Volume 15, No. 4 (2006), pp. Allerton Press, Inc.

MATHEMATICAL METHODS OF STATISTICS

LARGE VARIANCE GAUSSIAN PRIORS IN BAYESIAN NONPARAMETRIC ESTIMATION: A MAXISET APPROACH

F. Autin[1], D. Picard[2], and V. Rivoirard[3]

[1]CNRS-UMR 6632, Univ. Aix-Marseille 1, Centre de Mathématiques et Informatique
39 rue F. Joliot Curie, 13453 Marseille Cedex 13, France
E-mail: autin@cmi.univ-mrs.fr

[2]CNRS UMR7599, Université Paris VII - Denis Diderot
France
E-mail: picard@math.jussieu.fr

[3]Equipe Probab., Statist. et Modélisation, Labor. de mathématiques
CNRS-UMR 8628, Bâtiment 425, Université Paris 11– Orsay
15 rue Georges Clémenceau, 91405 Orsay Cedex, France
E-mail: Vincent.Rivoirard@math.u-psud.fr

In this paper we compare wavelet Bayesian rules taking into account the sparsity of the signal with priors which are combinations of a Dirac mass with a standard distribution properly normalized. To perform these comparisons, we take the maxiset point of view: i.e., we consider the set of functions which are well estimated (at a prescribed rate) by each procedure. We especially consider the standard cases of Gaussian and heavy-tailed priors. We show that while heavy-tailed priors have extremely good maxiset behavior compared to traditional Gaussian priors, considering large variance Gaussian priors (LVGP) leads to equally successful maxiset behavior. Moreover, these LVGP can be constructed in an adaptive way. We also show, using comparative simulations results that large variance Gaussian priors have very good numerical performance, confirming the maxiset prediction, and providing the advantage of high simplicity from the computational point of view.

Key words: Bayesian methods, minimax and maxiset approaches, nonparametric estimation.

2000 Mathematics Subject Classification: ???.

1. Introduction

Bayesian techniques have now become very popular to estimate signals decomposed in wavelet bases. Many authors have built Bayes estimates showing, from the practical point of view, impressive properties especially in estimation of inhomogeneous signals. Most of the simulations show that these procedures seriously outperform classical procedures and, in particular, thresholding procedures. See, for instance, Chipman *et al.* [5], Abramovich *et al.* [2], Clyde *et al.* [8], Johnstone and Silverman [15], Vidakovic [26] or Clyde and George [6], [7] who discussed the choice of the Bayes model to capture the sparsity of the signal to be estimated and the choice of the Bayes rule (among others, posterior mean or median). We also refer the reader to the very complete review paper of Antoniadis *et al.* [3] providing descriptions and comparisons of various Bayesian wavelet shrinkage and wavelet thresholding estimators.

To capture the sparsity of a signal f that is supposed to be decomposed in a wavelet basis $\{\psi_{jk}, j \geq -1, k \in \mathbb{Z}\}$ as $f = \sum_{j\geq -1}\sum_k \beta_{jk}\psi_{jk}$, where β_{jk} is the $\mathbb{L}_2$-scalar product between f and the wavelet function ψ_{jk}, the most common models introduce priors on the wavelet coefficients of the following form:

$$\beta_{jk} \sim \pi_{j,\varepsilon}\gamma_{j,\varepsilon} + (1-\pi_{j,\varepsilon})\delta(0), \tag{1.1}$$

where $0 \leq \pi_{j,\varepsilon} \leq 1$, $\delta(0)$ is a point mass at zero and the β_{jk}'s are independent. The nonzero part of the prior $\gamma_{j,\varepsilon}$ is assumed to be the dilation of a fixed symmetric, positive, unimodal and continuous density γ:

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

where the dilation parameter $\tau_{j,\varepsilon}$ is positive. The parameter $\pi_{j,\varepsilon}$ can be interpreted as the proportion of non-negligible coefficients. We also introduce the parameter

$$w_{j,\varepsilon} = \frac{\pi_{j,\varepsilon}}{1-\pi_{j,\varepsilon}}.$$

When the signal is sparse, most of the $w_{j,\varepsilon}$'s are small. These priors or their very close versions have been extensively used by the authors cited above and especially by Abramovich *et al.* [1], Johnstone and Silverman [14], [16]. To complete the definition of the prior model, we have to fix the hyperparameters $\tau_{j,\varepsilon}$ and $w_{j,\varepsilon}$. Finally the density γ will play a very important role. The most popular choice for γ is the normal density. It is also the density giving rise to the easiest procedures from the computational point of view. However heavy-tailed priors have proved also to work very well.

From the minimax point of view, recent works have studied these Bayes procedures and it was proved that Bayes rules can achieve optimal rates of convergence. Abramovich *et al.* [1] investigated theoretical performance of the procedures introduced by Abramovich *et al.* [2], considering priors of the form quoted above with some particular choice of the hyperparameters. 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 $\mathcal{B}^s_{p,q}$ when $p \geq 2$. When $p < 2$, these estimators show less impressive properties since they only behave as linear estimates. Recently, Abramovich *et al.* [1], Johnstone and Silverman [14], [16] investigated minimax properties of Bayes rules, with priors based on heavy-tailed distributions and they considered an empirical Bayes setting. In this case, the posterior mean and median turn out to be optimal for the whole scale of Besov spaces. Other more sophisticated results concerning minimax properties of Bayes rules have been established by Zhang [27].

Hence, summarizing the results cited above, the minimax results seem to indicate that Bayes procedures have comparable performance to thresholding estimates at least on the range of Besov spaces, but also seem to show a preference to heavy-tailed priors.

The main goal of this paper is to push a little further this type of comparison on 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 was to answer the following question: Are Gaussian priors always outperformed by heavy-tailed priors? And quite happily, one of our results will show that though some Bayesian procedures using Gaussian priors behave quite badly (in terms of maxisets as it was the case in terms of minimax rates) as compared to those with heavy tails, it is nevertheless possible to attain a very good maxiset behavior among procedures based on Gaussian priors. We prove that this can only be achieved provided that the hyperparameter $\tau_{j,\varepsilon}$ is "large". Under this assumption, the density $\gamma_{j,\varepsilon}$ 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.

As these Bayesian procedures with large variance Gaussian priors have not been much studied in the literature yet, we investigate their behavior also from a practical point of view and show a comparative simulations study with many standard and Bayesian procedures in the literature. As can be seen in our last section, such estimators turn out to have excellent numerical performance.

Let us only recall here that the maxiset point of view consists in determining the set of all functions which can be estimated at a specified rate of convergence for a specified procedure. Exhibiting maxisets of different estimation rules allows us to say that a procedure is more powerful than another one if its maxiset is larger.

The results that have been obtained up to now, using the maxiset point of view, are very promising since they generally show that the maxisets of well-known procedures are well understandable and easily interpretable sets. They have the advantage of being generally less pessimistic and seem also to enjoy the important advantage of giving theoretical claims which are often closer to the practical (simulations) situation, than other theoretical results (such as minimax rates).

The second section specifies the model and Bayesian rules we are going to consider. The third section recalls the definition of maxisets and briefly details some results obtained in the area, to allow a comparison with the results to be obtained for Bayesian rules. The forth section investigates maxisets of standard Bayesian rules: first the 'small variance' Gaussian priors, then the heavy-tailed priors. The fifth section is devoted to large variance Gaussian priors, and the last section presents the simulations results.

2. Model and Bayesian Rules

For the sake of simplicity, we will consider a white noise setting: $X_\varepsilon(\cdot)$ is a random measure satisfying on $[0, 1]$ the following equation:

$$X_\varepsilon(dt) = f(t)dt + \varepsilon W(dt),$$

where $0 < \varepsilon < 1$ is the noise level and f is a function defined on $[0,1]$, $W(\cdot)$ is a Brownian motion on $[0,1]$. As usual, to connect with the standard framework of sequences of experiments we put $\varepsilon = n^{-1/2}$.

Let $\{\psi_{jk}(\cdot),\ j \geq -1, k \in \mathbb{Z}\}$ be a compactly supported wavelet basis of $\mathbb{L}_2([0,1])$ such that any $f \in \mathbb{L}_2([0,1])$ can be represented as:

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

where $\beta_{jk} = \langle f, \psi_{jk} \rangle_{\mathbb{L}_2}$. As usual, the ψ_{-1k} denote the translations of the scaling function, and the ψ_{jk}, for $j \geq 0$, are the dilations and translations of the wavelet function. The model is reduced to a sequence space model if we put $y_{jk} = X_\varepsilon(\psi_{jk}) = \int f \psi_{jk} + \varepsilon Z_{jk}$, where the Z_{jk} are i.i.d. $\mathcal{N}(0,1)$. Let us note that at each level $j \geq 0$, the number of non-zero wavelet coefficients is less than or equal to $2^j + l_\psi - 1$, where l_ψ is the maximal size of the supports of the scaling function and the wavelet. So, there exists a constant S_ψ such that at each level $j \geq -1$, there are no more than $S_\psi \times 2^j$ coefficients to be estimated. In the sequel, we shall not distinguish between f and $\beta = (\beta_{jk})_{jk}$, its sequence of wavelet coefficients.

As explained in the Introduction, we consider the priors, where the β_{jk}'s are independent random variables with the following distribution:

$$\beta_{jk} \sim \pi_{j,\varepsilon} \gamma_{j,\varepsilon} + (1 - \pi_{j,\varepsilon}) \delta(0), \tag{2.1}$$

$$\gamma_{j,\varepsilon}(\beta_{jk}) = \frac{1}{\tau_{j,\varepsilon}} \gamma\left(\frac{\beta_{jk}}{\tau_{j,\varepsilon}}\right), \qquad w_{j,\varepsilon} = \frac{\pi_{j,\varepsilon}}{1 - \pi_{j,\varepsilon}}. \tag{2.2}$$

Here $0 \leq \pi_{j,\varepsilon} \leq 1$, $\delta(0)$ is the Dirac mass at 0, γ is a fixed symmetric, positive, unimodal, and continuous density, $\tau_{j,\varepsilon}$ is positive.

