Electronic Journal of Statistics
Vol. 3 (2009) 1084–1112
ISSN: 1935-7524
DOI: 10.1214/09-EJS466

Regression in random design and Bayesian warped wavelets estimators

Thanh Mai Pham Ngoc

Laboratoire de Probabilités et Modèles aléatoires
Université Paris 6, case 188, 4, Pl. Jussieu, 75252 Paris
France
e-mail: thanh.pham_ngoc@upmc.fr

Abstract: In this paper we deal with the regression problem in a random design setting. We investigate asymptotic optimality under minimax point of view of various Bayesian rules based on warped wavelets. We show that they nearly attain optimal minimax rates of convergence over the Besov smoothness class considered. Warped wavelets have been introduced recently, they offer very good computable and easy-to-implement properties while being well adapted to the statistical problem at hand. We particularly put emphasis on Bayesian rules leaning on small and large variance Gaussian priors and discuss their simulation performances, comparing them with a hard thresholding procedure.

AMS 2000 subject classifications: Primary 62G05, 62G08, 62G20, 62C10.
Keywords and phrases: Nonparametric regression, random design, warped wavelets, Bayesian methods.

Received August 2009.

Contents

1. Introduction

We observe independent pairs of variables (X_i, Y_i), for $i = 1, \dots, n$, under a random design regression model:

$$Y_i = f(X_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where f is an unknown regression function that we aim at estimating, and ε_i are independent normal errors with $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$. The design points X_i are assumed to be supported in the interval $[0, 1]$ and have a density g which will be supposed to be known. Furthermore we assume that the design density g is bounded from below, i.e. $0 < m \leq g$, where m is a constant. Many approaches have been proposed to tackle the problem of regression in random design, we mention among others the work of Hall and Turlach [17], Kovac and Silverman [22], Antoniadis et *al*. [4], Cai and Brown [8] and the model selection point of view adopted by Baraud [6].

The present paper provides a Bayesian approach to this problem based on *warped* wavelet basis. Warped wavelet basis $\{\psi_{jk}(G)\ j \geq -1, k \in \mathbb{Z}\}$ in regression with random design were recently introduced by Kerkyacharian and Picard in [20]. The authors proposed an approach which would depart as little as possible from standard wavelet thresholding procedures which enjoy optimality and adaptivity properties. These procedures have been largely investigated in the case of equispaced samples (see a series of pioneering articles by Donoho et al. [14], [15], [13]). Kerkyacharian and Picard actually pointed out that expanding the unknown regression function f in the warped basis instead of the standard wavelet basis could be very interesting. Of course, this basis has no longer the orthonormality property nonetheless it behaves under some conditions as standard wavelets. Kerkyacharian and Picard investigated the properties of this new basis. They showed that not only is it well adapted to the statistical problem at hand by avoiding unnecessary calculations, but it also offers very good theoretical features while being easily implemented. More recently Brutti [7] highlighted their easy-to-implement computational properties.

The novelty of our contribution lies in the combination of Bayesian techniques and *warped* wavelets to treat regression in random design. We actually want to investigate whether this yields optimal theoretical results and promising pratical performances, which will prove to be the case. We do not deal with the case of an unknown design density g which requires further machinery and will be the object of another paper.

Bayesian techniques for shrinking wavelet coefficients have become very popular in the last few years. The majority of them were devoted to fixed design regression scheme. Let us cite among others, papers of Abramovich et al. [1], [2], Clyde et al. [10], [11], [12], [5], Chipman et al. [9], Rivoirard [25], Pensky [24] in the case of i.i.d errors not necessarily Gaussian.

Most of those works are taking as distribution prior a mixture of Gaussian distributions. In particular, Abramovich et al. in [1] and [2] have explored optimality properties of Gaussian prior mixed with a point mass at zero and which may be viewed as an extreme case of a Gaussian mixture:

$$\beta_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta(0),$$

where β_{jk} are the wavelet coefficients of the unknown regression function, $\tau_j^2 = c_1 2^{-j\alpha}$ and $\pi_j = \min(1, c_2 2^{-j\beta})$ are the hyperparameters. This particular form was devised to capture the sparsity of the expansion of the signal in the wavelet basis.

Our approach will consist in a first time in using the same prior but in the context of warped wavelets. In Theorem 1, we show that the Bayesian estimator built, using warped wavelets with this prior and this form of hyperparameters achieves the optimal minimax rate within logarithmic term, on the considered Besov functional space. Unfortunately, the Bayesian estimator turns out not to be adaptive. Indeed, the hyperparameters depend on the Besov smoothness class index. In order to compensate this drawback, Autin et al. in [5] suggested to consider Bayesian procedures based on Gaussian prior with large variance. Following this suggestion, we will consider priors still specified in terms of a normal density mixed with a point mass at zero but with large variance Gaussian densities. In Theorem 2, we prove again that the Bayesian estimator built with this latter form of prior, still combined with warped wavelets achieves nearly optimal minimax rate of convergence while being adaptive. Eventually, our simulation results highlight the very good performances and behaviour of these Bayesian procedures, whatever the regularity of the test functions, the noise level and the design density which can be far from the uniform case.

This paper is organized as follows. In section 2 some necessary methodology is given: we start with a short review of wavelets and warped wavelets, explain the prior model and discuss the two hyperparameters form we consider. We give in section 3 some definitions of the functional spaces we consider. In section 4, we investigate the performances of our Bayesian estimators in terms of minimax rates in two cases: the first one when the Gaussian prior has small variance, the second case focuses on Gaussian prior with large variance. Section 5 is devoted to simulation results and discussion. Finally, all proofs of main results are given in the Proofs section.

2. Methodology

2.1. Warped bases

Wavelet series are generated by dilations and translations of a function ψ called the mother wavelet. Let ϕ denote the orthogonal father wavelet function. The function ϕ and ψ are compactly supported. Assume ψ has r vanishing moments. Let:

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k), \quad j, k \in \mathbb{Z}$$

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathbb{Z}.$$

For a given square-integrable function f in $\mathbb{L}_2[0,1]$, let us denote

$$\zeta_{j,k} =< f, \psi_{j,k} > .$$

In this paper, we use decompositions of 1- periodic functions on wavelet basis of $\mathbb{L}_2[0,1]$. We consider periodic orthonormal wavelet bases on $[0,1]$ which allow to have the following series representation of a function f:

$$f(x) = \sum_{j\geq -1} \sum_{k=0}^{2^j-1} \zeta_{jk}\psi_{jk}(x) \tag{2.1}$$

where we have denoted $\psi_{-1,k} = \phi_{0,k}$ the scaling function.

We are now going to give the essential background of *warped* wavelets which were introduced in details in [20]. First of all, let us define

$$G(x) = \int_0^x g(x)dx. \tag{2.2}$$

G is assumed to be a known function, continuous and strictly monotone from $[0,1]$ to $[0,1]$.

Let us expand the regression function f in the following sense:

$$f(G^{-1})(x) = \sum_{j\geq -1} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x)$$

or equivalently

$$f(x) = \sum_{j\geq -1} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(G(x))$$

where

$$\beta_{jk} = \int f(G^{-1})(x)\psi_{jk}(x)dx = \int f(x)\psi_{jk}(G(x))g(x)dx.$$

Hence, one immediately notices that expanding $f(G^{-1})$ in the standard basis is equivalent to expand f in the new *warped* wavelet basis $\{\psi_{jk}(G), j \geq -1, k \in \mathbb{Z}\}$. This may give a natural explanation that in the follow-on, regularity conditions will be expressed not for f but for $f(G^{-1})$.

We set $\hat{\beta}_{jk} = (1/n)\sum_{i=1}^n \psi_{jk}(G(X_i))Y_i$. $\hat{\beta}_{jk}$ is an unbiased estimate of β_{jk} since

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_{jk}) &= (1/n)\sum_{i=1}^n \mathbb{E}(\psi_{j,k}(G(X_i))(f(X_i)+\epsilon_i)) = \mathbb{E}(\psi_{j,k}(G(X))f(X) \\
&= \int f(x)\psi_{jk}(G(x))g(x)dx = \int f(G^{-1})(x)\psi_{jk}(x)dx = \beta_{jk}.
\end{aligned}$$

2.2. Priors and estimators

We set in the following

$$\gamma_{jk}^2 = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \psi_{jk}^2(G(X_i)). \tag{2.3}$$

As in Abramovich et al. (see [1], [2]), we use the following prior on the wavelet coefficients β_{jk} of the unknown function f with respect to the *warped* basis $\{\psi_{jk}(G), j \geq -1, k \in \mathbb{Z}\}$:

$$\beta_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta(0).$$

Considering the $\mathbb{L}_1$ loss, from this form of prior we derive the following Bayesian rule which is the posterior median:

$$\tilde{\beta}_{jk} = Med(\beta_{jk}|\hat{\beta}_{jk}) = \operatorname{sign}(\hat{\beta}_{jk}) \max(0, \zeta_{jk}) \tag{2.4}$$

where

$$\zeta_{jk} = \frac{\tau_j^2}{\gamma_{jk}^2 + \tau_j^2}|\hat{\beta}_{jk}| - \frac{\tau_j \gamma_{jk}}{\sqrt{\gamma_{jk}^2 + \tau_j^2}} \Phi^{-1}\left(\frac{1 + \min(\eta_{jk}, 1)}{2}\right) \tag{2.5}$$

where Φ is the normal cumulative distributive function and

$$\eta_{jk} = \frac{1 - \pi_j}{\pi_j} \frac{\sqrt{\tau_j^2 + \gamma_{jk}^2}}{\gamma_{jk}} \exp\left(- \frac{\tau_j^2 \hat{\beta}_{jk}^2}{2\gamma_{jk}^2(\tau_j^2 + \gamma_{jk}^2)}\right). \tag{2.6}$$

We set:

$$w_j(n) := \frac{\pi_j}{1 - \pi_j}. \tag{2.7}$$

We introduce now the estimator of the unknown regression f

$$\tilde{f}(x) = \sum_{j \leq J} \sum_{k=0}^{2^j - 1} \tilde{\beta}_{jk} \psi_{jk}(G(x)), \tag{2.8}$$

where J is a parameter which will be specified later.

