\Psi(C,\varepsilon)^{-1}\le\limsup_{\varepsilon\to0}R_\varepsilon\Psi(C,\varepsilon)^{-1}\le C_2,$$

where C_1 *and* C_2 *are positive constants depending on* α,p,α' *and* p'. *On* $\mathcal{C}$, C_2 *also depends on* η.

Theorem 2.1 generalizes the result established by Johnstone (1994) who considered the Gaussian sequence model

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

and evaluated the minimax risk for the weak ℓ_p balls (defined by the condition (1.23)) and the ℓ_2 loss. We naturally compare the rates of convergence of the minimax risk associated respectively with $\mathcal{WB}^{\alpha}_{p,p}(C)$ and $\mathcal{B}^{\alpha}_{p,p}(C)$. Theorem 2.1 and Theorem 1 of Donoho, Johnstone, Kerkyacharian and Picard (1997) show that the rates for $\mathcal{B}^{\alpha}_{p,p}(C)$ and $\mathcal{WB}^{\alpha}_{p,p}(C)$ are the same up to constants. Even if we observed that strong Besov spaces and weak Besov spaces are close, this result may seem surprising since $\mathcal{B}^{\alpha}_{p,p}(C)$ is strictly included into $\mathcal{WB}^{\alpha}_{p,p}(C)$.

In the framework of Poisson intensity estimation, minimax rates on weak Besov balls are also established in the critical case for the $\mathbb{L}_2$-loss. Under stronger assumptions, we obtain the same rates as in Theorem 2.1. See Theorem 6 of [R10] for further details.

Then, the next goal is to have an overview over typical enemies for classical non-linear procedures. For this purpose, we consider the Bayesian setting and prior distributions π_ε on $\beta=(\beta_{jk})_{jk}$. More precisely, as explained in introduction, we naturally use *least favorable priors* (see Johnstone (1994)) of a given weak Besov ball $\mathcal{WB}^{\alpha}_{p,p}(C)$ that provide a good idea of the worst functions of $\mathcal{WB}^{\alpha}_{p,p}(C)$ to be estimated. My results about least favorable priors are quite technical and I just give a short insight here. In the context of this chapter, I recall that, in particular, such a distribution π_ε has to satisfy following properties (see [R4] for more details):

- The Bayes risk of π_ε, denoted $B(\pi_\varepsilon)$ is such that

$$C_1B(\pi_\varepsilon)\le R_\varepsilon\le C_2B(\pi_\varepsilon), \tag{2.4}$$

where C_1 and C_2 are positive constants depending only on α,p,α',p' and

$$B(\pi_\varepsilon)=\inf_{\hat\beta}\ \mathbb{E}_{\pi_\varepsilon}\ \mathbb{E}_\beta\|\hat\beta-\beta\|^{p'}_{\mathcal{B}^{\alpha'}_{p',p'}}.$$

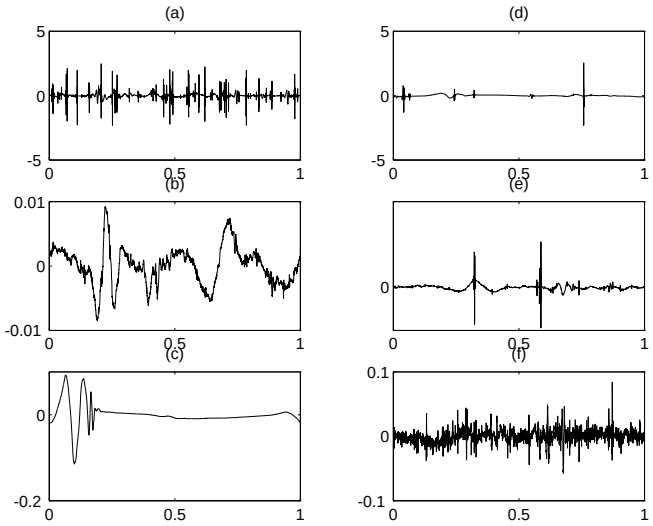

FIGURE 1. Realizations with various values of α and p; $\beta_{-10} = 0$; $n = 4096$ plotting points; (a): $p = 0.5$, $\alpha = 0.5$. (b): $p = 2$, $\alpha = 0.5$. (c): $p = 1$, $\alpha = 1.5$. (d): $p = 0.5$, $\alpha = 1.5$. (e): $p = 1$, $\alpha = 0.5$. (f): $p = 2$, $\alpha = 0$.

- The distribution π_ε is asymptotically concentrated on $\mathcal{WB}^\alpha_{p,p}(C)$:

$$\mathbb{P}_{\pi_\varepsilon}(\beta \notin \mathcal{WB}^\alpha_{p,p}(C)) \stackrel{\varepsilon\to 0}{\longrightarrow} 0. \tag{2.5}$$

Theorem 1 of [R4] shows that if π_ε is a least favorable prior, then under π_ε, the distribution of the wavelet coefficients β_{jk} takes the following form: the β_{jk}'s are independent, their distribution is symmetric with respect to 0 and $|\beta_{jk}|$ can be written:

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

where X_{jk} is a Pareto variable of parameter p, $j_* \in \mathbb{N}$, α_j and μ_j are non-negative real numbers. In addition, the convergence (2.5) occurs with an exponential rate. See Section 4 of [R4] for technical aspects. Properties (2.4) and (2.5) ensure that the typical enemies of weak Besov balls $\mathcal{WB}^\alpha_{p,p}(C)$ are well represented by simulations of least favorable priors π_ε. To shed lights on the role of the parameters p and α, some realizations of these enemies are displayed in Figure 1. As expected, the realizations are smoother when $p(\alpha + 1/2)$ is large. When the product $p(\alpha + 1/2)$ is fixed, the realizations have very high peaks when p is small with a regular behavior between them; whereas when p is great, the peaks are less high and realizations are less homogeneous between the peaks. To summary these results, we can say that when p decreases, the number of negligible coefficients increases, but the few non-negligible ones may be very large.

Since minimax risks for $\mathcal{B}^\alpha_{p,p}(C)$ and $\mathcal{WB}^\alpha_{p,p}(C)$ are the same, a natural question arises: are least favorable priors for Besov and weak Besov balls the same as well? The answer is no since we have the following result:

$$\mathbb{P}_{\pi_\varepsilon}(\beta \in \mathcal{B}^\alpha_{p,p}(C)) \stackrel{\varepsilon\to 0}{\longrightarrow} 0.$$

Actually, Johnstone (1994) pointed out least favorable priors for Besov spaces $\mathcal{B}^\alpha_{p,p}$ based on Gaussian distributions (when $p = 2$) or on two or three points distributions (when $p < 2$). Roughly speaking, we can claim that we have built a first Bayes model, based on Pareto distributions, typical of weak Besov spaces. By using an alternative approach, we further investigate this problem in the next section and strengthen these conclusions.

3. Bayesian modelling of sparsity

The goal of this section is to model sparsity by using a Bayesian approach. Remember that Definition 1.4 has introduced Lorentz spaces to measure the sparsity of signals. Here, we assume we are given an orthonormal basis $\psi = (\psi_k)_{k\in\mathbb{N}^*}$ of $\mathbb{L}_2(\mathbb{R}^d)$ and we consider signals f decomposed on ψ:

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

In this context, sparsity means that there is a relative small proportion of relative large entries of the coefficients θ_k's. So, in this section, we assume we are given some positive weights $\sigma = (\sigma_k)_{k\in\mathbb{N}^*}$ and we set for p and q two real numbers such that $0 < p < q$,

$$w\ell_{p,q}(\sigma) = \left\{ \theta = (\theta_k)_{k\in\mathbb{N}^*} : \quad \sup_{\lambda>0} \lambda^p \sum_k \mathbb{1}_{|\theta_k|>\lambda\sigma_k} \sigma_k^q < \infty \right\}. \tag{2.6}$$

Such sequence spaces naturally appear when maxisets in heteroscedastic white noise models are investigated (see Kerkyacharian and Picard (2000), [R3], [R5] or [R7]). According to Chapter 1, they naturally measure sparsity of the sequence $\theta = (\theta_k)_{k\in\mathbb{N}^*}$: if μ denotes the measure satisfying $\mu(k) = \sigma_k^q$, then

$$\theta \in w\ell_{p,q}(\sigma) \iff (\sigma_k^{-1}\theta_k)_k \in w\ell_p(\mu).$$

Sparsity is also naturally expressed by a Bayesian model. Most of the authors consider Bayesian models based upon Gaussian distributions. In the wavelet setting, 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) assumed the underlying signal to be composed of a piecewise-smooth deterministic part plus a zero-mean Gaussian part. Clyde, Parmigiani and Vidakovic (1995), Johnstone and Silverman (1998) and Abramovich, Sapatinas and Silverman (1998) have considered a mixture of a normal component and a point mass at zero for the wavelet coefficients. In the minimax approach, Johnstone and Silverman (2004, 2005) have shown the advantages in using heavy-tailed priors instead of Gaussian priors. In the wake of these works, we consider the following Bayes model: the distribution of θ is such that the θ_k's are independent and for any $k \geq 1$, there exist a fixed parameter $w_k \in (0,1)$ depending on k and a fixed symmetric density γ, such that, with probability $1 - w_k$, θ_k is equal to 0 and with probability w_k, the density of θ_k is γ_k, where for any $\theta \in \mathbb{R}$,

$$\gamma_k(\theta) = s_k \gamma(s_k \theta),$$

with $s_k > 0$. If δ_0 denotes the Dirac mass at 0, this model can be rewritten as follows:

$$\theta_k \sim (1 - w_k)\delta_0(\theta_k) + w_k \gamma_k(\theta_k), \quad k \geq 1. \tag{2.7}$$

So, roughly speaking, the first term models the negligible components and the second one non-negligible ones.

Now, the question is: can we connect the Bayesian and deterministic ways of capturing sparsity? More precisely, can we establish connections between the sequence spaces (2.6) and the model (2.7)? Actually, we would like to prove a result similar to the result proved by Abramovich, Sapatinas and Silverman (1998) and used by Abramovich, Amato and Angelini (2004). In the wavelet framework, Abramovich, Sapatinas and Silverman considered the previous Bayes model where γ is the density of a Gaussian variable with mean zero and unit variance. Then, they established a necessary and sufficient condition on the other hyperparameters of (2.7) to ensure that the signal built from the wavelet coefficients coming from (2.7) belongs, almost surely, to a prescribed Besov space (see Theorem 1 of Abramovich, Sapatinas and Silverman (1998)). Our goal is to do the same job with $w\ell_{p,q}(\sigma)$ spaces, but without fixing γ. We have the following result:

THEOREM 2.2. *Let us assume that for any $k \geq 1$ $s_k = \sigma_k^{-1}$ and we are given p and q such that $1 \leq q < \infty$ and $0 < p < q$. We note for any $\lambda \geq 0$,*

$$\tilde{F}(\lambda) = 2 \int_{\lambda}^{+\infty} \gamma(x)dx.$$

If there exists a constant C such that

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

then

$$\sup_{\lambda>0} \lambda^p \tilde{F}(\lambda) \sum_k w_k \sigma_k^q \leq C^p. \tag{2.8}$$

Conversely, if there exists a constant C such that

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

then

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

and $\theta \in w\ell_{p,q}(\sigma)$.

The previous result cannot be formulated as an equivalence because we have to fix the radius of the $w\ell_{p,q}(\sigma)$-space to obtain (2.8). But we point out the condition $\sup_{\lambda>0} \lambda^p \int_{\lambda}^{+\infty} \gamma(x)dx < \infty$, which means that the tails of γ cannot be heavier than those of a Pareto(p)-variable. Consequently, similarly to the result presented in Section 2, this result illustrates the strong connections between Pareto(p)-distributions and $w\ell_{p,q}(\sigma)$-spaces. This emphasizes the relevance of heavy-tailed distributions, and in particular of Pareto distributions, to build Bayes models designed to capture sparsity.

Now, it is of interest to study the performances of the Bayes rules associated with prior models. That is the goal of the next section.

4. Wavelet Bayesian thresholding procedures

4.1. From minimax procedures to constructive estimates. In this paragraph, we start from the wavelet procedure which achieves the minimax rates on weak Besov balls $\mathcal{WB}^{\alpha}_{p,p}(C)$. Section 4.2 of [R4] shows that the upper bound of the risk R_ε considered in Theorem 2.1 is obtained by using the following estimator based on the soft thresholding rule:

$$\hat{f}_\varepsilon = \sum_j \sum_k \text{sign}(y_{jk})(|y_{jk}| - \lambda_j)_+ \psi_{jk}, \tag{2.9}$$

with

$$\lambda_j = \begin{cases} \varepsilon\sqrt{-2\log(\alpha_j^p)} & \text{if } j \geq j_*, \\ 0 & \text{otherwise,} \end{cases} \tag{2.10}$$

where

$$j_* \overset{\varepsilon\to 0}{\sim} j^0(\varepsilon) := \min\{j \geq -1 : \quad \alpha_j < 1\}. \tag{2.11}$$

Let us note that the threshold λ_j depends on the parameters p and α_j of the least favorable priors π_ε. It is not surprising since a minimax estimator for a given function space Θ is well adapted to the worst functions of Θ that are modeled by least favorable priors. We mention that there exists j^* such that for $j > j^*$, $\alpha_j = 0$ and $\lambda_j = +\infty$. So, the sum in (2.9) is actually finite. We inspire from this minimax rule to build constructive estimators.

In the sequel, we consider the standard regression model

$$g_i = f\left(\frac{i}{n}\right) + \sigma\varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad 1 \leq i \leq n, \tag{2.12}$$

where $n = 2^N, N \in \mathbb{N}^*$. Using the discrete wavelet transform, the regression model (2.12) is reduced to the following one:

$$w_{jk} = d_{jk} + \sigma z_{jk}, \quad -1 \leq j \leq N-1, \quad k \in \mathcal{I}_j, \tag{2.13}$$

where $\mathcal{I}_j = \{k \in \mathbb{N} : \quad 0 \leq k < 2^j\}$. Since the discrete wavelet transform is an orthogonal transform, $z = (z_{jk})_{jk}$ 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)). The prior model considered in this paragraph is very close to π_ε: we suppose that the d_{jk}'s are independent and for any j and any k,

$$d_{jk} \sim F_j, \quad \text{where} \quad F_j = \frac{1}{2}(F_j^+ + F_j^-), \tag{2.14}$$

F_j^- is the reflection of F_j^+ about 0 and if X_j is a Pareto variable of parameter p, F_j^+ is the distribution of $\min(\alpha_j X_j - \alpha_j, \mu_j)$. Finally, α_j and μ_j are positive real numbers. Because of its improper nature, we place no prior on the scaling coefficient d_{-10}. The distribution F_j is a slight modification of the distribution of $\varepsilon^{-1}\beta_{jk}$ of Section 2. In particular, to avoid any discontinuity in the definition of the support of d_{jk}, we translate the variable $\alpha_j X_j$ by α_j. This slight modification enables us to capture very small values of d_{jk}. We suppose that the parameter α_j has the form

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

where C and δ are positive constants. The parameter μ_j, that tends to $+\infty$ when j goes to $+\infty$, does not play any role in the sequel. So, its value is not specified in this manuscript. To estimate d_{jk}, we propose

$$d_{jk}^* = \text{sign}(w_{jk})(|w_{jk}| - \lambda_j)_+,$$

where λ_j has the following form

$$\lambda_j = \begin{cases} \sigma\sqrt{-2\log(\alpha_j^p)} & \text{if } j \geq j_1, \\ 0 & \text{otherwise,} \end{cases}$$

where j_1 is the first integer j such that $\alpha_j < 1$. Then, the threshold λ_j can be rewritten as follows:

$$\lambda_j = \sigma\sqrt{\max\left(0, -2\log(\alpha_j^p)\right)}. \tag{2.15}$$

The function to be estimated is supposed to belong to the class of the worst functions to be estimated of an unknown weak Besov ball. So, under this assumption, we expect this thresholding procedure to achieve good performances since it is strongly inspired by the previous minimax procedure and more precisely by equations (2.10) and (2.11).

To apply our procedure, it is necessary to specify the values of C and δ that define α_j and the value of p. Of course, the weak Besov ball in which f lies is unknown, so we have to estimate these hyperparameters. The parameter p is not estimated and will be taken equal to 1 in Section 4.3 where numerical performances of this procedure are analyzed. For the estimation of (C, δ), we set

$$\hat{N}_j(\lambda^u) = \frac{1}{2^j}\sum_{k\in\mathcal{I}_j} \mathbb{1}_{\ |w_{jk}|>\lambda^u},$$

where λ^u is the universal threshold defined by $\lambda^u = \sigma\sqrt{2\log(n)}$, and we set

$$\hat{\alpha}_j = \lambda^u \hat{N}_j(\lambda^u)^{\frac{1}{p}}(1 - \hat{N}_j(\lambda^u)^{\frac{1}{p}})^{-1}. \tag{2.16}$$

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	6.93	0
j=5	0	0	8.07	0
j=6	0.46	0	8.62	1.22
j=7	3.53	2.08	8.68	3.61
j=8	5.00	3.45	8.69	4.92
j=9	6.13	4.41	8.69	5.95

TABLE 1. Values of $\hat{\lambda}_j$ $(p = 1)$ associated with 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'; n=1024; rsnr=3 ($\sigma = 7/3$); $\lambda^u = 8.69$.