2.1. Gaussian priors. Consider the case, where γ is the Gaussian density, which is the most classical choice. In this case, we easily derive that the Bayes estimation rules of β_{jk} associated with the l^1- and l^2-losses, respectively, are the posterior median and mean:

$$\breve{\beta}_{jk} = \mathrm{Med}(\beta_{jk} \mid y_{j,k}) = \mathrm{sign}(y_{j,k}) \max(0, \xi_{jk}), \tag{2.3}$$

$$\tilde{\beta}_{jk} = \mathbb{E}(\beta_{jk} \mid y_{j,k}) = \frac{b_j}{1 + \eta_{jk}} y_{j,k}, \tag{2.4}$$

where

$$\xi_{jk} = b_j |y_{j,k}| - \varepsilon \sqrt{b_j} \Phi^{-1}\left(\frac{1 + \min(\eta_{jk}, 1)}{2}\right),$$

$$b_j = \frac{\tau_{j,\varepsilon}^2}{\varepsilon^2 + \tau_{j,\varepsilon}^2},$$

$$\eta_{jk} = \frac{1}{w_{j,\varepsilon}} \frac{\sqrt{\varepsilon^2 + \tau_{j,\varepsilon}^2}}{\varepsilon} \exp\left(-\frac{\tau_{j,\varepsilon}^2 y_{j,k}^2}{2\varepsilon^2(\varepsilon^2 + \tau_{j,\varepsilon}^2)}\right),$$

and Φ is the normal cumulative distribution function.

To study the properties of such rules, it is interesting to make use of their shrinkage properties. Let us recall that $\hat{\beta}$ is said to be a shrinkage rule if $y_{j,k} \longrightarrow \hat{\beta}_{jk}$ is antisymmetric, increasing on $(-\infty, +\infty)$, and

$$0 \leq \hat{\beta}_{jk} \leq y_{j,k}, \qquad \forall\, y_{j,k} \geq 0.$$

Both rules quoted above obviously are shrinkage rules. We also note that $\breve{\beta}_{jk}$ is zero whenever $y_{j,k}$ falls in an implicitly defined interval $[-\lambda_{j,\varepsilon}, \lambda_{j,\varepsilon}]$.

We will first consider the following very classical form of the hyperparameters:

$$\tau^2_{j,\varepsilon} = c_1 2^{-\alpha j}, \qquad \pi_{j,\varepsilon} = \min(1, c_2 2^{-bj}), \tag{2.5}$$

where c_1, c_2, α, and b are positive constants. This particular form was suggested by Abramovich *et al.* [2] and then used by Abramovich *et al.* [1]. A nice interpretation was provided by these authors who explained how α, b, c_1, and c_2 can be derived in applications.

Our second part will be concerned with large variance rules. In this case, we will consider hyperparameters of the form

$$\tau_{j,\varepsilon} = \tau(\varepsilon) \qquad \text{and} \qquad w_{j,\varepsilon} = w(\varepsilon) \tag{2.6}$$

with specified conditions on the functions τ and w.

Remark 1. An alternative for eliciting these hyperparameters consists in using empirical Bayes methods and EM algorithm (see Clyde and George [6], [7] or Johnstone and Silverman [15]).

2.2. Heavy-tailed priors. For the sake of comparison, we will also consider priors, where the density γ is no longer Gaussian. We assume that there exist two positive constants M and M_1 such that

$$\sup_{\beta \geq M_1} \left| \frac{d}{d\beta} \log \gamma(\beta) \right| = M < \infty. \tag{2.7}$$

The hypothesis (2.7) means that the tails of γ have to be exponential or heavier. Indeed, under (2.7), we have:

$$\forall\, u \geq M_1, \quad \gamma(u) \geq \gamma(M_1) \exp(-M(u - M_1)).$$

In the minimax approach of Johnstone and Silverman [14], [16], the priors also satisfy (2.7). To complete the prior model, we assume that:

$$\tau_{j,\varepsilon} = \varepsilon, \qquad w_{j,\varepsilon} = w(\varepsilon) \to 0 \qquad \text{as} \quad \varepsilon \to 0, \tag{2.8}$$

and w is a positive continuous function. Using these assumptions, the following proposition describes the properties of the posterior median and mean:

Proposition 1. *Under the conditions* (2.7) *and* (2.8) *the estimates* $\breve{\beta}^{HT}_{jk} = \mathrm{Med}(\beta_{jk} \mid y_{j,k})$ *and* $\tilde{\beta}^{HT}_{jk} = \mathbb{E}(\beta_{jk} \mid y_{j,k})$ *are shrinkage rules. Moreover,* $\breve{\beta}^{HT}_{jk}$ *is a thresholding rule: there exists* $\breve{t}_\varepsilon$ *such that*

$$\breve{\beta}^{HT}_{jk} = 0 \iff |y_{j,k}| \leq \breve{t}_\varepsilon,$$

where the threshold $\breve{t}_\varepsilon \geq \varepsilon\sqrt{2\log(1/w(\varepsilon))}$ *for* ε *small enough and*

$$\lim_{\varepsilon \to 0} \frac{\breve{t}_\varepsilon}{\varepsilon\sqrt{2\log(1/w(\varepsilon))}} = 1.$$

Proof. The first assertion has been established by Johnstone and Silverman [14], [16]. The second assertion is an immediate consequence of Proposition 3 in Rivoirard [25]. □

3. Maxisets and Associated Functional Spaces

Let us first briefly recall the definition of maxisets. We consider a sequence of models $\mathcal{E}_n = \{P_\theta^n, \theta \in \Theta\}$, where the P_θ^n's are probability distributions on the measurable spaces Ω_n and Θ is the set of parameters. We also consider a sequence of estimates $\hat{q}_n$ of a quantity $q(\theta)$ associated with this sequence of models, a loss function $\rho(\hat{q}_n, q(\theta))$, and a rate of convergence α_n tending to 0. Then, we define the *maxiset* associated with the sequence $\hat{q}_n$, the loss function ρ, the rate α_n, and the constant T as the following set:

$$MS(\hat{q}_n, \rho, \alpha_n)(T) = \big\{\theta \in \Theta,\ \sup_n \mathbb{E}_\theta^n \rho(\hat{q}_n, q(\theta))(\alpha_n)^{-1} \leq T\big\}.$$

The focus in this domain has mainly been on the nonparametric situation. Let us briefly mention the differences from the minimax point of view. In this latter, we fix a set of functions and look at the worst performance of estimators. Here, instead of fixing a priori a (functional) set such as a Hölder, Sobolev or Besov ball, we settle the problem in a wider context: the parameter set Θ can be very large, such as the set of bounded measurable functions. Then, the maxiset is associated with the procedure in a more genuine way since it only depends on the model and the estimation rule at hand.

As explained in more detail later in this section, there already exist very interpretable results about maxisets. For instance, it has been established in Kerkyacharian and Picard [17] that 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 Cohen *et al.* [9]) are Lorentz spaces reflecting extremely well the practical observation that wavelet thresholding performs well when the number of wavelet coefficients is small. It has also been observed (see Kerkyacharian and Picard [19]) 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 with the oracle.

Although the two settings seem quite different, still there is a deep parallel between maxisets and minimax theory. For instance, facing a particular situation, the standard procedure to prove that a set B is the maxiset usually consists (exactly as in the minimax theory) of two steps: first showing that $B \subset MS(\hat{q}_n, \rho, \alpha_n)(T)$, but this is generally obtained using similar arguments as for proving upper bound inequalities in minimax setting, since one simply has to prove that $\theta \in B$ implies $\mathbb{E}_\theta^n \rho(\hat{q}_n, q(\theta)) \leq T\alpha_n$. The advantage of the maxiset setting is probably that the second inclusion $MS(\hat{q}_n, \rho, \alpha_n)(T) \subset B$ is often proved much simpler than proving the lower bound for minimax rates over complicated spaces.

3.1. Functional spaces. In this paper, for simplicity, we shall restrict ourselves to the case where ρ is the squared $\mathbb{L}_2$ norm, even though a large majority of the results can be extended to more general losses. For this study, we need to introduce the following classes of functions which are of typical use in maxiset theory.

We give definitions of the Besov and weak Besov spaces depending on the wavelet basis. However, as is established in Meyer [21] and Cohen *et al.* [9], most of them have also different definitions proving that this dependence in the basis is not crucial at all.

Definition 1. Let $s > 0$ and $R > 0$. A function $f = \sum_{j=-1}^{+\infty} \sum_k \beta_{jk}\psi_{jk} \in \mathbb{L}_2([0,1])$ belongs to the *Besov ball* $\mathcal{B}^s_{p,\infty}(R)$ if and only if

$$\left[\sup_{j\geq -1} 2^{j(s+\frac{1}{2}-\frac{1}{p})p} \sum_k |\beta_{jk}|^p\right]^{1/p} \leq R.$$

Note that, when $p = 2$, f belongs to $\mathcal{B}^s_{2,\infty}$ if and only if

$$\sup_{J\geq -1} 2^{2Js} \sum_{j\geq J} \sum_k \beta_{jk}^2 < +\infty. \tag{3.1}$$

This characterization is often used in the sequel. Recall that the class of Besov spaces $\mathcal{B}^s_{p,\infty}$ provides a useful tool to classify wavelet decomposed signals according to their regularity and sparsity properties (see Donoho *et al.* [12], Donoho and Johnstone [11] or Johnstone [13]). Roughly speaking, regularity increases when s increases, whereas sparsity increases when p decreases. Especially, the spaces with indices $p < 2$ are of particular interest, since they describe very wide classes of inhomogeneous but sparse functions. To model sparsity, a very convenient and natural tool consists in introducing the following particular class of Lorentz spaces that are in addition directly connected to the estimation procedures considered in this paper.