Note that in our case, the estimator resembles the usual ones in [5], [1] and [2], except that the deterministic noise variance has been replaced by a stochastic noise level γ_{jk}^2. Its expression is given by (2.3). This change will have a marked impact both on the proofs of theorems by using now large deviation inequalities and on simulation results.

Futhermore, such $\mathbb{L}_1$ rule is of thresholding type. Indeed, as underlined in [1] and [2], $\tilde{\beta}_{jk}$ is null whenever $\hat{\beta}_{jk}$ falls below a certain threshold λ_B. Some properties of the threshold λ_B that will be used in the sequel are given in lemma 1 in the Proofs section.

2.2.1. Gaussian priors with small variance

In this paper, two cases of hyperparameters will be considered. The first one involves Gaussian priors with small variances. We will state as suggested in Abramovich et al. (see [1], [2]):

$$\tau_j^2 = c_1 2^{-j\alpha} \quad \pi_j = \min(1, c_2 2^{-j\beta}), \tag{2.9}$$

where α and β are non-negative constants, $c_1, c_2 > 0$.

This choice of hyperparameters is exhaustively discussed in Abramovich et al. [2]. The authors stressed that this form of hyperparameters was actually designed in order to capture the sparsity of wavelet expansion. They pointed out the connection between Besov spaces parameters and this particular form of hyperparameters. They investigate various practical choices.

For this case of hyperparameters (2.9), the estimator of f will be denoted $\hat{f}$.

2.2.2. Gaussian priors with large variance

The second form of hyperparameters considered in the paper involves Gaussian priors with large variance as suggested in Autin et al. [5].

As a matter of fact, we suppose that the hyperparameters do not depend on j and we set:

$$\tau_j^2 := \tau(n)^2 = 1/\sqrt{n \log(n)}. \tag{2.10}$$

Besides, $w_j(n) := w(n)$. We suppose that there exist q_1 and q_2 such that for n large enough

$$n^{-q_1/2} \leq w(n) \leq n^{-q_2/2}. \tag{2.11}$$

This form of hyperparameters was emphasized in [5] in order to mimic heavy tailed priors such as Laplace or Cauchy distributions. Indeed, Johnstone and Silverman in [18], [19] showed that their empirical Bayes approach for regular regression setting, with a prior mixing a heavy-tailed density and a point mass at zero proved fruitful, both in theory and practice. Pensky in [24] also underlined the efficiency of this kind of hyperparameters.

We underscore that contrary to the first form of hyperparameters (2.9), these latter forms (2.10) and (2.11) lead to an adaptive Bayesian estimator.

For this case of hyperparameters (2.10) and (2.11), the estimator of f will be denoted $\check{f}$.

3. Functional spaces

In this paper, the functional classes of interest are Besov bodies and weak Besov bodies. Let us define them. Using the decomposition (2.1), we characterize the Besov spaces by using the following norm

$$\|f\|_{spq} = \begin{cases} \left[\sum_{j\geq -1} 2^{jq(s+1/2-1/p)} \|(\beta_{j,k})_k\|_{\ell_p}^q\right]^{1/q} & \text{if } q < \infty \\ \sup_{j\geq -1} 2^{j(s+1/2-1/p)} \|(\beta_{j,k})_k\|_{\ell_p} & \text{if } q = \infty. \end{cases}$$

If $\max(0, 1/p - 1/2) < s < r$ and $p, q \geq 1$

$$f \in B^s_{p,q} \Longleftrightarrow \|f\|_{spq} < \infty.$$

The Besov spaces have the following simple relationship

$$B^{s_1}_{p,q_1} \subset B^s_{p,q}, \quad \text{for } s_1 > s \text{ or } s_1 = s \text{ and } q_1 \leq q$$

and

$$B^s_{p,q} \subset B^{s_1}_{p_1,q}, \quad \text{for } p_1 > p \text{ and } s_1 \geq s - 1/p + 1/p_1.$$

The index s indicates the smoothness of the function. The Besov spaces capture a variety of smoothness features in a function including spatially inhomogeneous behavior when $p < 2$.

We recall and stress that in this paper as mentioned above, the regularity conditions will be expressed for the function $f(G^{-1})$ due to the *warped* basis context.

More precisely we shall focus on the space $B^s_{2,\infty}$. We have in particular

$$f \in B^s_{2,\infty} \Longleftrightarrow \sup_{J \geq -1} 2^{2Js} \sum_{j \geq J} \sum_k \beta^2_{jk} < \infty. \tag{3.1}$$

We define the Besov ball of some radius R as $B^s_{2,\infty}(R) = \{f : \|f\|_{s2\infty} \leq R\}$.

Let us define now the weak Besov space $W(r, 2)$

Definition 1. *Let $0 < r < 2$. We say that a function f belongs to the weak Besov body $W(r, 2)$ if and only if:*

$$\|f\|_{W_r} := [\sup_{\lambda > 0} \lambda^{r-2} \sum_{j \geq -1} \sum_k \beta^2_{jk} I\{|\beta_{jk} \leq \lambda|\}]^{1/2} < \infty. \tag{3.2}$$

And we have the following proposition

Proposition 1. *Let $0 < r < 2$ and $f \in W(r, 2)$. Then*

$$\sup_{\lambda > 0} \lambda^r \sum_{j \geq -1} \sum_k I\{|\beta_{jk}| > \lambda\} \leq \frac{2^{2-r} \|f\|^2_{W_r}}{1 - 2^{-r}}. \tag{3.3}$$

For the proof of this proposition see for instance [21].

To conclude this section, we have the following embedding

$$B^s_{2,\infty} \subset W_{2,2/(1+2s)},$$

which is not difficult to prove (see for instance [21]).

4. Minimax performances of the procedures

4.1. Bayesian estimators based on Gaussian priors with small variances

Theorem 1. *Assume that we observe model (1.1). We consider the hyperparameters defined by (2.9). Set* $J := J_\alpha$ *such that* $2^{J_\alpha} = (3/(2n))^{-1/\alpha}$.

Let $\alpha > 1$ *and* $\alpha \geq s$*, then we have the following upper bound:*

$$\sup_{f(G^{-1})\in B^s_{2,\infty}(R)} \mathbb{E}\|\hat{f} - f\|_2^2 = \mathcal{O}((1/n)^{1-1/\alpha}\log^2(n)) + \mathcal{O}((1/n)^{2s/\alpha}). \quad (4.1)$$

The optimal choice of the hyperparameter α in Theorem 1 should minimize the upper bound derived in (4.1). Consequently, let us choose now in (4.1) $\alpha = 2s + 1$, we immediately deduce the following corollary.

Corollary 1. *If one chooses for the prior parameter* $\alpha = 2s + 1$*, one gets*

$$\sup_{f(G^{-1})\in B^s_{2,\infty}(R)} \mathbb{E}\|\hat{f} - f\|_2^2 = \mathcal{O}(\log^2(n)n^{-2s/(2s+1)}).$$

This corollary shows that with this specific choice of hyperparameter α, one recovers the minimax rate of convergence up to a logarithmic factor that one achieves in a uniform design.

4.2. Bayesian estimators based on Gaussian priors with large variance

Theorem 2. *We consider the model (1.1). We assume that the hyperparameters are defined by (2.10) and (2.11). Set* $J := J_n$ *such that* $2^{J_n} = n/\log n$*, then we have:*

$$\sup_{f(G^{-1})\in B^s_{2,\infty}(R)} \mathbb{E}\|\check{f} - f\|_2^2 \leq C\left(\frac{\log(n)}{n}\right)^{2s/(2s+1)}.$$

It is worthwhile to make some comments about the results of Theorem 2. Here, the estimator turns out to be adaptive and contrary to the similar results in Proposition 2 in [20], we no longer have the limitation on the regularity index $s > 1/2$. Moreover, Kerkyacharian and Picard [20] had to stop the highest level J such that $2^J = (n/\log(n))^{1/2}$, here we stop at the usual level J_n such that $2^{J_n} = n/\log(n)$, one gets in standard thresholding.

5. Simulations and discussion

A simulation study is conducted in order to compare the numerical performances of the two Bayesian estimators based on warped wavelets and on Gaussian prior with small or large variance, described respectively in section 2.2.1 and 2.2.2 and the hard thresholding procedure using the universal threshold $\sigma\sqrt{2\log(n)}$, based on warped basis and introduced by Kerkyacharian and Picard [20] for the nonparametric regression model in a random design setting. For more details on Kerkyacharian and Picard procedure, the readers are referred to Willer [26], see also [16]. In fact, we have decided to concentrate on the procedure of Kerkyacharian and Picard. Indeed, it is interesting to point out differences and compare performances obtained by Bayesian procedures which apply local thresholds and a universal threshold procedure.

The main difficulties lie in implementing the Bayesian procedures with the stochastic variance (2.3). Note also the responses proposed by Amato et al. [3] and Kovac and Silverman [22].

All the simulations done in the present paper have been conducted with MATLAB and the Wavelet toolbox of MATLAB.