We estimate C and δ by using the linear regression:

$$(\hat{C}, \hat{\delta}) = \text{argmin}_{C,\delta} \sum_{j \in \mathcal{S}} (\log(\hat{\alpha}_j) - \log(C) + j\delta \log(2))^2,$$

where

$$\mathcal{S} = \{j \in \{1, \dots, N-1\} : \quad \hat{\alpha}_j \in (0, +\infty)\}$$

when $\text{card}(\mathcal{S}) \geq 2$. In this case, we set

$$\hat{\lambda}_j = \sigma \sqrt{\max\left(0, -2p \log(\hat{C} 2^{-j\hat{\delta}})\right)}. \tag{2.17}$$

If $\text{card}(\mathcal{S}) \leq 1$, we set

$$\hat{\lambda}_j = \begin{cases} 0 & \text{if } \hat{\alpha}_j = +\infty, \\ \sigma \sqrt{\max\left(0, -2p \log(\hat{\alpha}_j)\right)} & \text{for } j \in \mathcal{S}, \\ \lambda^u & \text{if } \hat{\alpha}_j = 0. \end{cases} \tag{2.18}$$

Table 1 gives the average over 100 replications of the values of the level-dependent threshold $\hat{\lambda}_j$ associated with the classical four test functions Blocks, Bumps, Heavisine and Doppler (see Donoho and Johnstone (1994)). Before going further, let us give a brief justification of this procedure (more details are given in [R2]): we notice that for all $\lambda < \mu_j$,

$$\mathbb{P}(|d_{jk}| > \lambda) = \left(\frac{\alpha_j}{\alpha_j + \lambda}\right)^p.$$

But, using extended Glivenko-Cantelli's Theorem,

$$\sup_{\lambda > 0} \left| \frac{1}{2^j} \sum_{k \in \mathcal{I}_j} \mathbb{1}_{|d_{jk}| > \lambda} - \mathbb{P}(|d_{jk}| > \lambda) \right| \overset{j \to \infty}{\longrightarrow} 0 \quad \text{a.e.}$$

Therefore, $\left(\frac{\alpha_j}{\alpha_j + \lambda}\right)^p$ 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 $\hat{N}_j(\lambda^u)$. So,

$$\left(\frac{\alpha_j}{\alpha_j + \lambda^u}\right)^p \approx \hat{N}_j(\lambda^u)$$

and

$$\alpha_j = C 2^{-j\delta} \approx \hat{\alpha}_j = \lambda^u \hat{N}_j(\lambda^u)^{\frac{1}{p}} (1 - \hat{N}_j(\lambda^u)^{\frac{1}{p}})^{-1}.$$

This provides a theoretical justification for (2.16). A numerical justification is given by Table 1 of [R2]. The pair of equations (2.17) and (2.18) are naturally justified by (2.15).
Now, we set

$$\hat{d}_{jk} = \text{sign}(w_{jk})(|w_{jk}| - \hat{\lambda}_j)_+,$$

for all $j \geq 0$, $k \in \mathcal{I}_j$, and $\hat{d}_{-10} = w_{-10}$. Finally, the estimate of the signal f is obtained by applying the inverse discrete wavelet transform to the vector $\hat{d} = (\hat{d}_{jk})_{jk}$. The performances of this Bayesian thresholding procedure, denoted *ParetoThresh*, are analyzed in Section 4.3.

4.2. Heavy-tailed and large variance Gaussian priors. This paragraph studies the theoretical performances of Bayesian procedures based on the commonly used prior model of the following form:

$$\beta_{jk} \sim p_{j,\epsilon}\gamma_{j,\epsilon} + (1 - p_{j,\epsilon})\delta(0). \tag{2.19}$$

We consider the statistical white noise model of Section 2 and the wavelet setting described there. Here, the β_{jk}'s are independent, δ_0 still denotes the Dirac mass at 0, and $p_{j,\epsilon} \in [0, 1]$ can be interpreted as the proportion of non-negligible coefficients. The non-zero part of the prior, $\gamma_{j,\epsilon}$, is assumed to be the dilation of a fixed symmetric, positive, unimodal and continuous density γ:

$$\gamma_{j,\epsilon}(\beta_{jk}) = \frac{1}{\tau_{j,\epsilon}}\gamma\left(\frac{\beta_{jk}}{\tau_{j,\epsilon}}\right),$$

where the dilation parameter $\tau_{j,\epsilon}$ is positive. The most popular choice for γ is the normal density. When the noise is Gaussian, it is also the density giving rise to the easiest procedures from a computational point of view (the prior and the noise are conjugate). From the minimax point of view, recent works have studied these Bayes procedures and it has been proved that Bayes rules can achieve optimal rates of convergence. Abramovich, Amato and Angelini (2004) investigated theoretical performance of the procedures introduced by Abramovich, Sapatinas and Silverman (1998) based on the Gaussian prior model (2.19) with

$$\tau^2_{j,\epsilon} = c_1 2^{-aj}, \quad p_{j,\epsilon} = \min(1, c_2 2^{-bj}), \tag{2.20}$$

where c_1, c_2, a and b are positive constants. For the mean squared error, they proved that the non adaptive posterior mean and posterior median achieve optimal rates up to a logarithmic factor on the Besov spaces $\mathcal{B}^{\alpha}_{p,q}$ when $p \geq 2$. When $p < 2$, these estimators only behave as linear estimates. Recently, Johnstone and Silverman (2004, 2005) investigated minimax properties of Bayes rules, with priors based on heavy-tailed distributions in an empirical Bayes setting. In this case, the posterior mean and median turn out to be optimal for the whole scale of Besov spaces.

The main goal of this paragraph is to push a little further comparisons of Bayesian procedures by adopting the maxiset point of view. In particular, since Gaussian priors have very interesting properties from the computational point of view, one of our motivations is to answer the following question: Are Gaussian priors always outperformed by heavy-tailed priors? And quite happily, one of our results is to show that if some Bayesian procedures using Gaussian priors behave quite unwell (in terms of maxisets as it was the case in terms of minimax rates) compared to those with heavy tails, it is nevertheless possible to attain a very good maxiset behavior. We prove that this can only be achieved under the condition that the hyperparameter $\tau_{j,\epsilon}$ is "large". Under this assumption, the density $\gamma_{j,\epsilon}$ is then more spread around 0, mimicking in some ways the behavior of a distribution with heavy-tails. Moreover, we prove that these procedures can be built in an adaptive way: their construction does not depend on the specified regularity or sparsity of the function at hand.

In the sequel, we consider either the posterior median or the posterior mean of each β_{jk} and we denote:

$$\breve{\beta}_{jk} = \text{Med}(\beta_{jk}|y_{jk}) \quad \text{and} \quad \tilde{\beta}_{jk} = \mathbb{E}(\beta_{jk}|y_{jk}).$$

If $\hat{f}_\varepsilon$ is an estimator and ρ_ε a rate of convergence, we set for any $R>0$:

$$MS(\hat{f}_\varepsilon,\rho_\varepsilon)(R)=\left\{f:\quad \sup_\varepsilon\left\{\rho_\varepsilon^{-2}\mathbb{E}\left[\|\hat{f}_\varepsilon-f\|_{\mathbb{L}_2}^2\right]\right\}\leq R^2\right\}$$

and we use Convention 1.1. Finally, we note

$$t_\varepsilon=\varepsilon\sqrt{|\log\varepsilon|}.$$

We first consider the procedure proposed by Abramovich, Sapatinas and Silverman (1998) and studied in the minimax setting by Abramovich, Amato and Angelini (2004). Consequently, we take the hyperparameters defined in (2.20) with $0\leq b<1$ and γ the density of $\mathcal{N}(0,1)$. We set

$$\breve{f}_\varepsilon(\tau,p)=\sum_{j,k}\breve{\beta}_{jk}\psi_{jk}\quad\text{and}\quad\tilde{f}_\varepsilon(\tau,p)=\sum_{j,k}\tilde{\beta}_{jk}\psi_{jk}.$$

As shown by Abramovich, Amato and Angelini (2004), the case $a<2\alpha+1$ is of no interest and optimal minimax properties are achieved with $a=2\alpha+1$. We obtain the following result.

THEOREM 2.3. *Let $\alpha>0$. We assume that $\hat{f}_\varepsilon$ is either $\breve{f}_\varepsilon(\tau,p)$ or $\tilde{f}_\varepsilon(\tau,p)$.*

(1) *For $a>2\alpha+1$, $\mathcal{B}^\alpha_{p,\infty}\not\subset MS(\hat{f}_\varepsilon,t_\varepsilon^{4\alpha/(1+2\alpha)})$ for any $1\leq p\leq\infty$.*
(2) *For $a=2\alpha+1$, $\mathcal{B}^\alpha_{p,\infty}\not\subset MS(\hat{f}_\varepsilon,t_\varepsilon^{4\alpha/(1+2\alpha)})$ if $p<2$.*
(3) *For $a=2\alpha+1$, $MS(\hat{f}_\varepsilon,\epsilon^{4\alpha/(1+2\alpha)})\subsetneq\mathcal{B}^\alpha_{2,\infty}$.*

Theorem 2.3 shows that the posterior median and mean associated with Gaussian priors and hyperparameters (2.20) do not achieve a suitable behavior. The last point even shows that they are outperformed by linear estimates (see [R1]).

We now consider heavy-tailed priors. We assume that there exist two positive constants M and M_1 such that

$$\sup_{x\geq M_1}\left|(\log\gamma)'(x)\right|=M<\infty. \tag{2.21}$$

Assumption (2.21) means that the tails of γ have to be exponential or heavier. Indeed, under (2.21), we have for any $x\geq M_1$:

$$\gamma(x)\geq\gamma(M_1)\exp(-M(x-M_1)).$$

To complete the prior model, we assume that:

$$\tau_{j,\epsilon}=\varepsilon\quad\text{and}\quad p_{j,\epsilon}=p_\varepsilon\overset{\varepsilon\to0}{\longrightarrow}0, \tag{2.22}$$

where $\varepsilon\longrightarrow p_\varepsilon$ a positive continuous function. We set

$$\breve{f}_\varepsilon(\tau,p,j_\varepsilon)=\sum_{j<j_\varepsilon}\sum_k\breve{\beta}_{jk}\psi_{jk}\quad\text{and}\quad\tilde{f}_\varepsilon(\tau,p,j_\varepsilon)=\sum_{j<j_\varepsilon}\sum_k\tilde{\beta}_{jk}\psi_{jk}, \tag{2.23}$$

where j_ε is such that $2^{j_\varepsilon}=\lfloor t_\varepsilon^{-2}\rfloor$. Proposition 1 of [R3] states that these estimates are shrinkage rules and $\breve{f}_\varepsilon(\tau,p,j_\varepsilon)$ is even a thresholding rule. So, we expect these procedures to mimic classical thresholding rules from the maxiset point of view, at least when the posterior median is considered. Indeed Theorems 2, 3, 4 and 5 in [R5] lead to the following result.

THEOREM 2.4. *Let $\alpha>0$. Under (2.22), we suppose that there exist two positive constants ρ_1 and ρ_2 such that for $\varepsilon>0$ small enough,*

$$\varepsilon^{\rho_1}\leq p_\varepsilon\leq\varepsilon^{\rho_2}.$$

Then, we have:

$$MS(\hat{f}_\varepsilon,t_\varepsilon^{4\alpha/(1+2\alpha)}):=:\mathcal{B}^{\frac{\alpha}{1+2\alpha}}_{2,\infty}\cap W_{\frac{2}{2\alpha+1}},$$

where $\hat{f}_\varepsilon$ is either $\breve{f}_\varepsilon(\tau,p,j_\varepsilon)$ or $\tilde{f}_\varepsilon(\tau,p,j_\varepsilon)$, as soon as $\rho_2\geq16$ for the posterior median and $\rho_2\geq64$ for the posterior mean.

So, the performances achieved by adaptive Bayesian procedures based on heavy-tailed prior densities are similar to those of classical non-linear procedures in the maxiset framework. In particular, they outperform the previous Gaussian Bayesian procedures from the maxiset point of view.

Now, we wonder whether heavy-tailed priors are unavoidable. Is it possible to build Gaussian priors leading to procedures with maxiset properties comparable to the properties of heavy-tailed methods discussed above? This is a very important issue since computations using Gaussian priors are mostly direct and obviously much easier than heavy-tailed priors. The answers are provided by the following theorem.

THEOREM 2.5. *Let $\alpha > 0$. We consider the prior model (2.19), where γ is the Gaussian density. As previously, we assume that $\tau_{j,\epsilon} = \tau_\varepsilon$ and $p_{j,\epsilon} = p_\varepsilon$ are independent of j. We consider estimates defined in (2.23) with following hyperparameters τ_ε and p_ε. If*

$$1 + \varepsilon^{-2}\tau_\varepsilon^2 = t_\varepsilon^{-1}$$

and there exist q_1 and q_2 such that for ε small enough

$$\varepsilon^{q_1} \leq p_\varepsilon \leq \varepsilon^{q_2},$$

we have:

$$MS(\hat{f}_\varepsilon, t_\varepsilon^{4\alpha/(1+2\alpha)}) := \mathcal{B}_{2,\infty}^{\frac{\alpha}{1+2\alpha}} \cap W_{\frac{2}{2\alpha+1}},$$

where $\hat{f}_\varepsilon$ is either $\tilde{f}_\varepsilon(\tau, p, j_\varepsilon)$ or $\breve{f}_\varepsilon(\tau, p, j_\varepsilon)$ as soon as $q_2 > 63/2$ for the posterior median and $q_2 \geq 65/2$ for the posterior mean.

Unlike the previous choice ($\tau_{j,\epsilon}^2 = \varepsilon^2$ or $\tau_{j,\epsilon}^2 = c_1 2^{-j\alpha}$), here we impose a "larger" variance:

$$\tau_{j,\epsilon}^2 \overset{\varepsilon \to 0}{\sim} \frac{\varepsilon}{\sqrt{|\log \varepsilon|}}.$$

It is the key point of the proof of Theorem 2.5. In a sense, we re-create the heavy tails by increasing the variance.

4.3. Numerical comparison study. We now investigate the behavior of Bayesian procedures from a practical point of view and show a comparative simulations study with several standard and Bayesian procedures of the literature. For this purpose, we consider the regression model (2.12) (with $n = 1024$ observations) considered in Section 4.1 and its transformation (2.13) by using the discrete wavelet transform:

$$w_{jk} = d_{jk} + \sigma z_{jk}, \quad -1 \leq j \leq 9, \quad k \in \mathcal{I}_j.$$

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. In the subsequent applications of ParetoThresh, we take $p = 1$ for every function, which provides quite good results. However, we shall discuss below the effect of varying p. To implement the Bayes rules based on Gaussian priors studied in Theorem 2.5 we reconstruct the d_{jk}'s, with the posterior median and the posterior mean of a prior having the following form:

$$d_{jk} \sim \frac{\omega_n}{1+\omega_n}\, \gamma_{j,n} + \frac{1}{1+\omega_n}\, \delta(0),$$

where $\omega_n = \frac{10\sigma}{\sqrt{n}}$, γ is the Gaussian density and

$$\gamma_{j,n}(d_{jk}) = \frac{1}{\tau_n}\, \gamma\left(\frac{d_{jk}}{\tau_n}\right),$$

with τ_n is such that $\frac{n\tau_n^2}{\sigma^2 + n\tau_n^2} = 0,999$. We respectively denote *GaussMedian* and *GaussMean* the posterior median and mean.

RSNR=3	Blocks	Bumps	Heavisine	Doppler
VisuShrink	3.88	5.63	0.32	1.59
SureShrink	1.93	2.15	0.31	0.89
ParetoThresh $(p = 1)$	1.56	1.78	0.30	0.78
GaussMedian	1.50	1.73	0.33	0.70
GaussMean	1.45	1.62	0.32	0.64
CauchyMean	1.41	1.52	0.27	0.60
BayesFactor	1.80	1.83	0.36	0.82
BayesThresh	1.47	1.62	0.31	0.67
RSNR=5	Blocks	Bumps	Heavisine	Doppler
VisuShrink	2.10	2.66	0.18	0.78
SureShrink	0.86	0.82	0.15	0.38
ParetoThresh $(p = 1)$	0.68	0.72	0.13	0.36
GaussMedian	0.72	0.76	0.20	0.30
GaussMean	0.62	0.68	0.19	0.29
CauchyMean	0.55	0.63	0.13	0.27
BayesFactor	0.67	0.70	0.24	0.34
BayesThresh	0.61	0.65	0.14	0.28
RSNR=8	Blocks	Bumps	Heavisine	Doppler
VisuShrink	1.07	1.26	0.12	0.40
SureShrink	0.34	0.37	0.09	0.19
ParetoThresh $(p = 1)$	0.32	0.35	0.07	0.17
GaussMedian	0.32	0.36	0.08	0.15
GaussMean	0.30	0.31	0.08	0.13
CauchyMean	0.25	0.28	0.07	0.14
BayesFactor	0.29	0.31	0.08	0.14
BayesThresh	0.27	0.30	0.07	0.13

TABLE 2. AMSEs for VisuShrink, SureShrink, ParetoThresh $(p = 1)$, GaussMedian, GaussMean, CauchyMean, BayesFactor and BayesThresh with various test functions and various values of the RSNR.

We compare our procedure to some classical procedures described in Section 4.2 of [R2] or Section 6 of [R5]: VisuShrink (Donoho and Johnstone (1994)), SureShrink (Donoho and Johnstone (1995)) for which we do not threshold the five coarsest levels, BayesFactor (Vidakovic (1998)), BayesThresh (proposed by Abramovich, Sapatinas and Silverman (1998) studied in Theorem 2.3) and CauchyMean (proposed by Johnstone and Silverman (2004, 2005) based on the heavy-tailed quasi-Cauchy prior that satisfies Assumption (2.21)). The Symmlet 8 wavelet filter is used for all the methods. The performance of each procedure is measured by using the mean-squared error associated to an estimator $\hat{f}$:

$$\text{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}\left(\frac{i}{n}\right) - f\left(\frac{i}{n}\right)\right)^2.$$

Table 2 shows the average mean-squared error (denoted AMSE) using 100 replications with different values for the root signal to noise ratio (RSNR). See [R2] and [R5] for graphs of reconstructions. Table 2 shows that 'purely Bayesian' procedures (GaussMedian, GaussMean, CauchyMean, BayesFactor, BayesThresh) are preferable to 'purely deterministic' ones (VisuShrink and SureShrink). Furthermore, CauchyMean provides the best behaviors here (its AMSEs are globally the smallest). We add that Table 2 and [R5] show that Bayesian rules using the posterior mean have better performances than those using the posterior median.

ParetoThresh has smaller mean-squared errors than SureShrink and VisuShrink but has generally larger mean-squared errors than the 'purely Bayesian procedures'. If the Bayesian methods achieve quite good performances under the AMSE approach, high-frequency artefacts may appear, whereas VisuShrink and SureShrink provide the best methods for removing the noise. But for ParetoThresh, these artefacts may partially disappear if we take small values of p. This effect could be expected taking into account the conclusions we have drawn from the realizations of section 4.1. We remark that except for 'Doppler' for which the AMSE (RSNR=3) attains its minimum for $p = 0.7$, this improvement has a cost: the AMSE increases. When p is larger than 1, the AMSEs are worse and there are more artefacts. Let us also mention that a possible alternative is to use the hard thresholding rule with $(\hat{\lambda}_j)_j$ or a Bayes rule (the posterior median or the posterior mean). But we do not obtain better results as far as the AMSE is concerned. Furthermore, the choice of the hard thresholding rule provides less regular reconstructions.