Definition 2. Let $0 < r < 2$ and $R > 0$. A function $f = \sum_{j=-1}^{+\infty} \sum_k \beta_{jk}\psi_{jk} \in \mathbb{L}_2([0,1])$ belongs to the *weak Besov ball* $W_r(R)$ if and only if

$$\left[\sup_{\lambda>0} \lambda^{r-2} \sum_{j\geq -1} \sum_k \beta_{jk}^2\, I\{|\beta_{jk}| \leq \lambda\}\right]^{1/2} \leq R.$$

It is not difficult to prove (see Cohen *et al.* [9]) that

$$f \in W_r \Leftrightarrow \sup_{\lambda>0} \lambda^r \sum_j I\{|\beta_{jk}| > \lambda\} < \infty.$$

We have, in particular,

$$\sup_{\lambda>0} \lambda^r \sum_{j,k} I\{|\beta_{jk}| > \lambda\} \leq \frac{2^{2-r}}{1-2^{-r}} \sup_{\lambda>0} \lambda^{r-2} \sum_{j\geq -1} \sum_k \beta_{jk}^2\, I\{|\beta_{jk}| \leq \lambda\}, \tag{3.2}$$

which shows the natural relationship between sparsity and weak Besov spaces and the connection with the regular Besov spaces introduced above. If $\subsetneq$ denotes the strict inclusion between two functional spaces, the following embeddings are not difficult to show (see, for instance, Meyer [21], Kerkyacharian and Picard [19] or Rivoirard [23]):

$$\mathcal{B}^s_{p,\infty} \subsetneq \mathcal{B}^s_{2,\infty} \subsetneq W_{\frac{2}{1+2s}} \quad \text{if} \quad s>0, \quad p>2, \tag{3.3}$$

$$\mathcal{B}^s_{p,\infty} \subsetneq W_{\frac{2}{1+2s}} \quad \text{if} \quad s>0, \quad p<2. \tag{3.4}$$

3.2. First connections between the spaces and maxiset results. In the present setting of white noise model, Rivoirard [23] proved that the maxisets of linear estimates for polynomial rates of convergence of the form $\varepsilon^{4s/(1+2s)}$ are Besov spaces $\mathcal{B}^s_{2,\infty}$. A similar result in the context of kernel estimates was established in Kerkyacharian and Picard [17]. We introduce the classical hard and soft thresholding rules:

$$\hat{f}^T = \sum_{-1\le j<j_\varepsilon} \sum_k y_{jk} I\{|y_{jk}| > mt_\varepsilon\}\psi_{jk},$$

$$\hat{f}^S = \sum_{-1\le j<j_\varepsilon} \sum_k \left(1 - \frac{mt_\varepsilon}{|y_{jk}|}\right) I\{|y_{jk}| > mt_\varepsilon\} y_{jk}\psi_{jk},$$

with m a positive constant, $j_\varepsilon \in \mathbb{N}$ such that

$$t_\varepsilon = \varepsilon\sqrt{\log(1/\varepsilon)}, \tag{3.5}$$

$$2^{-j_\varepsilon} \le t_\varepsilon^2 < 2^{1-j_\varepsilon} \tag{3.6}$$

(which will be denoted in the sequel by $2^{j_\varepsilon} \sim t_\varepsilon^{-2}$).

Under mild conditions, Kerkyacharian and Picard [18] proved:

$$MS\Big(\hat{f}^T,\, \|\cdot\|_2^2,\, (\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\Big) = \mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}}.$$

A similar result is obtained for the soft thresholding rule $\hat{f}^S$.

Remark 2. The embeddings (3.3) and (3.4) give clear information about the respective performance of linear procedures and thresholding rules, which have been extensively confirmed by practical results. In particular, one can observe that the spaces $\mathcal{B}^s_{p,\infty}$ for $p<2$ are never contained in the maxisets of the linear procedures ($\mathcal{B}^s_{2,\infty}$), while they are contained in the maxisets of thresholding procedures ($\mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}}$) under fairly wide conditions.

Notation. If $\subset$ denotes the inclusion between two spaces, for a given space $\mathcal{A}$ the notation

$$MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, \lambda_\varepsilon) \subset \mathcal{A}$$
$$\text{(resp.)} \quad \mathcal{A} \subset MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, \lambda_\varepsilon)$$

will mean in the sequel

$$\forall\, M \;\exists\, M', \quad MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, \lambda_\varepsilon)(M) \subset \mathcal{A}(M')$$
$$\text{(resp.)} \quad \forall\, M' \;\exists\, M, \quad \mathcal{A}(M') \subset MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, \lambda_\varepsilon)(M),$$

where M and M' denote the radii of balls of $MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, \lambda_\varepsilon)$ and $\mathcal{A}$, respectively.

4. Maxisets Results for 'Heavy-Tailed' and 'Small Variance Gaussian' Priors

4.1. Maxisets results for small variance Gaussian priors. Let us consider now the Bayesian rules with Gaussian priors as explained in Section 2.1, and especially those satisfying conditions (2.5), as introduced in Abramovich *et al.* [2] and studied in Abramovich *et al.* [1].

Theorem 1. *With the above choice of the hyperparameters, for $s > 0$ and $\beta^0 \in \{\breve{\beta}, \tilde{\beta}\}$,*

- $\alpha > 2s + 1$ *implies* $\mathcal{B}^s_{p,\infty} \not\subset MS(\beta^0, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)})$ *for any* $1 \le p \le \infty$,
- $\alpha = 2s + 1$ *implies* $\mathcal{B}^s_{p,\infty} \not\subset MS(\beta^0, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)})$ *if* $p < 2$.

Remark 3. Theorem 1 is established for the rate $t_\varepsilon^{4s/(1+2s)}$, but it can be generalized for any rate of convergence of the form $\varepsilon^{4s/(1+2s)}(\log(1/\varepsilon))^m$, with $m \ge 0$. The results established in Theorem 1 (if we, for example, refer to Remark 2) show that these rules are obviously outperformed by thresholding rules. It is worthwhile to notice in addition, that their behavior is (like that of linear procedures) highly non-robust regarding the tuning constant α. The behavior of these rules turns out to be very comparable to linear rules as is confirmed in the Appendix, where more details about the maxisets of these procedure are given.

The proof of Theorem 1 is based on the following result.

Proposition 2. *If $\beta \in MS(\beta^0, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)})$, then there exists a constant C such that, for ε small enough,*

$$\sum_{j,k} \beta_{jk}^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\} \le C t_\varepsilon^{\frac{4s}{1+2s}}. \tag{4.1}$$

Proof. Here we shall distinguish the cases of the posterior mean and median. The posterior median can be written as follows:

$$\breve{\beta}_{jk} = \operatorname{sign}(y_{j,k})\big(b_j|y_{j,k}| - g(\varepsilon, \tau_{j,\varepsilon}, y_{j,k})\big),$$

with $0 \le g(\varepsilon, \tau_{j,\varepsilon}, y_{j,k}) \le b_j|y_{j,k}|$.

Let us assume that $b_j|y_{j,k} - \beta_{jk}| \le (1 - b_j)|\beta_{jk}|/2$ and $\tau_{j,\varepsilon}^2 \le \varepsilon^2$, so $b_j \le 1/2$.

First, suppose that $y_{j,k} \ge 0$, so $\breve{\beta}_{jk} \ge 0$. If $\beta_{jk} \ge 0$, then

$$\begin{aligned} |\breve{\beta}_{jk} - \beta_{jk}| &= \big|b_j(y_{j,k} - \beta_{jk}) - (1 - b_j)\beta_{jk} - g(\varepsilon, \tau_{j,\varepsilon}, y_{j,k})\big| \\ &= (1 - b_j)\beta_{jk} - b_j(y_{j,k} - \beta_{jk}) + g(\varepsilon, \tau_{j,\varepsilon}, y_{j,k}) \ge \frac{1}{2}(1 - b_j)\beta_{jk} \ge \frac{1}{4}\beta_{jk}. \end{aligned}$$

If $\beta_{jk} \le 0$, then

$$|\breve{\beta}_{jk} - \beta_{jk}| \ge \frac{1}{4}|\beta_{jk}|.$$

The case $y_{j,k} \le 0$ is handled by using similar arguments and the particular form of the posterior median. So, we obtain:

$$\begin{aligned}
\mathbb{E}(\breve{\beta}_{jk} - \beta_{jk})^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} &\ge \frac{1}{16}\beta_{jk}^2\, \mathbb{P}\big(b_j|y_{j,k} - \beta_{jk}| \le (1-b_j)|\beta_{jk}|/2\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} \\
&\ge \frac{1}{16}\beta_{jk}^2\, \mathbb{P}\big(|y_{j,k} - \beta_{jk}| \le |\beta_{jk}|/2\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} \\
&\ge \frac{1}{16}\beta_{jk}^2 \big(1 - \mathbb{P}(|y_{j,k} - \beta_{jk}| > |\beta_{jk}|/2)\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\}.
\end{aligned}$$

Using the large deviations inequalities for the Gaussian variables, we obtain for ε small enough:

$$\begin{aligned}
&\mathbb{E}(\breve{\beta}_{jk} - \beta_{jk})^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\} \\
&\quad \ge \frac{1}{16}\beta_{jk}^2 \big(1 - \mathbb{P}(|y_{j,k} - \beta_{jk}| > t_\varepsilon/2)\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\} \\
&\quad \ge \frac{1}{32}\beta_{jk}^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\}.
\end{aligned}$$

This implies (4.1).

For the posterior mean, we have:

$$\begin{aligned}
\mathbb{E}(\tilde{\beta}_{jk} - \beta_{jk})^2 &= \mathbb{E}\left(\frac{b_j}{1+\eta_{jk}}(y_{j,k} - \beta_{jk}) - \Big(1 - \frac{b_j}{1+\eta_{jk}}\Big)\beta_{jk}\right)^2 \\
&\ge \frac{1}{4}\mathbb{E}\left(\Big(1 - \frac{b_j}{1+\eta_{jk}}\Big)\beta_{jk}\right)^2 I\left\{\frac{b_j}{1+\eta_{jk}}|y_{j,k} - \beta_{jk}| \le \Big(1 - \frac{b_j}{1+\eta_{jk}}\Big)|\beta_{jk}|/2\right\}.
\end{aligned}$$

So, we obtain:

$$\begin{aligned}
\mathbb{E}(\tilde{\beta}_{jk} - \beta_{jk})^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} &\ge \frac{1}{16}\beta_{jk}^2 \mathbb{P}\big(|y_{j,k} - \beta_{jk}| \le |\beta_{jk}|/2\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} \\
&\ge \frac{1}{16}\beta_{jk}^2 (1 - \mathbb{P}\big(|y_{j,k} - \beta_{jk}| > |\beta_{jk}|/2)\big) I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\}.
\end{aligned}$$

Finally, using similar arguments to those used for the posterior median, we obtain (4.1). Proposition 2 is proved. □

Proof of Theorem 1. Let us first investigate the case $\alpha > 2s+1$. Take β such that all the β_{jk}'s are zero, except 2^j coefficients at each level j that are equal to $2^{-j(s+\frac{1}{2})}$. Then, $\beta \in \mathcal{B}_{p,\infty}^s$. Since $\tau_{j,\varepsilon}^2 = c_1 2^{-j\alpha}$, if we put $2^{J_\alpha} \sim c_1^{\frac{1}{\alpha}} \varepsilon^{-\frac{2}{\alpha}}$ and $2^{J_s} \sim t_\varepsilon^{\frac{-2}{2s+1}}$, we observe that asymptotically $J_\alpha < J_s$. So, for ε small enough

$$\sum_{j,k} \beta_{jk}^2 I\{\tau_{j,\varepsilon}^2 \le \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\} = \sum_{J_\alpha \le j < J_s} 2^{-2js} \ge c\varepsilon^{\frac{4s}{\alpha}},$$

with c a positive constant. Using Proposition 2, we see that β does not belong to $MS(\beta^0, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)})$.

