Bernoulli **14**(4), 2008, 1089–1107
DOI: 10.3150/08-BEJ135

Mixing least-squares estimators when the variance is unknown

CHRISTOPHE GIRAUD

Université de Nice Sophia-Antipolis, Laboratoire J.-A. Dieudonné, Parc Valrose, 06108 Nice cedex 02, France. E-mail: cgiraud@math.unice.fr

We propose a procedure to handle the problem of Gaussian regression when the variance is unknown. We mix least-squares estimators from various models according to a procedure inspired by that of Leung and Barron [*IEEE Trans. Inform. Theory* **52** (2006) 3396–3410]. We show that in some cases, the resulting estimator is a simple shrinkage estimator. We then apply this procedure to perform adaptive estimation in Besov spaces. Our results provide non-asymptotic risk bounds for the Euclidean risk of the estimator.

Keywords: adaptive minimax estimation; Gibbs mixture; linear regression; oracle inequalities; shrinkage estimator

1. Introduction

We consider the regression framework, where we have noisy observations

$$Y_i = \mu_i + \sigma\varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

of an unknown vector $\mu = (\mu_1, \dots, \mu_n)' \in \mathbb{R}^n$. We assume that the ε_i's are i.i.d. standard Gaussian random variables and that the noise level $\sigma > 0$ is unknown. Our aim is to estimate μ.

In this direction, we introduce a finite collection $\{\mathcal{S}_m, m \in \mathcal{M}\}$ of linear spaces of $\mathbb{R}^n$, which we shall henceforth call *models*. To each model $\mathcal{S}_m$, we associate the least-squares estimator $\hat{\mu}_m = \Pi_{\mathcal{S}_m} Y$ of μ on $\mathcal{S}_m$, where $\Pi_{\mathcal{S}_m}$ denotes the orthogonal projector onto $\mathcal{S}_m$. The L^2-risk of the estimator $\hat{\mu}_m$ with respect to the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$ is

$$\mathbb{E}[\|\mu - \hat{\mu}_m\|^2] = \|\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + \dim(\mathcal{S}_m)\sigma^2. \tag{2}$$

Two strategies have emerged to handle the problem of the choice of an estimator of μ in this setting. One strategy is to select a model $\mathcal{S}_{\hat{m}}$ with a data-driven criterion and use $\hat{\mu}_{\hat{m}}$ to estimate μ. In the favorable cases, the order of risk of this estimator is the minimum over $\mathcal{M}$ of the risks (2). Model selection procedures have received a lot of attention in the literature, starting from the pioneering work of Akaike [1] and Mallows [19]. It is beyond the scope of this paper to provide a historical review of the topic. We simply mention, in the Gaussian setting, the papers of Birgé and Massart [7,8] (influenced by Barron and Cover [5] and Barron, Birgé and Massart [4]) which give non-asymptotic risk bounds for a selection criterion generalizing Mallows' C_p.

An alternative to model selection is mixing. One estimates μ by a convex (or linear) combination of the $\hat{\mu}_m$'s,

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m, \tag{3}$$

with weights w_m which are $\sigma(Y)$-measurable random variables. This strategy is not suitable when the goal is to select a single model $\mathcal{S}_{\hat{m}}$; nevertheless, it enjoys the nice property that $\hat{\mu}$ may perform better than the best of the $\hat{\mu}_m$'s. Various choices of weights w_m have been proposed, from an information-theoretic or Bayesian perspective. Risk bounds have been provided by Catoni [11], Yang [25,27], Tsybakov [23] and Bunea *et al.* [9] for regression on a random design, and by Barron [3], Catoni [10] and Yang [26] for density estimation. For the Gaussian regression framework we consider here, Leung and Barron [18] propose a mixing procedure for which they derive a precise non-asymptotic risk bound. When the collection of models is not too complex, this bound shows that the risk of their estimator $\hat{\mu}$ is close to the minimum over $\mathcal{M}$ of the risks (2). Another nice feature of their mixing procedure is that both the weights w_m and the estimators $\hat{\mu}_m$ are built on the same data set. This enables their procedure to handle cases where we cannot split the data into several sets, for example, in the fixed-design regression framework. Unfortunately, their choice of the weights w_m depends on the variance σ^2, which is usually unknown.

In the present paper, we consider the more practical situation where the variance σ^2 is unknown. Our mixing strategy is akin to that of Leung and Barron [18], but is not dependent on the variance σ^2. In addition, we show that both our estimator and the estimator of Leung and Barron are simple shrinkage estimators in some cases. From a theoretical point of view, we relate our weights w_m to a Gibbs measure on $\mathcal{M}$ and derive a sharp risk bound for the estimator $\hat{\mu}$. Roughly, this bound says that the risk of $\hat{\mu}$ is close to the minimum over $\mathcal{M}$ of the risks (2) in the favorable cases. We then discuss the choice of the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$ in various situations. Among others, we produce an estimation procedure which is rate-minimax adaptive over a large class of Besov balls.

