Universités de Paris 6 & Paris 7 - CNRS (UMR 7599)

PRÉPUBLICATIONS DU LABORATOIRE

DE PROBABILITÉS & MODÈLES ALÉATOIRES

4, place Jussieu - Case 188 - 75 252 Paris cedex 05
http://www.proba.jussieu.fr

Bayesian thresholding with priors based on Pareto distributions

V. RIVOIRARD

SEPTEMBRE 2001

Prépublication n^o 687

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris VI & Université Paris VII,
4, place Jussieu, Case 188, F-75252 Paris Cedex 05.

Bayesian thresholding with priors based on Pareto distributions.

Vincent Rivoirard*
Laboratoire de Probabilités et Modèles Aléatoires
CNRS-UMR 7599, Université Paris VI et Université Paris VII, France.

Abstract

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

Key Words: adaptive estimation, bayesian model, Pareto distribution, sparsity, wavelet thresholding, weak Besov spaces.
AMS subject classification: 62G05, 62G08, 62C12.

1 Introduction

1.1 Motivation

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

$$g_i = f(\frac{i}{n}) + \sigma\varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n. \tag{1.1}$$

We would like to build an efficient procedure that provides a data adaptive estimator of the signal f without having to assume any specific form about this signal. By expanding f on a wavelet basis which atoms are localized in both time and frequency, we expect a parsimonious representation of f: Only a small number of the wavelet coefficients are non negligible in which the main part of the information about f is contained. In the following,

*Correspondence to: Vincent Rivoirard. Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI et Université Paris VII, 16 rue de Clisson, F-75013 Paris, France. E-mail: rivoirar@math.jussieu.fr

we consider such *sparse* functions we estimate by using thresholding rules particularly appropriate to this framework. One goal of this paper is to discuss the choice of the threshold. Many authors have investigated this problem: Donoho and Johnstone (1994) proposed the VisuShrink procedure consisting in choosing the universal threshold $\lambda^u = \sigma\sqrt{2\log(n)}$ for each level. The choice of the threshold for Donoho and Johnstone (1995) is based on the minimization of Stein's unbiased estimate of risk for threshold estimates. Nason (1995) exploited the cross validation approach to choose the threshold. Let us also mention the methods based on the multiple hypotheses testing approach. See for instance Abramovich and Benjamini (1995) or Abramovich, Benjamini, Donoho and Johnstone (2000) who obtained thresholding rules by adapting the false discovery rate method developed by Benjamini and Hochberg (1995). Abramovich, Sapatinas and Silverman (1998) and Vidakovic (1998) considered thresholding within a bayesian framework. We adopt this approach and we place a prior model on the wavelet coefficients. But before this, we assume that the function f belongs to a weak Besov space.

The first motivation for the use of weak Besov spaces (denoted $\mathcal{W}^*(r,p)$ in the following) to obtain thresholding rules is provided by their definition: To decide whether f belongs to $\mathcal{W}^*(r,p)$, we introduce at each level j the number of its wavelet coefficients greater than a threshold λ, denoted $N_f(j,\lambda)$. We require a power-law bound $C(f) \times \lambda^{-p}$ on the sum over j of the $N_f(j,\lambda)$ penalized by a weight depending on r (see Definition 2.1). We note that if $p < r$, $\mathcal{W}^*(r,p)$ is very close to $\mathcal{B}_{s,p,p}$ ($s = \frac{r}{2p} - \frac{1}{2}$), a member of the class of the strong Besov spaces $\mathcal{B}_{s,p,q}$ often considered by the statisticians. We have the natural inclusion $\mathcal{B}_{s,p,p} \subset \mathcal{W}^*(r,p)$.

Furthermore, weak Besov spaces appeared in statistics to evaluate the performance of classical estimation procedures. Cohen, DeVore, Kerkyacharian and Picard (2000) and Kerkyacharian and Picard (2000) wondered what is the maximal space over which a procedure attains a prescribed rate of convergence. Kerkyacharian and Picard (2000) roughly proved that if a procedure verifies an oracle inequality then its maxiset contains a weak Besov space. Wavelet thresholding is an example of such a procedure. In section 2.2, Proposition 2.1 gives a concrete example of the nature of maxisets associated with wavelet thresholding rules. This result provides the natural relationship between weak Besov spaces and wavelet thresholding rules.

Finally, we note that $\mathcal{W}^*(r,p)$ appears as a generalization of the weak l_p space (denoted wl_p in the following), often considered in approxima-

tion theory (see Johnstone (1994), Donoho (1996), Donoho and Johnstone (1996), or Cohen, DeVore and Hochmuth (2000)). Abramovich, Benjamini, Donoho and Johnstone (2000) used weak l_p spaces to define more precisely the notion of sparsity we mentioned previously. For them, sparsity means that there is a relatively small proportion of relatively large coefficients and they introduce a weak l_p constraint to control this proportion. So, as we shall see in section 2.2, weak Besov spaces may appear as natural spaces to capture signals in function of their regularity properties and their sparsity. We shall discuss the roles of the parameters r and p in this framework.
But another way to capture the sparsity of a signal is throughout the use of a bayesian model. Most of the authors consider bayesian models based upon gaussian distributions: For instance, both Clyde, Parmigiani and Vidakovic (1998) and Abramovich, Sapatinas and Silverman (1998) consider a mixture of a normal component and a point mass at zero for the wavelet coefficients. Chipman, Kolaczyk and McCulloch (1997) impose a mixture of two gaussian distributions with different variances for negligible and non negligible wavelet coefficients. Huang and Cressie (2000) assume the underlying signal to be composed of a piecewise-smooth deterministic part plus a zero-mean gaussian part.
As for us, rather than adopting the gaussian point of view, we consider priors based upon Pareto distributions. The reasons of this choice are the following: Rivoirard (2001) investigated the problem of estimation over weak Besov balls, by using the Bayes method. This approach enables us to exhibit a least favorable prior $\pi^{r,p,C}$ for the wavelet coefficients β_{jk} of a function f lying in the weak Besov ball $\mathcal{W}^*(r,p)(C)$. It takes the following form: The β_{jk}'s are independent and the distribution of each β_{jk} is built from the distribution of $\tilde{\alpha}_j X_{jk}$, where X_{jk} is a Pareto(p) variable and $\tilde{\alpha}_j$ is a level-dependent dilation parameter (see section 2.3). The realizations built from $\pi^{r,p,C}$ can be viewed as the worst functions to be estimated and lying in $\mathcal{W}^*(r,p)(C)$. What is more, $\pi^{r,p,C}$ is typical of the ball $\mathcal{W}^*(r,p)(C)$: For instance, we can prove that it cannot be a least favorable prior for the problem of estimation over $\mathcal{B}_{s,p,p}(C)$ ($s = \frac{r}{2p} - \frac{1}{2}$), the strong Besov ball naturally associated with $\mathcal{W}^*(r,p)(C)$.