Now we investigate the case $\alpha = 2s+1$. Take β such that all the β_{jk}'s are zero, except one coefficient at each level j that is equal to $2^{-j(s+\frac{1}{2}-\frac{1}{p})}$. Then, $\beta \in \mathcal{B}^s_{p,\infty}$. Similarly, we put $2^{J_\alpha} \sim c_1^{\frac{1}{\alpha}} \varepsilon^{-\frac{2}{\alpha}}$ and $2^{\tilde{J}_s} \sim t_\varepsilon^{-1/(s+\frac{1}{2}-\frac{1}{p})}$, then we observe that asymptotically $J_\alpha < \tilde{J}_s$. So, for ε small enough

$$\sum_{j,k} \beta_{jk}^2 I\{\tau_{j,\varepsilon}^2 \leq \varepsilon^2\} I\{|\beta_{jk}| > t_\varepsilon\} = \sum_{J_\alpha \leq j < \tilde{J}_s} 2^{-2j(s+\frac{1}{2}-\frac{1}{p})} \geq \tilde{c}\varepsilon^{4(s+\frac{1}{2}-\frac{1}{p})/\alpha},$$

with $\tilde{c}$ a positive constant. Using Proposition 2, we see that β does not belong to $MS(\beta^0, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)})$, since $p < 2$. □

4.2. Heavy-tailed priors. Consider now the case of priors satisfying conditions (2.7) and (2.8). If we set

$$\breve{f}_\varepsilon^{HT} = \sum_{j<j_\varepsilon} \sum_k \breve{\beta}_{jk}^{HT} \psi_{jk}, \qquad \breve{\beta}_{jk}^{HT} = \mathrm{Med}(\beta_{jk} \mid y_{j,k}), \tag{4.2}$$

and

$$\tilde{f}_\varepsilon^{HT} = \sum_{j<j_\varepsilon} \sum_k \tilde{\beta}_{jk}^{HT} \psi_{jk}, \qquad \tilde{\beta}_{jk}^{HT} = \mathbb{E}(\beta_{jk} \mid y_{j,k}), \tag{4.3}$$

where j_ε is such that $2^{j_\varepsilon} \sim t_\varepsilon^{-2}$, then using the results of Proposition 1, 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 established by Rivoirard [25] lead to the following result.

Theorem 2. *Let $s > 0$. We suppose that there exist two positive constants ρ_1 and ρ_2 such that for $\varepsilon > 0$ small enough,*

$$\varepsilon^{\rho_1} \leq w(\varepsilon) \leq \varepsilon^{\rho_2}.$$

Then, we have

$$MS\big(f_\varepsilon^0, \|\cdot\|_2^2, (\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big) = \mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}},$$

where $f_\varepsilon^0 \in \{\tilde{f}_\varepsilon^{HT}, \breve{f}_\varepsilon^{HT}\}$ as soon as $\rho_2 \geq 16$ for the posterior median and $\rho_2 \geq 64$ for the posterior mean.

So, the performance of adaptive Bayesian procedures based on heavy-tailed prior densities is similar to that of classical nonlinear procedures in the maxiset framework. In particular, they obviously outperform the above small-variance Bayesian procedures from the maxiset point of view.

5. Gaussian Priors with Large Variance

The previous section has shown the power of the Bayes procedures built from heavy-tailed prior models in the maxiset setting. The goal of this section is to answer the following questions. Are heavy-tailed priors unavoidable? Is it possible to build Gaussian priors leading to procedures with maxiset properties comparable to the heavy-tailed methods discussed above? Moreover, can we construct these procedures in such a way that they automatically adapt to the regularity of the function (adaptivity property). In other words, if γ is the Gaussian density, does there exist an adaptive choice of the hyperparameters $\pi_{j,\varepsilon}$ and $w_{j,\varepsilon}$ such that

$$MS\big(f^0_\varepsilon, \|\cdot\|_2^2, (\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big) = \mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}} \,?$$

This is a very important issue, since calculations using Gaussian priors are mostly direct and much easier than for heavy-tailed priors. The answers are provided by the following Theorem 3.

Consider the following estimates:

$$\breve{f}^{LV}_\varepsilon = \sum_{j<j_\varepsilon}\sum_k \breve{\beta}_{jk}\psi_{jk}, \qquad \breve{\beta}_{jk} = \mathrm{Med}(\beta_{jk} \mid y_{j,k}), \tag{5.1}$$

and

$$\tilde{f}^{LV}_\varepsilon = \sum_{j<j_\varepsilon}\sum_k \tilde{\beta}_{jk}\psi_{jk}, \qquad \tilde{\beta}_{jk} = \mathbb{E}(\beta_{jk} \mid y_{j,k}) \tag{5.2}$$

(recall that the posterior mean and median are given in (2.4) and (2.3)), with the following choice of hyperparameters:

$$\tau_{j,\varepsilon} = \tau(\varepsilon) \qquad \text{and} \qquad w_{j,\varepsilon} = w(\varepsilon). \tag{5.3}$$

Theorem 3. *We consider the prior model* (1.1)*, where γ is the Gaussian density. We assume that $\tau_{j,\varepsilon} = \tau(\varepsilon)$ and $w_{j,\varepsilon} = w(\varepsilon)$ are independent of j with w a continuous positive function. We consider $\breve{f}_\varepsilon$ and $\tilde{f}_\varepsilon$ introduced in* (5.1) *and* (5.2)*. If*

$$1 + \varepsilon^{-2}\tau(\varepsilon)^2 = t_\varepsilon^{-1}$$

and there exist q_1 and q_2 such that

$$\varepsilon^{q_1} \le w(\varepsilon) \le \varepsilon^{q_2}$$

for ε small enough, then we have:

$$MS\big(f^0_\varepsilon, \|\cdot\|_2^2, (\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big) = \mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}},$$

where $f^0_\varepsilon \in \{\tilde{f}_\varepsilon, \breve{f}_\varepsilon\}$ as soon as $q_2 > 63/2$ for the posterior median and $q_2 \ge 65/2$ for the posterior mean.

Unlike the previous choice ($\tau^2_{j,\varepsilon} = \varepsilon^2$ or $\tau^2_{j,\varepsilon} = 2^{-j\alpha}$), here we impose a “larger” variance. It is the key point of the proof of Theorem 3. In a sense, we reconstruct heavy tails by increasing the variance. The proof of Theorem 3 essentially relies on the following proposition.

Proposition 3. *Let $s > 0$ and let $\varpi_{jk}(\varepsilon)$ be a sequence of random weights lying in $[0,1]$. We assume that there exist positive constants c, m, and $K(\varpi)$ such that for any $\varepsilon > 0$,*

$$\hat{\beta}(\varepsilon) = (\varpi_{jk}(\varepsilon) y_{j,k})_{jk}$$

is a shrinkage rule satisfying for any ε,

$$\varpi_{jk}(\varepsilon) = 0, \quad a.e. \quad \forall\, j \geq j_\varepsilon \quad with \quad 2^{j_\varepsilon} \sim t_\varepsilon^{-2}, \quad \forall\, k, \tag{5.4}$$

$$|y_{jk}| \leq m t_\varepsilon \Rightarrow \varpi_{jk}(\varepsilon) \leq c t_\varepsilon, \quad a.e. \quad \forall\, j < j_\varepsilon, \quad \forall\, k, \tag{5.5}$$

$$(1 - \varpi_{jk}(\varepsilon)) \leq K(\varpi)\Big(\frac{t_\varepsilon}{|y_{j,k}|} + t_\varepsilon\Big), \quad a.e. \quad \forall\, j < j_\varepsilon, \quad \forall\, k, \tag{5.6}$$

and let

$$\hat{f}_\varepsilon = \sum_{j<j_\varepsilon} \sum_k \varpi_{jk}(\varepsilon) y_{j,k} \psi_{jk}.$$

Let $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}} \cap W_{\frac{2}{1+2s}}$ and note that

$$\|f\|^2_{B_{2,\infty}^{\frac{s}{1+2s}}} = \sup_{J \geq -1} 2^{2Js} \sum_{j \geq J} \sum_k \beta_{jk}^2 < \infty$$

and

$$\|f\|^2_{W_{\frac{2}{1+2s}}} = \sup_{\lambda > 0} \lambda^{r-2} \sum_{j \geq -1} \sum_k \beta_{jk}^2 I\{|\beta_{jk}| \leq \lambda\} < \infty.$$

Then, as soon as $m \geq 8$, we have the following inequality:

$$\begin{aligned}
\mathbb{E}\|\hat{f}_\varepsilon - f\|_2^2 \leq \Bigg[& 4c^2 S_\psi + 4(1 + K(\varpi)^2)\|f\|_2^2 + 4\sqrt{3} S_\psi \\
& + 2(2^{\frac{4s}{1+2s}} + 2^{\frac{-4s}{1+2s}}) m^{\frac{4s}{1+2s}} \|f\|^2_{W_{\frac{2}{1+2s}}} \\
& + \frac{8m^{-2/1+2s}}{(1 - 2^{-2/1+2s})}(1 + 8K(\varpi)^2)\|f\|^2_{W_{\frac{2}{1+2s}}} + \|f\|^2_{B_{2,\infty}^{\frac{s}{1+2s}}} \Bigg] t_\varepsilon^{\frac{4s}{1+2s}},
\end{aligned}$$

and

$$\mathcal{B}_{2,\infty}^{\frac{s}{1+2s}} \cap W_{\frac{2}{1+2s}} \subset MS(\hat{f}_\varepsilon, \|\cdot\|_2^2, t_\varepsilon^{4s/(1+2s)}).$$

Proof. Using (5.4), we have

$$\mathbb{E}\|\hat{f}_\varepsilon - f\|_2^2 = \mathbb{E}\| \sum_{j<j_\varepsilon, k} \big(\varpi_{jk}(\varepsilon) y_{j,k} - \beta_{jk}\big)\psi_{j,k}\|_2^2 + \sum_{j \geq j_\varepsilon, k} \beta_{jk}^2.$$

The second term is a bias term bounded by $t_\varepsilon^{\frac{4s}{1+2s}} \|f\|^2_{B_{2,\infty}^{\frac{s}{1+2s}}}$.