To conclude, if, as said previously, 'purely Bayesian' procedures are preferable to 'purely deterministic' ones from the AMSE point of view, they may have a drawback not shared by VisuShrink and SureShrink: their high computational time. ParetoThresh built by using a prior, without being 'purely Bayesian', does not have to cope with this handicap: the parameters of the prior model that define the level-dependent thresholds are easily computed. So, ParetoThresh appears as a good compromise between 'purely Bayesian' and 'purely deterministic' procedures.

5. Conclusions

The Bayesian approach of sparsity enhances heavy-tailed priors and Pareto distributions. And, roughly speaking, various points of view have shown that the Pareto parameter provides the parameter of the Lorentz space containing the realizations of the prior model. These studies lie in the wake of other works that have popularized the use of heavy-tailed priors in many frameworks. As an illustration of this statement, I just mention the paper by Dalalyan and Tsybakov (2009) who established sparse oracle inequalities for aggregation in the PAC-Bayesian setting. Of course, the numerical difficulties raised by such Bayes models (that can nevertheless be overcome by Monte-Carlo methods) explain why some reluctance for the use of heavy-tailed priors remains. In this case, we have observed that Gaussian distributions with large variance constitute an alternative.

The possible extensions of the results presented earlier have naturally evolved with time, and are not the same as in the early 2000s when most of previous works have been written. In particular, recent contributions have answered many questions related to heavy-tailed priors. For instance, minimax optimality of the associated Bayes rules has been established by Johnstone and Silverman (2004, 2005), Pensky (2006) and Pham Ngoc (2009) in different settings. More generally, recent works have provided major advances for the frequentist study of Bayes rules either from the methodological and practical points of view (in the spirit of [R2]) or from the theoretical point of view (in the spirit of [R5]). I just mention the most significant papers among the recent published ones: Abramovich, Grinshtein and Pensky (2007), Abramovich, Angelini and de Canditiis (2007), Pensky and Sapatinas (2007), Bochkina and Sapatinas (2005, 2006, 2009), Silverman (2007) and Cutillo, Jung, Ruggeri and Vidakovic (2008). Of course, some problems remain open as, for instance, the sharp study of Bayes rules associated with block prior models built to model possible dependency between coefficients.

As far as I am concerned, one of the most exciting topic for further research consists in the study of asymptotic properties of posterior distributions in the non-parametric or semi-parametric settings, where there is still little work. Most of existing works deal with consistency or, more recently, with concentration rates. We refer the reader to Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001), Ghosal and van der Vaart (2007) and Gayraud and Rousseau (2005). There is also a growing literature on more specific cases, i.e. when specific families of non-parametric prior distribution are considered, see for instance van der Vaart and van Zanten (2008), Scricciolo (2006) and Castillo (2008) who obtained non adaptive minimax posterior concentration rates based on Dirichlet mixtures and Gaussian process priors. Recently some adaptive minimax

concentration rates have been obtained (see for instance Belitser and Ghosal (2003), Huang (2004), Kruijer, Rousseau and van der Vaart (2009), van der Vaart and van Zanten (2009) or Rousseau (2009)). Beyond their own interest, deriving non-parametric posterior rates often constitutes the first step to obtain Bernstein Von Mises theorems. The latter consists in the asymptotic normality of the posterior distribution centered at some kind of maximum likelihood estimator with variance equal to the asymptotic frequentist variance of the centering point. This kind of results has many interesting applications. To my opinion its major application is in ensuring that Bayesian credible sets have good frequentist coverage. I mention that in a joint work [R14] with Judith Rousseau, we are investigating Bernstein-Von Mises properties in semi-parametric models where the parameter of interest is a linear functional of the density of the observations such as the cumulative distribution function when the prior puts positive mass on absolutely continuous densities with respect to the Lebesgue measure. For this purpose, we establish concentration rates for posterior distributions on Sobolev and Besov balls in the Fourier and wavelet settings (note that we only consider Besov balls $\mathcal{B}^{\alpha}_{p,q}(C)$ with $p \geq 2$). This work is in progress. A natural extension of this work is the study of Bernstein-Von Mises properties for non-linear functionals such as integrated quadratic functionals studied in Section 3 of Chapter 1. This problem is of particular interest since the minimax rate is not always parametric and varies according to the regularity of the underlying function f. Another interesting problem is the computation of concentration rates for posterior distributions on Besov balls $\mathcal{B}^{\alpha}_{p,q}(C)$ with $p < 2$.

As seen previously, from the frequentist point of view, the power of the Bayes approach is to provide a complete methodology to construct estimation rules: we first choose a Bayes model depending on hyperparameters, then we use simple heuristics to estimate these hyperparameters. Finally, we rely on natural posterior Bayes rules, such as the posterior median or the posterior mean for instance, to obtain the final estimate. In particular, in the wavelet setting, this approach provides an automatic way of calibrating shrinkage or thresholding procedures. This methodology has successfully been applied in many works but in particular by Abramovich, Sapatinas and Silverman (1998), Vidakovic (1998) and Johnstone and Silverman (2004, 2005), illustrating the good performances of Bayes rules in practice. However, elicitation of the hyperparameters (for instance by using the empirical Bayes approach) makes the theoretical study of these estimators difficult which remains in general an open question (except for the procedure proposed by Johnstone and Silverman (2004, 2005)). So, this constitutes a wide quasi-unexplored research field. In the sequel, we shall deal with this calibration problem for thresholding rules from both practical and theoretical points of view. It will be studied in the standard frequentist approach and will rely on sharp concentration inequalities. Before this, we revisit very classical non-parametric estimation problems handled in the next chapter outside the framework provided by standard assumptions.

CHAPTER 3

Assumption-free non-parametric estimation

1. Introduction

This chapter deals with the classical problem of density or Poisson intensity estimation for unidimensional data. Such statistical problems are questions that lie at the core of many data preprocessing. From this point of view, no assumption should be made on the underlying function to estimate. So, our aim is to provide an adaptive method which requires as few assumptions as possible on the underlying signal. In particular, we do not want to have any assumption on the density support or on the radius of the ball of $\mathbb{L}_\infty$ the signal belongs to. Moreover, this method should be quite easy to implement and should have good theoretical performances as well.

As noted in [R13], in our setting, methodologies based on kernel methods are the most widespread in practice and most of them are intensively based on cross-validation. These methods do not require in general the preliminary knowledge of the support but do not provide theoretical guarantees from the minimax point of view. Concerning wavelet thresholding, the DWT algorithm due to Mallat (1989) combined with a keep or kill rule on each coefficient makes these methods as one of the easiest adaptive methods to implement, once the threshold is known. After rescaling and binning the data as in Antoniadis, Grégoire and Nason (1999) for instance, one can reasonably think that the number of observations in a "not too small" interval is Gaussian. So basically the thresholding rules adapted to the Gaussian regression setting should work here. But, Herrick, Nason and Silverman (2001) have observed that in practice the basic Gaussian approximation for general wavelet bases was quite poor. Furthermore, we shall see in Section 2.2 that this method relies heavily on the precise knowledge of the support so that the size of the bins has to be adequately chosen. Finally, model selection estimates fundamentally depend on the a priori knowledge of the support to choose the model collections (see Section 1 of [R13] for more details).

The goal of Section 2 is to suggest a wavelet thresholding procedure based on data-driven thresholds. If the signal to be estimated is denoted f, we just assume that $f \in \mathbb{L}_1(\mathbb{R})$ and $f \in \mathbb{L}_2(\mathbb{R})$. The first assumption is natural since f is either a density or an intensity. We use the second one to allow $\mathbb{L}_2$-decompositions of f on an orthonormal basis. We emphasize that in particular, in the sequel, f can be unbounded and nothing is said about its support which can be unknown or even infinite. Our results are based on the very general Theorem 3.1. Its purpose is to give general conditions under which our estimate satisfies oracle inequalities. Since it does not depend on the statistical model at hand, it can be viewed as en extension of Theorem 3.1 of Kerkyacharian Picard (2000). It is then applied in the density and Poisson models.

Then, in Section 3, we extend this setting by using a dictionary of functions (instead of a single basis) in the recently widely investigated framework of the curse of dimension (that can be denoted by using subsequent notations '$M \gg n$'). We aim at establishing sharp oracle inequalities under very mild assumptions on the dictionary. Our starting point is that most of the papers in the literature assume that the functions of the dictionary are bounded by a constant independent of M and n, which constitutes a strong limitation, in particular for dictionaries based on histograms or wavelets (see for instance Bunea, Tsybakov and Wegkamp (2006, 2007a, 2007b, 2007c), Bunea (2008) or van de Geer (2008)). We propose a data-driven Dantzig procedure for which such assumptions on the functions of the dictionary will not be considered. Furthermore, I mention that, in contrast with what Bunea, Tsybakov and Wegkamp (2009) did, we obtain oracle inequalities with leading constant 1, that are established under much weaker assumptions on the dictionary.

2. Theoretical and numerical studies of a data-driven wavelet thresholding procedure

As said in Introduction, we first state a general result that can be viewed as a generalization of Theorem 3.1 of Kerkyacharian and Picard (2000). The following theorem is self-contained and can be used in various settings, as done in the sequel. This is the main reason for the following very abstract formulation.

THEOREM 3.1. *To estimate a countable family $\beta = (\beta_\lambda)_{\lambda\in\Lambda}$, such that $\|\beta\|_{\ell_2} < \infty$, we assume that a family of coefficient estimators $(\hat\beta_\lambda)_{\lambda\in\Gamma}$, where Γ is a known deterministic subset of Λ, and a family of possibly random thresholds $(\eta_\lambda)_{\lambda\in\Gamma}$ are available and we consider the thresholding rule $\tilde\beta = (\hat\beta_\lambda 1_{|\hat\beta_\lambda|\geq\eta_\lambda} 1_{\lambda\in\Gamma})_{\lambda\in\Lambda}$. Let $\varepsilon > 0$ be fixed. Assume that there exist a deterministic family $(F_\lambda)_{\lambda\in\Gamma}$ and three constants $\kappa \in [0,1[$, $\omega \in [0,1]$ and $\mu > 0$ (that may depend on ε but not on λ) with the following properties.*

(A1) *For all λ in Γ,*

$$\mathbb{P}(|\hat\beta_\lambda - \beta_\lambda| > \kappa\eta_\lambda) \leq \omega.$$

(A2) *There exist $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$ and a constant $R > 0$ such that for all λ in Γ,*

$$\left(\mathbb{E}(|\hat\beta_\lambda - \beta_\lambda|^{2p})\right)^{\frac{1}{p}} \leq R\max(F_\lambda, F_\lambda^{\frac{1}{p}}\varepsilon^{\frac{1}{q}}).$$

(A3) *There exists a constant θ such that for all λ in Γ such that $F_\lambda < \theta\varepsilon$*

$$\mathbb{P}(|\hat\beta_\lambda - \beta_\lambda| > \kappa\eta_\lambda, |\hat\beta_\lambda| > \eta_\lambda) \leq F_\lambda\mu.$$

Then the estimator $\tilde\beta$ satisfies

$$\frac{1-\kappa^2}{1+\kappa^2}\mathbb{E}\|\tilde\beta-\beta\|_{\ell_2}^2 \leq \mathbb{E}\inf_{m\subset\Gamma}\left\{\frac{1+\kappa^2}{1-\kappa^2}\sum_{\lambda\notin m}\beta_\lambda^2 + \frac{1-\kappa^2}{\kappa^2}\sum_{\lambda\in m}(\hat\beta_\lambda-\beta_\lambda)^2 + \sum_{\lambda\in m}\eta_\lambda^2\right\} + LD\sum_{\lambda\in\Gamma}F_\lambda$$

with

$$LD = \frac{R}{\kappa^2}\left(\left(1+\theta^{-1/q}\right)\omega^{1/q} + (1+\theta^{1/q})\varepsilon^{1/q}\mu^{1/q}\right).$$

Observe that this result makes sense only when $\sum_{\lambda\in\Gamma}F_\lambda < \infty$ and in this case, if LD (which stands for large deviation inequalities) is small enough, the main term of the right hand side is given by the first term. Comments of Assumptions (A1), (A2) and (A3) can be found in Section 4.1 of [R10]. In the sequel, this result is applied in the Poisson intensity estimation and density estimation settings.

The Poisson intensity model (see [R10]): We just give notations for this statistical model. More details can be found in [R10]. In the sequel, we consider a Poisson process on the real line, denoted N, whose mean measure μ is finite and absolutely continuous with respect to the Lebesgue measure. Given n a positive integer, we introduce $f_p \in \mathbb{L}_1(\mathbb{R})$ the intensity of N as

$$f_p(x) = \frac{d\mu_x}{ndx}.$$

Since f_p belongs to $\mathbb{L}_1(\mathbb{R})$, the total number of points of the process N, denoted $N_\mathbb{R}$, satisfies $\mathbb{E}(N_\mathbb{R}) = n\|f_p\|_1$ and $N_\mathbb{R} < \infty$ almost surely. In the sequel, f_p will be held fixed and n will go to $+\infty$. The introduction of n could seem artificial, but it allows to present our asymptotic theoretical results in a meaningful way. We denote by dN the discrete random measure $\sum_{T\in N}\delta_T$. Hence we have for any compactly supported function g, $\int g(x)dN_x = \sum_{T\in N} g(T)$. In this framework, our goal is to estimate f_p by using the realizations of N.

The density model (see [R13]): In this model, the goal is to estimate a density f_d from the observations of an iid n-sample of density f_d, denoted $X_1, \ldots, X_n$.

In both cases, to apply Theorem 3.1, we specify the coefficients $\beta = (\beta_\lambda)_{\lambda\in\Lambda}$, the estimates $(\hat\beta_\lambda)_{\lambda\in\Gamma}$ and the random thresholds $(\eta_\lambda)_{\lambda\in\Gamma}$. In the sequel, we consider the signal f that can be either f_p or f_d and is assumed to belong to $\mathbb{L}_2(\mathbb{R})$. We consider a special biorthogonal wavelet basis and the decomposition of f on this basis takes the following form:

$$f = \sum_{k\in\mathbb{Z}} \beta_{-1k}\tilde\phi_k + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}\tilde\psi_{jk}, \tag{3.1}$$

where for any $j \geq 0$ and any $k \in \mathbb{Z}$,

$$\beta_{-1k} = \int_{\mathbb{R}} f(x)\phi_k(x)dx, \quad \beta_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)dx.$$

The most basic example of biorthogonal wavelet basis is the Haar basis where the father wavelets are given by

$$\phi_k = \tilde\phi_k = 1_{[k;k+1]}$$

and the mother wavelets are given by

$$\psi_{jk} = \tilde\psi_{jk} = 2^{j/2}\left(1_{[k2^{-j};(k+1/2)2^{-j})} - 1_{[(k+1/2)2^{-j};(k+1)2^{-j}]}\right).$$

The other examples we consider are more precisely described in Section 2.2 of [R10] and Appendix A of [R13] but we just mention that they are obtained by the standard dilations and translations of four compactly supported functions $\phi, \psi, \tilde\phi, \tilde\psi$. The essential feature is that it is possible to use, on the one hand, decomposition wavelets ϕ and ψ that are piecewise constants, and, on the other hand, smooth reconstruction wavelets $\tilde\phi$ and $\tilde\psi$. In particular, except for the Haar basis, decomposition and reconstruction wavelets are different. To shorten mathematical expressions, we set

$$\Lambda = \{\lambda = (j,k) : \quad j \geq -1, k \in \mathbb{Z}\}$$

and for any $\lambda \in \Lambda$, $\varphi_\lambda = \phi_k$ (respectively $\tilde\varphi_\lambda = \tilde\phi_k$) if $\lambda = (-1,k)$ and $\varphi_\lambda = \psi_{j,k}$ (respectively $\tilde\varphi_\lambda = \tilde\psi_{j,k}$) if $\lambda = (j,k)$ with $j \geq 0$. Similarly, $\beta_\lambda = \alpha_k$ if $\lambda = (-1,k)$ and $\beta_\lambda = \beta_{j,k}$ if $\lambda = (j,k)$ with $j \geq 0$. Now, (3.1) can be rewritten as

$$f = \sum_{\lambda\in\Lambda} \beta_\lambda\tilde\varphi_\lambda \quad \text{with} \quad \beta_\lambda = \int \varphi_\lambda(x)f(x)dx.$$

Now, we introduce for any $\lambda \in \Lambda$, the natural estimator of β_λ defined by

$$\hat\beta_\lambda = \frac{1}{n}\int \varphi_\lambda(x)dN_x \tag{3.2}$$

for the Poisson model and

$$\hat\beta_\lambda = \frac{1}{n}\sum_{i=1}^{n} \varphi_\lambda(X_i) \tag{3.3}$$

for the density model, that satisfies $\mathbb{E}(\hat\beta_\lambda) = \beta_\lambda$. We denote $\hat V_{\lambda,n}$ an unbiased estimate of $V_{\lambda,n} = \mathrm{var}(\hat\beta_\lambda)$, the variance of $\hat\beta_\lambda$, defined by

$$\hat V_{\lambda,n} = \frac{1}{n^2}\int \varphi_\lambda^2(x)dN_x$$

for the Poisson model and

$$\hat V_{\lambda,n} = \frac{1}{n^2(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\varphi_\lambda(X_i) - \varphi_{j,k}(X_j))^2$$

for the density model. Then, given some parameter $\gamma > 0$, we define the threshold

$$\eta_{\lambda,\gamma} = \sqrt{2\gamma\tilde V_{\lambda,n}\log n} + \frac{c_1\|\varphi_\lambda\|_\infty\gamma\log n}{3n}, \tag{3.4}$$

with

$$\tilde{V}_{\lambda,n} = \hat{V}_{\lambda,n} + \frac{c_1\|\varphi_\lambda\|_\infty}{n}\sqrt{2\gamma \log n \hat{V}_{\lambda,n}} + \frac{c_2\|\varphi_\lambda\|_\infty^2 \gamma \log n}{n^2}$$

where

$$\begin{cases} c_1 = 1 \text{ and } c_2 = 3 & \text{for the Poisson model,} \\ c_1 = 2 \text{ and } c_2 = 8 & \text{for the density model.} \end{cases}$$

Finally given some subset Γ_n of Λ of the form

$$\Gamma_n = \{\lambda = (j,k) \in \Lambda : \quad j \leq j_0\},$$

where $j_0 = j_0(n)$ is an integer, we set for any $\lambda \in \Lambda$,

$$\tilde{\beta}_\lambda = \hat{\beta}_\lambda 1_{\{|\hat{\beta}_\lambda| \geq \eta_{\lambda,\gamma}\}} 1_{\{\lambda \in \Gamma_n\}}$$

and we set $\tilde{\beta} = (\tilde{\beta}_\lambda)_{\lambda \in \Lambda}$. The estimator of f is

$$\tilde{f}_{n,\gamma} = \sum_{\lambda \in \Lambda} \tilde{\beta}_\lambda \tilde{\varphi}_\lambda$$

and only depends on the choice of γ and j_0 fixed later.