We consider here four test functions of Donoho and Johnstone [13] representing different level of spatial variability. The test functions are plotted in Fig. 1. For each of the four objects under study, we compare the three estimators at two noise levels, one with signal-to-noise ratio $RSNR = 4$ and another with $RSNR = 7$. As in Willer [26], we also consider different cases of design density which are plotted in Fig. 2. The first two densities are uniform or slightly varying whereas the last two ones aim at depicting the case where a hole occurs in the density design. The sample size is equal to $n = 1024$ and the wavelet we used is the Symmlet8.

In order to compare the behaviors of the estimators, the RMSE criterion was retained. More precisely, if $\hat{f}(X_i)$ is the estimated function value at X_i and n is the sample size, then

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2}. \tag{5.1}$$

The RMSE displayed in Tab. 1 are computed as the average over 100 runs of expression (5.1). In each run, we hold all factors constant, except the design points (random design) and the noise process that were regenerated.

Table 1
Values of RMSE over 100 runs

		RSNR=4			RSNR=7		
	design density	E1	E2	E3	E1	E2	E3
Blocks	Sine	0.0194	0.0219	0.0227	0.0113	0.0161	0.0129
	Hole2	0.0196	0.0220	0.0226	0.0114	0.0163	0.0130
Bumps	Sine	0.0243	0.240	0.259	0.0156	0.0167	0.0172
	Hole2	0.0241	0.0237	0.0253	0.0155	0.0167	0.0169
HeaviSine	Sine	0.0164	0.0141	0.0133	0.0103	0.0092	0.0093
	Hole2	0.0169	0.0146	0.0138	0.0107	0.0097	0.0096
Doppler	Sine	0.0236	0.0231	0.0236	0.0157	0.0238	0.0248
	Hole2	0.0244	0.0238	0.0248	0.0166	0.0172	0.0176

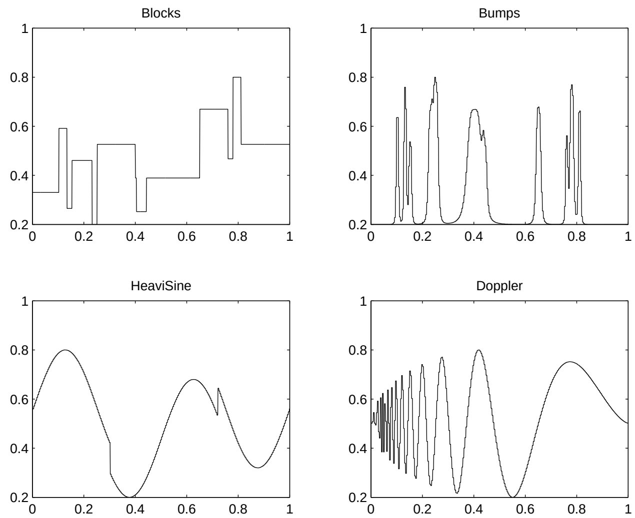

FIGURE 1. *Test functions.*

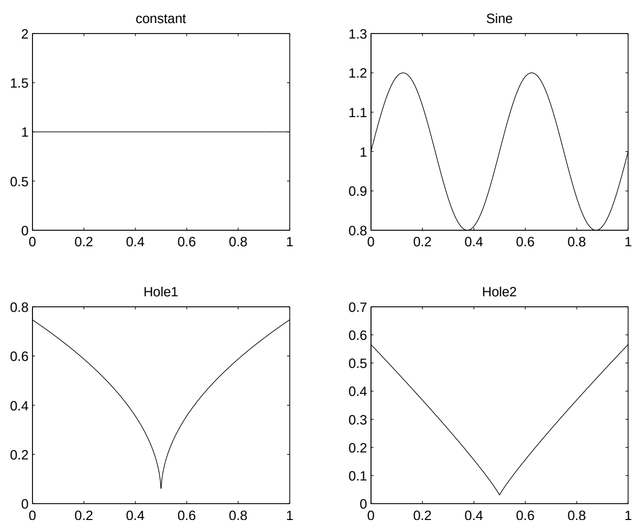

FIGURE 2. *Design density.*

E1 corresponds to the Bayesian estimator based on Gaussian prior with large variance, E2 to the Bayesian estimator based on Gaussian prior with small variance and E3 to the estimator built following the Kerkyacharian and Picard procedure in [20].

In order to implement E1, we made the following choices of hyperparameters described in section 2.2.2: in (2.11), $q_1 = q_2 = q = 1$ proved to be a good compromise whatever the function of interest to be estimated while leading to good graphic reconstructions. We set $w(n) = 20 \times n^{-q/2}$ and $\tau(n) = 20 \times$

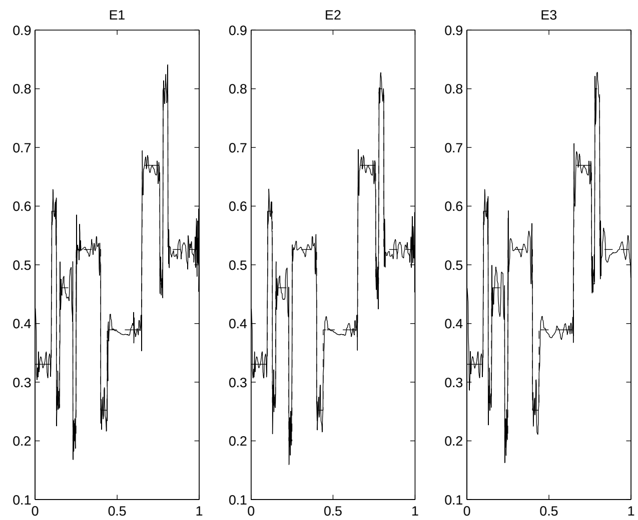

Figure 3. *Blocks target and Sine density, RSNR = 4.*

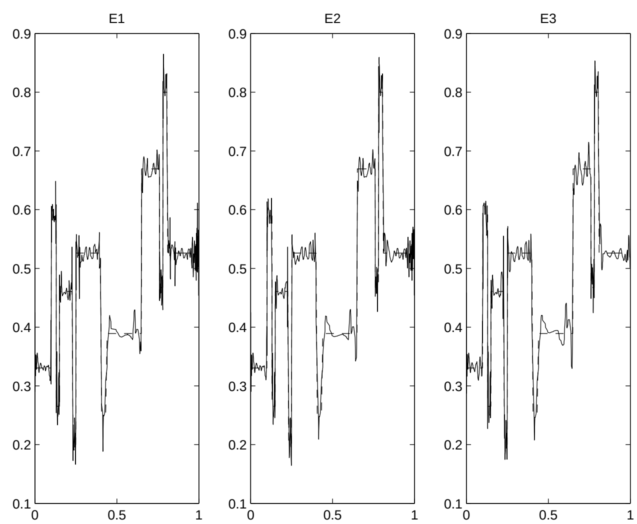

Figure 4. *Blocks target and Hole2 design density, RSNR = 4.*

$\sigma^2/(n\log(n))$. To implement E2, we set $c_1 = 1$, $c_2 = 2$, $\alpha = 0.5$ and $\beta = 1$, following the choices recommended in [2].

The following plots compare the visual quality of the reconstructions (see Fig. 3. to Fig. 8). The solid line is the estimator and the dotted line is the true function.

We shall now comment and discuss the results displayed in Tab. 1 as well as the various visual reconstructions.

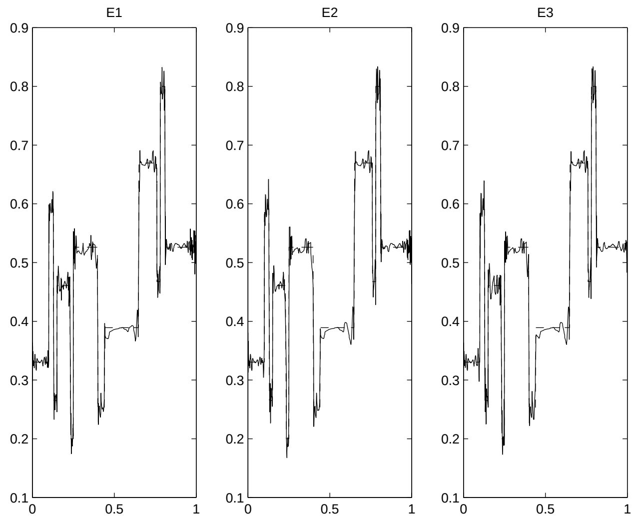

FIGURE 5. *Blocks target and Hole2 design density, RSNR = 7.*

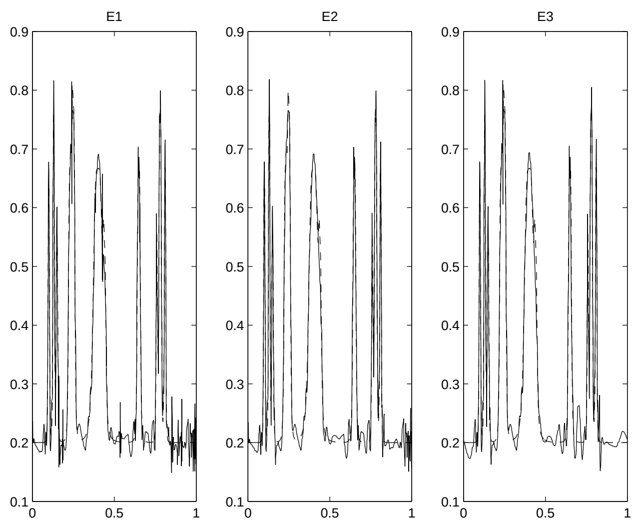

FIGURE 6. *Bumps target and Sine design density, RSNR = 4.*

The performances are always better for the Bayesian estimators except for the case of the HeaviSine test function. More precisely, the RMSE for Blocks whatever the noise level and design densities are smaller for Estimator 1. Moreover the RMSE are almost equal for Estimator 1 and 2 in the case of Bumps test function, whatever the design densities and for a noise level RSNR=4. This may be due to the irregularity of the Bumps, Blocks and Doppler test functions which are much rougher than the HeaviSine which is more regular. Indeed, Estimator