We split $\mathbb{E}\sum_{j<j_\varepsilon,k}(\varpi_{jk}(\varepsilon) y_{j,k} - \beta_{jk})^2$ into $2(A+B)$ with

$$A = \mathbb{E} \sum_{j<j_\varepsilon,k} \big[\varpi_{jk}(\varepsilon)^2 (y_{j,k} - \beta_{jk})^2 + (1 - \varpi_{jk}(\varepsilon))^2 \beta_{jk}^2\big] I\{|y_{j,k}| \leq m t_\varepsilon\},$$

$$B = \mathbb{E} \sum_{j<j_\varepsilon,k} \big[\varpi_{jk}(\varepsilon)^2 (y_{j,k} - \beta_{jk})^2 + (1 - \varpi_{jk}(\varepsilon))^2 \beta_{jk}^2\big] I\{|y_{j,k}| > m t_\varepsilon\}.$$

Again, we split A into $A_1 + A_2$, and using (5.5)

$$\begin{aligned}
A_1 &= \mathbb{E} \sum_{j<j_\varepsilon,k} \varpi_{jk}(\varepsilon)^2 (y_{j,k} - \beta_{jk})^2\, I\{|y_{j,k}| \le mt_\varepsilon\} \\
&\le c^2 S_\psi 2^{j_\varepsilon} t_\varepsilon^2 \varepsilon^2 \le 2c^2 S_\psi t_\varepsilon^2, \\
A_2 &= \mathbb{E} \sum_{j<j_\varepsilon,k} (1 - \varpi_{jk}(\varepsilon))^2 \beta_{jk}^2\, I\{|y_{j,k}| \le mt_\varepsilon\} \\
&\le \mathbb{E} \sum_{j<j_\varepsilon,k} \beta_{jk}^2\ I\{|y_{j,k}| \le mt_\varepsilon\} \big[I\{|\beta_{jk}| \le 2mt_\varepsilon\} + I\{|\beta_{jk}| > 2mt_\varepsilon\} \big] \\
&\le (2mt_\varepsilon)^{4s/1+2s} \|f\|^2_{W_{\frac{2}{1+2s}}} + \sum_{j<j_\varepsilon,k} \beta_{jk}^2 \mathbb{P}(|\beta_{jk} - y_{j,k}| \ge mt_\varepsilon) \\
&\le (2mt_\varepsilon)^{4s/1+2s} \|f\|^2_{W_{\frac{2}{1+2s}}} + \|f\|_2^2 \varepsilon^{m^2/2} \\
&\le (2mt_\varepsilon)^{4s/1+2s} \|f\|^2_{W_{\frac{2}{1+2s}}} + \|f\|_2^2 t_\varepsilon^2.
\end{aligned}$$

We have used here the concentration property of the Gaussian distribution and the fact that $m^2 \ge 4$.

Next,

$$\begin{aligned}
B := B_1 + B_2 = \mathbb{E} \sum_{j<j_\varepsilon,k} &\big[\varpi_{jk}(\varepsilon)^2 (y_{j,k} - \beta_{jk})^2 + (1 - \varpi_{jk}(\varepsilon))^2 \beta_{jk}^2 \big]\, I\{|y_{j,k}| > mt_\varepsilon\} \\
&\times \big[I\{|\beta_{jk}| \le mt_\varepsilon/2\} + I\{|\beta_{jk}| > mt_\varepsilon/2\} \big].
\end{aligned}$$

For B_1 we use the Schwarz inequality:

$$\begin{aligned}
&\mathbb{E}(y_{j,k} - \beta_{jk})^2 I\{|y_{j,k} - \beta_{jk}| > mt_\varepsilon/2\} \\
&\qquad \le \big(\mathbb{P}(|y_{j,k} - \beta_{jk}| > mt_\varepsilon/2) \big)^{1/2} (\mathbb{E}(y_{j,k} - \beta_{jk})^4)^{1/2}.
\end{aligned}$$

Now, observing that $\mathbb{E}(y_{j,k} - \beta_{jk})^4 = 3\varepsilon^4$ and that $\mathbb{P}(|y_{j,k} - \beta_{jk}| > mt_\varepsilon/2) \le \varepsilon^{\frac{m^2}{8}}$, we have for $m^2 \ge 32$:

$$\begin{aligned}
B_1 &\le \sqrt{3} \sum_{j<j_\varepsilon,k} \varepsilon^2 I\{|\beta_{jk}| \le mt_\varepsilon/2\} \varepsilon^{\frac{m^2}{16}} + \sum_{j<j_\varepsilon,k} \beta_{jk}^2 I\{|\beta_{jk}| \le mt_\varepsilon/2\} \\
&\le 2\sqrt{3} S_\psi t_\varepsilon^2 + \Big(\frac{m}{2} t_\varepsilon \Big)^{4s/1+2s} \|f\|^2_{W_{\frac{s}{1+2s}}}.
\end{aligned}$$

For B_2, we use (3.2) to obtain

$$\begin{aligned}
B_2 &= \mathbb{E} \sum_{j<j_\varepsilon,k} \big[\varpi_{jk}(\varepsilon)^2 (y_{j,k} - \beta_{jk})^2 + (1 - \varpi_{jk}(\varepsilon))^2 \beta_{jk}^2 \big] \\
&\quad \times I\{|y_{j,k}| > mt_\varepsilon\} I\{|\beta_{jk}| > mt_\varepsilon/2\} \\
&\le \sum_{j<j_\varepsilon,k} \varepsilon^2 I\{|\beta_{jk}| > mt_\varepsilon/2\} + B_3 \le \frac{4m^{-2/1+2s}}{(1 - 2^{-2/1+2s})} \|f\|^2_{W_{\frac{2}{1+2s}}} t_\varepsilon^{4s/1+2s} + B_3,
\end{aligned}$$

where

$$B_3 := \sum_{j<j_\varepsilon,k} \mathbb{E}(1-\varpi_{jk}(\varepsilon))^2\beta_{jk}^2\ I\{|y_{j,k}|>mt_\varepsilon\}I\{|\beta_{jk}|>mt_\varepsilon/2\}$$
$$\times\big[I\{|y_{j,k}|\geq|\beta_{jk}|/2\}+I\{|y_{j,k}|<|\beta_{jk}|/2\}\big] := B_3'+B_3'',$$
$$B_3'' \leq \sum_{j<j_\varepsilon,k}\beta_{jk}^2\mathbb{P}\big(|y_{j,k}-\beta_{jk}|\geq mt_\varepsilon/4\big)\leq \|f\|_2^2t_\varepsilon^2,$$

since $m^2\geq 64$. We have used in the line above the concentration property of the Gaussian distribution. Now using (5.6) and (3.2), we get,

$$\begin{aligned}B_3' &\leq \sum_{j<j_\varepsilon,k}\mathbb{E}\beta_{jk}^2(1-\varpi_{jk}(\varepsilon))^2 I\{|y_{j,k}|\geq|\beta_{jk}|/2\}I\{|\beta_{jk}|>mt_\varepsilon/2\}I\{|y_{j,k}|\geq mt_\varepsilon\}\\ &\leq \sum_{j<j_\varepsilon,k}\mathbb{E}\beta_{jk}^2K(\varpi)^2\left(\frac{t_\varepsilon}{|y_{j,k}|}+t_\varepsilon\right)^2 I\{|y_{j,k}|\geq|\beta_{jk}|/2\}I\{|\beta_{jk}|>mt_\varepsilon/2\}\\ &\leq K(\varpi)^2\frac{32m^{-2/1+2s}}{1-2^{-2/1+2s}}\|f\|^2_{W_{\frac{2}{1+2s}}}t_\varepsilon^{4s/1+2s}+2K(\varpi)^2\|f\|_2^2t_\varepsilon^2. \qquad\square\end{aligned}$$

Proof of Theorem 3. We shall prove that under our assumption the LVGP rules satisfy Assumptions (5.4), (5.5), and (5.6). Assumption (5.4) is checked obviously. Note that we already remarked in Sec. 2.1 that they are shrinkage rules. Now, fix $m\geq 8$ and assume that $|y_{j,k}|\leq mt_\varepsilon$. Then,

$$\begin{aligned}\eta_{jk} &= \frac{1}{w(\varepsilon)}\frac{\sqrt{\varepsilon^2+\tau(\varepsilon)^2}}{\varepsilon}\exp\left(-\frac{\tau(\varepsilon)^2y_{j,k}^2}{2\varepsilon^2(\varepsilon^2+\tau(\varepsilon)^2)}\right)\\ &\geq \frac{1}{w(\varepsilon)}t_\varepsilon^{-1/2}\exp\left(-\frac{m^2t_\varepsilon^2}{2\varepsilon^2}\right)\geq \varepsilon^{\frac{m^2}{2}-\frac{1}{2}}\frac{1}{w(\varepsilon)}(\log(1/\varepsilon))^{-1/4}.\end{aligned}$$

- If $q_2>\frac{m^2-1}{2}$, for ε small enough, $\eta_{jk}\geq 1$ and $\breve{\beta}_{jk}=0$.
- If $q_2\geq\frac{m^2+1}{2}$, for ε small enough, $\eta_{jk}\geq t_\varepsilon^{-1}$ and $\frac{b_j}{1+\eta_{jk}}\leq t_\varepsilon$.