The definition of $\eta_{\lambda,\gamma}$ is based on the following heuristics. Given $\lambda = (j,k) \in \Lambda$, when there exists a constant $c_0 > 0$ such that $f(x) \geq c_0$ for x in the support of φ_λ satisfying $\|\varphi_\lambda\|_\infty^2 = o_n(n(\log n)^{-1})$, then, with large probability, the deterministic term of (3.4) is negligible with respect to the random one. In this case, the random term is the main one and we asymptotically derive

$$\eta_{\lambda,\gamma} \approx \sqrt{2\gamma \tilde{V}_{\lambda,n} \log n}. \tag{3.5}$$

Having in mind that $\tilde{V}_{\lambda,n}$ is a convenient estimate for $\text{var}(\hat{\beta}_\lambda)$, the shape of the right hand term of the formula (3.5) is classical. In fact, it strongly looks like the threshold proposed by Juditsky and Lambert-Lacroix (2004) in the density estimation framework or the universal threshold η^U proposed by Donoho and Johnstone (1994) in the Gaussian regression framework. Indeed, we recall that

$$\eta^U = \sqrt{2\sigma^2 \log n},$$

where σ^2 (assumed to be known in the Gaussian framework) is the variance of each noisy wavelet coefficient. Actually, the deterministic term of (3.4) constitutes the main difference with the threshold defined in Juditsky and Lambert-Lacroix (2004). It allows to consider γ close to 1, which is essential for the calibration issue handled in Section 4. In addition, it allows to control large deviations terms for high resolution levels. As often suggested in the literature, instead of estimating $\text{var}(\hat{\beta}_\lambda)$, we could use the inequality

$$\text{var}(\hat{\beta}_\lambda) \leq \frac{\|f\|_\infty}{n}$$

and we could use this upper bound in the definition of the threshold. But this requires a strong assumption: f is bounded and $\|f\|_\infty$ is known. In this chapter, $\text{var}(\hat{\beta}_\lambda)$ is estimated, which allows not to impose these conditions. But unlike Juditsky and Lambert-Lacroix (2004) who propose to use $\hat{V}_{\lambda,n}$, we slightly overestimate $\text{var}(\hat{\beta}_\lambda)$ to control large deviation terms and this is the reason why we introduce $\tilde{V}_{\lambda,n}$. Finally, observe that the constants c_1 and c_2 differ according to the statistical model. Actually, proofs and computations are more involved for density estimation because sharp upper and lower bounds for the variance of the noisy wavelet coefficients $\hat{\beta}_\lambda$ are more intricate.

In the next section, once fixed j_0 and γ, we apply Theorem 3.1 with $\hat{\beta}_\lambda$ defined in (3.2) or (3.3), $\eta_\lambda = \eta_{\lambda,\gamma}$ defined in (3.4) and

$$\Gamma = \Gamma_n = \{\lambda = (j,k) \in \Lambda : \quad -1 \leq j \leq j_0\}.$$

The other quantities are specified in [R10] and [R13].

2.1. Theoretical results. This section gives the theoretical properties of the estimate $\tilde{f}_{n,\gamma}$.

2.1.1. *Oracle results.* We refer to [R10] and [R13] for the detailed presentation of the oracle point of view, proposed by Donoho and Johnstone (1994), in our settings. We just mention that, here, the oracle is

$$\bar{f} = \sum_{(j,k)\in\Gamma_n} \bar{\beta}_\lambda \tilde{\varphi}_\lambda,$$

where $\bar{\beta}_\lambda = \hat{\beta}_\lambda 1_{\left\{\beta_\lambda^2 > V_{\lambda,n}\right\}}$ satisfies

$$\mathbb{E}\left[(\bar{\beta}_\lambda - \beta_\lambda)^2\right] = \min\left(\beta_\lambda^2, V_{\lambda,n}\right).$$

By keeping the coefficients $\hat{\beta}_\lambda$ larger than the thresholds defined in (3.4), our estimator satisfies the following oracle inequality.

THEOREM 3.2. *Let us consider a biorthogonal wavelet basis satisfying the properties described in Section 2.2 of [R10]. Let us fix two constants $c \geq 1$ and $c' \in \mathbb{R}$ and let us define for any n, $j_0 = j_0(n)$ the integer such that $2^{j_0} \leq n^c (\log n)^{c'} < 2^{j_0+1}$. If $\gamma > c$, then $\tilde{f}_{n,\gamma}$ satisfies the following oracle inequality: for n large enough*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C_1 \left[\sum_{(j,k)\in\Gamma_n} \min\left(\beta_\lambda^2, V_{\lambda,n}\log n\right) + \sum_{(j,k)\notin\Gamma_n} \beta_\lambda^2\right] + \frac{C_2 \log n}{n} \tag{3.6}$$

where C_1 is a positive constant depending only on γ, c and the functions that generate the biorthogonal wavelet basis. C_2 is also a positive constant depending on γ, c c', $\|f\|_1$, $\|f\|_2$ and the functions that generate the basis.

Note that Theorem 3.2 holds with $c = 1$ and $\gamma > 1$ and in the Poisson setting the negligible term $\log n/n$ can be replaced with $1/n$. Following the oracle point of view of Donoho and Johnstone, Theorem 3.2 shows that our procedure is optimal up to the logarithmic factor. This logarithmic term is in some sense unavoidable. It is the price we pay for adaptivity (i.e. for not knowing the coefficients that we must keep). Our result is true provided $f \in \mathbb{L}_1(\mathbb{R}) \cap \mathbb{L}_2(\mathbb{R})$. So, assumptions on f are very mild here. From this key result, we can deduce maxiset results for our procedure and then minimax rates for $\tilde{f}_{n,\gamma}$. See Section 3.1 of [R10] that is devoted to maxiset results of our procedure.

2.1.2. *Minimax results.* The goal of this section is to derive the minimax rates on the whole class of Besov spaces. The subsequent results will constitute generalizations of the results derived by Juditsky and Lambert-Lacroix (2004) who pointed out minimax rates for density estimation on the class of Hölder spaces. For this purpose, $j_0 = j_0(n)$ is the integer such that

$$2^{j_0} \leq \left(\frac{n}{\log n}\right)^c < 2^{j_0+1}.$$

So, in this subsection, $c' = -c$ but the real number c is chosen later. Unfortunately, in some situations, it will be necessary to strengthen our assumptions. More precisely, sometimes, we assume that f is bounded. So, for any $R \geq 1$, we consider the following set of functions:

$$\mathcal{L}_{1,2,\infty}(R) = \left\{f \text{ is such that } \|f\|_1 \leq R, \|f\|_2 \leq R \text{ and } \|f\|_\infty \leq R\right\}.$$

Note that, when $f = f_d$, then the condition $\|f\|_1 \leq R$ is automatically satisfied since $R \geq 1$. In the Poisson setting, we could take $R > 0$. Now, let us state the upper bound of the $\mathbb{L}_2$-risk of $\tilde{f}_{n,\gamma}$.

THEOREM 3.3. *Let $R > 0$, $R' \geq 1$, $1 \leq p, q \leq \infty$ and $\alpha \in \mathbb{R}$ such that $\max\left(0, \frac{1}{p} - \frac{1}{2}\right) < \alpha < r + 1$, where r denotes the wavelet smoothness parameter introduced in Section 2.2 of [R10]. Let $c \geq 1$ such that*

$$\alpha\left(1 - \frac{1}{c(1+2\alpha)}\right) \geq \frac{1}{p} - \frac{1}{2} \tag{3.7}$$

and $\gamma > c$. Then, there exists a constant C depending on R', γ, c, on the parameters of the Besov ball and on $\Phi = \{\phi, \psi, \tilde{\phi}, \tilde{\psi}\}$ such that for any n,

- *if $p \leq 2$,*

$$\sup_{f \in \mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{1,2,\infty}(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C \left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \tag{3.8}$$

- *if $p > 2$,*

$$\sup_{f \in \mathcal{B}^{\alpha}_{p,q}(R) \cap \mathbb{L}_1(R') \cap \mathbb{L}_2(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C \left(\frac{\log n}{n}\right)^{\frac{\alpha}{\alpha+1-\frac{1}{p}}}. \tag{3.9}$$

First, let us briefly comment assumptions of these results. When $p \leq 2$, the result is true as soon as c is large enough to satisfy Condition (3.7). But when $p > 2$, Condition (3.7) is satisfied and the result is true for any $c \geq 1$ and $0 < \alpha < r + 1$. In addition, we do not need to restrict ourselves to the set of bounded functions. But, when $p \leq 2$, we establish (3.8) only for bounded functions. Actually, this assumption is in some sense unavoidable as proved in Section 6.4 of Birgé (2008).

When $p \leq 2$, the rate of the risk of $\tilde{f}_{n,\gamma}$ corresponds to the classical minimax rate (up to the logarithmic term) for estimating a compactly supported density (see Donoho, Johnstone, Kerkyacharian and Picard (1996)). When $p > 2$, the upper bound of the risk deteriorates. Note that when $p = \infty$, the risk is bounded by $(\log n/n)^{\alpha/(1+\alpha)}$ up to a constant. This rate was also derived by Juditsky and Lambert-Lacroix (2004) for estimation on balls of $\mathcal{B}^{\alpha}_{\infty,\infty}$. Now, combining upper bounds (3.8) and (3.9), under assumptions of Theorem 3.3, we point out the following rate for our procedure:

$$\sup_{f \in \mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{1,2,\infty}(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C \left(\frac{\log n}{n}\right)^{\frac{\alpha}{\alpha+\frac{1}{2}+\left(\frac{1}{2}-\frac{1}{p}\right)_+}}.$$

The following result derives lower bounds of the minimax risk showing that this rate is the optimal rate up to a logarithmic term. So, the next result establishes the optimality properties of $\tilde{f}_{n,\gamma}$ under the minimax approach.

THEOREM 3.4. *Let $R > 0$, $R' \geq 1$, $1 \leq p, q \leq \infty$ and $\alpha \in \mathbb{R}$ such that $\max\left(0, \frac{1}{p} - \frac{1}{2}\right) < \alpha < r + 1$. Then, there exists a positive constant $\tilde{C}$ depending on R', γ, c, on the parameters of the Besov ball and on Φ such that*

$$\lim \inf_{n \to +\infty} n^{\frac{\alpha}{\alpha+\frac{1}{2}+\left(\frac{1}{2}-\frac{1}{p}\right)_+}} \inf_{\hat{f}} \sup_{f \in \mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{1,2,\infty}(R')} \mathbb{E}\left[\|\hat{f}_n - f\|_2^2\right] \geq \tilde{C}.$$

Furthermore, let $p^* \geq 1$ and $\alpha^* > 0$ such that

$$\alpha^* \left(1 - \frac{1}{c(1+2\alpha^*)}\right) \geq \frac{1}{p^*} - \frac{1}{2}. \tag{3.10}$$

Then, $\tilde{f}_{n,\gamma}$ is adaptive minimax up to a logarithmic term on

$$\left\{\mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{1,2,\infty}(R') : \quad \alpha^* \leq \alpha < r + 1,\ p^* \leq p \leq +\infty,\ 1 \leq q \leq \infty\right\}.$$

Note that when restricting on compactly supported signals, when $p > 2$, $\mathcal{B}^{\alpha}_{p,\infty}(R) \subset \mathcal{B}^{\alpha}_{2,\infty}(\tilde{R})$ for $\tilde{R}$ large enough and in this case, the rate does not depend on p.

Our results show the role played by the support of the functions to be estimated on minimax rates. As already observed, when $p \leq 2$, the support has no influence since the rate exponent remains unchanged whatever the size of the support (finite or not). Roughly speaking, it means that it is not harder to estimate bounded non-compactly supported functions than bounded compactly supported functions from the minimax point of view. It is not the case when non-compactly supported signals are considered. Actually, we note an elbow phenomenon at $p = 2$ and the rate

deteriorates when p increases. Let us give an interpretation of this observation in terms of sparsity. When $p < 2$, functions of the Besov spaces $\mathcal{B}^{\alpha}_{p,q}$ are sparse where at each level, a very few number of the wavelet coefficients are non-negligible. But these coefficients can be very large. When $p > 2$, $\mathcal{B}^{\alpha}_{p,q}$-spaces typically model dense signals where the wavelet coefficients are not large but most of them can be non-negligible. This explains why the size of the support plays a role for minimax rates when $p > 2$: when the support is larger, the number of wavelet coefficients to be estimated increases dramatically.

Finally, we note that our procedure achieves the minimax rate, up to a logarithmic term: $\tilde{f}_{n,\gamma}$ is near rate-optimal without knowing the regularity and the support of the underlying signal to be estimated.

2.2. Numerical study. We give here a brief overview of the numerical performances of our procedure. We just present the results for the density model. The simulation study for the Poisson model can be found in Section 5.2 of [R11] where comparisons with methodologies proposed by Rudemo (1982), Kolaczyk (1999), Reynaud-Bouret (2003), Willet and Nowak (2007), Birgé (2006), Baraud and Birgé (2006) and Figueroa-López and Houdré (2006) are discussed. Our goal is essentially to detect the existence of a curse of support from the numerical point of view. We first provide a simulation study illustrating the distortion of the most classical support-dependent estimators when the support or the tail is increasing. Next we provide an application of our method to famous real data sets.

We compare our method to representative methods of each main trend in density estimation, namely kernel, binning plus thresholding and model selection. The considered methods are the following. The first one is the kernel method, denoted **K**, consisting in a basic cross-validation choice of a global bandwidth with a Gaussian kernel. The second method requires a complex preprocessing of the data based on binning. Observations $X_1, \ldots, X_n$ are first rescaled and centered by an affine transformation denoted T such that $T(X_1), \ldots, T(X_n)$ lie in $[0, 1]$. We denote f_T the density of the data induced by the transformation T. We divide the interval $[0, 1]$ into 2^{b_n} small intervals of size 2^{-b_n}, where b_n is an integer, and count the number of observations in each interval. We apply the root transform due to Brown, Cai, Zhang, Zhao and Zhou (2007) and the universal hard individual thresholding rule on the coefficients computed with the DWT Coiflet-basis filter. We finally apply the unroot transform to obtain an estimate of f_T and the final estimate of the density is obtained by applying T^{-1} combined with a spline interpolation. This method is denoted **RU**. The last method is also support-dependent. After rescaling the data as previously, we estimate f_T by the algorithm of Willett and Nowak (2007). It consists in a complex selection of a grid and of polynomials on that grid that minimizes a penalized loglikelihood criterion. The final estimate of the density is obtained by applying T^{-1}. This method is denoted **WN**. Our practical method has been implemented in the Haar basis (method **H**) and in the Spline basis (method **S**). Moreover, we have also implemented the choice $\gamma = 0.5$ in the Spline basis. We denote this method **S***. See Section 4.1 of [R13] for more details.

We have generated n-samples according to very different densities g_d and h_k, with $n = 1024$. Both signals are supported by the whole real line. We have computed for each estimator its integrated squared error (ISE).

The first signal, g_d, consists in a mixture of two standard Gaussian densities:

$$g_d = \frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{2}\mathcal{N}(d, 1),$$

where $\mathcal{N}(\mu, \sigma)$ represents the density of a Gaussian variable with mean μ and standard deviation σ. The parameter d varies in $\{10, 30, 50, 70\}$ so that we can see the curse of support on the quality of estimation.