1.2 Outline

In this paper, we assume we are given a prior model directly inspired by the least favorable priors $\pi^{r,p,C}$ of the weak Besov balls $\mathcal{W}^*(r,p)(C)$: We

suppose that the wavelet coefficients β_{jk} of f are independent, each β_{jk} has a symmetric distribution and $|\beta_{jk}| \sim \min(\alpha_j X_{jk} - \alpha_j, \mu_j)$, where X_{jk} is a Pareto(p) variable, α_j and μ_j are level-dependent parameters precisely defined in section 3.1. In section 3.2, we establish a relationship between these parameters and the parameters of the weak Besov space in which the function f lies (see Theorem 3.1). We present various realizations that give insight into the meaning of the weak Besov space parameters. In particular, we note that if f is typical of $\mathcal{W}^*(r,p)$, then the regularity of f increases with r. When p is small, f presents very high peaks with a regular behavior between them. When p is great, the peaks are less high and between them, the behavior is less regular.
The rest of the paper is devoted to the construction of thresholding rules. We consider the model (1.1) and we translate it into the wavelet domain by using the discrete wavelet transform. We assume that the discrete wavelet coefficients of f are provided by the prior model defined in section 4.2 and roughly described previously. To estimate each discrete wavelet coefficient, we use the soft thresholding rule. To choose the threshold, we take into account a result proved by Rivoirard (2001). In this paper, a minimax thresholding rule is exhibited. At large resolution levels j, the threshold $\tilde{\lambda}_j$ is proportional to $\sqrt{-2\log(\tilde{\alpha}_j^p)}$, where $\tilde{\alpha}_j$ is the dilation parameter that appears in the definition of the least favorable prior $\pi^{r,p,C}$ associated with $\mathcal{W}^*(r,p)(C)$ (see section 2.3). This minimax point of view suggests to use:

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

In section 4.2, we give more precise justifications for this choice. We propose a method to estimate the parameters appearing in the definition of λ_j. In section 4.3, we measure the performances of the resulting procedure, called ParetoThresh, by using the four test signals: 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'. It is compared to the non bayesian procedures VisuShrink and SureShrink we have described previously. For this, we use the mean-squared error. Under this criterion, Table 3 shows that for the estimation of 'Blocks', 'Bumps' and 'Doppler', ParetoThresh substantially improves VisuShrink and SureShrink. But this is not true any more for 'Heavisine'. Taking into account the properties of 'Heavisine', we explain why, to some extent, this result could have been expected.

1.3 Contents

The paper is organized as follows: Section 2 is devoted to weak Besov spaces. After giving their definition, we recall the results obtained by Rivoirard (2001) for the problem of estimation over weak Besov balls. In section 3, we define a bayesian model and we investigate the conditions for the resulting functions to belong to weak Besov spaces. Section 4 is devoted to the construction of ParetoThresh, the data adaptive procedure we propose. Finally, in section 5, we give the proof of Theorem 3.1.

2 Estimation over weak Besov spaces

After an overview of wavelet bases (see section 2.1), we introduce in section 2.2 weak Besov spaces. Finally, in section 2.3, we recall some relevant statistical aspects of weak Besov spaces. From now on, we note $X \sim \mathcal{P}(p)$ $(p > 0)$ to mean that X is a Pareto(p) variable, i.e. X has the density $g(t) = \mathbf{1}_{t \geq 1} p t^{-p-1}$.

2.1 Wavelet series representation

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

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

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

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

Actually, for all $m \in \mathbb{N}$, we can build the functions ϕ and ψ to be of 'regularity m': ϕ and ψ are of class C^m, each of them and their derivatives

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

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

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

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

where C_1 and C_2 are constants not depending on f.
Often, for practical reasons, the functions considered in the literature are only defined on a compact set, the interval $[0,1]$ for instance. Cohen, Daubechies and Vial (1993) have described the necessary corrections to adapt wavelets to a bounded interval. As for us, in sections 3 and 4, we focus on periodic functions f with unit period, and we work with periodic wavelets, we still note ψ_{jk}. This modification, described by Daubechies (1992), implies that the wavelet coefficients are restricted to the indices $\{j \geq -1, k \in \mathcal{I}_j\}$, where

$$\mathcal{I}_j = \{k \in \mathbb{N} : \quad 0 \leq k < 2^j\}. \tag{2.1}$$

2.2 Sparsity and definition of weak Besov spaces

Abramovich, Benjamini, Donoho and Johnstone (2000) introduced the notion of sparsity of an infinite vector $\theta \in \mathbb{R}^{\mathbb{N}}$ through the following approach: The vector θ is said to be sparse if there is a small proportion of relatively

large entries. Therefore, they order the components of θ according to their size:

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

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

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

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

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

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

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

We define weak Besov spaces as follows:

Definition 2.1. For all $j \geq -1$ and $\lambda > 0$, we consider $N_f(j, \lambda)$ the number of the wavelet coefficients of f at level j greater than λ:

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

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

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

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

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

The weak Besov space $\mathcal{W}^*(r,p)$ can be viewed as a weighted weak l_p space. The weights penalize the counting of the β_{jk}'s greater than λ for the large scales according to the sign of $r-2$. Therefore, the use of weak Besov spaces may appear as a good device to measure the sparsity of a wavelet expanded signal. Using Markov's inequality, it is easy to prove that for $p<r$, the strong Besov space $\mathcal{B}_{\frac{r}{2p}-\frac{1}{2},p,p}$ is included into $\mathcal{W}^*(r,p)$.
Finally, we recall that Cohen, DeVore, Kerkyacharian and Picard (2000) introduced weak Besov spaces to characterize maxisets for the wavelet thresholding procedure. Among others, they proved the following result:

Proposition 2.1. *Let $1<r<\infty$ and $\alpha\in(0,1)$. We suppose that $f\in L_r([0,1])$. Under the white noise model*

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

we consider the following thresholding estimator

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

with

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

We have

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

2.3 Minimax risk and least favorable priors

Rivoirard (2001) evaluated the minimax risk over weak Besov balls $\mathcal{W}^*(r,p)(C)$ for $\mathcal{B}_{s',p',p'}$ norms by using a bayesian approach. This enables us to exhibit least favorable priors (noted LFP) which will inspire the prior model chosen in section 3.1. Let us recall here the main results we obtain: We restrict

our attention to functions f supported by the interval $[0,1]$. They can be written:

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

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

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

We consider the white noise model

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

which means that $\varepsilon > 0$ is known, $f \in L_2([0,1])$ is unknown and for all $\phi \in L_2([0,1])$, $\int_{[0,1]} \phi(t) dY_t = \int_{[0,1]} \phi(t) f(t) dt + \varepsilon \int_{[0,1]} \phi(t) dW_t$ is observable. Among non parametric situations, this statistical model is one of the simplest, at least technically. What is more, it arises as an appropriate large sample limit for more general non parametric models, such as the regression model (1.1) considered previously. (cf. Brown and Low (1996)). These are the reasons why this model is often considered in non parametric situations. We study the asymptotic behavior of the minimax risk

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

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

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

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

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

Under suitable conditions, Theorem 1 of Rivoirard (2001) proves that on $\mathcal{R}$, the rate of convergence of R_ε is $\varepsilon^{p'\alpha}$, where $\alpha = (s - s')/(s + \frac{1}{2})$ and

$s = \frac{r}{2p} - \frac{1}{2}$. On $\mathcal{C}$, the rate of convergence is $\varepsilon^{p'\alpha} \log(\varepsilon^{-1})^{\frac{p'\alpha}{2}}$.
To identify LFP, we first reduce to $M_{r,p}(C)$, the natural set of probability measures associated to $\mathcal{W}^*(r,p)(C)$ and defined as follows:

$$M_{r,p}(C) = \left\{ \pi(d\beta) : \quad \sum_{j=-1}^{\infty} 2^{j(\frac{r}{2}-1)} \mathbb{E}_\pi \sum_{k \in \tilde{\mathcal{I}}_j} \mathbb{1}_{|\beta_{jk}|>\lambda} \leq \left(\frac{C}{\lambda}\right)^p, \quad \forall \lambda > 0 \right\}. \tag{2.2}$$

For the definition of the LFP, Rivoirard (2001) exhibited two sequences of real numbers $(\tilde{\alpha}_j)_{j \geq -1}$ and $(\tilde{\mu}_j)_{j \geq -1}$ depending on the zone. We do not recall their exact definitions, which would just add useless technical aspects here, but we briefly recall their properties:

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

$$\sum_{j=j_1}^{+\infty} 2^{j\frac{r}{2}} \tilde{\alpha}_j^p = \left(\frac{C}{\varepsilon}\right)^p, \tag{2.3}$$

  where j_1 is an integer depending on the zone. For instance, on $\mathcal{R}$, j_1 is the first integer j such that $\tilde{\alpha}_j < 1$.