So, Assumption (5.5) is checked for both rules. Now, let us prove Assumption (5.6). Fix a constant $M\geq\sqrt{6+4q_1}$. We assume $|y_{j,k}|>Mt_\varepsilon$. Then, for ε small enough,

$$\begin{aligned}\eta_{jk} &= \frac{1}{w(\varepsilon)}\frac{\sqrt{\varepsilon^2+\tau(\varepsilon)^2}}{\varepsilon}\exp\left(-\frac{\tau(\varepsilon)^2y_{j,k}^2}{2\varepsilon^2(\varepsilon^2+\tau(\varepsilon)^2)}\right)\\ &\leq \frac{1}{w(\varepsilon)}\frac{\sqrt{\varepsilon^2+\tau(\varepsilon)^2}}{\varepsilon}\varepsilon^{\frac{M^2}{4}}\leq\frac{1}{w(\varepsilon)}t_\varepsilon^{-1/2}\varepsilon^{\frac{M^2}{4}}\leq t_\varepsilon.\end{aligned}$$

Consider first the posterior median. Using the above inequality, we have for ε small enough, and for any $j<j_\varepsilon$ and any k,

$$\varepsilon\sqrt{b_j}\Phi^{-1}\left(\frac{1+\min(\eta_{jk},1)}{2}\right)\leq t_\varepsilon.$$

So,

$$\begin{aligned}|y_{j,k}-\breve\beta_{jk}| &= |y_{j,k}-\breve\beta_{jk}|I\{|y_{j,k}|>Mt_\varepsilon\}+|y_{j,k}-\breve\beta_{jk}|I\{|y_{j,k}|\le Mt_\varepsilon\}\\ &\le \big((1-b_j)|y_{j,k}|+t_\varepsilon\big)I\{|y_{j,k}|>Mt_\varepsilon\}+2|y_{j,k}|I\{|y_{j,k}|\le Mt_\varepsilon\}\\ &\le t_\varepsilon|y_{j,k}|+(1+2M)t_\varepsilon,\end{aligned}$$

which implies (5.6) for the posterior median. Now, let us deal with the posterior mean. For ε small enough, and for any $j<j_\varepsilon$ and any k,

$$\begin{aligned}|y_{j,k}-\tilde\beta_{jk}| &= |y_{j,k}-\tilde\beta_{jk}|I\{|y_{j,k}|>Mt_\varepsilon\}+|y_{j,k}-\tilde\beta_{jk}|I\{|y_{j,k}|\le Mt_\varepsilon\}\\ &\le \left(1-\frac{b_j}{1+\eta_{jk}}\right)|y_{j,k}|I\{|y_{j,k}|>Mt_\varepsilon\}+2|y_{j,k}|I\{|y_{j,k}|\le Mt_\varepsilon\}\\ &\le (1-b_j+\eta_{jk})|y_{j,k}|I\{|y_{j,k}|>Mt_\varepsilon\}+2|y_{j,k}|I\{|y_{j,k}|\le Mt_\varepsilon\}\\ &\le 2t_\varepsilon|y_{j,k}|+2Mt_\varepsilon,\end{aligned}$$

which implies (5.6) for the posterior mean.

Assumptions (5.4), (5.5), and (5.6) are checked for both rules, which finally proves that their maxiset contains $\mathcal{B}_{2,\infty}^{\frac{s}{1+2s}}\cap W_{\frac{2}{1+2s}}$ for the rate

$$t_\varepsilon^{4s/(1+2s)}=(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}.$$

We prove now the reverse inclusion:

$$MS\big(f_\varepsilon^0,\|\cdot\|_2^2,(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big)\subset\mathcal{B}_{2,\infty}^{s/(2s+1)}\cap W_{\frac{2}{2s+1}}.$$

Observe that $\beta^0_{jk}=0$ when $j\ge j_\varepsilon$, which implies

$$\sum_{j>j_\varepsilon,k}\beta_{jk}^2\le\mathbb{E}\|f_\varepsilon^0-f\|_2^2\le ct_\varepsilon^{\frac{4s}{1+2s}}\le c2^{-j_\varepsilon\frac{2s}{1+2s}}.$$

Letting ε vary, we obtain the characterization (3.1), which proves that

$$MS\big(f_\varepsilon^0,\|\cdot\|_2^2,(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big)\subset\mathcal{B}_{2,\infty}^{s/(2s+1)}.$$

If we remember that $|y_{jk}|\le mt_\varepsilon$ implies $0\le\beta^0_{jk}/y_{jk}\le ct_\varepsilon$ (Assumption (5.5)), we have for $f\in MS(f_\varepsilon^0,\|\cdot\|_2^2,(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)})(M)$:

$$\begin{aligned}&(1-ct_\varepsilon)^2\sum_{j,k}\beta_{jk}^2I\{|\beta_{jk}|\le mt_\varepsilon\}\\ &\quad=2(1-ct_\varepsilon)^2\sum_{j,k}\beta_{jk}^2\Big[\mathbb{P}(y_{jk}-\beta_{jk}<0)I\{\beta_{jk}\ge0\}+\mathbb{P}(y_{jk}-\beta_{jk}>0)I\{\beta_{jk}<0\}\Big]\\ &\qquad\qquad\times I\{|\beta_{jk}|\le mt_\varepsilon\}\\ &\quad\le 2\mathbb{E}\sum_{j,k}\Big[(\beta_{jk}-\beta^0_{jk})^2I\{\beta_{jk}\ge0\}+(\beta_{jk}-\beta^0_{jk})^2I\{\beta_{jk}<0\}\Big]I\{|\beta_{jk}|\le mt_\varepsilon\}\\ &\quad\le 2\mathbb{E}\sum_{j,k}(\beta_{jk}-\beta^0_{jk})^2\le 2M\,(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}.\end{aligned}$$

Hence $\sup_{\lambda>0}\lambda^{-\frac{4s}{2s+1}}\sum_{j\ge-1}\sum_k\beta_{jk}^2I\{|\beta_{jk}|\le\lambda\}<\infty$ and f belongs to $W_{\frac{2}{2s+1}}$. $\square$

6. Simulations

Dealing with the prior model (1.1), we compare in this section the performance of both LVGP rules described in the previous section, in (5.1) and (5.2), with many other procedures: the thresholding rules of Donoho and Johnstone [11] called VisuShrink and of Nason [22] called GlobalSure, the ParetoThresh (with $p = 1.3$) proposed by Rivoirard [24] built using Pareto priors and hyperparameters as well as the Bayesian procedures of Abramovich *et al.* [2] denoted as BayesThresh and those proposed by Johnstone et Silverman [14] and implemented by Antoniadis *et al.* [4] built with the heavy-tailed Laplace prior with scale factor $\alpha = 0.5$ (LaplaceBayesMedian, LaplaceBayesMean) and with the heavy-tailed quasi-Cauchy prior (CauchyBayesMedian, CauchyBayesMean). For this purpose, we use the mean-squared error in the following regression model.

6.1. Model and discrete wavelet transform. Consider the standard regression problem:

$$g_i = f\left(\frac{i}{n}\right) + \sigma\varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \qquad 1 \le i \le n, \tag{6.1}$$

where $n = 1024$. We introduce the discrete wavelet transform $d := \mathcal{W}f^0$ (denoted DWT) of the vector $f_0 = (f(\frac{i}{n}),\, 1 \le i \le n)^T$. 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 [20]. The DWT provides n discrete wavelet coefficients d_{jk}, $-1 \le j \le N-1$, $k \in \mathcal{I}_j$. They are related to the wavelet coefficients β_{jk} of f by the simple relation

$$d_{jk} \approx \beta_{jk}\sqrt{n}.$$

Using the DWT, the regression model (6.1) is reduced to the following one:

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

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

In the sequel, we suppose that σ is known. Nevertheless, it could be robustly 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 [11]).

To implement the LVGP rules, we reconstruct the d_{jk}'s, as 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 $\omega_n = \omega^* = 10(\frac{\sigma}{\sqrt{n}})^q$ $(q > 0)$, $\delta(0)$ is a point mass at zero, 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 such that $\frac{n\tau_n^2}{\sigma^2+n\tau_n^2} = 0.999$.

Dealing with this prior model, we denote *GaussMedian* and *GaussMean* the LVGP rules described in (5.1) and (5.2), respectively.

The Symmlet 8 wavelet basis (as described on p. 198 of Daubechies [10]) is used for all the methods of reconstruction.

6.2. Simulations and discussion. In Table 6.1 we measure the performance of both estimators by using the four test functions: "Blocks", "Bumps", "Heavisine" and "Doppler" by using the mean-squared error defined by:

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

Remark. Recall that the test functions have been chosen by Donoho *et al.* [12] to represent a large variety of inhomogeneous signals.

Table 6.1 shows the average mean-squared error (denoted AMSE) using 100 replications for VisuShrink, GlobalSure, ParetoThresh, BayesThresh, GaussMedian, GaussMean (for $q = 1$), LaplaceBayesMedian, LaplaceBayesMean, CauchyBayesMedian and CauchyBayesMean, with different values for the root signal to noise ratio (RSNR).

The results provided below can be summarized as follows:

- According to Table 6.1, we remark that "purely Bayesian" procedures (BayesThresh, GaussMedian, GaussMean, CauchyBayesMedian, CauchyBayesMean, LaplaceBayesMedian and LaplaceBayesMean) are preferable to "purely deterministic" ones (VisuShrink and GlobalSure) under the AMSE approach for inhomogeneous signals.
- We observe that Bayesian rules using the posterior mean (GaussMean, LaplaceBayesMean and CauchyBayesMean) have better performance than those using the posterior median (GaussMedian, LaplaceBayesMedian and CauchyBayesMedian).
- CauchyBayesMean provides the best behavior here since its AMSEs are globally the smallest (11 times out of 12).
- GaussMean shows the performance which is rather close to CauchyBayesMean. It outperforms BayesThresh 11 times out of 12. This confirms our maxiset previous results, and shows that GaussMean is an excellent choice if we take into account the performance as well as the computation time.

In the sequel, we present some simulations of the Bayesian rules using Gaussian priors (Figs. 6.1 and 6.2) and heavy-tailed priors (Figs. 6.3 and 6.4) when RSNR = 5.

In Figure 6.1, we note that in both Bayesian procedures some high-frequency artefacts appear. However, these artefacts disappear if we take large values of q. Figure 6.2 shows an example of reconstructions using GaussMedian and GaussMean when the RSNR is equal to 5 ($\sigma = 7/5$) for different values of q.

As we can see in Figure 6.2, the artefacts are less numerous when q increases. But this improvement has a cost: in general the AMSE increases when q is close to 0 or strictly greater than 1. Consequently, the value $q = 1$ appears to be a good compromise to obtain good reconstruction and good AMSE with the GaussMedian and GaussMean procedures.

7. More on Maxisets of 'Small Variance Gaussian Priors'

In a minimax setting, Abramovich *et al.* [1] obtained the following result.