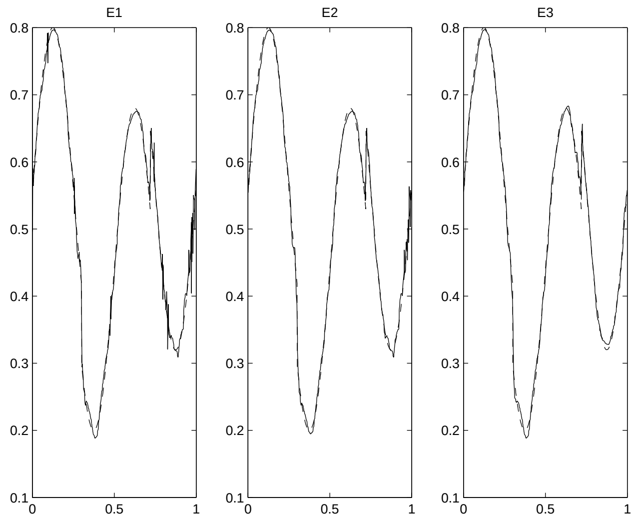

FIGURE 7. *HeaviSine target and Sine design density, RSNR = 7.*

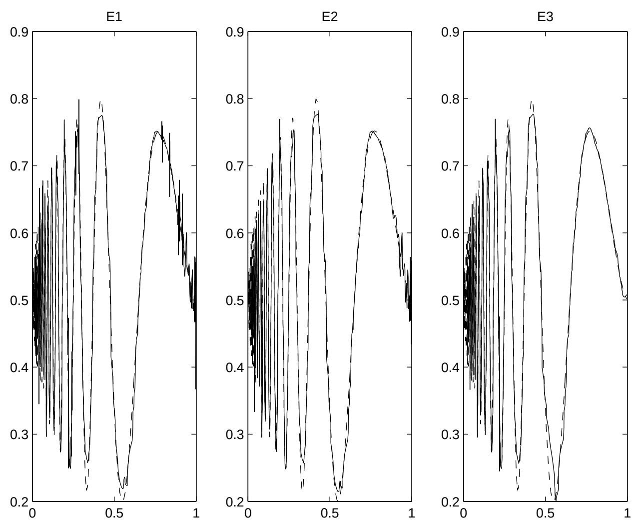

FIGURE 8. *Doppler target and Hole2 design density, SNR = 4.*

1 and 2 tend to detect better the corner of Blocks, the high peaks in Bumps, and the high frequency parts of Doppler as the graphics show it. We may explain this by the fact that Estimators 1 and 2 have level-dependent thresholds whereas Estimator 3 has a hard universal threshold.

As for the reconstructions, one can see that they are slightly better in the case of Sine density and small noise, whereas there are small deteriorations when a hole occurs in the design density, but this change does not affect the visual quality in too big proportions. This fact highlights the interest of "warping" the wavelet basis. Warping the basis allows the estimators to behave still correctly when the design densities are far from the uniform density such as in the case of Hole2.

6. Proofs

In the sequel C denotes some positive constant which may change from one line to another line. We also assume without loss of generality that $\sigma = 1$ in model (1.1).

We have that

$$\mathbb{E}(\psi_{jk}^2(G(X))) = \int \psi_{jk}^2(G(x))g(x)dx = \int \psi_{jk}^2(y)dy = 1.$$

hence we get $\mathbb{E}(\gamma_{jk}^2) = 1/n$, the expression of γ_{jk}^2 being given by (2.3).

Let us define the following event:

$$\Omega_n^\delta = \{|\gamma_{jk}^2 - 1/n| \leq \delta\}. \tag{6.1}$$

To make proofs clearer, we recall the Bernstein inequality that we will use in the sequel. (see in [23] Proposition 2.8 and formula (2.16))

Proposition 2. *Let $Z_1, \dots, Z_n$ be independent and square integrable random variables such that for some nonnegative constant b, $Z_i \leq b$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])$$

and $v = \sum_{i=1}^n \mathbb{E}(Z_i^2)$. Then for any positive x, we have

$$\mathbb{P}[S \geq x] \leq \exp\left(\frac{-v}{b^2} h(\frac{bx}{v})\right)$$

where $h(u) = (1+u)\log(1+u) - u$.

It is easy to prove that

$$h(u) \leq \frac{u^2}{2(1+u/3)}$$

which immediately yields

$$\mathbb{P}[S \geq x] \leq \exp\left(\frac{-x^2}{2(v+bx/3)}\right).$$

Lemma 1. *Let ς be some positive constant. We have*

$$\mathbb{P}(|\gamma_{jk}^2 - 1/n| > \varsigma/n) \leq 2e^{-n^{1-1/\alpha}\frac{\varsigma^2}{2C(1+\varsigma/3)}} \quad \forall\, j \leq J_\alpha \tag{6.2}$$

$$\mathbb{P}(|\gamma_{jk}^2 - 1/n| > \varsigma/n) \leq 2e^{-\varsigma^2 \log(n)/(C\|\psi\|_4^4 + \varsigma\|\psi\|_\infty^2)} \quad \forall\, j \leq J_n. \tag{6.3}$$

Proof of Lemma 1
Let us deal with the first case $j \le J_\alpha$. To bound $\mathbb{P}(|\gamma_{jk}^2 - 1/n| > \varsigma/n)$ we will use the Bernstein inequality and apply Proposition 2. In the present situation $Z_i = (1/n^2)\psi_{jk}^2(G(X_i))$.

First of all, in order to apply the Bernstein inequality, we need the value of the sum

$$v = \sum_{i=1}^{n} \mathbb{E}[((1/n^2)\psi_{j,k}^2(G(X_i)))^2]$$

we have

$$\begin{aligned} \mathbb{E}\psi_{j,k}^4(G(X)) &= \int_0^1 \psi_{j,k}^4(G(x))g(x)dx = \int_0^1 \psi_{j,k}^4(y)dy \\ &\le \int_0^1 2^{2j}\psi^4(2^j y - k)dy \le 2^j \int \psi^4(y)dy \le C\|\psi\|_4^4 2^j \end{aligned} \tag{6.4}$$

hence

$$(1/n^4)\sum_{i=1}^{n} \mathbb{E}\psi_{j,k}^4(G(X_i)) \le (C/n^3)2^{J_\alpha} = \frac{C}{n^{3-1/\alpha}}$$

moreover

$$\psi_{jk}^2(G(X)) \le \|\psi\|_\infty^2 2^j \le Cn^{1/\alpha} \quad j \le J_\alpha \quad a.s$$

so

$$\mathbb{P}(|\gamma_{jk}^2 - 1/n)| > \varsigma/n) \le 2\exp(-\frac{\varsigma^2}{2C(1+\varsigma/3)}\frac{n^{-2}}{n^{-3+1/\alpha}}).$$

Let us now deal with the second case $j \le J_n$. To bound $\mathbb{P}(|\gamma_{jk}^2 - 1/n| > \varsigma/n)$ we will follow the lines of the proof of the first case. Here again

$$Z_i = 1/n^2\psi_{jk}^2(G(X_i)).$$

According to (6.4), we have

$$\mathbb{E}(1/n^4\psi_{jk}^4(G(X))) \le C2^j/n^4 \le C/(n^3\log(n)), \quad j \le J_n$$

and

$$v = \sum_{i=1}^{n} \mathbb{E}(1/n^4\psi_{jk}^4(G(X))) \le C\|\psi\|_4^4/(n^2\log(n))$$

and

$$1/n^2\psi_{jk}^2(G(X))) \le \|\psi\|_\infty^2 2^j/(n^2) \le \|\psi\|_\infty^2/(n\log(n)), \quad j \le J_n \quad a.s$$

consequently

$$\mathbb{P}(|\gamma_{jk}^2 - 1/n| > \varsigma/n) \le 2e^{-\varsigma^2\log(n)/(C\|\psi\|_4^4 + \varsigma\|\psi\|_\infty^2)}.$$

The following lemma shows that the properties of the Bayesian estimators $\check{f}$ and $\hat{f}$ can be controlled on the event Ω_n^δ. To lighten the notations for the proof of this lemma, we will denote Ω_n for Ω_n^δ and Ω_n^c the complementary of Ω_n.