Figure 1 shows the reconstructions for $d = 10$ and Figure 2 for $d = 70$. In the sequel, the method **RU** is implemented with $b_n = 5$, which is the best choice for the reconstruction with $d = 10$. All the methods give satisfying results for $d = 10$. When d is large, the rescaling and

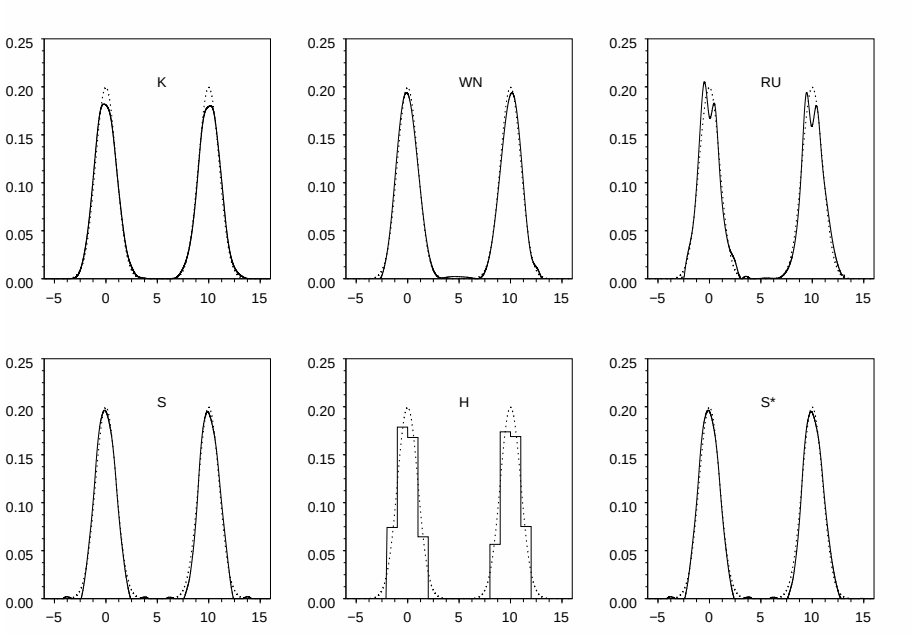

FIGURE 1. Reconstruction of g_d (true: dotted line, estimate: solid line) for the 6 different methods for $d = 10$

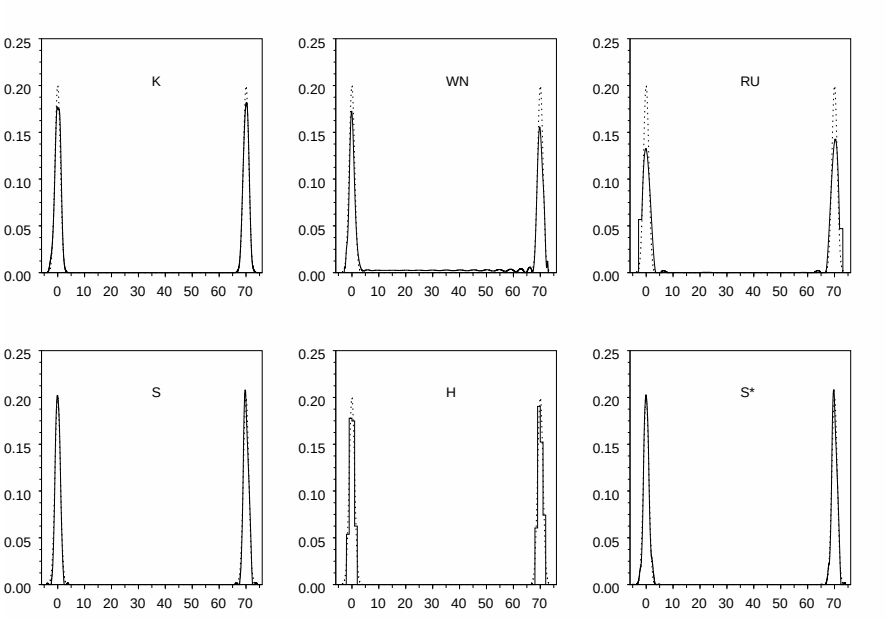

FIGURE 2. Reconstruction of g_d (true: dotted line, estimate: solid line) for the 6 different methods for $d = 70$

binning preprocessing leads to a poor regression signal which makes the regression thresholding rules non convenient, as illustrated by the method **RU** with $d = 70$. Reconstructions for **K, WN, S** and **S*** seem satisfying but a study of the ISE of each method (see Figure 3) reveals that both support-dependent methods (**RU** and **WN**) have a risk that increases with d. On the contrary, methods **K** and **S** are the best ones and more interestingly their performances do not vary with d. This robustness is also true for **H** and **S***. **S*** is a bit undersmoothing which explains the variability

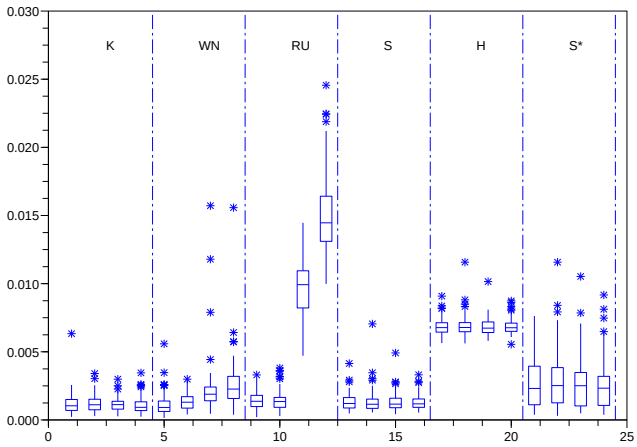

FIGURE 3. Boxplot of the ISE for g_d over 100 simulations for the 6 methods and the 4 different values of d. A column, delimited by dashed lines, corresponds to one method (respectively **K, WN, RU, S, H, S***). Inside this column, from left to right, one can find for the same method the boxplot of the ISE for respectively $d = 10, 30, 50$ and 70.

of its ISE. Finally note that **H** is even better than **RU** despite the inappropriate choice of the Haar basis.

The other signal, h_k, is both heavy-tailed and irregular. It consists in a mixture of 4 Gaussian densities and one Student density:

$$h_k = 0.45T(k) + 0.15\mathcal{N}(-1, 0.05) + 0.1\mathcal{N}(-0.7, 0.005) + 0.25\mathcal{N}(1, 0.025) + 0.15\mathcal{N}(2, 0.05),$$

where $T(k)$ denotes the density of a Student variable with k degrees of freedom. The parameter k varies in $\{2, 4, 8, 16\}$. The smaller k, the heavier the tail is and this without changing the shape of the main part that has to be estimated. Figure 4 shows the reconstruction for $k = 2$. Clearly **RU** is not suitable. The kernel method **K** suffers from a lack of spatial adaptivity, as expected. The four remaining methods seem satisfying. In particular for this very irregular signal it is not clear that the Haar basis is a bad choice. Note however that to represent reconstructions, we have focused on the area where the spikes are located. In particular the support-dependent method **WN** is non-zero on a very large interval which tends to deteriorate its ISE. Indeed, Figure 5 shows that the ISE of the support-dependent methods (**RU, WN**) increases when the tail becomes heavier, whereas the other methods have remarkable stable ISE. Methods **S** and **H** are more robust and better than **WN** for $k = 2$. The ISE may be improved for this irregular signal by taking $\gamma = 0.5$ as noted with the performances of **S***.

To illustrate our procedure on real data, we consider two real data sets named, respectively in our study, "Old Faithful geyser" and "Suicide" taken from Weisberg (1980) and Copas and Fryer (1980). These data are well known and have been widely used elsewhere. This allows to compare our procedure with other methods. To estimate the function f, we apply our practical methodology with the spline basis and the parameter γ equal to 1. Figures 7 and and 8 of [R13] represent, respectively, the resulting estimate for the "Old Faithful geyser" set and for the "Suicide" one. Respectively two or three peaks are detected providing multimodal reconstructions. So, in comparison with the ones performed in Silverman (1986) and Sain and Scott (1996), our estimate detects significant events and not artefacts. More interestingly, both estimates equal zero on an interval located between the last two peaks. This cannot occur with the Gaussian kernel estimate mentioned previously. Of course, this has a strong impact for practical purposes, so this point is

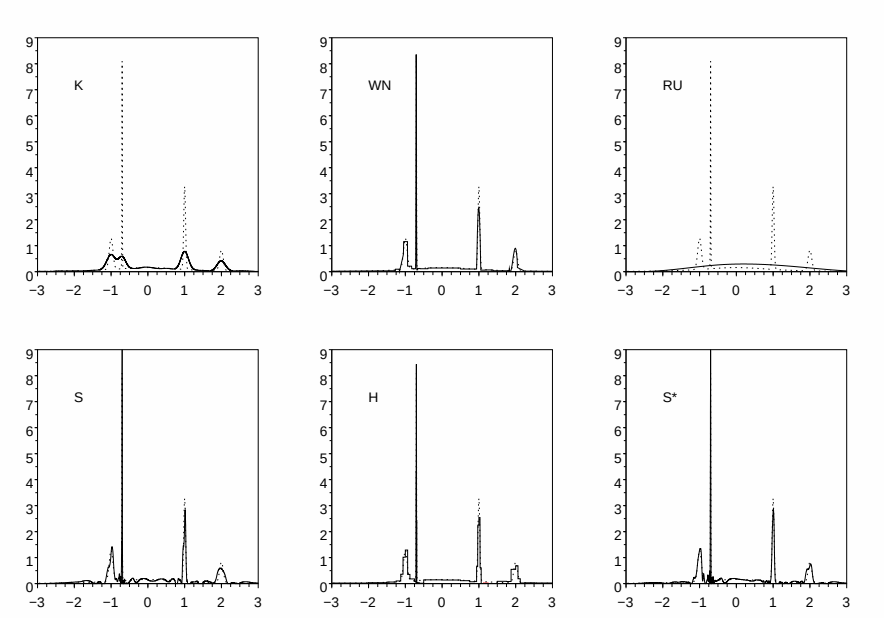

FIGURE 4. Reconstruction of h_k (true: dotted line, estimate: solid line) for the 6 different methods for $k = 2$

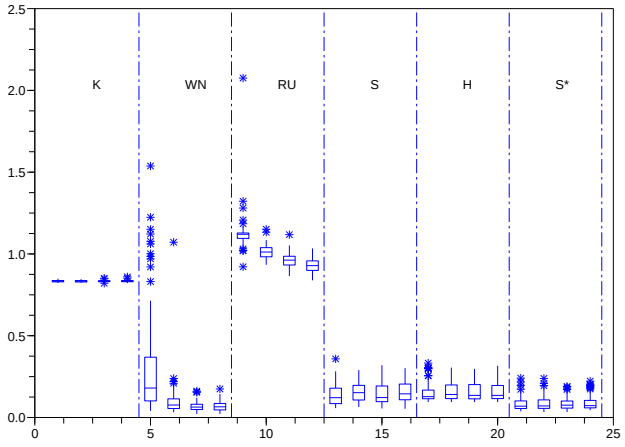

FIGURE 5. Boxplot of the ISE for h_k over 100 simulations for the 6 methods and the 4 different values of k. A column, delimited by dashed lines, corresponds to one method (respectively **K, WN, RU, S, H, S***). Inside this column, from left to right, one can find for the same method the boxplot of the ISE for respectively $k = 2, 4, 8$ and 16.

crucial. This tends to show that the proposed procedure is relevant for real data, even for relatively small sample size.

3. A data-driven Dantzig procedure for density estimation

In this section, we still consider the density model already studied for the Lasso estimate by Bunea, Tsybakov and Wegkamp (2007a, 2009) and van de Geer (2008). By using the observations

of a n-sample of variables $X_1, \ldots, X_n$ of density $f \in \mathbb{L}_1(\mathbb{R}) \cap \mathbb{L}_2(\mathbb{R})$, we aim at building a data-driven Dantzig estimate of f inspired by the data-driven wavelet thresholding rule of the previous section. As in Section 4 of Chapter 1, we consider a dictionary of functions denoted $\Upsilon = (\varphi_\lambda)_{\lambda=1,\ldots,M}$, with

$$n \leq M \leq \exp(n^\delta) \tag{3.11}$$

for $\delta < 1$. Assumption (3.11) can be relaxed and we can take $M < n$ provided slight modifications of the subsequent quantities. We search estimates of f as linear combinations f_μ of the dictionary functions:

$$f_\mu = \sum_{\lambda=1}^{M} \mu_\lambda \varphi_\lambda,$$

with $\mu = (\mu_\lambda)_{\lambda=1,\ldots,M} \in \mathbb{R}^M$. In the sequel, we assume without any loss of generality that, for any λ, $\|\varphi_\lambda\|_2 = 1$. But we emphasize that, unlike Section 2, the functions φ_λ's do not necessarily constitute a basis. We still denote, for $\lambda \in \{1, \ldots, M\}$,

$$\beta_\lambda = \int \varphi_\lambda(x) f(x) dx \quad \text{and} \quad \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(X_i).$$

Note also that for any μ and any λ, the $\mathbb{L}_2$-scalar product between f_μ and φ_λ can be easily computed:

$$\int \varphi_\lambda(x) f_\mu(x) dx = \sum_{\lambda'=1}^{M} \mu_{\lambda'} \int \varphi_{\lambda'}(x) \varphi_\lambda(x) dx = (G\mu)_\lambda$$

where G is the Gram matrix associated to the dictionary Υ defined for any $1 \leq \lambda, \lambda' \leq M$ by

$$G_{\lambda,\lambda'} = \int \varphi_\lambda(x) \varphi_{\lambda'}(x) dx.$$

Any reasonable choice of μ should ensure that the coefficients $(G\mu)_\lambda$ are close to $\hat{\beta}_\lambda$ for all λ. Therefore, using Candès and Tao's approach (see Candès and Tao (2007)), we define the Dantzig constraint:

$$\forall\, \lambda \in \{1, \ldots . M\}, \quad |(G\mu)_\lambda - \hat{\beta}_\lambda| \leq \eta_{\lambda,\gamma} \tag{3.12}$$

and the Dantzig estimate $\hat{f}^D$ by $\hat{f}^D = f_{\hat{\mu}^{D,\gamma}}$ with

$$\hat{\mu}^{D,\gamma} = \operatorname{argmin}_{\mu \in \mathbb{R}^M} \|\mu\|_{\ell_1} \quad \text{such that } \mu \text{ satisfies the Dantzig constraint (3.12)},$$

where for $\gamma > 0$ and $\lambda \in \{1, \ldots, M\}$, $\eta_{\lambda,\gamma}$ is defined in (3.4) (with $c_1 = 2$ and $c_2 = 8$). The constraint (3.12) will be referred as the *adaptive Dantzig constraint* in the sequel. To justify the introduction of the density estimate $\hat{f}^D$, let us set $\mu_0 = (\mu_{0,\lambda})_{\lambda=1,\ldots,M} \in \mathbb{R}^M$ such that

$$P_\Upsilon f = \sum_{\lambda=1}^{M} \mu_{0,\lambda} \varphi_\lambda$$

where P_Υ is the projection on the space spanned by Υ. We have

$$(G\mu_0)_\lambda = \int (P_\Upsilon f) \varphi_\lambda = \int f \varphi_\lambda = \beta_\lambda.$$

If $\gamma > 1$, for any $0 < \varepsilon < \gamma - 1$, Theorem 1 of [R12] proves that μ_0 satisfies the adaptive Dantzig constraint (3.12) with probability larger than $1 - C_1(\varepsilon, \delta, \gamma) M^{1 - \frac{\gamma}{1+\varepsilon}}$, where $C_1(\varepsilon, \delta, \gamma)$ is a constant only depending on ε, δ and γ (the probability $1 - C_1(\varepsilon, \delta, \gamma) M^{1 - \frac{\gamma}{1+\varepsilon}}$ will be used in the sequel). Actually, we force the set of parameters μ satisfying the adaptive Dantzig constraint to contain μ_0 with large probability and to be as small as possible. Therefore, $\hat{f}^D$ is a good candidate among sparse estimates linearly decomposed on Υ for estimating f.

3.1. Theoretical results. Let us state the main result of this chapter. We introduce the vector $\eta_\gamma = (\eta_{\lambda,\gamma})_{\lambda=1,\dots,M}$ considered with the tuning parameter $\gamma > 1$. For any $J \subset \{1,\dots,M\}$, we set $J^C = \{1,\dots,M\} \setminus J$ and define μ_J the vector which has the same coordinates as μ on J and zero coordinates on J^C. We introduce a local assumption indexed by a subset J_0.

- **Local Assumption** Given $J_0 \subset \{1,\dots,M\}$, for some constants $\kappa_{J_0} > 0$ and $\nu_{J_0} > 0$ depending on J_0, we have for any μ,

$$(LA(J_0,\kappa_{J_0},\nu_{J_0})) \qquad \|f_\mu\|_2 \geq \kappa_{J_0}\|\mu_{J_0}\|_{\ell_2} - \nu_{J_0}\left(\|\mu_{J_0^C}\|_{\ell_1} - \|\mu_{J_0}\|_{\ell_1}\right)_+.$$

We obtain the following oracle type inequality.

THEOREM 3.5. *Let $J_0 \subset \{1,\dots,M\}$ be fixed. We suppose that $(LA(J_0,\kappa_{J_0},\nu_{J_0}))$ holds. Then, with probability at least $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\|\hat{f}^D - f\|_2^2 \leq \inf_{\mu\in\mathbb{R}^M}\left\{\|f_\mu - f\|_2^2 + \beta\frac{\Lambda(\mu,J_0^c)^2}{|J_0|}\left(1 + \frac{2\nu_{J_0}\sqrt{|J_0|}}{\kappa_{J_0}}\right)^2 + 16|J_0|\left(\frac{1}{\beta} + \frac{1}{\kappa_{J_0}^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2\right\},$$

with

$$\Lambda(\mu,J_0^c) = \|\mu_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\mu}^{D,\gamma}\|_{\ell_1} - \|\mu\|_{\ell_1}\right)_+}{2}.$$

Let us comment each term of the right hand side of the oracle inequality. The first term is an approximation term which measures the closeness between f and f_μ. This term can vanish if f can be decomposed on the dictionary. The second term is a price to pay when either μ is not supported by the subset J_0 considered or it does not satisfy the condition $\|\hat{\mu}^{D,\gamma}\|_{\ell_1} \leq \|\mu\|_{\ell_1}$ which holds as soon as μ satisfy the adaptive Dantzig constraint. Finally, the last term, which does not depend on μ, can be viewed as a variance term corresponding to the estimation on the subset J_0. Indeed, remember that $\eta_{\lambda,\gamma}$ relies on an estimate of the variance of $\hat{\beta}_\lambda$. Furthermore, we have with high probability (see Theorem 1 of [R12]):

$$\|\eta_\gamma\|_{\ell_\infty}^2 \quad \leq \quad 2\left(16\mathrm{var}(\hat{\beta}_\lambda)\gamma\log M + \left(\frac{10\|\varphi_\lambda\|_\infty\gamma\log M}{n}\right)^2\right).$$

So, if $\|f\|_\infty < \infty$ and if there exists a constant c_1 such that for any λ,

$$\|\varphi_\lambda\|_\infty^2 \leq c_1\left(\frac{n}{\log M}\right)\|f\|_\infty, \tag{3.13}$$

(which is true for instance for a bounded dictionary), then

$$\|\eta_\gamma\|_{\ell_\infty}^2 \leq C\|f\|_\infty\frac{\log M}{n},$$

(where C is a constant depending on γ and c_1) and tends to 0 when n goes to ∞. Furthermore, if $f = f_{\mu_0}$ and if $(LA(J_0,\kappa_{J_0},\nu_{J_0}))$ holds with $J_0 = J_{\mu_0}$, under (3.13), the proof of Theorem 3.5 yields the more classical inequality: with at least the probability $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$,

$$\|\hat{f}^D - f\|_2^2 \leq C'|J_0|\|f\|_\infty\frac{\log M}{n},$$

where C' is a constant.

Assumption $(LA(J_0,\kappa_{J_0},\nu_{J_0}))$ is local, in the sense that the constants κ_{J_0} and ν_{J_0} (or their mere existence) may highly depend on the subset J_0. For a given μ, the best choice for J_0 in the oracle inequality of Theorem 3.5 depends thus on the interaction between these constants and the value of μ itself.