TABLE 6.1. AMSEs for VisuShrink, GlobalSure, ParetoThresh, BayesThresh, GaussMedian, GaussMean, LaplaceBayesMedian, LaplaceBayesMean, CauchyBayesMedian and CauchyBayesMean, with various test functions and various values of the RSNR

RSNR=5	Blocks	Bumps	Heavisine	Doppler
VisuShrink	2.08	2.99	0.17	0.77
GlobalSure	0.82	0.92	0.18	0.59
ParetoThresh	0.73	0.85	0.15	0.36
BayesThresh	0.67	0.74	0.15	0.30
GaussMedian	0.72	0.76	0.20	0.30
GaussMean	0.62	0.68	0.19	0.29
LaplaceBayesMedian	0.59	0.69	0.14	0.30
LaplaceBayesMean	0.56	0.65	**0.13**	0.28
CauchyBayesMedian	0.60	0.67	0.14	0.29
CauchyBayesMean	**0.55**	**0.63**	**0.13**	**0.27**
RSNR=7	Blocks	Bumps	Heavisine	Doppler
VisuShrink	1.29	1.77	0.12	0.47
GlobalSure	0.42	0.48	0.12	0.21
ParetoThresh	0.40	0.46	0.09	0.21
BayesThresh	0.38	0.45	0.10	0.16
GaussMedian	0.41	0.42	0.12	**0.15**
GaussMean	0.35	0.38	0.11	**0.15**
LaplaceBayesMedian	0.33	0.37	0.09	0.17
LaplaceBayesMean	0.31	0.36	**0.08**	0.16
CauchyBayesMedian	0.32	0.36	0.09	0.17
CauchyBayesMean	**0.29**	**0.34**	**0.08**	**0.15**
RSNR=10	Blocks	Bumps	Heavisine	Doppler
VisuShrink	0.77	1.04	0.08	0.27
GlobalSure	0.25	0.29	0.08	0.11
ParetoThresh	0.21	0.25	0.06	0.12
BayesThresh	0.22	0.25	0.06	0.09
GaussMedian	0.21	0.23	0.06	0.08
GaussMean	0.18	0.20	0.06	**0.07**
LaplaceBayesMedian	0.17	0.20	**0.05**	0.09
LaplaceBayesMean	0.17	0.19	**0.05**	0.09
CauchyBayesMedian	0.17	0.19	**0.05**	0.09
CauchyBayesMean	**0.16**	**0.18**	**0.05**	0.09

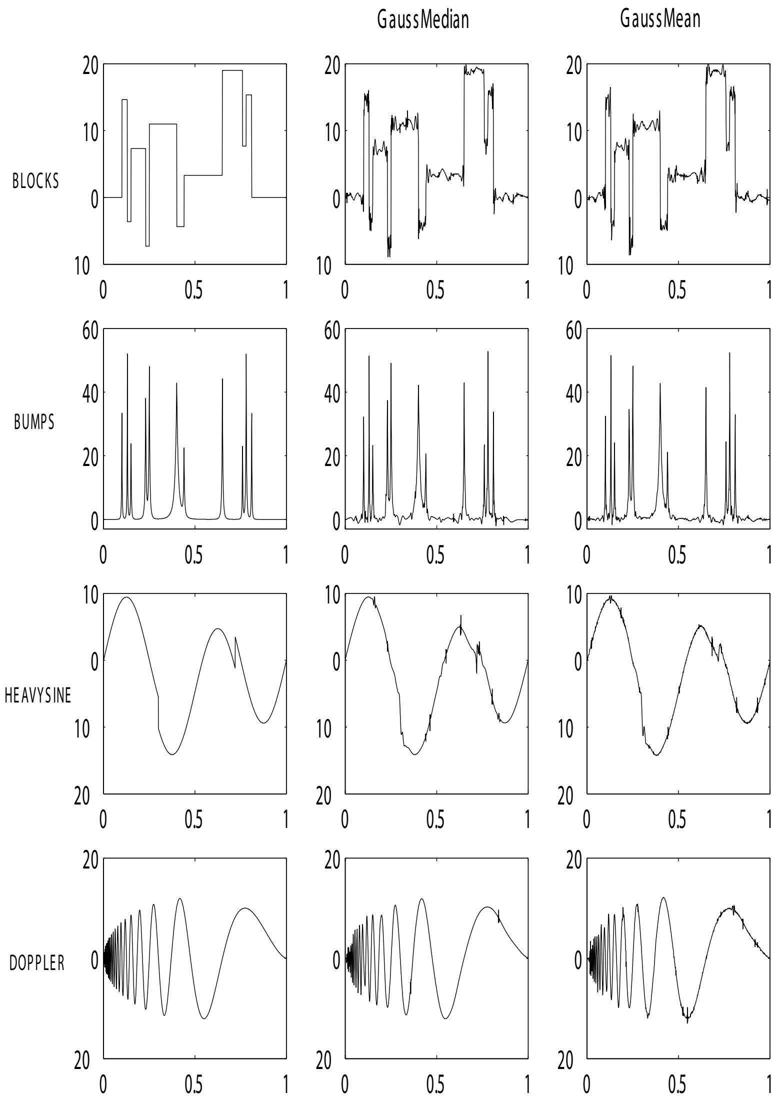

FIGURE 6.1. Original test functions and reconstructions using GaussMedian and GaussMean with $q = 1$ (RSNR=5)

Theorem 4. *Let β^0 be $\breve{\beta}$ or $\tilde{\beta}$. With $\alpha = 2s+1$ and any $0 \leq b < 1$, there exist two positive constants C_1 and C_2 such that $\forall\, \varepsilon > 0$,*

$$C_1(\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(2s+1)} \leq \sup_{\beta \in \mathcal{B}^s_{2,\infty}(M)} \mathbb{E}\|\beta^0 - \beta\|_2^2 \leq C_2 \log(1/\varepsilon)\varepsilon^{4s/(2s+1)}.$$

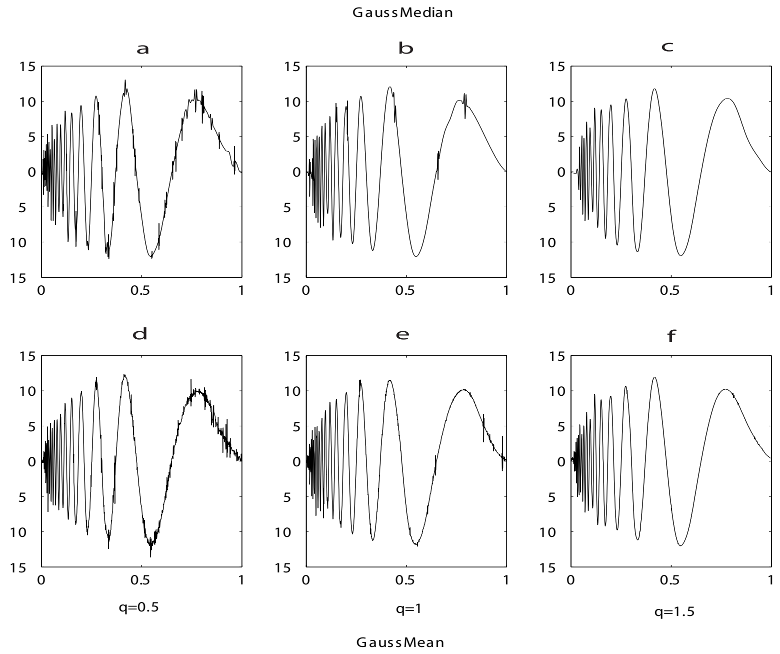

FIGURE 6.2. Reconstructions with GaussMedian (schemes a, b, c) and GaussMean (schemes d, e, f) for various values of q when RSNR=5; a: AMSE=0.37; b: AMSE=0.30; c: AMSE=0.33; d: AMSE=0.39; e: AMSE=0.29; f: AMSE=0.30

So, the posterior mean and median achieve the optimal rate up to an unavoidable logarithmic term. Now, let us consider the maxiset setting.

Theorem 5. *For $s > 0$, $\alpha = 2s + 1$, any $0 \le b < 1$, and if β^0 is $\breve{\beta}$ or $\tilde{\beta}$,*

1. *for the rate $\varepsilon^{4s/(1+2s)}$,*

$$MS\big(\beta^0, \|\cdot\|_2^2, \varepsilon^{4s/(1+2s)}\big) \subsetneq \mathcal{B}^s_{2,\infty}.$$

2. *For the rate $\big(\varepsilon\sqrt{\log(1/\varepsilon)}\big)^{4s/(1+2s)}$,*

$$MS\big(\beta^0, \|\cdot\|_2^2, (\varepsilon\sqrt{\log(1/\varepsilon)})^{4s/(1+2s)}\big) \subset \mathcal{B}^{*s}_{2,\infty},$$

with

$$\mathcal{B}^{*s}_{2,\infty} = \Big\{f \in \mathbb{L}^2 \colon \sup_{J>0} 2^{2Js} J^{-2s/(1+2s)} \sum_{j \ge J} \sum_k \beta_{jk}^2 < \infty\Big\}.$$

3. *For the rate $\varepsilon^{4s/(1+2s)} \log(1/\varepsilon)$,*

$$\mathcal{B}^s_{2,\infty} \subset MS\big(\beta^0, \|\cdot\|_2^2, \varepsilon^{4s/(1+2s)} \log(1/\varepsilon)\big).$$

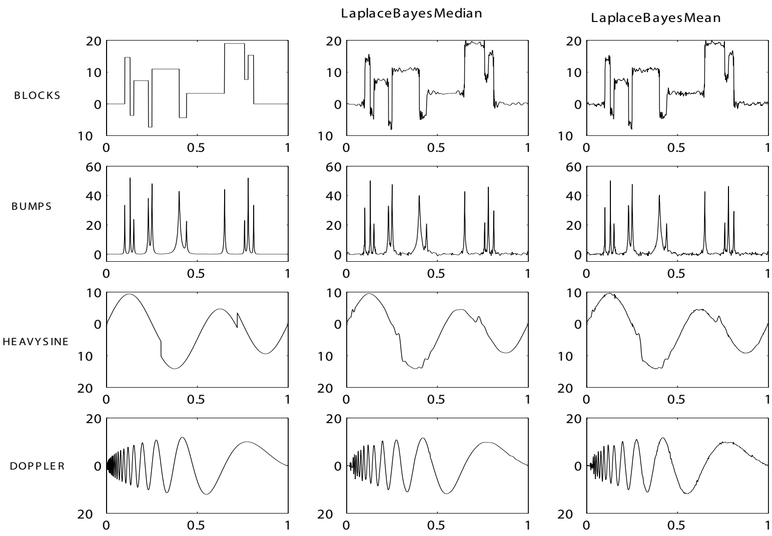

FIGURE 6.3. Original test functions and reconstructions using LaplaceBayesMedian and LaplaceBayesMean (RSNR=5)