Lemma 2. *We have*

$$\mathbb{E}[I(\Omega_n^c)\|\check{f}-f\|_2^2] = o((\log(n)/n)^{2s/(2s+1)}$$

$$\mathbb{E}[I(\Omega_n^c)\|\hat{f}-f\|_2^2] = o((1/n)^{1-1/\alpha}\log(n)).$$

Proof of Lemma 2.
We have

$$\begin{aligned}\mathbb{E}\left[I(\Omega_n^c)\|\check{f}-f\|_2^2\right] &\leq CJ_n\mathbb{E}\left[\sum_{j\leq J_n}\sum_k(\tilde{\beta}_{jk}-\beta_{jk})^2 I(\Omega_n^c)\right] + \mathbb{P}(\Omega_n^c)\sum_{j>J_n}\sum_k \beta_{jk}^2 \\ &\leq V+B.\end{aligned}$$

Let us first deal with the variance term V. The estimator $\tilde{\beta}_{jk}$ can be written as $\tilde{\beta}_{jk} = w_{jk}\hat{\beta}_{jk}$ with $0 \leq w_{jk} \leq 1$. We have

$$\begin{aligned}V &\leq CJ_n\mathbb{E}\left[\sum_{j\leq J_n,k}\Big(w_{jk}(\hat{\beta}_{jk}-\beta_{jk})-(1-w_{jk})\beta_{jk}\Big)^2 I(\Omega_n^c)\right] \\ &\leq 2CJ_n\mathbb{E}\left[\sum_{j\leq J_n}\sum_k w_{jk}^2(\hat{\beta}_{jk}-\beta_{jk})^2 I(\Omega_n^c)\right] \\ &\quad + 2CJ_n\sum_{j\leq J_n}\sum_k \mathbb{E}\left[(1-w_{jk})^2\beta_{jk}^2 I(\Omega_n^c)\right] \\ &\leq 2CJ_n\mathbb{E}\left[\sum_{j\leq J_n}\sum_k(\hat{\beta}_{jk}-\beta_{jk})^2 I(\Omega_n^c)\right] + 2CJ_n\sum_{j\leq J_n}\sum_k \mathbb{E}\left[\beta_{jk}^2 I(\Omega_n^c)\right]\end{aligned}$$

because $0 \leq w_{jk} \leq 1$. Then, using Cauchy Scharwz inequality we get

$$V \leq 2CJ_n\sum_{j\leq J_n}\sum_k\left[\mathbb{E}(\hat{\beta}_{jk}-\beta_{jk})^4\right]^{\frac{1}{2}}\mathbb{P}(\Omega_n^c)^{\frac{1}{2}} + 2CJ_n\sum_{j\leq J_n}\sum_k\beta_{jk}^2\mathbb{P}(\Omega_n^c).$$

Using (6.3) and (6.23) we have

$$\begin{aligned}V &\leq 2CJ_n2^{J_n}e^{-\varsigma^2\log(n)/(2C\|\psi\|_4^4/n+\varsigma\|\psi\|_\infty^2)} \\ &\quad + 2CJ_n\|f(G^{-1})\|_2^2 e^{-\varsigma^2\log(n)/(C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2)}.\end{aligned}$$

We recall that $2^{J_n} = n/\log(n)$, accordingly by choosing ς large enough we have

$$V = o((\log(n)/n)^{2s/(2s+1)}$$

As for the term B

$$B \leq C2^{-2J_ns}e^{-\varsigma^2\log(n)/(C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2)}$$

which completes the proof for $\check{f}$.

The proof for $\hat{f}$ is similar, all inequalities hold a fortiori since, in the case of the estimator $\hat{f}$ we have $\mathbb{P}(\Omega_n^c) \leq e^{-Cn^{1-1/\alpha}}$ (see (6.2)).

We consider the setting of Theorem 1. We recall that $\tilde{\beta}_{jk}$ is zero whenever $|\hat{\beta}_{jk}|$ falls below a threshold λ_B and we have the following lemma concerning the behavior of λ_B.

Lemma 3. *On the event Ω_n^δ defined by (6.1) with $\delta = 1/(2n)$, for $\alpha > 1$ we have*

$$\lambda_B \approx \sqrt{\frac{\log(n)}{n}}, \quad j < J_\alpha \tag{6.5}$$

and J_α is taken such that $2^{J_\alpha} = (\frac{3}{2n})^{-1/\alpha}$.

Proof of Lemma 3.
We follow the lines of the proof of lemma 1. in [1].

On the one hand we have (see proof of lemma 1. in [1] page 228)

$$\lambda_B{}^2 \leq \frac{2\gamma_{jk}^2(\gamma_{jk}^2+\tau_j^2)}{\tau_j^2}\log\left(\frac{1-\pi_j}{\pi_j}\frac{\sqrt{\gamma_{jk}^2+\tau_j^2}}{\gamma_{jk}}+c\right)$$

where c is some suitable positive constant. Besides, we have $1/(2n) \leq \gamma_{jk}^2 \leq 3/(2n)$, therefore

$$\begin{aligned}\lambda_B{}^2 \quad \leq \quad & \frac{2(3/(2n))((3/(2n))+c_1(3/(2n)))}{c_1(3/(2n))} \\ & \times \log\left(\frac{1-c_2(3/(2n))^{\beta/\alpha}}{c_2(3/(2n))^{\beta/\alpha}}\frac{\sqrt{(1+c_1)(3/(2n))}}{\sqrt{1/(2n)}}+c\right)\end{aligned}$$

hence we get

$$\lambda_B{}^2 \leq \tilde{c}(1/n)\log(\tilde{c}(1/n)^{(-\frac{\beta}{\alpha})}+c)$$

where $\tilde{c}$ denotes a positive constant depending on c_1 and c_2 and which may be different at different places. Since

$$\tilde{c}(1/n)\log(\tilde{c}(1/n)^{(-\frac{\beta}{\alpha})}+c) \approx -\tilde{c}(\beta/\alpha)(1/n)\log(1/n)$$

we finally get

$$\lambda_B{}^2 \leq -\tilde{c}(\beta/\alpha)(1/n)\log(1/n).$$

On the other hand, for the reverse inequality, we have (see proof of lemma 1. in [1] page 228 and formula (14) in [1] page 221)

$$\lambda_B{}^2 \geq \frac{2\gamma_{jk}^2(\gamma_{jk}^2+\tau_j^2)}{\tau_j^2}\log\left(\frac{1-\pi_j}{\pi_j}\frac{\sqrt{\gamma_{jk}^2+\tau_j^2}}{\gamma_{jk}}\right)$$

but $|\gamma_{jk}^2 - 1/n| \leq \ 1/(2n)$ consequently one has

$$\lambda_B{}^2 \geq -\tilde{c}(\beta/\alpha)(1/n)(\log(1/n))$$

which completes the proof.

Proof of Theorem 1.
On the event Ω_n^δ defined by (6.1) with $\delta = 1/(2n)$, by the usual decomposition of the MISE into a variance and a bias term, we get

$$\begin{aligned}\mathbb{E}\|\hat{f} - f\|_2^2 &\leq 2\big[\mathbb{E}\| \sum_{j\leq J_\alpha}\sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2 + \| \sum_{j>J_\alpha}\sum_k \beta_{jk}\psi_{j,k}(G)\|_2^2\big] \\ &\leq 2(V+B)\end{aligned}$$

with

$$V = \mathbb{E}\| \sum_{j\leq J_\alpha}\sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2$$

$$B = \| \sum_{j>J_\alpha}\sum_k \beta_{jk}\psi_{j,k}(G)\|_2^2.$$

We first deal with the term V. We have

$$\| \sum_{j\leq J_\alpha}\sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2 \leq J_\alpha \sum_{j\leq J_\alpha} \| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2.$$

We want to show that

$$\| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2 \leq C \sum_k (\tilde{\beta}_{jk} - \beta_{jk})^2.$$

For this purpose we have

$$\begin{aligned}\| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}(G)\|_2^2 &= \int | \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{jk}(G(x))|^2 dx \\ &= \int | \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{jk}(x)|^2 \frac{1}{g(G^{-1}(x))} dx \\ &= \| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}\|_{\mathbb{L}_2(\varrho)}^2\end{aligned}$$

where $\varrho(x) = 1/(g(G^{-1}))(x)$.

Now using inequality (44) p. 1075 in [20] we have

$$\| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}\|_{\mathbb{L}_2(\varrho)}^2 \leq C2^j \sum_k |\tilde{\beta}_{jk} - \beta_{jk}|^2 \varrho(I_{j,k})$$

where $I_{j,k}$ denotes the interval $[\frac{k}{2^j}, \frac{k+1}{2^j}]$ and $\varrho(I_{jk}) = \int_{I_{jk}} \varrho(x)dx$. But the design density g is bounded from below by m. Hence we get

$$\varrho(I_{j,k}) \leq 2^{-j}/m$$

and consequently

$$\| \sum_k (\tilde{\beta}_{jk} - \beta_{jk})\psi_{j,k}\|_{\mathbb{L}_2(\varrho)}^2 \leq C \sum_k (\tilde{\beta}_{jk} - \beta_{jk})^2.$$

We decompose now V into three terms

$$V \leq CJ_\alpha \mathbb{E} \sum_{j \leq J_\alpha} \sum_k [(\tilde{\beta}_{jk} - \beta^{'}_{jk})^2 + (\beta^{'}_{jk} - \beta^{''}_{jk})^2 + (\beta^{''}_{jk} - \beta_{jk})^2]$$

where

$$\beta^{'}_{jk} = b_j \hat{\beta}_{jk} I\{|\hat{\beta}_{jk}| \geq \kappa \lambda_B\}$$

with κ a positive constant and

$$b_j = \frac{\tau_j^2}{\tau_j^2 + \gamma_{jk}^2}$$

$$\beta^{''}_{jk} = b_j \beta_{jk}.$$

As a consequence we have

$$V \leq CJ_\alpha(A_1 + A_2 + A_3).$$

We are now going to upperbound each term A_1, A_2 and A_3. We start with A_1

$$A_1 = \sum_{j \leq J_\alpha} \sum_k \mathbb{E}[(\tilde{\beta}_{jk} - \beta^{'}_{jk})^2].$$