- $(\tilde{\mu}_j)_{j \geq -1}$ is an increasing sequence of positive real numbers such that $\tilde{\mu}_j \overset{j \to \infty}{\sim} \sqrt{-2\log(\tilde{\alpha}_j^p)}$.

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

- The distribution of β_{jk} (denoted $\tilde{F}_j$) is symmetric about 0,
- $|\beta_{jk}| = \begin{cases} \varepsilon \min(\tilde{\alpha}_j X_{jk}, \tilde{\mu}_j), & \text{where } X_{jk} \sim \mathcal{P}(p), \quad \text{if } j \geq j_1 \\ 0 & \text{otherwise,} \end{cases}$

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

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

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

$$K_1\, B(\pi_\varepsilon^{r,p,C}) \leq R_\varepsilon \leq K_2\, B(\pi_\varepsilon^{r,p,C}), \tag{2.4}$$

and

$$\sup_{\pi \in M_{r,p}(C)} B(\pi) \leq K_3\, B(\pi_\varepsilon^{r,p,C}). \tag{2.5}$$

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

$$\pi_\varepsilon^{r,p,(\gamma_\varepsilon^{-1}C)}\, (\beta \in \mathcal{W}^*(r,p)(C)) \stackrel{\varepsilon\to 0}{\longrightarrow} 1. \tag{2.6}$$

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

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

with

$$\tilde{\lambda}_j = \begin{cases} \varepsilon\sqrt{-2\log(\tilde{\alpha}_j^p)} & \text{if } j \geq j_1, \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

attains the minimax rate of convergence up to constants. We shall inspire from this minimax rule to build constructive estimators in section 4.2.

3 Construction of functions typical of weak Besov spaces

From now on, following section 2.1, we consider a periodic signal

$$f(t) = \sum_{j=-1}^{\infty} \sum_{k \in \mathcal{I}_j} \beta_{jk} \psi_{jk}(t), \tag{3.1}$$

where $\mathcal{I}_j$ is given in (2.1). As pointed out by Johnstone (1994), this computational simplification affects only a fixed number of wavelet coefficients at each level j. We place a prior model on the wavelet coefficients of f to capture its sparsity. But section 2.2 pointed out that the sparsity of a signal can be revealed by the weak Besov space $\mathcal{W}^*(r,p)$ in which the signal lies. So, through Theorem 3.1, we connect these two approaches of sparsity by giving a relationship between the parameters of the prior model and $\mathcal{W}^*(r,p)$. This enables us to build functions typical of $\mathcal{W}^*(r,p)$.

3.1 The prior model

To fix a prior model, we exploit the LFP $\pi_\varepsilon^{r,p,C}$ defined in section 2.3 and that is naturally connected to the weak Besov ball $\mathcal{W}^*(r,p)(C)$: We suppose that the β_{jk}'s are independent and for $j \geq 0$ and $k \in \mathcal{I}_j$, the distribution of each β_{jk} is $F_j^{\alpha_j,\mu_j,p}$, where

- $F_j^{\alpha_j,\mu_j,p} = \frac{1}{2}(F_j^+ + F_j^-)$,
- F_j^- is the reflection of F_j^+ about 0,
- F_j^+ is the distribution of $\min(\alpha_j X_j - \alpha_j, \mu_j)$, where $X_j \sim \mathcal{P}(p)$,
- α_j and μ_j are positive real numbers.

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

$$\alpha_j = C 2^{-j\delta}$$

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

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

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

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

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

3.2 The main result and simulations

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

Theorem 3.1. *We consider the function f given in (3.1). Let $0 < p < \infty$, $0 < r < \infty$. Given three positive real numbers δ, C and M, we define for all $j \geq 0$, $\alpha_j = C2^{-j\delta}$, and $\mu_j = \sqrt{\max\left(M^2, -2\log(\alpha_j^p)\right)}$. Let us assume that the wavelet coefficients β_{jk} of f are independent and for $j \geq 0$ and $k \in \mathcal{I}_j$, β_{jk} has the distribution $F_j^{\alpha_j,\mu_j,p}$ given in section 3.1 . For all fixed value of β_{-10},*

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

The proof of this theorem is given in section 5.

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

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

This can be connected to the condition (2.3) verified by the sequences $(\tilde{\alpha}_j)_{j\geq j_1}$ used to define the LFP for the weak Besov balls $\mathcal{W}^*(r,p)(C)$.

Theorem 3.1 gives us a help to have a good understanding of weak Besov spaces. Figure 1 presents various typical realizations with different values for the parameters δ and p. Since the values of C and M do not play a role in the shape of the realizations we get, we set $C = 0.1$ and $M = 2$ for each realization. We used Daubechies's least asymmetric wavelet of order 8.

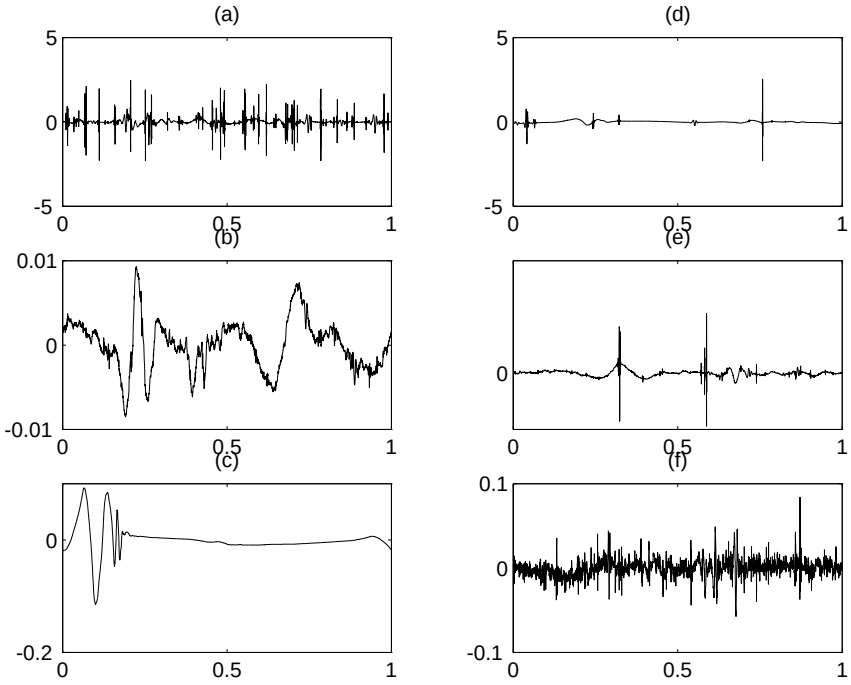

Figure 1: *Realizations with various values of δ and p; $\beta_{-10} = 0$; n=4096 plotting points; (a): $p = 0.5$, $\delta = 1$. (b): $p = 2$, $\delta = 1$. (c): $p = 1$, $\delta = 2$. (d): $p = 0.5$, $\delta = 2$. (e): $p = 1$, $\delta = 1$. (f): $p = 2$, $\delta = 0.5$.*