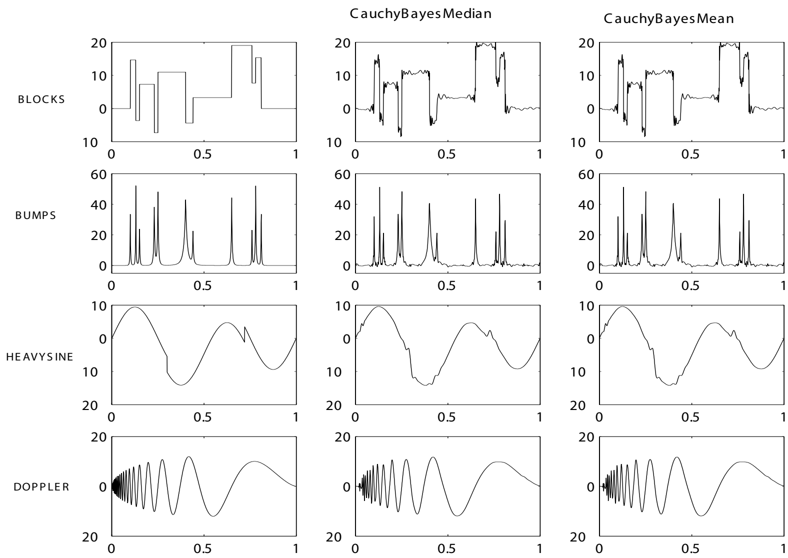

FIGURE 6.4. Original test functions and reconstructions using CauchyBayesMedian and CauchyBayesMean (RSNR=5)

Proof. Let us first prove the inclusion

$$MS\big(\beta^0, \|\cdot\|_2^2, \varepsilon^{4s/(1+2s)}\big) \subset \mathcal{B}^s_{2,\infty}.$$

For this, note that $\lambda_\varepsilon = (c_1^{-1}\varepsilon^2)^{1/\alpha}$. We observe that if $2^{-j} \leq \lambda_\varepsilon$, then $b_j \leq 1/2$ and

$$|\beta^0_{jk}| \leq \frac{1}{2}|y_{jk}|.$$

Since $y_{jk}\beta^0_{jk} \geq 0$, if $2^{-j} \leq \lambda_\varepsilon$,

$$\mathbb{E}\beta_{jk}^2 I\{\beta_{jk} \geq 0\} I\{y_{jk} < \beta_{jk}\} \leq 4\mathbb{E}(\beta^0_{jk} - \beta_{jk})^2 I\{\beta_{jk} \geq 0\} I\{y_{jk} < \beta_{jk}\},$$

and

$$\mathbb{E}\beta_{jk}^2 I\{\beta_{jk} < 0\} I\{y_{jk} > \beta_{jk}\} \leq 4\mathbb{E}(\beta^0_{jk} - \beta_{jk})^2 I\{\beta_{jk} < 0\} I\{y_{jk} > \beta_{jk}\}.$$

Therefore, since $\mathbb{P}(y_{j,k} - \beta_{jk} < 0) = \mathbb{P}(y_{j,k} - \beta_{jk} > 0) = 1/2$, whenever $f \in MS(\beta^0, \|\cdot\|_2^2, \varepsilon^{4s/(1+2s)})(M)$, we have:

$$\begin{aligned}
&\sum_{j,k} \beta_{jk}^2 I\{2^{-j} \leq \lambda_\varepsilon\} \\
&\quad = 2\sum_{j,k} \beta_{jk}^2 \Big[\mathbb{P}(y_{j,k} - \beta_{jk} < 0) I\{\beta_{jk} \geq 0\} + \mathbb{P}(y_{j,k} - \beta_{jk} > 0) I\{\beta_{jk} < 0\}\Big] \\
&\qquad\qquad \times I\{2^{-j} \leq \lambda_\varepsilon\} \\
&\quad \leq 8\mathbb{E}\sum_{j,k} \Big[(\beta^0_{jk} - \beta_{jk})^2 I\{\beta_{jk} \geq 0\} + (\beta^0_{jk} - \beta_{jk})^2 I\{\beta_{jk} < 0\}\Big] I\{2^{-j} \leq \lambda_\varepsilon\} \\
&\quad \leq 8\mathbb{E}\sum_{j,k} (\beta^0_{jk} - \beta_{jk})^2 \leq 8M\ \varepsilon^{4s/(1+2s)}.
\end{aligned}$$

Since $\alpha = 2s + 1$, we deduce

$$\sup_{J \geq -1} 2^{2Js} \sum_{j \geq J} \sum_k \beta_{jk}^2 \leq 8Mc_1^{2s/(1+2s)},$$

and f belongs to $\mathcal{B}^s_{2,\infty}$. To prove that the inclusion is strict, we just use Theorem 4. The second inclusion is easily obtained by using similar arguments. Finally, the proof of the last one is provided by Theorem 4. □

As recalled in Sec. 3.2, for the rates $\varepsilon^{4s/(1+2s)}$, the maxisets of linear estimates are exactly Besov spaces $\mathcal{B}^s_{2,\infty}$. So Theorem 5 shows that the Bayesian procedures built by Abramovich *et al.* [1] are outperformed by linear estimates for polynomial rates of convergence. Furthermore, these procedures cannot achieve the same performance as classical nonlinear procedures, since we have the following result.

Proposition 4. *For any $s > 0$,*

$$\mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}} \not\subset \mathcal{B}^{*s}_{2,\infty}.$$

Proof. To prove this result, we build a *sparse* function belonging to $\mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}}$ but not to $\mathcal{B}^{*s}_{2,\infty}$. Let us consider $f = \sum_{j,k} \beta_{jk}\psi_{jk}$, where at each level j, 2^{jn} wavelet coefficients take the value $2^{-j\beta}$, whereas the other ones are equal to 0, with $0 \le n \le 1$ and $\beta > n/2$ (so $f \in \mathbb{L}_2$). For any $J \ge 1$,

$$\begin{aligned} 2^{2Js} J^{-2s/(1+2s)} \sum_{j\ge J}\sum_k \beta_{jk}^2 &= 2^{2Js} J^{-2s/(1+2s)} \sum_{j \ge J} 2^{nj} 2^{-2j\beta} \\ &\ge 2^{J(2s+n-2\beta)} J^{-2s/(1+2s)}. \end{aligned}$$

So,

$$n - 2\beta + 2s > 0 \Rightarrow f \notin \mathcal{B}^{*s}_{2,\infty}. \tag{7.1}$$

Similarly,

$$n - 2\beta + 2s/(1+2s) \le 0 \Rightarrow f \in \mathcal{B}^{s/(2s+1)}_{2,\infty}. \tag{7.2}$$

And

$$\begin{aligned} \lambda^{-4s/(1+2s)} \sum_{j,k} \beta_{jk}^2 I\{|\beta_{jk}| \le \lambda\} &= \lambda^{-4s/(1+2s)} \sum_j 2^{jn} 2^{-2j\beta} I\{2^{-j\beta} \le \lambda\} \\ &\le \lambda^{-4s/(1+2s) - n/\beta + 2}. \end{aligned}$$

So,

$$n - 2\beta + 2ns \le 0 \Rightarrow f \in W_{\frac{2}{2s+1}}. \tag{7.3}$$

As soon as $n < 1$ (which yields that the signal is sparse), it is then possible to choose $\beta > n/2$ such that (7.1), (7.2), and (7.3) hold. So, f belongs to $\mathcal{B}^{s/(2s+1)}_{2,\infty} \cap W_{\frac{2}{2s+1}}$ but not to $\mathcal{B}^{*s}_{2,\infty}$. □

References

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

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

[3] A. Antoniadis, J. Bigot, and Sapatinas, *Wavelet estimators in nonparametric regression: a comparative simulation study*, J. Statist. Software, **6** (2001), 1–83.

[4] A. Antoniadis, M. Jansen, I. Johnstone, and B. Silverman, Ebayesthresh: Matlab software for empirical Bayes thresholding, 2000
http://www-lmc.imag.fr/lmc-sms/Anestis.Antoniadis/EBayesThresh.

[5] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, *Adaptive Bayesian wavelet shrinkage*, J. Amer. Statist. Assoc., **92** (1997), 1413–1421.

[6] M. Clyde and E. I. George, *Robust empirical Bayes estimation in wavelets. Restricted nonlinear approximation*, discussion paper, 1998.

[7] M. Clyde and E. I. George, *Flexible empirical Bayes estimation for wavelets*, J. Roy. Statist. Soc. Ser. B. Statist. Methodol., **62** (2000), 681–698.

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

[9] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard, *Maximal spaces with given rate of convergence for thresholding algorithms*, Appl. Comput. Harmon. Anal., **11** (2001), 167–191.

[10] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[11] D. I. Donoho and I. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, **81** (1994), 425–455.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: asymptotia*? J. Roy. Statist. Soc. Ser. B., **57** (1995), 301–369. With discussion and a reply by the authors.

[13] I. Johnstone, *Minimax Bayes, asymptotic minimax and sparse wavelet priors.* In: *Statistical Decision Theory and Related Topics*, V (West Lafayette, IN, 1992), pp. 303-326. Springer, New-York, 1994.

[14] I. M. Johnstone and B. W. Silverman, *Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences*, Ann. Statist., **32** (2004), 1594–1649.

[15] I. M. Johnstone and B. W. Silverman, *Empirical Bayes approaches to mixture problems and wavelet regression*, Ann. Statist. (2005). (To be published.)

[16] I. M. Johnstone and B. W. Silverman, *Empirical Bayes selection of wavelet thresholds*, Ann. Statist., **33** (2005), 1700–1752.

[17] G. Kerkyacharian and D. Picard, *Density estimation by kernel and wavelets methods: optimality of Besov spaces*, Statist. Probab. Lett., **18** (1993), 327–336.

[18] G. Kerkyacharian and D. Picard, *Thresholding algorithms, maxisets and well concentrated bases*, Test, **9** (2000), 283–344. With comments and a rejoinder by the authors.

[19] G. Kerkyacharian and D. Picard, *Minimax or maxisets*? Bernoulli, **8** (2002), 219–253.

[20] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press Inc., San Diego, CA, 1998.

[21] Y. Meyer, *Ondelettes et opérateurs*, Actualités Mathématiques, Hermann, Paris, 1990.

[22] G. P. Nason, *Wavelet shrinkage using cross-validation*, J. Roy. Statist. Syst. Sci. B, **23** (1996), 1–11.

[23] V. Rivoirard, *Maxisets for linear procedures*, Statist. Probab. Lett., **67** (2004), 267–275.

[24] V. Rivoirard, *Thresholding Procedure with Priors Based on Pareto Distributions*, Test, **13** (2004), 213–246.

[25] V. Rivoirard, *Bayesian modelling of sparse sequences and maxisets for Bayes rules*, Math. Methods Statist., **14** (2005), 346–376.

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

[27] C.-H. Zhang, *General empirical Bayes wavelet methods and exactly adaptive minimax estimation*, Ann. Statist., **33** (2005), 54–100.

[*Received June* 2006]