As precised in the beginning of section 2.2, $\tilde{\beta}_{jk} = 0$ for $|\hat{\beta}_{jk}| < \lambda_B$. As well, $\beta^{'}_{jk} = 0$ for $|\hat{\beta}_{jk}| < \kappa\lambda_B$ and $\tilde{\beta}_{jk} - \beta^{'}_{jk} \to 0$ monotonically as $\hat{\beta}_{jk} \to \infty$. Hence

$$\max_{\hat{\beta}_{jk}} |\tilde{\beta}_{jk} - \beta^{'}_{jk}| = b_j \lambda_B$$

which implies

$$A_1 \leq C \sum_{j < J_\alpha} \sum_k \mathbb{E}(b_j^2 \lambda_B^2).$$

We have $\lambda_B \approx \sqrt{\frac{\log n}{n}}$ and $b_j \leq 1$ for $j \leq J_\alpha$ hence we get

$$A_1 \leq C \sum_{j \leq J_\alpha} \sum_{k=0}^{2^j-1} \frac{\log(n)}{n}$$

so

$$\begin{aligned} A_1 &\leq C \frac{\log(n)}{n} \sum_{j \leq J_\alpha} 2^j && (6.6) \\ &\leq C \frac{\log(n)}{n} \left(\frac{1}{n}\right)^{-1/\alpha} && (6.7) \end{aligned}$$

finally

$$A_1 = \mathcal{O}\left(\log(n) \left(\frac{1}{n}\right)^{1-1/\alpha} \right)$$

Let us now consider the second term A_2

$$\begin{aligned} A_2 &= \sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\mathbb{E}(\beta_{jk}^{'}-\beta_{jk}^{''})^2 \\ &= \sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\mathbb{E}(b_j\hat{\beta}_{jk}I\{|\hat{\beta}_{jk}|\geq\kappa\lambda_B\}-b_j\beta_{jk})^2 \end{aligned}$$

We have that $b_j \leq 1$, consequently it follows

$$\begin{aligned} A_2 &= \sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\mathbb{E}((\hat{\beta}_{jk}-\beta_{jk})^2 I\{|\hat{\beta}_{jk}|\geq\kappa\lambda_B\})+\mathbb{E}\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2 I\{|\hat{\beta}_{jk}|<\kappa\lambda_B\} \\ &= A_2^{'}+A_2^{''} \end{aligned}$$

We have

$$A_2^{'} \leq \sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\mathbb{E}(\hat{\beta}_{jk}-\beta_{jk})^2.$$

Using inequality (64) in [20] p. 1086 we have

$$\mathbb{E}(\hat{\beta}_{jk}-\beta_{jk})^2 \leq C\frac{1+\|f\|_\infty^2}{n} \tag{6.8}$$

hence

$$A_2^{'} = \mathcal{O}((1/n)^{1-1/\alpha}).$$

We now bound the term $A_2^{''}$:

$$\begin{aligned} A_2^{''} &= \mathbb{E}\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2 I\{|\hat{\beta}_{jk}|<\kappa\lambda_B\}(I\{|\beta_{jk}|<2\kappa\lambda_B\}+I\{|\beta_{jk}|>2\kappa\lambda_B\}) \\ &\leq \mathbb{E}\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2 I\{|\beta_{jk}|<2\kappa\lambda_B\}+\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2\mathbb{P}(|\hat{\beta}_{jk}-\beta_{jk}|>\kappa\lambda_B) \\ &= T_3+T_4 \end{aligned} \tag{6.9}$$

We have

$$T_3 \leq C\sum_{j\leq J_\alpha}\lambda_B^2 2^j \leq C\frac{\log(n)}{n}n^{1/\alpha} = C\log(n)n^{-1+1/\alpha}.$$

Let us focus on T_4, we have

$$\hat{\beta}_{jk}-\beta_{jk} = 1/n\sum_{i=1}^n\psi_{j,k}(G(X_i))(f(X_i)+\varepsilon_i)-\mathbb{E}\psi_{j,k}(G(X))f(X)$$

Hence

$$\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > \kappa\sqrt{\log(n)/n}) \leq \mathbb{P}_1 + \mathbb{P}_2$$

where

$$\mathbb{P}_1 = \mathbb{P}(|1/n\sum_{i=1}^{n}\psi_{j,k}(G(X_i))(f(X_i)) - \mathbb{E}\psi_{j,k}(G(X))f(X)| > \kappa/2\sqrt{(\log(n)/n)}) \tag{6.10}$$

and

$$\mathbb{P}_2 = \mathbb{P}(|1/n\sum_{i=1}^{n}\psi_{j,k}(G(X_i))\varepsilon_i| > \kappa/2\sqrt{(\log(n)/n)}) \tag{6.11}$$

Kerkyacharian and Picard in [20] in order to prove inequality (65) in [20] showed p. 1088 that

$$\mathbb{P}_1 \leq 2\exp\left(-\frac{3\kappa^2\log(n)}{4\|f\|_\infty(3+\kappa)}\right) \tag{6.12}$$

if $2^j \leq n/\log(n)$. As for $\mathbb{P}_2$, conditionally on $(X_1,\dots,X_n)$ we have

$$1/n\sum_{i=1}^{n}\psi_{j,k}(G(X_i))\varepsilon_i \sim N(0,\gamma_{jk}^2),$$

where γ_{jk}^2 has been defined in (2.3). Using exponential inequality for Gaussian random variable we have

$$\begin{aligned}
\mathbb{P}_2 &\leq \mathbb{E}(\exp(-\frac{\kappa^2\log(n)}{8n\gamma_{jk}^2})) \\
&= \mathbb{E}e^{-\frac{\kappa^2\log(n)}{8n\gamma_{jk}^2}}(I(|\gamma_{jk}^2 - 1/n| \leq 1/2n) + I(|\gamma_{jk}^2 - 1/n| > 1/(2n))) \\
&\leq e^{-\frac{\kappa^2\log(n)}{12}} + \mathbb{P}(|\gamma_{jk}^2 - 1/n| > 1/(2n)). \qquad (6.13)
\end{aligned}$$

Using (6.2) with $\varsigma = 1/2$, we have for $\alpha > 1$

$$\begin{aligned}
T_4 &\leq \left(2e^{(-Cn^{1-1/\alpha})} + e^{-\frac{\kappa^2\log(n)}{12}} + 2\exp\left(\frac{-3\kappa^2\log(n)}{4\|f\|_\infty(3+\kappa)}\right)\right)\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2 \\
&\leq \left(2e^{(-Cn^{1-1/\alpha})} + e^{-\frac{\kappa^2\log(n)}{12}} + 2\exp\left(\frac{-3\kappa^2\log(n)}{4\|f\|_\infty(3+\kappa)}\right)\right)\|f(G^{-1})\|_2^2
\end{aligned}$$

It remains to fix κ large enough so that we get

$$T_4 = \mathcal{O}(\log(n)n^{-1+1/\alpha}).$$

So we have for $A_2^{''}$, with $\alpha > 1$,

$$A_2^{''} = \mathcal{O}\left(\frac{\log(n)}{n^{1-1/\alpha}}\right)$$

Finally we get for A_2

$$A_2 = \mathcal{O}\left(\log(n)\left(\frac{1}{n}\right)^{1-1/\alpha}\right).$$

Let us consider now the term A_3

$$\begin{aligned} A_3 &\leq C\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\mathbb{E}(\beta_{jk}^{"}-\beta_{jk})^2 \\ &= C\sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\beta_{jk}^2(1-b_j)^2 = \sum_{j\leq J_\alpha}\sum_{k=0}^{2^j-1}\left(\frac{\gamma_{jk}^2}{\tau_j^2+\gamma_{jk}^2}\right)^2\beta_{jk}^2. \end{aligned}$$

Since $|\gamma_{jk}^2 - 1/n| \leq 1/(2n)$, we get

$$A_3 \leq \sum_{j\leq J_\alpha}\left(\frac{3/(2n)}{c_1 2^{-j\alpha}+1/(2n)}\right)^2\sum_{k=0}^{2^j-1}\beta_{jk}^2$$

but $f(G^{-1})$ belongs to the Besov ball $B_{2,\infty}^s(R)$ which entails

$$\sum_{k=0}^{2^j-1}\beta_{jk}^2 \leq M2^{-2js}$$

hence

$$A_3 \leq C/n^2\sum_{j\leq J_\alpha}2^{2j(-s+\alpha)}$$

We have

$$A_3 \leq C/n^2(1/n)^{\frac{-2(-s+\alpha)}{\alpha}} = \mathcal{O}(1/n)^{2s/\alpha}.$$

We are now in position to give an upper bound for the variance term V namely

$$V \leq CJ_\alpha(\log(n)(1/n)^{1-1/\alpha}+(1/n)^{2s/\alpha}).$$

It remains to bound the bias term B. In [20] p.1083 using inequality (44) the authors have proved that for any l we get

$$\begin{aligned} \|\sum_{j\geq l}\sum_k\beta_{jk}\psi_{j,k}(G)\|_2 &\leq \sum_{j\geq l}\|\sum_k\beta_{jk}\psi_{j,k}(G)\|_2 \\ &\leq C\sum_{j\geq l}2^{j/2}\left(\sum_k|\beta_{jk}|^2\varrho(I_{j,k})\right)^{1/2}. \qquad (6.14) \end{aligned}$$

Applying (6.14) with in our case of a bounded from below design density, $\varrho(I_{j,k}) \leq 2^{-j}/m$ and $l = J_\alpha$, it follows

$$\begin{aligned}\| \sum_{j \geq J_\alpha} \sum_k \beta_{jk} \psi_{j,k}(G) \|_2 &\leq C \sum_{j \geq J_\alpha} \left(\sum_k |\beta_{jk}|^2 \right)^{1/2} \\ &\leq C \sum_{j \geq J_\alpha} 2^{-js} \leq C 2^{-J_\alpha s}\end{aligned}$$

hence

$$B = \| \sum_{j > J_\alpha} \sum_{k=0}^{2^j - 1} \beta_{jk} \psi_{j,k}(G(x)) \|_2^2 \leq C 2^{-2J_\alpha s} = C(1/n)^{2s/\alpha}$$

which completes the proof of Theorem 1.