Naturally, we note that when p is fixed, the realizations are more regular when δ is great (compare (b) and (f) or (c) and (e) or (a) and (d)). The same conclusion is true when δ is fixed and p is great (compare (a), (b) and (e) or (c) and (d)). In fact, as expected, the realizations are smoother when the product δp is great. It is interesting to wonder what happens when the product δp is fixed, and when we take different values for δ and p. Comparing (d), (e) and (f) or (b) and (c), we notice that when p is small, the realizations show very high peaks with a regular behavior between

the peaks. When p is great, the peaks are less high, and between the peaks the behavior is less homogenous. To sum up, we can say that when p decreases, the number of negligible coefficients increases, but the few remaining coefficients may be very large. Rivoirard (2001) drew the same conclusions by using a different approach that exploits the LFP associated with the weak Besov balls $\mathcal{W}^*(r,p)(C)$. Note that these two approaches complement one another:

- The first one produces typical functions of weak Besov spaces but we do not control the radius of the weak Besov ball that contains a function f provided by our prior model. This radius may be very great, which may pollute our perception of the regularity of the function f.
- The second one controls the radius of the weak Besov balls, but each LFP is typical of a set of probability measures. For instance, in section 2.3, we chose $M_{r,p}(C)$ defined by (2.2). To some extent, this choice was judicious since we obtained the properties (2.4), (2.5) and (2.6), but we could have made another choice.

4 Thresholding rules

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

4.1 Model and discrete wavelet transform

Let us consider the standard regression problem:

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

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

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

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

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

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

$$d_{jk} \approx \beta_{jk} \times \sqrt{n}. \tag{4.2}$$

Using the DWT, the regression model (4.1) is reduced to the following one:

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

where

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

and

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

Since $\mathcal{W}$ is orthogonal, z is a vector of independent $\mathcal{N}(0,1)$ variables. Now, instead of estimating f, we estimate the d_{jk}'s.
We suppose in the following that σ is known. Nevertheless, it could robustly be estimated by the median absolute deviation of the $(d_{N-1,k})_{k \in \mathcal{I}_{N-1}}$ divided by 0.6745 (see Donoho and Johnstone (1994)).

4.2 Choice of the threshold

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

$$\eta^{HT}(y,\lambda) = y\mathbb{1}_{|y|>\lambda},$$

$$\eta^{ST}(y,\lambda) = \text{sign}(y)\,(|y|-\lambda)_+ .$$

This paper considers soft thresholding although hard thresholding is a possible alternative. However, the soft thresholding rule is smoother ($y \longrightarrow \eta^{ST}(y,\lambda)$ is continuous) and it makes it possible to build minimax estimators over weak Besov spaces, as recalled in section 2.3. The choice of the threshold is then crucial. If λ_j is too small (respectively too large) then the estimator tends to overfit (respectively underfit) the data. Let us describe two non bayesian procedures that have minimax properties:

Donoho and Johnstone (1994) proposed their VisuShrink procedure with $\lambda_j = \lambda^u := \sigma\sqrt{2\log(n)}$. If λ^u often underfits the data, it 'guarantees' a noise-free reconstruction since

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

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

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

Otherwise, we choose the threshold $\lambda_j = \sigma\sqrt{2\log(2^j)}$.
Under a Bayes model, a natural approach to build estimators could be to use the mean of the posterior distribution which is the Bayes rule under the squared error loss. But the posterior mean does not involve in general thresholding rules. That is the reason why Abramovich, Sapatinas and Silverman (1998) focus on the posterior median within the following framework: They consider a prior model having the following form:

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

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

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

He constructs a procedure that mimics the hard thresholding rule. He estimates d_{jk} by $y_{jk}\mathbb{1}_{\eta_{jk}<1}$ where $\eta_{jk} = \mathbb{P}(d_{jk} = 0 \mid y_{jk})/\mathbb{P}(d_{jk} \neq 0 \mid y_{jk})$.
As for us, we place the following prior on the discrete wavelet coefficients: We suppose that the d_{jk}'s are independent and for all $j \geq 0$, $k \in \mathcal{I}_j$, $d_{jk} \sim F_j^{\alpha_j,\mu_j,p}$, where $F_j^{\alpha_j,\mu_j,p}$ is given in section 3.1. The parameters α_j and μ_j are given by $\alpha_j = C2^{-j\delta}$ and $\mu_j = \sqrt{\max\left(M^2, -2\log(\alpha_j^p)\right)}$, where δ, C and M are positive constants. To estimate d_{jk}, we propose

$$d_{jk}^* = \eta^{ST}(y_{jk}, \lambda_j),$$

where λ_j has the following form

$$\lambda_j = \begin{cases} \sigma\sqrt{-2\log(\alpha_j^p)} & \text{if } j \geq j_1, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

as suggested by (2.7). Indeed, since our prior model is very close to the LFP over weak Besov balls $\mathcal{W}^*(r,p)(C)$, using (4.3) to define λ_j where j_1 is the first integer j such that $\alpha_j < 1$ seems judicious. Then, the threshold λ_j can be rewritten as follows:

$$\lambda_j = \sigma\sqrt{\max\left(0, -2\log(\alpha_j^p)\right)}. \tag{4.4}$$

To apply this procedure, it is necessary to specify the values of C and δ that define α_j and the value of p. If we know the weak Besov space in which the function to be estimated lies and if an efficient method is provided to estimate δ or p, by using Theorem 3.1, it is easy to estimate the other parameter. However, we shall ignore this strategy and the choice of the value of p will be made in section 4.3. For the estimation of (C,δ), we set

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

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

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

We estimate C and δ by using the linear regression:

$$(\hat{C},\hat{\delta}) = \arg\min_{C,\delta}\sum_{j\in\mathcal{S}}(\log(\hat{\alpha}_j) - \log(C) + j\delta\log(2))^2, \tag{4.5}$$

where

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

But $(\hat{C},\hat{\delta})$ are well defined only if $\mathrm{card}(\mathcal{S}) \geq 2$. So, when $\mathrm{card}(\mathcal{S}) \geq 2$, we set

$$\hat{\lambda}_j = \sigma\sqrt{\max\left(0, -2p\log(\hat{C}2^{-j\hat{\delta}})\right)}. \tag{4.6}$$

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

$$\hat{\lambda}_j = \begin{cases} 0 & \text{if } \hat{\alpha}_j = +\infty, \\ \sigma\sqrt{\max\left(0, -2p\log(\hat{\alpha}_j)\right)} & \text{for } j \in \mathcal{S}, \\ \lambda^u & \text{if } \hat{\alpha}_j = 0. \end{cases} \tag{4.7}$$

Before going further, let us give a precise justification for this procedure: 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,