Before presenting our mixing procedure, we briefly recall that of Leung and Barron [18]. Assuming that the variance σ^2 is known, they use the weights

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp\bigl(-\beta[\|Y - \hat{\mu}_m\|^2/\sigma^2 + 2\dim(\mathcal{S}_m) - n]\bigr), \qquad m \in \mathcal{M}, \tag{4}$$

where $\{\pi_m,\ m \in \mathcal{M}\}$ is a given prior distribution on $\mathcal{M}$ and $\mathcal{Z}$ normalizes the sum of the w_m's to one. These weights have a Bayesian flavor. Indeed, they appear with $\beta = 1/2$ in Hartigan [16], which considers the Bayes procedure with the following (improper) prior distribution: pick an m in $\mathcal{M}$ according to π_m and then sample μ 'uniformly' on $\mathcal{S}_m$. Nevertheless, in Leung and Barron [18], the role of the prior distribution $\{\pi_m,\ m \in \mathcal{M}\}$ is to favor models with low complexity. Therefore, the choice of π_m is driven by the complexity of the model $\mathcal{S}_m$ rather than from a prior knowledge of μ. In this sense their approach differs from the classical Bayesian point of view. Note that the term $\|Y - \hat{\mu}_m\|^2/\sigma^2 + 2\dim(\mathcal{S}_m) - n$ appearing in the weights (4) is an unbiased estimator of the risk (2) rescaled by σ^2. The size of the weight w_m then depends on the difference between this estimator of the risk (2) and $-\log(\pi_m)$, which can be thought as a complexity-driven penalty (in the spirit of Barron and Cover [5] or Barron *et al.* [4]). The

parameter β tunes the balance between these two terms. For $\beta \leq 1/4$, Theorem 5 of [18] provides a sharp risk bound for the procedure.

The rest of the paper is organized as follows. We present our mixing strategy in the next section and express in some cases the resulting estimator $\hat{\mu}$ as a shrinkage estimator. In Section 3, we state non-asymptotic risk bounds for the procedure and discuss the choice of the tuning parameters. Finally, in Section 4, we propose some weighting strategies for estimating BV functions or for adaptive regression over Besov balls. Section 6 is devoted to the proofs.

We end this section with some notation which we shall use throughout this paper. We write $|m|$ for the cardinality of a finite set m and $\langle x, y\rangle$ for the inner product of two vectors x and y in $\mathbb{R}^n$. For any real number x, we denote by $(x)_+$ its positive part and by $\lfloor x \rfloor$ its integer part.

2. The estimation procedure

We assume henceforth that $n \geq 3$.

2.1. The estimator

We start with a finite collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$ and to each model $\mathcal{S}_m$, we associate the least-squares estimator $\hat{\mu}_m = \Pi_{\mathcal{S}_m} Y$ of μ on $\mathcal{S}_m$. We also introduce a probability distribution $\{\pi_m,\ m \in \mathcal{M}\}$ on $\mathcal{M}$, which is meant to take into account the complexity of the family and to favor models with low dimension. For example, if the collection $\{\mathcal{S}_m, m \in \mathcal{M}\}$ has (at most) e^{ad} models per dimension d, we suggest the choise $\pi_m \propto \mathrm{e}^{(a+1/2)\dim(\mathcal{S}_m)}$; see the example at the end of Section 3.1. As mentioned before, the quantity $-\log(\pi_m)$ can be interpreted as a complexity-driven penalty associated to the model $\mathcal{S}_m$ (in the sense of Barron *et al.* [4]). The performance of our estimation procedure strongly depends on the choice of the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$ and the probability distribution $\{\pi_m, m \in \mathcal{M}\}$. In Section 4, we discuss some suitable choices of these families for estimating BV or Besov functions.

Hereafter, we assume that there exists a linear space $\mathcal{S}_* \subset \mathbb{R}^n$ of dimension $p < n$ such that $\mathcal{S}_m \subset \mathcal{S}_*$ for all $m \in \mathcal{M}$. We will then roughly estimate the variance of the noise by

$$\hat{\sigma}^2 = \frac{\|Y - \Pi_{\mathcal{S}_*} Y\|^2}{N_*}, \tag{5}$$

where $N_* = n - p$. We emphasize that we *do not* assume that $\mu \in \mathcal{S}_*$ and the estimator $\hat{\sigma}^2$ is (positively) biased in general. It turns out that our estimation procedure does not need a precise estimation of the variance σ^2 and the choice (5) gives good results; see the discussion in the next section.