Lemma 4. *Let w_{jk} a sequence of random weights lying in $[0,1]$. We assume that there exist positive constants c, m and K such that for any $\varepsilon > 0$,*

$$\check{\beta}_n = (w_{jk} \hat{\beta}_{jk})_{jk}$$

is a shrinkage rule verifying for any n,

$$w_{jk}(n) = 0, \quad a.e. \quad \forall\, j \geq J_n \quad with \quad 2^{J_n} \sim n/\log(n) := t_n^2, \,\forall\, k \tag{6.15}$$

$$|\hat{\beta}_{jk}| \leq m t_n \Rightarrow w_{jk} \leq c t_n, \quad \forall\, j \leq J_n, \forall\, k, \tag{6.16}$$

$$(1 - w_{jk}(n)) \leq K \left(\frac{t_n}{|\hat{\beta}_{jk}|} + t_n \right) \quad a.e. \quad \forall\, j \leq J_n, \forall\, k. \tag{6.17}$$

and let

$$\check{f} = \sum_{j < J_n} \sum_k w_{jk} \hat{\beta}_{jk} \psi_{jk}(G(x))$$

Then

$$\sup_{f(G^{-1}) \in B^s_{2,\infty}(R)} \mathbb{E} \| \check{f} - f \|_2^2 \leq (\log(n)/n)^{2s/(2s+1)}.$$

Proof of Lemma 4.

$$\begin{aligned}\mathbb{E}\|\check{f} - f\|_2^2 &\leq 2C \left(J_n \sum_{j \leq J_n} \sum_k \mathbb{E}(\check{\beta}_{jk} - \beta_{jk})^2 + \| \sum_{j > J_n} \sum_k \beta_{jk}^2 \psi_{jk}(G(x)) \|_2^2 \right) \\ &\leq V_1 + B_1.\end{aligned}$$

We first consider the term V_1

$$\begin{aligned}V_1 &\leq 2J_n \mathbb{E} \sum_{j \leq J_n} \sum_k (w_{jk}^2 (\hat{\beta}_{jk} - \beta_{jk})^2 + (1 - w_{jk})^2 \beta_{jk}^2) I\{|\hat{\beta}_{jk}| \leq m t_n\} \\ &+ J_n \mathbb{E} \sum_{j \leq J_n} \sum_k (w_{jk}^2 (\hat{\beta}_{jk} - \beta_{jk})^2 + (1 - w_{jk})^2 \beta_{jk}^2) I\{|\hat{\beta}_{jk}| > m t_n\} \\ &= V_1^{'} + V_1^{"}\end{aligned}$$

$$V_1^{'} = J_n(T_5 + T_6)$$
$$T_5 = \mathbb{E}\sum_{j\le J_n}\sum_k w_{jk}^2(\hat{\beta}_{jk} - \beta_{jk})^2 I\{|\hat{\beta}_{jk}| \le mt_n\}$$

but according to (6.8) we have for $2^j \le \log(n)/n$

$$\mathbb{E}(\hat{\beta}_{jk} - \beta_{jk})^2 \le C\frac{1 + \|f\|_\infty^2}{n}$$

hence using (6.16) it follows

$$T_5 \le Ct_n^2 2^{J_n} 1/n.$$

As for T_6

$$\begin{aligned} T_6 &= \mathbb{E}\sum_{j\le J_n}\sum_k (1 - w_{jk})^2\beta_{jk}^2 I\{|\hat{\beta}_{jk}| \le mt_n\} \\ &\le \mathbb{E}\sum_{j\le J_n}\sum_k (1 - w_{jk})^2\beta_{jk}^2 I\{|\hat{\beta}_{jk}| \le mt_n\} \\ &\quad \times [I\{|\beta_{jk}| \le 2mt_n\} + I\{|\beta_{jk} > 2mt_n|\}]. \end{aligned}$$

By (3.3) we get

$$T_6 \le \quad 2(mt_n)^{2s/(2s+1)}\|f\|^2_{W_{2/(1+2s)}} + \textstyle\sum_{j\le J_n}\sum_k \beta_{jk}^2\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > mt_n).$$

We are going to bound $\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > mt_n)$. We have

$$\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| \ge m\sqrt{\log(n)/n}) \le \mathbb{P}_3 + \mathbb{P}_4$$

where

$$\mathbb{P}_3 = \mathbb{P}(|1/n\sum_{i=1}^n \psi_{j,k}(G(X_i))(f(X_i) - \mathbb{E}\psi_{j,k}(G(X))f(X)| \ge m/2\sqrt{\log(n)/n}) \tag{6.18}$$

and

$$\mathbb{P}_4 = \mathbb{P}(|1/n\sum_{i=1}^n \psi_{j,k}(G(X_i))\varepsilon_i| > m/2\sqrt{(\log(n)/n)}). \tag{6.19}$$

Kerkyacharian and Picard in [20] in order to prove inequality (65) in [20] showed p. 1088 that

$$\mathbb{P}_3 \le 2\exp\left(-\frac{3m^2\log(n)}{4\|f\|_\infty(3+m)}\right) \tag{6.20}$$

if $2^j \le n/\log(n)$. As for $\mathbb{P}_4$, conditionally on $(X_1, \ldots, X_n)$ we have

$$1/n\sum_{i=1}^n \psi_{j,k}(G(X_i))\varepsilon_i \sim N(0, \gamma_{jk}^2)$$

where γ_{jk}^2 has been defined in (2.3).

$$\begin{aligned}\mathbb{P}_4 &\leq \mathbb{E}\left(\exp\left(-\frac{m^2\log(n)}{8n\gamma_{jk}^2}\right)\right)\\ &= \mathbb{E}e^{-\frac{m^2\log(n)}{8n\gamma_{jk}^2}}(I(|\gamma_{jk}^2-1/n|\leq\varsigma/n)+I(|\gamma_{jk}^2-1/n|>\varsigma/n))\\ &\leq e^{-\frac{m^2\log(n)}{8(\varsigma+1)}}+\mathbb{P}(|\gamma_{jk}^2-1/n|>\varsigma/n). \quad (6.21)\end{aligned}$$

Using (6.3) to bound $\mathbb{P}(|\gamma_{jk}^2-1/n|>\varsigma/n)$ we get

$$\mathbb{P}(|\hat\beta_{jk}-\beta_{jk}|>mt_n)\leq 2e^{-\varsigma^2\log(n)/(C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2)}+e^{-\frac{m^2\log(n)}{8(\varsigma+1)}}+2e^{(-\frac{3m^2\log(n)}{4\|f\|_\infty(3+m)})}$$

thus

$$\mathbb{P}(|\hat\beta_{jk}-\beta_{jk}|>mt_n)\leq 2n^{\frac{-\varsigma^2}{C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2}}+n^{\frac{-m^2}{8(\varsigma+1)}}+2n^{\frac{-3m^2}{4\|f\|_\infty(3+m)}} \quad (6.22)$$

which entails by fixing m and ς large enough

$$\begin{aligned}T_6 &\leq 2(mt_n)^{4s/(2s+1)}\|f\|^2_{W_{2/(1+2s)}}+t_n^2\sum_{j\leq J_n}\sum_k\beta_{jk}^2\\ &\leq 2(mt_n)^{4s/(2s+1)}\|f\|^2_{W_{2/(1+2s)}}+\|f(G^{-1})\|_2^2t_n^2.\end{aligned}$$

Let us look at the term $V_1^{"}$

$$V_1^{"}=\mathbb{E}\sum_{j\leq J_n}\sum_k(w_{jk}^2(\hat\beta_{jk}-\beta_{jk})^2+(1-w_{jk})^2\beta_{jk}^2)I\{|\hat\beta_{jk}|>mt_n\}$$

$$\begin{aligned}V_1^{"} &= \mathbb{E}\sum_{j\leq J_n}\sum_k(w_{jk}^2(\hat\beta_{jk}-\beta_{jk})^2+(1-w_{jk})^2\beta_{jk}^2)I\{|\hat\beta_{jk}|>mt_n\}\\ &\quad\times[I\{|\beta_{jk}|\leq mt_n/2\}+I\{|\beta_{jk}>mt_n/2|\}]\\ &= T_7+T_8\end{aligned}$$

for the term T_7, we use the Cauchy Scharwz inequality

$$\begin{aligned}T_7 &\leq \sum_{j\leq J_n}\sum_k(\mathbb{E}(\hat\beta_{jk}-\beta_{jk})^4)^{1/2}(\mathbb{P}(|\hat\beta_{jk}-\beta_{jk}|>mt_n/2))^{1/2}\\ &+ \sum_{j\leq J_n}\sum_k\beta_{jk}^2I\{|\hat\beta_{jk}|>mt_n\}I\{|\beta_{jk}|\leq mt_n/2\}.\end{aligned}$$