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

Therefore, for all $\lambda > 0$, $\left(\frac{\alpha_j}{\alpha_j+\lambda}\right)^p$ 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 $\hat{N}_j(\lambda^u)$. Using the four test functions ('Blocks', 'Bumps', 'Heavisine', and 'Doppler'), Table 1 compares for $j \in \{0, \ldots, 9\}$ the values of $N_j(\lambda^u)$ and the average over 100 replications of the values of $\hat{N}_j(\lambda^u)$. It shows that our approximation is acceptable.
So,

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

and

$$\alpha_j = C2^{-j\delta} \approx \hat{\alpha}_j = \lambda^u \hat{N}_j(\lambda^u)^{\frac{1}{p}}(1 - \hat{N}_j(\lambda^u)^{\frac{1}{p}})^{-1}.$$

This provides a justification for (4.5). The pair of equations (4.6) and (4.7) are naturally justified by (4.4).
Now, we set

$$\hat{d}_{jk} = \eta^{ST}(y_{jk}, \hat{\lambda}_j),$$

for all $j \geq 0$, $k \in \mathcal{I}_j$, and $\hat{d}_{-10} = y_{-10}$. Finally, to estimate the signal, we use the reconstruction formula and an estimator of f^0 is provided by:

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

Level	Blocks		Bumps		Heavisine		Doppler	
j	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$	N_j	$\hat{N}_j$
j=0	1	1	1	0.76	1	1	1	0.99
j=1	2	2	1	1	2	2	2	2
j=2	4	4	3	3	4	4	3	3
j=3	6	6.02	4	4.67	6	5.44	7	6.62
j=4	9	9.50	10	9.71	0	0.29	6	6.06
j=5	8	9.07	10	11.03	0	0.06	7	6.61
j=6	6	6.45	16	16.16	0	0.11	7	7.15
j=7	5	4.87	16	14.74	0	0.02	5	4.17
j=8	4	3.34	7	7.80	0	0.05	0	0.39
j=9	0	0.60	2	2.79	0	0.13	0	0.15

Table 1: Comparison of the values of $N_j = N_j(\lambda^u)$ and $\hat{N}_j = \hat{N}_j(\lambda^u)$ for 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'; n=1024; rsnr=3 ($\sigma = 7/3$).

The performances of this bayesian thresholding procedure denoted from now on as ParetoThresh, are analyzed in the next section. Table 2 gives the average over 100 replications of the values of the level-dependent threshold $\hat{\lambda}_j$ associated with the four test functions.
Donoho and Johnstone (1994) noted that for the coarsest levels the coefficients should not be shrunk to 0. Huang and Cressie (2000) and Abramovich and Benjamini (1995) showed that the choice of these levels is essential for VisuShrink and SureShrink. Let us note that ParetoThresh automatically provides the levels where the coefficients are not shrunk, and this, with a data adaptive method.

4.3 Examples and discussion

In this section, we apply our ParetoThresh procedure to one-dimensional signal processing. We use the four test functions: 'Blocks', 'Bumps', 'Heavisine' and 'Doppler'. These functions have been chosen by Donoho and Johnstone (1994) to represent a large variety of inhomogeneous signals. More precisely, our procedure deals with the 1024 equally spaced values on $[0, 1]$ of these signals. In the subsequent applications of ParetoThresh, we take $p = 1$ for every function, which provides quite good results. However, we shall discuss below the effect of varying p. We compare our procedure

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

to VisuShrink and SureShrink, described in section 4.2 for which we do not threshold the five coarsest levels. Daubechies's least asymmetric wavelet of order 8 is used for all the methods. The performance of each procedure is measured by using the mean-squared error associated to an estimator $\hat{f}$:

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

Table 3 shows the average mean-squared error (denoted AMSE) using 100 replications for VisuShrink, SureShrink and ParetoThresh with different values for the root signal to noise ratio (RSNR).
Figures 2, 3, 4 and 5 show the reconstructions we obtain for these three methods when the RSNR is equal to 3.

Table 3 shows that ParetoThresh has generally smaller mean-squared error over the three test functions 'Blocks', 'Bumps', and 'Doppler', than SureShrink second and VisuShrink third in the rankings. For 'Heavisine', we recall that the estimation of the level-dependent thresholds for ParetoThresh is based upon a small number of data passing the threshold λ^u (see Table 1). Furthermore, to some extent, by its regularity properties, 'Heavisine' can not be viewed as belonging to the class of 'the worst functions to be estimated'. Consequently, the prior model we adopt does not seem to be a good model for this signal. If ParetoThresh behaves well under

RSNR	Signal	VisuShrink	SureShrink	ParetoThresh ($p = 1$)
RSNR=3	Blocks	3.3143	1.7850	*1.4406*
	Bumps	5.6100	2.0378	*1.8082*
	Heavisine	0.3136	*0.3042*	0.3136
	Doppler	2.1588	1.0911	*0.9508*
RSNR=5	Blocks	1.8624	0.7645	*0.6963*
	Bumps	2.7345	0.8523	*0.7224*
	Heavisine	0.1946	*0.1816*	0.1843
	Doppler	1.0358	0.4378	*0.4317*
RSNR=8	Blocks	0.9745	0.3449	*0.3207*
	Bumps	1.3139	*0.3032*	0.3266
	Heavisine	0.1312	0.1028	*0.0855*
	Doppler	0.5374	0.2434	*0.2329*

Table 3: AMSEs for VisuShrink, SureShrink and ParetoThresh ($p = 1$) with various test functions and various values of the RSNR.

the AMSE approach, we note that high-frequency artefacts appear, whereas VisuShrink provides the best method for removing the noise. SureShrink lies in between. But these artefacts may partially disappear if we take small values of p as illustrated by Figure 6. This effect may be expected taking into account the conclusions we have drawn from the realizations of section 3.2. 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 greater than 1, the AMSEs are worse and the artefacts are more numerous. Finally, let us mention that a possible alternative is to use the hard thresholding rule with $(\hat{\lambda}_j)_j$. However, the resulting constructions are less regular. Another alternative is to use a Bayes rule. Let us note that it is easy to implement the hard and soft thresholding rules. This is not necessarily the case for a Bayes rule since it results from the minimization of the Bayes risk, even if under a good choice of the loss function we can exhibit the explicit form of the Bayes rule (for instance, the posterior mean, or the posterior median). We can add that neither the use of the hard thresholding rule with $(\hat{\lambda}_j)_j$, nor the use of a Bayes rule (the posterior mean or the posterior median) provides better results as far as the AMSE is concerned.