As usual, when $M > n$, properties of the Dantzig estimate can be derived from global assumptions on the structure of the dictionary Υ. For $l \in \mathbb{N}$, we denote

$$\phi_{\min}(l) = \min_{|J|\leq l} \min_{\substack{\mu\in\mathbb{R}^M \\ \mu_J\neq 0}} \frac{\|f_{\mu_J}\|_2^2}{\|\mu_J\|_{\ell_2}^2} \qquad \text{and} \qquad \phi_{\max}(l) = \max_{|J|\leq l} \max_{\substack{\mu\in\mathbb{R}^M \\ \mu_J\neq 0}} \frac{\|f_{\mu_J}\|_2^2}{\|\mu_J\|_{\ell_2}^2}.$$

These quantities correspond to the 'restricted eigenvalues' of the Gram matrix G. We also consider the 'restricted correlations'

$$\theta_{l,l'} = \max_{\substack{|J|\leq l \\ |J'|\leq l' \\ J\cap J'=\emptyset}} \max_{\substack{\mu,\mu'\in\mathbb{R}^M \\ \mu_J\neq 0, \mu'_{J'}\neq 0}} \frac{\langle f_{\mu_J}, f_{\mu'_{J'}}\rangle}{\|\mu_J\|_{\ell_2}\|\mu'_{J'}\|_{\ell_2}}.$$

We will use one of the following assumptions considered in Bickel, Ritov and Tsybakov (2009) (see [R12] for a discusion of these assumptions).

- **Assumption 1** For some integer $1 \leq s \leq M/2$, we have

(A1(s)) $$\phi_{\min}(2s) > \theta_{s,2s}.$$

.

- **Assumption 2** For some integers s and l such that

(3.14) $$1 \leq s \leq \frac{M}{2}, \quad l \geq s \quad \text{and} \quad s + l \leq M,$$

we have

(A2(s,l)) $$l\phi_{\min}(s + l) > s\phi_{\max}(l).$$

In the sequel, we set, under Assumption 1,

$$\kappa_1(s) = \sqrt{\phi_{\min}(2s)}\left(1 - \frac{\theta_{s,2s}}{\phi_{\min}(2s)}\right) > 0, \quad \nu_1(s) = \frac{\theta_{s,2s}}{\sqrt{s\phi_{\min}(2s)}}$$

and under Assumption 2,

$$\kappa_2(s,l) = \sqrt{\phi_{\min}(s+l)}\left(1 - \sqrt{\frac{s\phi_{\max}(l)}{l\phi_{\min}(s+l)}}\right) > 0, \quad \nu_2(s,l) = \sqrt{\frac{\phi_{\max}(l)}{l}}.$$

To shorten notations, we set $\kappa = \kappa_1(s)$ and $\nu = \nu_1(s)$ under (A1(s)) (respectively $\kappa = \kappa_2(s,l)$ and $\nu = \nu_2(s,l)$ under (A2(s,l)). If (A1(s)) and (A2(s,l)) are both satisfied, $\kappa = \max(\kappa_1(s), \kappa_2(s,l))$ and $\nu = \min(\nu_1(s), \nu_2(s,l))$. Roughly speaking, Proposition 1 of [R12] proves that either Assumption 1 or Assumption 2 implies $(LA(J_0, \kappa_{J_0}, \nu_{J_0}))$. We deduce the following result.

THEOREM 3.6. *Let s and l two integers satisfying (3.14). We suppose that (A1(s)) or (A2(s,l)) is true. Then, with probability at least $1 - C_1(\varepsilon,\delta,\gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\|\hat{f}^D - f\|_2^2 \leq \inf_{\mu\in\mathbb{R}^M} \inf_{\substack{J_0\subset\{1,\ldots,M\} \\ |J_0|=s}} \left\{\|f_\mu - f\|_2^2 + \beta\frac{\Lambda(\mu, J_0^c)^2}{s}\left(1 + \frac{2\nu\sqrt{s}}{\kappa}\right)^2 + 16s\left(\frac{1}{\beta} + \frac{1}{\kappa^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2\right\}$$

where

$$\Lambda(\mu, J_0^c) = \|\mu_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\mu}^{D,\gamma}\|_{\ell_1} - \|\mu\|_{\ell_1}\right)_+}{2}.$$

Remark that the best subset J_0 of cardinal s in Theorem 3.6 can be easily chosen for a given μ: it is given by the set of the s largest coordinates of μ. This was not necessarily the case in Theorem 3.5 for which a different subset may give a better local condition and then may provide a smaller bound. If we further assume the mild assumption (3.13) on the sup norm of the dictionary, we deduce the following result.

COROLLARY 3.1. *Let s and l two integers satisfying (3.14). We suppose that (A1(s)) or (A2(s,l)) is true. If (3.13) is satisfied, with probability at least $1 - C_1(\varepsilon, \delta, \gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$, any μ that satisfies the adaptive Dantzig constraint and for the best subset J_0 of cardinal s (that corresponds to the s largest coordinates of μ in absolute value),*

$$\|\hat{f}^D - f\|_2^2 \leq \|f_\mu - f\|_2^2 + \beta c_2(1 + \kappa^{-2}\nu^2 s)\frac{\|\mu_{J_0^C}\|_{\ell_1}^2}{s} + c_3(\beta^{-1} + \kappa^{-2})s\|f_0\|_\infty \frac{\log M}{n}, \tag{3.15}$$

where c_2 is an absolute constant and c_3 depends on c_1 and γ.

Note that, when μ is s-sparse so that $\mu_{J_0^C} = 0$, the oracle inequality (3.15) corresponds to the classical oracle inequality obtained in parametric frameworks (see Candès and Plan (2007) or Candès and Tao (2007) for instance) or in non-parametric settings. See, for instance Bunea, Tsybakov and Wegkamp (2006, 2007a, 2007b, 2007c), Bunea (2008) or van de Geer (2008) but in these works, the functions of the dictionary are assumed to be bounded by a constant independent of M and n. So, the adaptive Dantzig estimate requires weaker conditions since under (3.13), $\|\varphi_\lambda\|_\infty$ can go to ∞ when n grows. This point is capital for practical purposes, in particular when wavelet bases are considered.

We end this theoretical study by briefly showing the strong connections between Lasso and Dantzig estimates, which has already been illustrated in Bickel, Ritov and Tsybakov (2009) for non-parametric regression models (see Section 4 of [R12] for more details). We consider the Lasso estimator given by the solution of the following minimization problem

$$\hat{\mu}^{L,\gamma} = \operatorname{argmin}_{\mu \in \mathbb{R}^M} \left\{ R(\mu) + 2\sum_{\lambda=1}^{M} \eta_{\lambda,\gamma}|\mu_\lambda| \right\}, \tag{3.16}$$

where

$$R(\mu) = \|f_\mu\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_\mu(X_i)$$

and $\hat{\mu}^{L,\gamma}$ appears as a data-driven version of classical Lasso estimates. We denote $\hat{f}^L = f_{\hat{\mu}^{L,\gamma}}$.

The first order condition for the minimization of the expression given in (3.16) corresponds exactly to the adaptive Dantzig constraint and thus Theorem 3.6 always applies to $\hat{\mu}^{L,\gamma}$. Actually, one can prove a slightly stronger result.

THEOREM 3.7. *Let us assume that assumptions of Theorem 3.6 are true. Let $J_0 \subset \{1, \ldots, M\}$ of size $|J_0| = s$. Then, with probability at least $1 - C_1(\varepsilon, \delta, \gamma)M^{1-\frac{\gamma}{1+\varepsilon}}$, we have for any $\beta > 0$,*

$$\left| \|\hat{f}^D - f\|_2^2 - \|\hat{f}^L - f\|_2^2 \right| \leq \beta \frac{\|\hat{\mu}^{L,\gamma}_{J_0^C}\|_{\ell_1}^2}{s}\left(1 + \frac{2\nu\sqrt{s}}{\kappa}\right)^2 + 16s\left(\frac{1}{\beta} + \frac{1}{\kappa^2}\right)\|\eta_\gamma\|_{\ell_\infty}^2.$$

3.2. Numerical study. We end this section by very briefly describing the numerical performances of the data-driven Dantzig and Lasso procedures respectively computed with the homotopy-path-following method proposed by Asif and Romberg (2009) and the LARS algorithm. Following recommendations of the next chapter, we take $\gamma = 1.01$. I refer the reader to Section 5.2 of [R12] for more details where the simulation study is performed with a collection of 6 dictionaries described below, 4 densities and for 2 sample sizes. We compare our procedures with a non adaptive Dantzig estimator where $\text{var}(\hat{\beta}_\lambda)$ is replaced with $\|f\|_\infty/n$ and we consider a two-step estimation procedure proposed by Candès and Tao (2007) which consists in the following additional least-square step. Let $\hat{J}^{D,\gamma}$ be the support of the estimate $\hat{\mu}^{D,\gamma}$. This defines a subset of the dictionary on which the density is regressed

$$\left(\hat{\mu}^{D+LS,\gamma}\right)_{\hat{J}^{D,\gamma}} = G^{-1}_{\hat{J}^{D,\gamma}}(\hat{\beta}_\lambda)_{\hat{J}^{D,\gamma}}$$

where $G_{\hat{J}^{D,\gamma}}$ is the submatrix of G corresponding to the subset chosen. The values of $\hat{\mu}^{D+LS,\gamma}$ outside $\hat{J}^{D,\gamma}$ are set to 0 and $\hat{f}^{D+LS,\gamma}$ is set accordingly. We describe now the dictionaries we

consider. We focus numerically on densities defined on the interval $[0,1]$ so we use dictionaries adapted to this setting. The first four are orthonormal systems, which are used as a benchmark, while the last two are "real" dictionaries. More precisely, our dictionaries are

(1) the Fourier basis with $M = n + 1$ elements (denoted "Fou"),
(2) the histogram collection with the classical number $\sqrt{n}/2 \leq M = 2^{j_0} < \sqrt{n}$ of bins (denoted "Hist"),
(3) the Haar wavelet basis with maximal resolution $n/2 < M = 2^{j_1} < n$ and thus $M = 2^{j_1}$ elements (denoted "Haar"),
(4) the more regular Daubechies 6 wavelet basis with maximal resolution $n/2 \leq M = 2^{j_1} < n$ and thus $M = 2^{j_1}$ elements (denoted "Wav"),
(5) the dictionary made of the union of the Fourier basis and the histogram collection and thus comprising $M = n + 1 + 2^{j_0}$ elements. (denoted "Mix"),
(6) the dictionary which is the union of the Fourier basis, the histogram collection and the Haar wavelets of resolution greater than 2^{j_0} comprising $M = n + 1 + 2^{j_1}$ elements (denoted "Mix2").

The orthonormal families we have chosen are often used by practitioners. Our dictionaries combine very different orthonormal families, sine and cosine with bins or Haar wavelets, which ensures a sufficiently incoherent design.

Boxplots of Figures 3 and 4 of [R12] summarize the numerical experiments for $n = 500$ and $n = 2000$ and 100 repetitions of the procedures. As expected, Dantzig and Lasso estimators are strictly equivalent when the dictionary is orthonormal and very close otherwise. For both algorithms and most of the densities, the best solution appears to be the "Mix2" dictionary. This shows that the dictionary approach yields an improvement over the classical basis approach. One observes also that the "Mix" dictionary is better than the best of its constituent, namely the Fourier basis and the histogram family, which corroborates our theoretical results. The adaptive constraints are much tighter than their non adaptive counterparts and yield to much better numerical results. Our last series of experiments shows the significant improvement obtained with the least-square step. As hinted by Candès and Tao (2007), this can be explained by the bias common to ℓ_1 methods which is partially removed by this final least-square adjustment.

4. Conclusions

This chapter has revisited the very classical problems of estimating a density and a Poisson intensity in a setting where the unknown signal or its support is unbounded. We have shown that we can build wavelet thresholding estimation procedures that achieve minimax rates (up to a logarithmic term) and are adaptive with respect to the support and the unknown regularity. We have in particular shown that rates can deteriorate according to the sparsity of the signal and have detected an elbow phenomenon for the value $p = 2$ (see Section 2.1.2). Previous results are stated for the $\mathbb{L}_2$-loss and natural extentions should be to consider $\mathbb{L}_{p'}$-losses, with $p' \neq 2$, in particular to analyze how the rates and the elbow phenomenon depend on the loss function. We could also investigate whether the logarithmic term appearing in the upper bound is necessary or not.

It seems to me that another interesting research field consists in the building of data-driven coarsest and finest resolution levels for reconstruction to take into account spatial features of the signal. This scaling problem is a key issue and I do not know theoretical or practical methodologies concerning this point. More generally, given a statistical framework, the question of the ideal dictionary for signal reconstructions remains a topic which is still to a great extent unexplored. Section 3 has shown that such an issue is crucial.

The Lasso and wavelet thresholding procedures proposed in this chapter, that could be used or adapted for very various unidimensional and multidimensional frameworks, are further studied from the calibration point of view in the next chapter.

CHAPTER 4

Calibration

1. Introduction

This chapter constitutes a natural extension of the previous chapter and we still use the notations introduced there. The topic of this chapter is the study of the calibration of the tuning parameter γ of the wavelet thresholding and Dantzig estimates proposed in Chapter 3. We wonder how should this parameter be chosen to obtain good results in both theory and practice.

Previously, we have proved that $\tilde{f}_{n,\gamma}$ achieves optimal theoretical results provided γ is large enough. Such an assumption is very classical in the wavelet thresholding literature (see for instance Cavalier and Koo (2002), Donoho, Johnstone, Kerkyacharian and Picard (1996) or Juditsky and Lambert-Lacroix (2004)). Unfortunately, most of the time, the theoretical choice of the threshold parameter is not suitable for practical issues. More precisely, this choice is often too conservative. See for instance Juditsky and Lambert-Lacroix (2004) who illustrate this statement in Remark 5 of their paper: the tuning parameter of their threshold has to be larger than 14 to obtain theoretical results, but they suggest to take it in the interval $[\sqrt{2}, 2]$ for practical issues. So, one of the main goals of this chapter is to fill the gap between the optimal parameter choice provided by theoretical results on the one hand and by a simulation study on the other hand.

For Lasso-type estimators, the regularization parameter is, most of the time, of the form $a\sqrt{\log M/n}$ with a a positive constant (see Bickel, Ritov and Tsybakov (2009), Bunea, Tsybakov and Wegkamp (2006, 2007a, 2007b), Candès and Plan (2007), Lounici (2008) or Meinhausen and Yu (2009) for instance). Then, one can derive oracle inequalities that are satisfied with large probability that depends on the tuning parameter a that is hard to calibrate in practice.

Only a few papers have been devoted to theoretical calibration of statistical procedures. In the model selection setting, the issue of calibration has been addressed by Birgé and Massart (2007). They considered penalized estimators in a Gaussian homoscedastic regression framework with known variance and calibration of penalty constants is based on the following methodology. They showed that there exists a minimal penalty since taking smaller penalties leads to estimation procedures with suboptimal convergence rates. Under some conditions, they further prove that the optimal penalty is twice the minimal penalty. This so-called 'slope heuristic' method has been successfully applied for practical purposes by Lebarbier (2005) for change points detection or Maugis and Michel (2008) in mixture models. Baraud, Giraud and Huet (2008) (respectively Arlot and Massart (2009)) generalized these results when the variance is unknown (respectively for non-Gaussian or heteroscedastic data). These approaches constitute alternatives to popular cross-validation methods whose computational cost can be a drawback.

The next section describes the framework in which the theoretical study of calibration is lead for the wavelet thresholding and Dantzig estimates. We prove the existence of a minimal value for the tuning parameter γ. The numerical study of Section 3 allows to go further by presenting the situations for which theoretical results seem to remain valid. This section also provides a simple guide on how to select a convenient tuning parameter in practice.

Finally, I mention that some calibration results for kernel rules can also be found in [R8] but they are not presented in this manuscript.

2. A theoretical approach of calibration

2.1. Wavelet thresholding in the Poisson setting. In this theoretical section, we consider the estimator $\tilde{f}_{n,\gamma}$ proposed in (3.5) built with the Haar basis in the Poisson model. In the sequel, j_0 is the integer such that $2^{j_0} \leq n < 2^{j_0+1}$ and we discuss the choice of γ. The calibration study is restricted to the class $\mathcal{F}$ defined as the set of positive functions that can be decomposed on a finite combination of $(\tilde{\varphi}_\lambda)_{\lambda\in\Lambda}$:

$$\mathcal{F} = \left\{ f = \sum_{\lambda\in\Lambda} \beta_\lambda \tilde{\varphi}_\lambda \geq 0 : \quad \text{card}\{\lambda \in \Lambda :\ \beta_\lambda \neq 0\} < \infty \right\}.$$

To study sharp performances of our procedure, we introduce a subclass of the class $\mathcal{F}$: for any n and any radius R, we define:

$$\mathcal{F}_n(R) = \left\{ f \geq 0 : \quad f \in \mathbb{L}_1(R) \cap \mathbb{L}_2(R) \cap \mathbb{L}_\infty(R),\ F_\lambda \geq \frac{(\log n)(\log\log n)}{n} 1_{\beta_\lambda \neq 0},\ \forall\, \lambda \in \Lambda \right\},$$

where for any λ, we set

$$F_\lambda = \int_{\text{supp}(\varphi_\lambda)} f(x)dx \quad \text{and} \quad \text{supp}(\varphi_\lambda) = \{x \in \mathbb{R} :\ \varphi_\lambda(x) \neq 0\},$$

which allows to establish a decomposition of $\mathcal{F}$. Indeed, we have the following result proved in Section 3 of [R11]:

PROPOSITION 4.1. *When n (or R) increases, $(\mathcal{F}_n(R))_{n,R}$ is a non-decreasing sequence of sets. In addition, we have:*

$$\bigcup_n \bigcup_R \mathcal{F}_n(R) = \mathcal{F}.$$

The definition of $\mathcal{F}_n(R)$ especially relies on the technical condition

$$F_\lambda \geq \frac{(\log n)(\log\log n)}{n} 1_{\beta_\lambda \neq 0}. \tag{4.1}$$

Remember that the distribution of the number of points of N that lies in $\text{supp}(\varphi_\lambda)$ is the Poisson distribution with mean nF_λ. So, the previous condition ensures that we have a significant number of points of N to estimate non-zero wavelet coefficients. Another main point is that under (4.1),

$$\sqrt{V_{\lambda,n}\log n} \geq \frac{\log n \|\varphi_\lambda\|_\infty}{n} \times \sqrt{\log\log n},$$

so (3.5) is true with large probability. The term $\frac{(\log n)(\log\log n)}{n}$ appears for technical reasons but could be replaced by any term u_n such that

$$\lim_{n\to\infty} u_n = 0 \quad \text{and} \quad \lim_{n\to\infty} u_n^{-1} \left(\frac{\log n}{n} \right) = 0.$$

In practice, many interesting signals are well approximated by a function of $\mathcal{F}$. So, using Proposition 4.1, a convenient estimate is an estimate with a good behavior on $\mathcal{F}_n(R)$, at least for large values of n and R. We now focus on $\tilde{f}_{n,\gamma}$ with the special value $\gamma = 1 + \sqrt{2}$ and we study its oracle properties on $\mathcal{F}_n(R)$. Roughly speaking, the following result can be viewed as a complement of Theorem 3.2 for which the constants C_1 and C_2 are specified at the price of the restriction to $\mathcal{F}_n(R)$.