Furthermore, using inequality (64) p. 1086 in [20] we get for $2^j\leq n/\log(n)$

$$\mathbb{E}(\hat\beta_{jk}-\beta_{jk})^4\leq C\frac{1+\|f\|_\infty^4}{n^2} \quad (6.23)$$

and by (6.22)

$$\mathbb{P}(|\hat{\beta}_{jk}-\beta_{jk}|>mt_n/2)\leq 2n^{\frac{-\varsigma^2}{C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2}}+n^{\frac{-m^2}{32(\varsigma+1)}}+2n^{\frac{-3m^2}{16\|f\|_\infty(3+m)}}$$

from which follows by fixing again m and ς large enough

$$\begin{aligned} T_7 &\leq C/n.2^{J_n}(n^{\frac{-\varsigma^2}{C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2}}+n^{\frac{-m^2}{32(\varsigma+1)}}+2n^{\frac{-3m^2}{16\|f\|_\infty(3+m)}})^{1/2} \\ &\quad+\sum_{j\leq J_n}\sum_k \beta_{jk}^2 I\{|\beta_{jk}|\leq mt_n/2\} \\ &\leq t_n^2+((m/2)t_n)^{4s/(1+2s)}\|f\|^2_{W_{s/(2s+1)}}. \end{aligned}$$

For the term T_8

$$\begin{aligned} T_8 &= \mathbb{E}\sum_{j\leq J_n}\sum_k (w_{jk}^2(\hat{\beta}_{jk}-\beta_{jk})^2+(1-w_{jk})^2\beta_{jk}^2) \\ &\quad\times I\{|\hat{\beta}_{jk}|>mt_n\}I\{|\beta_{jk}>mt_n/2|\} \\ &\leq \frac{4m^{-2/(2s+1)}}{(1-2^{-2/(1+2s)})}\|f\|^2_{W_{2/(1+2s)}}(t_n)^{4s/(1+2s)} \\ &\quad+\mathbb{E}\sum_{j\leq J_n}\sum_k (1-w_{jk})^2\beta_{jk}^2 I\{|\hat{\beta}_{jk}|>mt_n\}I\{|\beta_{jk}>mt_n/2|\} \\ &\quad\times[I\{|\hat{\beta}_{jk}|\geq|\beta_{jk}/2|\}+I\{|\hat{\beta}_{jk}|<|\beta_{jk}/2|\}]. \end{aligned}$$

Hereafter we decompose

$$\begin{aligned} &\mathbb{E}\sum_{j\leq J_n}\sum_k (1-w_{jk})^2\beta_{jk}^2)I\{|\hat{\beta}_{jk}|>mt_n\}I\{|\beta_{jk}>mt_n/2|\} \\ &\qquad\times[I\{|\hat{\beta}_{jk}|\geq|\beta_{jk}/2|\}+I\{|\hat{\beta}_{jk}|<|\beta_{jk}/2|\}] \\ &\quad=T_8'+T_8^{"} \end{aligned}$$

$$T_8^{"}\leq\sum_{j\leq J_n}\sum_k\beta_{jk}^2\mathbb{P}(|\hat{\beta}_{jk}-\beta_{jk}|>mt_n/4)$$

using (6.22) we get for m and ς large enough

$$T_8^{"}\leq(2n^{\frac{-\varsigma^2}{C\|\psi\|_4^4+\varsigma\|\psi\|_\infty^2}}+n^{\frac{-m^2}{128(\varsigma+1)}}+2n^{\frac{-3m^2}{64\|f\|_\infty(3+m)}})\sum_{j\leq J_n}\sum_k\beta_{jk}^2\leq t_n^2$$

as for $T_8^{'}$

$$T_8^{'}=\mathbb{E}\sum_{j\leq J_n}\sum_k(1-w_{jk})^2\beta_{jk}^2 I\{|\hat{\beta}_{jk}|>mt_n\}I\{|\beta_{jk}|>mt_n/2\}I\{|\hat{\beta}_{jk}|\geq|\beta_{jk}|/2\}$$

using (6.17) we get

$$\begin{aligned} T_8^{'} &\leq \mathbb{E}\sum_{j\leq J_n}\sum_k K^2\beta_{jk}^2\left(\frac{t_n}{|\hat{\beta}_{jk}|}+t_n\right)^2 I\{|\hat{\beta}_{jk}|\geq|\beta_{jk}|/2\}I\{|\beta_{jk}|>mt_n/2\} \\ &\leq K^2\sum_{j\leq J_n}\sum_k\beta_{jk}^2\left(\frac{2t_n}{|\beta_{jk}|}+t_n\right)^2 I\{|\beta_{jk}|>mt_n/2\} \\ &\leq 2K^2\sum_{j\leq J_n}\sum_k\beta_{jk}^2\left(\frac{4t_n^2}{|\beta_{jk}|^2}+t_n^2\right)I\{|\beta_{jk}|>mt_n/2\} \\ &= 8K^2t_n^2\sum_{j\leq J_n}\sum_k I\{|\beta_{jk}|>mt_n/2\}+2K^2t_n^2\|f(G^{-1})\|_2^2 \end{aligned}$$

using (3.3) it follows

$$\begin{aligned} T_8^{'} &\leq 8K^2t_n^2\left(\frac{mt_n}{2}\right)^{-2/(1+2s)}\frac{2^{2-2/(1+2s)}}{1-2^{-2/(1+2s)}}\|f\|^2_{W_{2/(1+2s)}}+2K^2t_n^2\|f(G^{-1})\|_2^2 \\ &\leq 32K^2\frac{m^{-2/(1+2s)}}{1-2^{-2/(1+2s)}}t_n^{4s/(1+2s)}+2K^2t_n^2\|f(G^{-1})\|_2^2. \end{aligned}$$

It remains to bound the bias term B_1. To this purpose we use the fact that $f\in B^s_{2,\infty}$

$$B_{-1}=\|\sum_{j>J_n}\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{j,k}(G(x))\|_2^2\leq C2^{-2J_ns}=Ct_n^{2s}\leq Ct_n^{4s/(2s+1)}$$

which completes the proof.

Proof of Theorem 2.
In order to prove the Theorem 2., we have to prove that the Bayesian estimators (2.4) based on Gaussian priors with large variance (2.10) and (2.11) satisfy the conditions of Lemma 4.

We will not get into details of the proof because this latter is identical to the proof of Theorem 3. in [5], with the sole exception that here, the proof is carried over the event Ω_n^δ with $\delta=\varsigma/n$, ς some positive constant. Indeed, as precised above in section 2.2, a key observation is that instead of having a deterministic noise $\varepsilon=1/\sqrt{n}$ like in [5], here we have to deal with a stochastic noise γ_{jk}^2 which expression is given by (2.3).

Acknowledgements

The author wishes to thank her advisor Dominique Picard and Vincent Rivoirard for interesting discussions and suggestions.

References

[1] Abramovich, F., Amato, U. and C. Angelini. On optimality of Bayesian wavelet estimators. *Scand. J. Statist.*, 31(2):217–234, 2004. MR2066250

[2] Abramovich F., Sapatinas, T. and Silverman, B. W. Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):725–749, 1998. MR1649547

[3] Amato, U., Antoniadis, A. and Pensky, M. Wavelet kernel penalized estimation for non-equispaced design regression. *Stat. Comput.*, 16(1):37–55, 2006. MR2224188

[4] Antoniadis, A., Grégoire, G. and Vial, P. Random design wavelet curve smoothing. 35:225–232, 1997. MR1484959

[5] Autin, F., Picard, D. and Rivoirard, V. Large variance Gaussian priors in Bayesian nonparametric estimation: a maxiset approach. *Math. Methods of Statist.*, 15(4):349–373, 2006. MR2301657

[6] Baraud, Y. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146, 2002. MR1918295

[7] Brutti, P. Warped wavelets and vertical thresholding. Preprint. arxiv:0801.3319v1.

[8] Cai, T. T. and Brown, L. D. Wavelet shrinkage for nonequispaced samples. *Ann. Statist.*, 26:1783–1799, 1998. MR1673278

[9] Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. Adaptive Bayesian Wavelet Shrinkage. *J. Amer. Statist. Assoc.*, 92:1413–1421, 1997.

[10] Clyde, M. and George, E. I. Flexible Empirical Bayes Estimation for Wavelets. *J. Roy. Statist. Soc., Ser. B.*, 62(4):681–698, 2000. MR1796285

[11] Clyde, M., Parmigiani, G. and Vidakovic, B. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2):391–401, 1998. MR1649120

[12] Clyde, M. A. and George, E. I. Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian inference in wavelet-based models*, volume 141 of *Lecture Notes in Statist.*, pages 309–322. Springer, New York, 1999. MR1699849

[13] Donoho, D. L. and Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. MR1311089

[14] Donoho, D. L. and Johnstone, I. M. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995. MR1379464

[15] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors. MR1323344

[16] Gannaz, I. *Estimation par ondelettes dans les modèles partiellement linéaires*. Ph.D. Thesis, Université Joseph Fourier, 2007.

[17] Hall, P. and Turlach, B. A. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Statist.*, 25:1912–1925, 1997. MR1474074

[18] Johnstone, I. M. and Silverman, B. W. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, 2004. MR2089135

[19] Johnstone, I. M. and Silverman, B. W. Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, 33(4):1700–1752, 2005. MR2166560

[20] Kerkyacharian, G. and Picard, D. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004. MR2108043

[21] Kerkyacharian, G. and Picard, D. Thresholding algorithms, maxisets and well concentrated basis. *Test*, 9(2):283–344, 2004. MR1821645

[22] Kovac, A. and Silverman, B. W. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.*, 95:172–183, 2000.

[23] Massart, P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. MR2319879

[24] Pensky, M. Frequentist optimality of Bayesian wavelet shrinkage rules for Gaussian and non-Gaussian noise. *Ann. Statist.*, 34(2):769–807, 2006. MR2283392

[25] Rivoirard, V. Bayesian modeling of sparse sequences and maxisets for Bayes rules. *Math. Methods Statist.*, 14(3):346–376, 2005. MR2195330

[26] Willer, T. *Estimation non paramétrique et problèmes inverses*. Ph.D. Thesis, Université Paris VII, 2006.