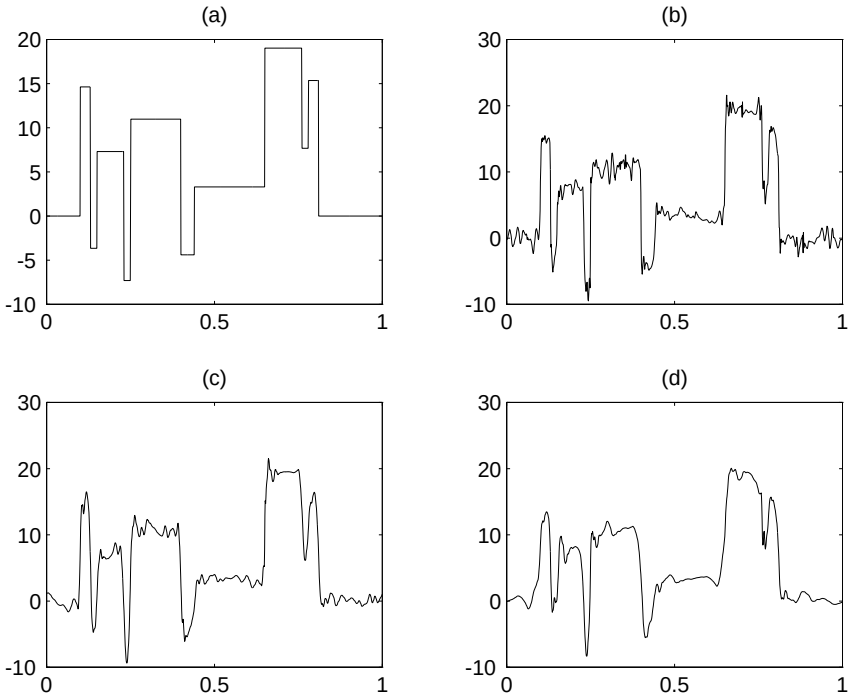

Figure 2: Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink; (a): 'Blocks' (b): ParetoThresh ($p = 1$) (c): SureShrink (d): VisuShrink

5 Appendix : Proof of Theorem 3.1

Proof of necessity: Let us assume that $f \in \mathcal{W}^*(r,p)$ a.e. For all $\lambda \in]0, \mu_0[$, we consider,

$$\forall\, n \in \mathbb{N}, \quad U_n(\lambda) = \frac{1}{\sqrt{c_n}} \left| \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbf{1}_{|\beta_{jk}|>\lambda} - \mathbb{P}(|\beta_{jk}| > \lambda) \right) \right|,$$

where

$$c_n = \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(r-2)}.$$

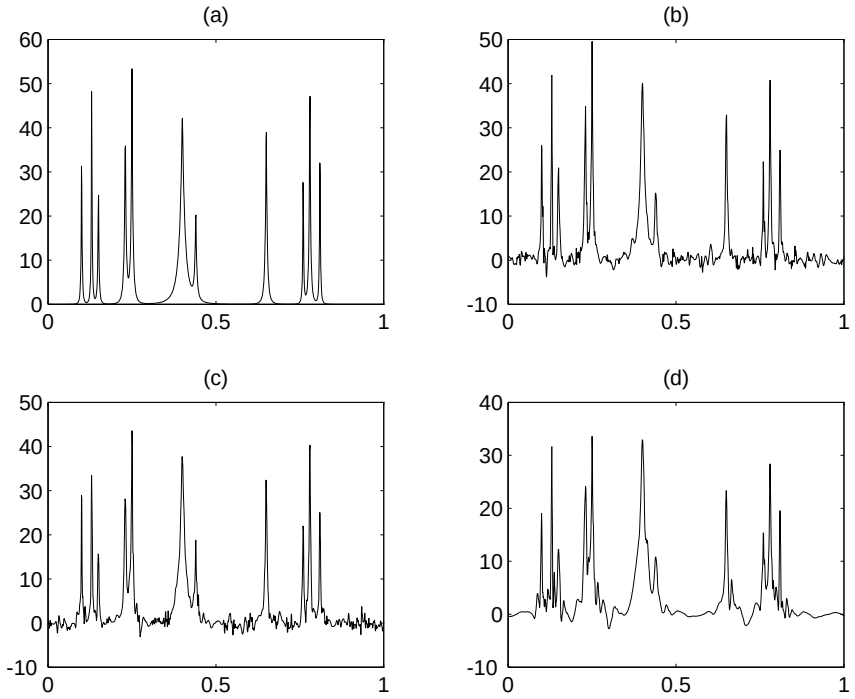

Figure 3: *Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink; (a): 'Bumps' (b): ParetoThresh* ($p = 1$) *(c): SureShrink (d): VisuShrink*

Since $\lambda < \mu_0$,

$$U_n(\lambda) = \frac{1}{\sqrt{c_n}} \left| \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{X^*_{jk} \leq \lambda} - F^*_j(\lambda) \right) \right|,$$

where $X^*_{jk} = \alpha_j X_{jk} - \alpha_j$, F^*_j is its continuous distribution function and $X_{jk} \sim \mathcal{P}(p)$. As in Shorack and Wellner (1986) (p 117), we set

$$\overline{F}_n = \frac{1}{c_n} \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(r-2)} F^*_j,$$

$$T_{jk} = \overline{F}_n(X^*_{jk}),$$

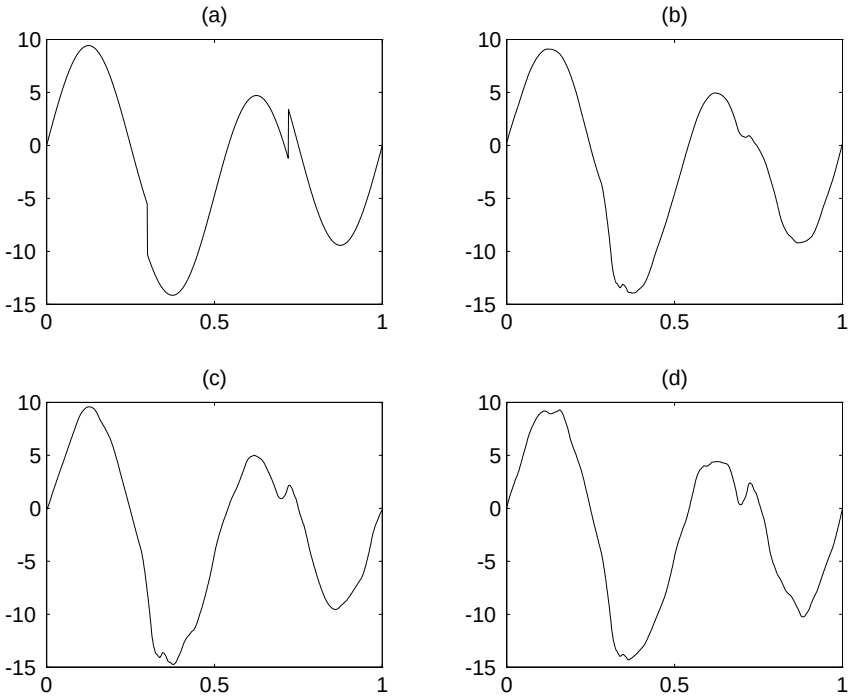

Figure 4: Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink; (a): 'Heavisine' (b): ParetoThresh ($p = 1$) (c): SureShrink (d): VisuShrink

and we consider the weighted empirical process of the T_{jk}'s,

$$Z_n(\lambda) = \frac{1}{\sqrt{c_n}} \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{T_{jk} \leq \lambda} - G_j(\lambda) \right),$$

where $G_j = F_j^* \circ \overline{F}_n^{-1}$ is the distribution function of T_{jk}. So $U_n(\lambda)$ can be written