THEOREM 4.1. *Let $R > 0$ be fixed. Let $\gamma = 1 + \sqrt{2}$ and let $\eta_{\lambda,\gamma}$ be as in (3.4) (with $c_1 = 1$ and $c_2 = 3$). Then $\tilde{f}_{n,\gamma}$ achieves the following oracle inequality: for n large enough, for any $f \in \mathcal{F}_n(R)$,*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2 \right] \leq 12 \log n \left[\sum_{\lambda\in\Gamma_n} \min(\beta_\lambda^2, V_{\lambda,n}) + \frac{1}{n} \right]. \tag{4.2}$$

Inequality (4.2) shows that on $\mathcal{F}_n(R)$, our estimate achieves the oracle risk up to the term $12\log n$ and the negligible term $\frac{1}{n}$ (see Paragraph 2.1.1 of Chapter 3). Finally, let us mention that when $f \in \mathcal{F}_n(R)$,

$$\sum_{\lambda \notin \Gamma_n} \beta_\lambda^2 = 0.$$

Our result is stated with $\gamma = 1 + \sqrt{2}$. This value comes from optimizations of upper bounds given by technical arguments of the proofs of Theorem 4.1. So, the value $\gamma = 1 + \sqrt{2}$ should not be seen as the optimal one. But, Theorem 4.1 constitutes a first theoretical calibration result and this is the first step for choosing the parameter γ in an optimal way.

Now, we are ready to lead the calibration study. Theorem 3.2 has established that for any signal, we achieve the oracle estimator up to a logarithmic term provided $\gamma > 1$. So, our primary interest is to wonder what happens, from the theoretical point of view, when $\gamma \leq 1$? To handle this problem, we consider the simplest signal in our setting, namely

$$f = 1_{[0,1]}.$$

Applying Theorem 3.2 with $\gamma > 1$ gives

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C\frac{\log n}{n},$$

where C is a constant. The following result shows that this rate cannot be achieved for this particular signal when $\gamma < 1$.

THEOREM 4.2. *Let $f = 1_{[0,1]}$. If $\gamma < 1$ then there exists $\delta < 1$ not dependent of n such that*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \geq \frac{c}{n^\delta},$$

where c is a constant.

Theorem 4.2 establishes that, asymptotically, $\tilde{f}_{n,\gamma}$ with $\gamma < 1$ cannot estimate a very simple signal at a convenient rate of convergence. This provides a lower bound for the threshold parameter γ: we have to take $\gamma \geq 1$.

Now, let us study the upper bound for the parameter γ. For this purpose, we do not consider a particular signal, but we use the worst oracle ratio on the whole class $\mathcal{F}_n(R)$. When $\gamma = 1 + \sqrt{2}$, Theorem 4.1 shows that for n large enough,

$$\sup_{f \in \mathcal{F}_n(R)} \frac{\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right]}{\sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_{\lambda,n}) + \frac{1}{n}} \leq 12\log n.$$

Our aim is to establish that the oracle ratio on $\mathcal{F}_n(R)$ for the estimator $\tilde{f}_{n,\gamma}$ where γ is large, is larger than the previous upper bound. This goal is reached in the following theorem.

THEOREM 4.3. *Let $\gamma_{\min} > 1$ be fixed and let $\gamma > \gamma_{\min}$. Then, for any $R \geq 2$,*

$$\sup_{f \in \mathcal{F}_n(R)} \frac{\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right]}{\sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_{\lambda,n}) + \frac{1}{n}} \geq 2(\sqrt{\gamma} - \sqrt{\gamma_{\min}})^2 \log n \times (1 + o_n(1)).$$

Now, if we choose $\gamma > (1 + \sqrt{6})^2 \approx 11.9$, we can take $\gamma_{\min} > 1$ such that the resulting maximal oracle ratio of $\tilde{f}_{n,\gamma}$ is larger than $12\log n$ for n large enough. So, taking $\gamma > 12$ is a bad choice for estimation on the whole class $\mathcal{F}_n(R)$.

Note that the function $1_{[0,1]}$ belongs to $\mathcal{F}_n(2)$, for all $n \geq 2$. So, combining Theorems 4.1, 4.2 and 4.3 proves that the convenient choice for γ belongs to the interval $[1, 12]$. Finally, observe that the rate exponent deteriorates for $\gamma < 1$ whereas we only prove that the choice $\gamma > 12$ leads to worse rates constants.

We do not present the calibration results of $\tilde{f}_{n,\gamma}$ in the density model and we refer the reader to Section 2.2 of [R13]. Let us just mention that the analogue of Theorem 4.2 can be established

at the cost of more involved computations. The rest of this section is devoted to the problem of calibrating the Dantzig and Lasso estimates.

2.2. Dantzig and Lasso estimates in the density setting. We consider the estimates $\hat{f}^D = f_{\hat{\mu}^{D,\gamma}}$ and $\hat{f}^L = f_{\hat{\mu}^{L,\gamma}}$ defined in Section 3 of Chapter 3, and as previously, we prove that the sufficient condition $\gamma > 1$ is 'almost' a necessary condition since we derive a special and very simple framework in which Lasso and Dantzig estimates cannot achieve the optimal rate if $\gamma < 1$ ('almost' means that the case $\gamma = 1$ remains an open question). Let us describe this simple framework. The dictionary Υ considered in this section is again the orthonormal Haar system. In this case, $M = n$. In this setting, since functions of Υ are orthonormal, the Gram matrix G is the identity. Thus, the Lasso and Dantzig estimates both correspond to the soft thresholding rule:

$$\hat{f}^D = \hat{f}^L = \sum_{\lambda=1}^{M} \mathrm{sign}(\hat{\beta}_\lambda) \left(|\hat{\beta}_\lambda| - \eta_{\lambda,\gamma} \right) 1_{\{|\hat{\beta}_\lambda| > \eta_{\lambda,\gamma}\}} \tilde{\varphi}_\lambda.$$

Now, our goal is to estimate $f = 1_{[0,1]}$ by using $\hat{f}^D$ depending on γ and to show the influence of this tuning parameter.

THEOREM 4.4. *On the one hand, if $\gamma > 1$, there exists a constant C such that*

$$\mathbb{E}\left[\|\hat{f}^D - f\|_2^2 \right] \leq \frac{C \log n}{n}. \tag{4.3}$$

On the other hand, if $\gamma < 1$, there exists a constant c and $\delta < 1$ such that

$$\mathbb{E}\left[\|\hat{f}^D - f\|_2^2 \right] \geq \frac{c}{n^\delta}. \tag{4.4}$$

This result shows again that choosing $\gamma < 1$ is a bad choice in our setting.

3. A numerical approach of calibration

In this section, we discuss the ideal numerical choice for the parameter γ keeping in mind that the value $\gamma = 1$ constitutes a border for the theoretical results (see Theorems 3.2 and 4.2 and 4.4). For this purpose, we first consider the Poisson setting. We consider either the Haar basis or a special case of spline systems given in Figure 1 of [R11]. The latter, called hereafter the spline basis, has the following properties. First, the support of ϕ, ψ, $\tilde{\phi}$ and $\tilde{\psi}$ is included in $[-4, 5]$. The reconstruction wavelets $\tilde{\phi}$ and $\tilde{\psi}$ belong to $C^{1.272}$. Finally, the wavelet ψ is a piecewise constant function orthogonal to polynomials of degree 4. We still consider the thresholding rule $\tilde{f}_{n,\gamma}$ with $\tilde{f}_{n,\gamma}$ defined in (3.5) with

$$\eta_{\lambda,\gamma} = \sqrt{2\gamma \log(n) \hat{V}_{\lambda,n}} + \frac{\gamma \log n}{3n} \|\varphi_\lambda\|_\infty.$$

Observe that $\eta_{\lambda,\gamma}$ slightly differs from the threshold defined in (3.4) since $\tilde{V}_{\lambda,n}$ is now replaced with $\hat{V}_{\lambda,n}$. It allows to derive the parameter γ as an explicit function of the threshold which is necessary to draw figures without using a discretization of γ, which is crucial in the sequel. The performances of our thresholding rule associated with the threshold $\eta_{\lambda,\gamma}$ defined in (3.4) are probably equivalent. Given n and a function f, we denote $R_n(\gamma)$ the ratio between the ℓ_2-performance of our procedure (depending on γ) and the oracle risk where the wavelet coefficients at levels $j > j_0$ are omitted. We have:

$$R_n(\gamma) = \frac{\sum_{\lambda \in \Gamma_n} (\tilde{\beta}_\lambda - \beta_\lambda)^2}{\sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_{\lambda,n})} = \frac{\sum_{\lambda \in \Gamma_n} (\hat{\beta}_\lambda 1_{|\hat{\beta}_\lambda| \geq \eta_{\lambda,\gamma}} - \beta_\lambda)^2}{\sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_{\lambda,n})}.$$

Of course, R_n is a stepwise function and the change points of R_n correspond to the values of γ such that there exists λ with $\eta_{\lambda,\gamma} = |\hat{\beta}_\lambda|$. The average over 1000 simulations of $R_n(\gamma)$ is computed providing an estimation of $\mathbb{E}(R_n(\gamma))$. This average ratio, denoted $\overline{R_n}(\gamma)$ and viewed as a function of γ, is plotted for $n \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ and for three very

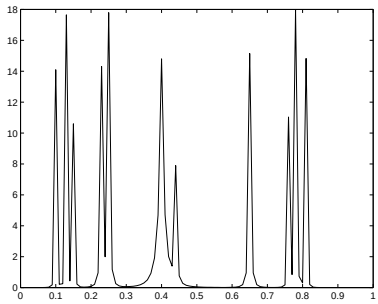

FIGURE 1. The function 'Bumps'.

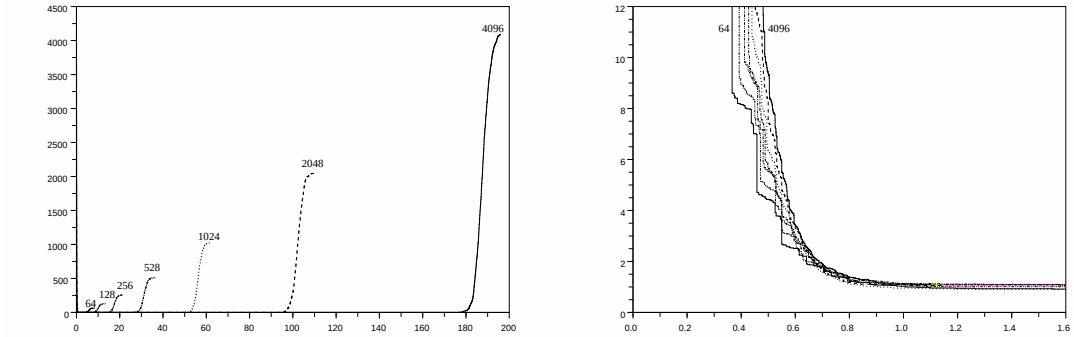

FIGURE 2. The function $\gamma \to \overline{R_n}(\gamma)$ at two scales for 'Haar1' decomposed on the Haar basis and for $n \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ with $j_0 = \log_2(n)$.

different signals. The signal 'Haar1' is $1_{[0,1]}$, the signal 'Gauss1' is the density of a centered Gaussian variable with variance 0.0625. The signal 'Bumps' is given in Figure 1 (see Section 8 of [R11] for a precise definition of Bumps). For non compactly supported signals, we need to compute an infinite number of wavelet coefficients to determine this ratio. To overcome this problem, we omit the tails of the signals and we focus our attention on an interval that contains all observations. Of course, we ensure that this approximation is negligible with respect to the values of R_n. As previously, we take $j_0 = \log_2(n)$. Figure 2 displays $\overline{R_n}$ for 'Haar1' decomposed on the Haar basis. The left side of Figure 2 gives a general idea of the shape of $\overline{R_n}$, while the right side focuses on small values of γ. Similarly, Figures 3 and 4 display $\overline{R_n}$ for 'Gauss1' decomposed on the spline basis and for 'Bumps' decomposed on the Haar and the spline bases.

To discuss our results, we introduce

$$\gamma_{\min}(n) = \operatorname{argmin}_{\gamma>0} \overline{R_n}(\gamma).$$

For 'Haar1', $\gamma_{\min}(n) \geq 1$ for any value of n and taking $\gamma < 1$ deteriorates the performances of the estimate. The larger n, the stronger the deterioration is. Such a result was established from the theoretical point of view in Theorem 4.2. In fact, Figure 2 allows to draw the following major conclusion for 'Haar1':

$$\overline{R_n}(\gamma) \approx \overline{R_n}(\gamma_{\min}(n)) \approx 1 \tag{4.5}$$

for γ belonging to a large interval that contains the value $\gamma = 1$. For instance, when $n = 4096$, the function $\overline{R_n}$ is close to 1 for any value of the interval $[1, 177]$. So, we observe a kind of "plateau

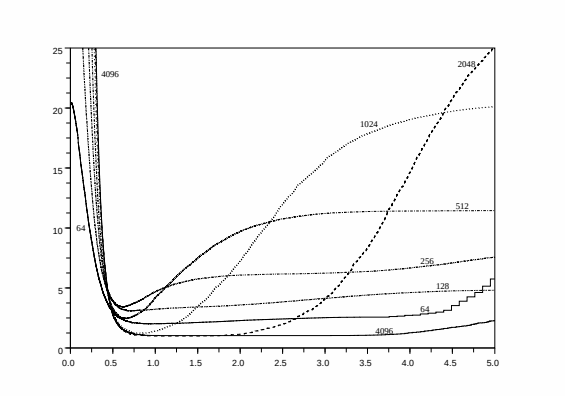

FIGURE 3. The function $\gamma \rightarrow \overline{R_n}(\gamma)$ for 'Gauss1' decomposed on the spline basis and for $n \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ with $j_0 = \log_2(n)$.

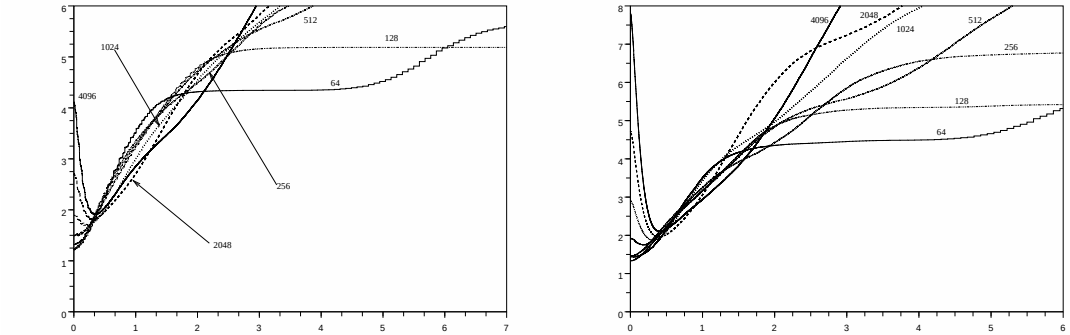

FIGURE 4. The function $\gamma \rightarrow \overline{R_n}(\gamma)$ for 'Bumps' decomposed on the Haar and the spline bases and for $n \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ with $j_0 = \log_2(n)$.

phenomenon". Finally, we conclude that our thresholding rule with $\gamma = 1$ performs very well since it achieves the same performance as the oracle estimator.

For 'Gauss1', $\gamma_{\min}(n) \geq 0.5$ for any value of n. Moreover, as soon as n is large enough, the oracle ratio for $\gamma_{\min}(n)$ is close to 1. Besides, when $n \geq 2048$, as for 'Haar1', $\gamma_{\min}(n)$ is larger than 1. We observe the "plateau phenomenon" as well and as for 'Haar1', the size of the plateau increases when n increases. This can be explained by the following important property of 'Gauss1': 'Gauss1' can be well approximated by a finite combination of the atoms of the spline basis. So, we have the strong impression that the asymptotic result of Theorem 4.2 could be generalized for the spline basis.

Conclusions for 'Bumps' are very different. Remark that this irregular signal has many significant wavelet coefficients at high resolution levels whatever the basis. We have $\gamma_{\min}(n) < 0.5$ for each value of n. Besides, $\gamma_{\min}(n) \approx 0$ when $n \leq 256$, which means that all the coefficients until $j = j_0$ have to be kept to obtain the best estimate. So, the parameter j_0 plays an essential role and has to be well calibrated to ensure that there are no non-negligible wavelet coefficients for $j > j_0$. Other differences between Figure 2 (or Figure 3) and Figure 4 have to be emphasized. For 'Bumps', when $n \geq 512$, the minimum of $\overline{R_n}$ is well localized, there is no plateau anymore and $\overline{R_n}(1) > 2$. Note that $\overline{R_n}(\gamma_{\min}(n))$ is larger than 1.

Previous preliminary conclusions show that the ideal choice for γ and the performance of the thresholding rule highly depend on the decomposition of the signal on the wavelet basis. This study

has been pursued for six other signals in Section 5.2 of [R11], which allows to draw consistent conclusions with respect to the issue of calibrating γ from the numerical point of view. To present them, let us introduce two classes of functions.

The first class is the class of signals that only have negligible coefficients at high levels of resolution. The wavelet basis is well adapted to the signals of this class that contains 'Haar1' for the Haar basis and 'Gauss1' for the spline basis. For such signals, the estimation problem is close to a parametric problem. In this case, the performance of the oracle estimate can be achieved at least for n large enough and (4.5) is true for γ belonging to a large interval that contains the value $\gamma = 1$. These numerical conclusions strengthen and generalize theoretical conclusions of Section 2.1.

The second class of functions is the class of irregular signals with significant wavelet coefficients at high resolution levels. For such signals $\gamma_{\min}(n) < 0.8$ and there is no "plateau" phenomenon (in particular, we do not have $\overline{R_n}(1) \simeq \overline{R_n}(\gamma_{\min}(n))$).

Of course, estimation is easier and performances of our procedure are better when the signal belongs to the first class. But in practice, it is hard to choose a wavelet system such that the intensity to be estimated satisfies this property. However, our study allows to use the following simple rule. If the practitioner has no idea of the ideal wavelet basis to use, he should perform the thresholding rule with $\gamma = 1$ (or γ slightly larger than 1) that leads to convenient results whatever the class the signal belongs to.