$$U_n(\lambda) = \left| Z_n(\overline{F}_n(\lambda)) \right|.$$

Now, we suppose that

$$r - 1 - 2\delta p < 0, \tag{5.1}$$

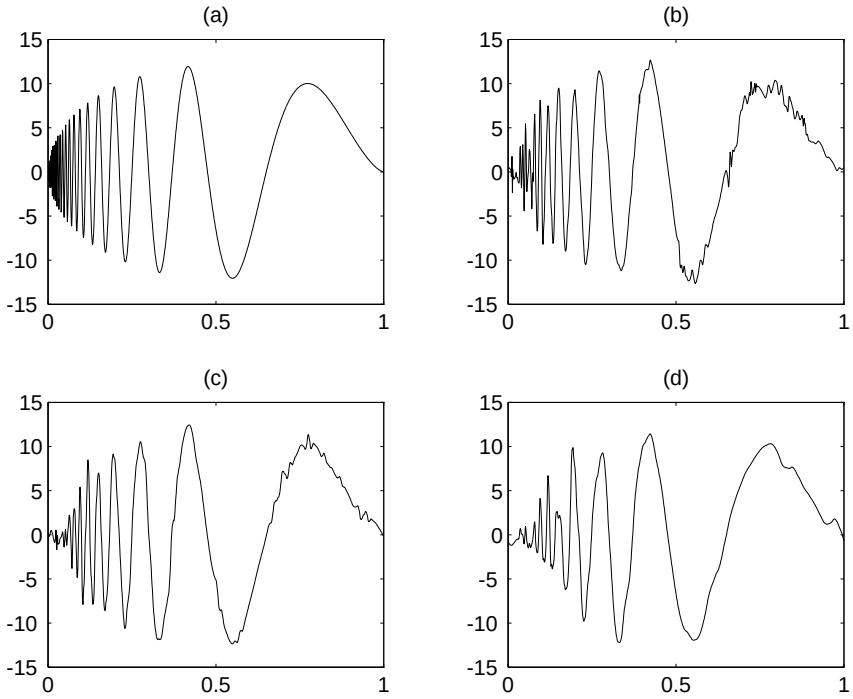

Figure 5: *Original test function and various reconstructions using ParetoThresh, SureShrink and VisuShrink; (a): 'Doppler' (b): ParetoThresh* ($p = 1$) *(c): SureShrink (d): VisuShrink*

and we consider δ' a real number smaller than δ such that $r - 1 - 2\delta' p < 0$. We consider the sequence $(\lambda_n)_{n \in \mathbb{N}}$ defined by

$$\forall\, n \in \mathbb{N}, \quad \lambda_n = \min(C\ 2^{-\delta' n}, \frac{\mu_0}{2}).$$

We have $\forall\, n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{P}\left(\lambda_n^p \sqrt{c_n} U_n(\lambda_n) \geq 1\right) &= \mathbb{P}\left(\left|Z_n(\overline{F}_n(\lambda_n))\right| \geq \lambda_n^{-p} c_n^{-\frac{1}{2}}\right) \\ &\leq \mathbb{P}\left(w_n(1) \geq \lambda_n^{-p} c_n^{-\frac{1}{2}}\right), \end{aligned}$$

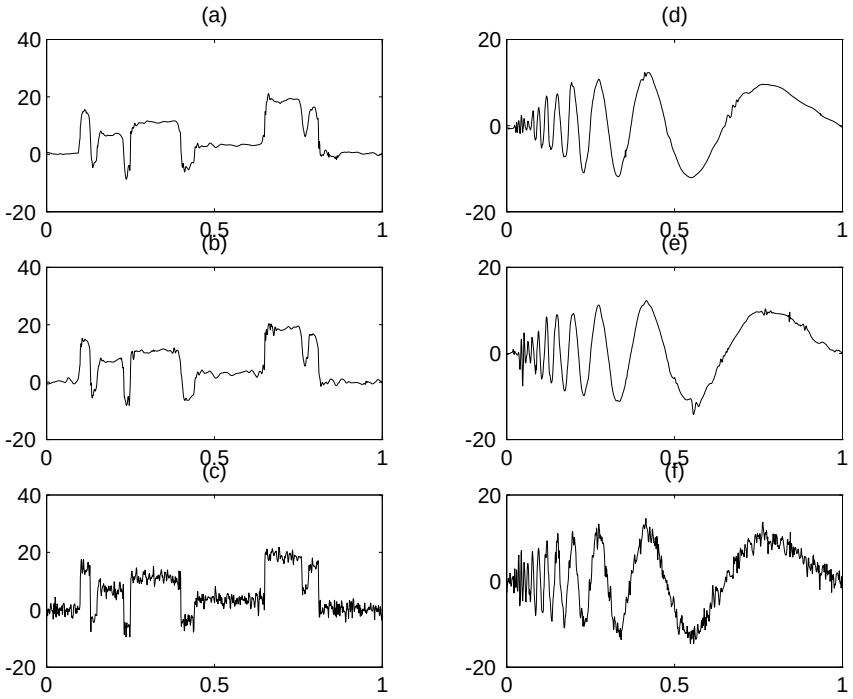

Figure 6: *Various reconstructions of 'Blocks' and 'Doppler' using ParetoThresh with different values of p; RSNR=3; (a): $p = 0.5$, AMSE=1.6340. (b): $p = 0.7$, AMSE=1.5420. (c): $p = 2$, AMSE=2.4473. (d): $p = 0.5$, AMSE=0.9646. (e): $p = 0.7$, AMSE=0.9186. (f): $p = 2$, AMSE=1.9665.*

where w_n is the modulus of continuity of Z_n. So, using Shorack and Wellner (1986) (p 119),

$$\mathbb{P}\left(\lambda_n^p \sqrt{c_n} U_n(\lambda_n) \geq 1\right) \leq K \left(\lambda_n^{-p} c_n^{-\frac{1}{2}}\right)^{-4},$$

where K is a universal constant. Since $r - 1 - 2\delta' p < 0$,

$$\sum_n \mathbb{P}\left(\lambda_n^p \sqrt{c_n} U_n(\lambda_n) \geq 1\right) < \infty,$$

and

$$\sup_n |\lambda_n^p \sqrt{c_n} U_n(\lambda_n)| < \infty \text{ a.e.},$$

which means that

$$\sup_n \left| \lambda_n^p \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \left(\mathbb{1}_{|\beta_{jk}|>\lambda_n} - \mathbb{P}(|\beta_{jk}| > \lambda_n) \right) \right| < \infty \text{ a.e.}$$

But, using the definition of $\mathcal{W}^*(r,p)$,

$$\begin{aligned} \sup_n \lambda_n^p \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda_n} &\leq \sup_{\lambda>0} \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda} \\ &< \infty \text{ a.e.} \end{aligned}$$

Therefore,

$$\begin{aligned} & \sup_n \lambda_n^p \sum_{j=0}^{n} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{P}(|\beta_{jk}| > \lambda_n) < \infty \\ \Rightarrow \quad & \sup_n \lambda_n^p \sum_{j=0}^{n} 2^{j\frac{r}{2}} \left(\frac{\alpha_j}{\lambda_n + \alpha_j} \right)^p < \infty \\ \Rightarrow \quad & \sup_n \sum_{j=\lfloor \frac{n\delta'}{\delta} \rfloor + 1}^{n} 2^{j\frac{r}{2}} \alpha_j^p \left(1 + \alpha_j \lambda_n^{-1} \right)^{-p} < \infty. \end{aligned}$$

But for all $j \geq \frac{n\delta'}{\delta}$, $\alpha_j \lambda_n^{-1} \leq 1$. Consequently, the last inequality implies that

$$\sup_n \sum_{j=\lfloor \frac{n\delta'}{\delta} \rfloor + 1}^{n} 2^{j\frac{r}{2}} \alpha_j^p < \infty,$$

which means that $\frac{r}{2} < \delta p$.
Using the embedding $\mathcal{W}^*(r,p) \subset \mathcal{W}^*(r',p)$, when $r' < r$, we can omit the hypothesis (5.1) and assert that if f is in $\mathcal{W}^*(r,p)$ almost everywhere then $\frac{r}{2} < \delta p$.

□

Proof of sufficiency: Here K denotes a constant, eventually depending on r, p, δ, C and that may be different at each line. We suppose that $\frac{r}{2} < \delta p$ and we consider for all $\lambda > 0$,

$$Y_\lambda = \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda}.$$

We are going to prove that

$$\sup_{\lambda>0} Y_\lambda < \infty \text{ a.e.} \tag{5.2}$$

First, we note that using the three-series theorem, the condition $\frac{r}{2} < \delta p$ implies $Y_\lambda < \infty$ a.e. for all $\lambda > 0$.
Let us begin by studying Y_λ when λ tends to 0. For all $n \in \mathbb{N}$, we set $Y^n = Y_{\lambda_n}$, where $\lambda_n = 2^{-n}$. We consider $j_0(n)$ the greatest integer j such that $\alpha_j \geq \lambda_n$.

$$\begin{aligned}
\mathbb{E}(Y^n) &= \lambda_n^p \sum_{j=0}^{\infty} 2^{j\frac{r}{2}} \left(\frac{\alpha_j}{\lambda_n + \alpha_j}\right)^p \\
&\leq \lambda_n^p \sum_{j=0}^{j_0-1} 2^{j\frac{r}{2}} + \sum_{j=j_0}^{\infty} 2^{j\frac{r}{2}} \alpha_j^p \\
&\leq K\ \lambda_n^p\ 2^{j_0\frac{r}{2}} \\
&\leq K\ \lambda_n^{p-\frac{r}{2\delta}}.
\end{aligned}$$

Therefore, $\sum_n \mathbb{E}(Y^n) < \infty$, which implies $\mathbb{P}(Y^n > 1 \quad \text{i.o.}) = 0$ and

$$\sup_n Y^n < \infty \text{ a.e.}$$

Now, let $0 < \lambda \leq 1$ a fixed real number. There exists an integer n such that $\lambda_{n+1} < \lambda \leq \lambda_n$.

$$\begin{aligned}
Y_\lambda &= \lambda^p \sum_{j=0}^{\infty} \sum_{k\in\mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda} \\
&\leq \lambda_n^p \sum_{j=0}^{\infty} \sum_{k\in\mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda_{n+1}} \\
&\leq \left(\frac{\lambda_n}{\lambda_{n+1}}\right)^p Y^{n+1} \\
&\leq 2^p \sup_n Y^n.
\end{aligned}$$

Therefore

$$\sup_{0<\lambda\leq 1} Y_\lambda < \infty \text{ a.e.}$$

Finally, we study Y_λ when λ is large. Let n be a fixed integer.

$$\mathbb{E}(Y_n) = n^p \sum_{j=0}^{\infty} 2^{j\frac{r}{2}} \, \mathbb{P}(|\beta_{jk}| > n).$$

But

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

If $m_n = \left[\frac{n^2+2p\log(C)}{2p\delta\log(2)}\right] + 1$, for $n \geq M$,

$$\begin{aligned} \mathbb{E}(Y_n) &= n^p \sum_{j=m_n}^{\infty} 2^{j\frac{r}{2}} \left(\frac{\alpha_j}{n+\alpha_j}\right)^p \\ &\leq \sum_{j=m_n}^{\infty} 2^{j\frac{r}{2}} \alpha_j^p \\ &\leq K 2^{m_n(\frac{r}{2}-\delta p)}. \end{aligned}$$

As previously, $\sum_n \mathbb{E}(Y_n) < \infty$, and

$$\sup_n Y_n < \infty \text{ a.e.}$$

Now, let $\lambda \geq 1$ a fixed real number. There exists $n \in \mathbb{N}^*$ such that $n \leq \lambda < n+1$, and

$$\begin{aligned} Y_\lambda &= \lambda^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>\lambda} \\ &\leq (n+1)^p \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}_j} 2^{j(\frac{r}{2}-1)} \mathbb{1}_{|\beta_{jk}|>n} \\ &\leq \left(\frac{n+1}{n}\right)^p Y_n \\ &\leq 2^p \sup_n Y_n. \end{aligned}$$

Therefore,

$$\sup_{\lambda \geq 1} Y_\lambda < \infty \text{ a.e.}$$

and (5.2) is true. Theorem 3.1 is proved.

□

Acknowledgments

I would like to thank professor Dominique Picard for a wealth of advice and encouragement.

References

Abramovich, F. and Y. Benjamini (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. *Wavelets and Statistics. Lecture Notes in Statistics 103* (A. Antoniadis and G. Oppenheim, eds.) Springer-Verlag, 5–14

Abramovich, F., Y. Benjamini, D.L. Donoho and I.M. Johnstone (2000). Adapting to unknown sparsity by controlling the false discovery rate. To appear.

Abramovich, F., T. Sapatinas and B.W. Silverman (1998). Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society, B*, **60**, 725–749.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B*, **57**, 289–300.

Brown, L.D. and M.G. Low (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, **24**, 2384–2398.

Chipman, H.A., E.D. Kolaczyk and R.E. McCulloch (1997). Adaptive bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.

Clyde, M., G. Parmigiani and B. Vidakovic (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.

Cohen, A., I. Daubechies and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, **1**, 54–81.

Cohen, A., R.A. DeVore and R. Hochmuth (2000). Restricted nonlinear approximation. *Constructive Approximation*, **16**, 85-113.

Cohen, A., R.A. DeVore, G. Kerkyacharian and D. Picard (2000). Maximal spaces with given rate of convergence for thresholding algorithms. *Applied and Computational Harmonic Analysis*, **11**, 167–191.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

DeVore, R.A. (1989). Degree of nonlinear approximation. *Approximation Theory*

VI, Volume 1 (C.K. Chui, L.L. Schumaker and J.D. Wards, eds.) Academic Press, 175–201.

DeVore, R.A. and G.G. Lorentz (1993). *Constructive Approximation.* Springer-Verlag, New York.

Donoho, D.L. (1996). Unconditional bases and bit-level compression. *Applied and Computational Harmonic Analysis*, **3**, 388–392.

Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D.L. and I.M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.

Donoho, D.L. and I.M. Johnstone (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli*, **2**, 39–62.

Härdle, W., G. Kerkyacharian, D. Picard and A. Tsybakov (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics 129.* Springer-Verlag, New York.

Huang, H.C. and N. Cressie (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, **42**, 262-276.

Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. *Statistical Decision Theory and Related Topics, Volume 5* (S.S. Gupta and J.O. Berger, eds.) Springer-Verlag, 303–326.

Kerkyacharian, G. and D. Picard (2000). Minimax or maxisets?. Technical Report. Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing.* Academic Press, San Diego.

Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

Nason, G.P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, B*, **58**, 463–479.

Peetre, J. (1976). *New Thoughts on Besov Spaces.* Duke University Press, Durham.

Rivoirard, V. (2001). Non linear estimation over weak Besov spaces and minimax Bayes method. Preprint Laboratoire de Probabilités et Modèles aléatoires, Universités Paris VI et Paris VII. Submitted to *Bernoulli.*

Shorack, G.R. and J.A. Wellner (1986). *Empirical Processes with Applications to*

Statistics. Wiley, New York.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, **93**, 173–179.