For the Dantzig and Lasso estimates, a small simulation study is also carried out to strengthen theoretical asymptotic results. Performing our estimation procedure 100 times, we compute the average risk for several values of the tuning parameter γ and several values of n. This computation is summarized in Figure 1 of [R12]. Still denoting $\gamma_{\min}(n)$ the value of γ that minimizes the average risk, we note that $1/2 \leq \gamma_{\min}(n) \leq 1$ for all values of n, with $\gamma_{\min}(n)$ getting closer to 1 as n increases. Taking γ too small strongly deteriorates the performance while a value close to 1 ensures a risk withing a factor 2 of the optimal risk. The assumption $\gamma > 1$ giving a theoretical control on the quadratic error is thus not too conservative. Following these results, we have taken $\gamma = 1.01$ in our numerical experiments in the previous chapter.

4. Conclusions

Theoretical calibration issues have not been widely investigated yet, although it constitutes a major concern for practitioners. This topic is a very exciting research field, and, in the wake of works by Massart and Arlot, can be considered for very various estimation procedures or statistical models. We can note that this chapter does not address all the theoretical issues raised previously. Indeed, if Section 3 seems to show that Theorems 4.2 and 4.4 can be extended outside the Haar basis setting, we are, at this stage, unable to prove it. Furthermore, Section 3 presents some situations in which the choice $\gamma > 1$ is not convenient and the question of the existence of a theoretical minimal value for γ remains open. More dramatically, we have actually no guarantee that the shape of the threshold is convenient for such situations. I emphasize that, of course, calibration issues can be handled only if the form of tuning parameter is suitable. Finally, observe that if we have pointed out minimal tuning parameters in some cases, the question of the optimal ones has not been addressed in our theoretical setting. In particular, the problem of the existence of a relationship analog to the remarkable magic formula derived by Birgé and Massart :

$$\text{the optimal penalty} = 2 \times \text{the minimal penalty},$$

valid in very special cases, remains an open problem hard to solve from both theoretical and practical points of view.

Conclusion

To end this manuscript, I present Laure Sansonnet's PhD dissertation topic entitled 'Adaptive estimation of Poisson interactions'. Co-supervised with Patricia Reynaud-Bouret, this topic constitutes a natural extension of some results presented in the previous chapters.

"The subject of this thesis is the study of some non-parametric statistical problems in the framework of a Poisson interactions model. Such models are used for instance in genetics, to study favored distances between patterns on a strand of DNA. In this setting, we naturally introduce a so-called reproduction function that allows to quantify the favored positions of the patterns and can be modeled as the intensity of a Poisson process. Our primary interests are the estimation of this function and some associated problems on tests. Besides, it is natural to assume that the reproduction function is localized. Therefore, natural tools to handle these issues are wavelet thresholding. Such algorithms have proved efficient in a very simple Poisson setting, both from a theoretical and a practical point of view. The task of Laure Sansonnet will consist in extending these methods in the setting mentioned above, in which the basic model has different versions. Very few non-parametric statistical results have been established in this field, where applications are manifold. This gives the opportunity to Laure to consider various research directions."

List of my papers

Publications:

[R1] RIVOIRARD V. (2004). Maxisets for linear procedures. *Statistics and Probability Letters* **67**(3), 267–275.

[R2] RIVOIRARD V. (2004). Thresholding procedure with priors based on Pareto distributions. *Test* **13**(1), 213–246.

[R3] RIVOIRARD V. (2005). Bayesian modelling of sparse sequences and maxisets for Bayes rules. *Mathematical Methods of Statistics* **14**(3), 346–376.

[R4] RIVOIRARD V. (2006). Non linear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli* **12**(4), 609–632.

[R5] AUTIN F., PICARD D. and RIVOIRARD V. (2006). Large variance Gaussian priors in Bayesian nonparametric estimation: a maxiset approach. *Mathematical Methods of Statistics* **15**(4), 349–373.

[R6] RIVOIRARD V. and TRIBOULEY K. (2008). The maxiset point of view for estimating integrated quadratic functionals. *Statistica Sinica* **18**(1), 255–279.

[R7] LOUBES J.M. and RIVOIRARD V. (2009). Review of rates of convergence and regularity conditions for inverse problems. *International Journal of Tomography & Statistics* **11**(S09), 61–82.

[R8] BERTIN K. and RIVOIRARD V. (2009). Maxiset in sup-norm for kernel estimators. To appear in *Test*.

[R9] AUTIN F., LE PENNEC E., LOUBES J.M. and RIVOIRARD V. (2009). Maxisets for model selection. To appear in *Constructive approximation.*

Submitted articles:

[R10] REYNAUD-BOURET P. and RIVOIRARD V. (2009). Near optimal thresholding estimation of a Poisson intensity on the real line.

[R11] REYNAUD-BOURET P. and RIVOIRARD V. (2009). Calibration of thresholding rules for Poisson intensity estimation.

[R12] BERTIN K., LE PENNEC E. and RIVOIRARD V. (2009). Adaptive Dantzig density estimation.

[R13] REYNAUD-BOURET P., RIVOIRARD V. and TULEAU-MALOT C. (2009). Adaptive density estimation: a curse of support?

Paper in preparation:

[R14] RIVOIRARD V. and ROUSSEAU J. Bernstein Von-Mises Theorem for linear functionals of the density.

Bibliography

[1] Abramovich F., Amato U. and Angelini C. (2004). On optimality of Bayesian wavelet estimators. *Scand. J. Stat.* **31**(2), 217–234.

[2] Abramovich F., Angelini C. and De Canditiis D. (2007). Pointwise optimality of Bayesian wavelet estimators. *Annals of the Institute of Statistical Mathematics* **59**, 425–434.

[3] Abramovich F. and Benjamini Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In Wavelets and Statistics, Lecture Notes in Statistics. **103** A. Antoniadis and G. Oppenheim (Eds), 5–14.

[4] Abramovich F, Benjamini Y, Donoho D.L. and Johnstone I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann Stat* **34**, 584–653.

[5] Abramovich F., Grinshtein V. and Pensky M. (2007). On optimality of Bayesian testimation in the normal means problem. *Annals of Statistics* **35**, 2261–2286.

[6] Abramovich F., Sapatinas T. and Silverman B.W. (1998). Wavelet thresholding via a Bayesian approach *J. Roy. Statist. Soc. B* **60**, 725–749.

[7] Antoniadis A., Grégoire G. and Nason G. (1999). Density and hazard rate estimation for right censored data using wavelet methods. *Journ. Royal Statist. Soc. B* **61**(1), 63–84.

[8] Arlot S. and Massart P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research* **10**, 245–279.

[9] Asif M. S. and Romberg J. (2009). Dantzig selector homotopy with dynamic measurements. *Proceedings of SPIE Computational Imaging VII.* **7246**(1), 72460E.

[10] Autin F. (2006). Maxiset for density estimation on $\mathbb{R}$. *Math. Methods Statist* **15**(2), 123–145.

[11] Autin F. (2008). Maxisets for mu-thresholding rules. *Test* **17**(2), 332–349.

[12] Baraud Y. and Birgé L. (2006). Estimating the intensity of a random measure by histogram type estimators. To appear in *Probab. Theory Related Fields.*

[13] Baraud Y., Giraud C. and Huet S. (2008). Gaussian model selection with unknown variance. To appear in *The Annals of Statistics.*

[14] Belitser E. and Ghosal S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31**, 536–559.

[15] Bertin K. (2005). Sharp adaptive estimation in sup-norm for d-dimensional Hölder classes, *Math. Methods Statist.* **14**(3), 267–298.

[16] Bickel P.J. and Ritov Y. (1988). Estimating integrated squared density derivatives. *Sankhya A* **50**, 381–393.

[17] Bickel P.J., Ritov Y. and Tsybakov A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**(4), 1705–1732.

[18] Birgé L. (2006). Model selection for Poisson processes. Manuscript.

[19] Birgé L. (2008). Model selection for density estimation with $\mathbb{L}_2$-loss. Submitted.

[20] Birgé L. and Massart P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* **138**(1-2), 33–73.

[21] Bochkina N. and Sapatinas T. (2005). On the Posterior Median Estimators of Possibly Sparse Sequences. *Annals of Institute of Statistical Mathematics* **57**, 315–351.

[22] Bochkina N. and Sapatinas T. (2006). On pointwise optimality of Bayes Factor wavelet regression estimators. *Sankhya* **68**, 513–541.

[23] Bochkina N. and Sapatinas T. (2009). Minimax rates of convergence and optimality of Bayes factor wavelet regression estimators under pointwise risks. *Statistica Sinica* **19** (To appear).

[24] Brown L., Cai T., Zhang R., Zhao L. and Zhou H. (2007). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*. To appear.

[25] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006),* Lecture Notes in Artificial Intelligence **4005** (Lugosi, G. and Simon, H.U.,eds.), Springer-Verlag, Berlin-Heidelberg.

[26] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007a). Sparse density estimation with ℓ_1 penalties. *Lecture Notes in Artificial Intelligence* **4539**, 530–543.
[27] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007b). Aggregation for Gaussian regression. *Annals of Statistics* **35**(4), 1674–1697.
[28] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007c). Sparsity Oracle Inequalities for the LASSO. *Electronic Journal of Statistics* **1**, 169–194.
[29] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2009). Spades and Mixture Models. Submitted.
[30] Bunea F. (2008). Consistent selection via the Lasso for high dimensional approximating regression models. *IMS Lecture notes-Monograph Series* **3**, 122–137.
[31] Cai T.T. and Low M.G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33**, 2930–2956.
[32] Candès E.J. and Plan Y. (2007). Near-ideal model selection by ℓ_1-minimization. To appear in *The Annals of Statistics.*
[33] Candès E.J. and Tao T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. With comments and a rejoinder by the authors. *The Annals of Statistics* **35**(6), 2313–2351.
[34] Castillo I. (2008). Electronic Journal of Statistics **2**, 1281–1299.
[35] Cavalier L. and Koo J.Y. (2002). Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. *IEEE Trans. Inform. Theory* **48**(10), 2794–2802.
[36] Chipman H.A., Kolaczyk E.D. and McCulloch R.E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Stat. Assoc.* **92**, 1413–1421.
[37] Clyde M., Parmigiani G. and Vidakovic B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–402.
[38] Cohen A. (2000). Wavelet methods in numerical analysis, Handbook of Numerical Analysis, vol. VII, P.G. Ciarlet and J.L. Lions, eds.
[39] Cohen A., DeVore R., Kerkyacharian G. and Picard D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.* **11**(2), 167–191.
[40] Cohen A., DeVore R.A. and Hochmuth R. (2000). Restricted nonlinear approximation. *Constr. Approx.* **16**, 85–113.
[41] Copas J.B. and Fryer M.J. (1980). Density estimation and suicide risks in psychiatric treatment. *J. Roy. Statist. Soc. A* **143**, 167–176.
[42] Cutillo L., Jung Y.Y., Ruggeri F. and Vidakovic B. (2008). Larger Posterior Mode Wavelet Thresholding and Applications, *Journal of Statistical Planning and Inference* **138**, 3758–3773.
[43] Dalalyan A. and Tsybakov A.B. (2009). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. Submitted.
[44] DeVore R.A. (1989). Degree of nonlinear approximation. In Approximation theory VI, 1(College Station, TX, 1989), 175-201. Academic Press, Boston, MA.
[45] DeVore R.A, Konyagin S. and Temlyakov V. (1998). Hyperbolic wavelet approximation. *Constr. Approx.* **14**, 1–26.
[46] DeVore R.A. and Lorentz G.G. (1993). *Constructive approximation*, Springer-Verlag, Berlin.
[47] Donoho D.L. (1993). *Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data*, Different perspectives on wavelets (San Antonio, TX, 1993), 173–205,Proc. Sympos. Appl. Math., **47**, Amer. Math. Soc., Providence, RI.
[48] Donoho D.L. and Johnstone I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3), 425–455.
[49] Donoho D.L. and Johnstone I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200–1224.
[50] Donoho D.L. and Johnstone I.M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2**, 39–62.
[51] Donoho D.L., Johnstone I.M., Kerkyacharian G. and Picard D. (1996). Density estimation by wavelet thresholding. *Annals of Statistics* **24**(2), 508–539.
[52] Donoho D.L., Johnstone I.M., Kerkyacharian G. and Picard D. (1997). Universal near minimaxity of wavelet shrinkage. Festschrift for Lucien Le Cam, 183–218, Springer, New York.
[53] Donoho D.L. and Nussbaum M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6**(3), 290–323.
[54] Efromovich S. and Low M.G. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24**, 1106–1125.
[55] Fan J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19**, 1273–1294.
[56] Figueroa-López J.E. and Houdré C. (2006). Risk bounds for the non-parametric estimation of Lévy processes. IMS Lecture Notes-Monograph series High Dimensional Probability **51**, 96–116.
[57] Gayraud G. and Rousseau J. (2005). Rates of convergence for a Bayesian level set estimation. *Scand. Journ. Statist.* **14**(1), 75–94.
[58] Ghosal S., Ghosh J.K. and van der Vaart A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist* **28**, 500–531.

[59] Ghosal S. and van der Vaart A.W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35**, 697–723.
[60] Herrick D.R.M., Nason G.P. and Silverman B.W. (2001). Some new methods for wavelet density estimation. *Sankhya Ser. A* **63** (3), 394–411.
[61] Huang T.Z. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32**, 1556–1593.
[62] Huang H.C. and Cressie N. (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics* **42**(3), 262–276.
[63] Ibragimov I.A. and Khas'minskii R.Z. (1980) Some estimation problems for stochasitc differential equations *Lecture Notes in Control and Inform. Sci. B* **25**, 1–12. Springer, Berlin.
[64] Johnstone I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. Statistical decision theory and related topics, V (West Lafayette, IN, 1992), 303–326, Springer, New York.
[65] Johnstone I.M. and Silverman B.W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. Technical report.
[66] Johnstone I.M. and Silverman B. W. (2004). Needles and hay in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32**, 1594–1649.
[67] Johnstone I.M. and Silverman B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* **33**, 1700-1752.
[68] Juditsky A. and Lambert-Lacroix S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10**(2), 187–220.
[69] Kerkyacharian G. and Picard D. (1993). Density estimation by kernel and wavelets methods: optimality of Besov spaces. *Statist. Probab. Lett.* **18**(4), 327–336.
[70] Kerkyacharian G. and Picard D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. With comments and a rejoinder by the authors. *Test* **9**(2), 283–344.
[71] Kerkyacharian G. and Picard D. (2002). Minimax or maxisets? *Bernoulli* **8**(2), 219–253.
[72] Kolaczyk, E.D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9**(1), 119–135.
[73] Kruijer W., Rousseau J., Van Der Vaart A.W. (2009). Adaptive Bayesian Density Estimation with Location-Scale Mixtures. Submitted.
[74] Laurent B. and Massart P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics* **28**(5), 1302–1338.
[75] Lebarbier E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85**(4), 717–736.
[76] Lepskii O. V. (1992). On problems of adaptive estimation in white Gaussian noise, In: Topics in nonparametric estimation, Adv. Soviet Math. **12** Amer. Math. Soc.Providence, RI, 87–106.
[77] Lorentz G.G. (1950). Some new functional spaces. *Ann. of Math.* **51**(2), 37–55.
[78] Lorentz G.G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.* **72**, 903–937
[79] Lounici K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of statistics* **2**, 90–102.
[80] Mallat S. (1989). Multiresolution approximations and wavelet orthonormal bases of $\mathbb{L}_2(\mathbb{R})$. *Trans. Amer. Math. Soc.* **315**(1), 69–87.
[81] Massart, P. (2007). Concentration inequalities and model selection. *Lectures on probability theory and statistics (Saint-Flour, 2003)*, Lecture Notes in Math., 1896, Springer, Berlin.
[82] Maugis C. and Michel B. (2008). Data-driven penalty calibration: A case study for Gaussian mixture model selection. Submitted.
[83] Meinhausen N. and Yu B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**(1), 246–270.
[84] Meyer Y. (1990). *Ondelettes et opérateurs. I.*, Actualités Mathématiques, Hermann, Paris.
[85] Nason G.P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society Series B* **58**, 463–479.
[86] Pensky M. (2006). Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. *Annals of Statistics* **34**, 769–807.
[87] Pensky M. and Sapatinas T. (2007). Frequentist optimality of Bayes factor estimators in wavelet regression models. *Statistica Sinica* **17**, 599–633.
[88] Pham Ngoc T.M. (2009). Regression in random design and Bayesian warped wavelets estimators. Submitted.
[89] Reynaud-Bouret P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probability Theory and Related Fields* **126**(1), 103–153.
[90] Robert C.P. (2006). *Le choix bayésien. Principes et pratique*. Springer. Berlin.
[91] Rousseau J. (2009). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparamatric estimation of the density. To appear in *Annals of Statist.*
[92] Rudemo M. (1982). Empirical choice of histograms and density estimators. *Scand. J. Statist.* **9**(2), 65–78.

[93] Sain S.R. and Scott D.W. (1996). On locally adaptive density estimation. *J. Amer. Statist. Assoc.* **91**(436), 1525–1534.

[94] Scricciolo C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34**, 2897–2920.

[95] Shen X. and Wasserman L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**, 687–714.

[96] Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Monograph on Statistics and Applied Probability **26** Chapman & Hall.

[97] Silverman B.W. (2007). Empirical Bayes thresholding: adapting to sparsity when it advantageous to do so. *J. Korean Stat. Soc.* **36**(1), 1–29.

[98] Temlyakov V. (1999). Greedy algorithms and m-term approximation with regard to redundant dictionaries. *J. Approx. Theory* **98**, 117–145.

[99] Tribouley K. (2000). Adaptive Estimation of Integrated Functionals. *Mathematical Methods of Statistics* **9**, 19–36.

[100] van de Geer S. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.*, **36**(2), 614–645.

[101] van der Vaart A.W. and van Zanten H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics* **36**(3), 1435–1463.

[102] van der Vaart A.W. and van Zanten H. (2009). Adapative Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. To appear in *Annals of Statist.*

[103] Vidakovic B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173–179.

[104] Weisberg S. (1980). *Applied Linear Regression* New-York: Wiley.

[105] Willett R.M. and Nowak R.D. (2007). Multiscale Poisson Intensity and Density Estimation. *IEEE Transactions on Information Theory* **53**(9), 3171–3187.