Finally, we associate to the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$ a collection $\{L_m, m \in \mathcal{M}\}$ of non-negative weights. We recommend setting $L_m = \dim(\mathcal{S}_m)/2$, but any (sharp) upper bound of this quantity may also be appropriate; see the discussion after Theorem 1. Then, for a given positive constant β, we define the estimator $\hat{\mu}$ by

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m \qquad \text{with } w_m = \frac{\pi_m}{\mathcal{Z}} \exp\left(\beta \frac{\|\hat{\mu}_m\|^2}{\hat{\sigma}^2} - L_m\right), \tag{6}$$

where $\mathcal{Z}$ is a constant that normalizes the sum of the w_m's to one. An alternative formula for w_m is $w_m = \pi_m \exp(-\beta\|Y - \hat{\mu}_m\|^2/\hat{\sigma}^2 - L_m)/\mathcal{Z}'$ with $\mathcal{Z}' = \mathrm{e}^{-\beta\|Y\|^2/\hat{\sigma}^2}\mathcal{Z}$. We can interpret the term $\|Y - \hat{\mu}_m\|^2/\hat{\sigma}^2 + L_m/\beta$ appearing in the exponential as a (biased) estimate of the risk (2) rescaled by σ^2. As in (4), the balance in the weight w_m between this estimate of the risk and the penalty $-\log(\pi_m)$ is tuned by β. We refer to the discussion after Theorem 1 for the choice of this parameter. The weights $\{w_m, m \in \mathcal{M}\}$ can be viewed as a Gibbs measure on $\mathcal{M}$ and we will use this property to assess the performance of the procedure. We will also see in Section 2.3 that $\hat{\mu}$ reduces to a simple shrinkage estimator in some cases.

2.2. On the choice of $\hat{\sigma}^2$

In our estimation procedure, we *do not* assume that $\mu \in \mathcal{S}_*$, so $\hat{\sigma}^2$ has a bias $\|\mu - \Pi_{\mathcal{S}_*}\mu\|^2/N_*$. In particular, if $\Pi_{\mathcal{S}_*}\mu$ is a poor approximation of μ, then both $\hat{\sigma}^2$ and the $\{\hat{\mu}_m,\ m \in \mathcal{M}\}$ behave poorly. As explained in Section 3, we can actually consider spaces $\mathcal{S}_*$ with codimension N_* of order only $\log n$, which means that $\mathcal{S}_*$ can be very large. A clever choice of $\mathcal{S}_*$ should then prevent a large bias. For example, let us consider the case where $\mu_i = f(x_i)$ and the models $\mathcal{S}_m$ are built on wavelets. The variance is then estimated by regressing the observations on the wavelets of high order. The result of Section 4.2 shows that when the f belongs to some Besov ball $\mathcal{B}^{\alpha}_{p,\infty}(R)$, this estimation of the variance is reasonable enough to obtain an adaptive rate-minimax estimation of the signal.

Finally, we may, in practice, replace the residual estimator $\hat{\sigma}^2$ by a difference-based estimator (Rice [21], Hall *et al.* [15], Munk *et al.* [20], Tong and Wang [22], Wang *et al.* [28], etc.) or by any nonparametric estimator (e.g., Lenth [17]). These choices should give good results, but we are not able to prove any bound similar to (11) or (12) when using one of these estimators.

2.3. A simple shrinkage estimator

In this section, we focus on the case where $\mathcal{M}$ consists of all the subsets of $\{1, \dots, p\}$, for some $p < n$ and $\mathcal{S}_m = \mathrm{span}\{v_j, j \in m\}$ with $\{v_1, \dots, v_p\}$ an orthonormal family of vectors in $\mathbb{R}^n$. We use the convention $\mathcal{S}_\emptyset = \{0\}$ and $\mathcal{S}_*$ corresponds here to $\mathcal{S}_{\{1,\dots,p\}}$. An example of such a setting is given in Section 4.1.

To favor models with small dimensions, we choose the probability distribution π,

$$\pi_m = \left(1 + \frac{1}{p^\alpha}\right)^{-p} p^{-\alpha|m|}, \qquad m \in \mathcal{M}, \tag{7}$$

with $\alpha > 0$. We also set $L_m = b|m|$ for some $b \geq 0$.

Proposition 1. *Under the above assumptions, we have the following expression for* $\hat{\mu}$:

$$\hat{\mu} = \sum_{j=1}^{p} (c_j Z_j) v_j, \qquad \textit{with } Z_j = \langle Y, v_j \rangle \textit{ and } c_j = \frac{\exp(\beta Z_j^2/\hat{\sigma}^2)}{p^\alpha \exp(b) + \exp(\beta Z_j^2/\hat{\sigma}^2)}. \tag{8}$$

The proof of this proposition is postponed to Section 6.1. The main interest of (8) is to allow a fast computation of $\hat{\mu}$. Indeed, we only need to compute the p coefficients c_j instead of the 2^p weights w_m of (6).

The coefficients c_j are shrinkage coefficients taking values in $[0, 1]$. They are close to one when Z_j is large and close to zero when Z_j is small. The transition from 0 to 1 occurs when $Z_j^2 \approx \beta^{-1}(b + \alpha \log p)\hat{\sigma}^2$. The choice of the tuning parameters α, β and b will be discussed in Section 3.2.

Remark 1. Other choices are possible for $\{\pi_m,\ m \in \mathcal{M}\}$ and they lead to different c_j's. Let us mention the choice $\pi_m = ((p+1)\binom{p}{|m|})^{-1}$, for which the c_j's are given by

$$c_j = \frac{\int_0^1 q \prod_{k \neq j}[q + (1-q)\exp(-\beta Z_k^2/\hat{\sigma}^2 + b)]\,\mathrm{d}q}{\int_0^1 \prod_{k=1}^p [q + (1-q)\exp(-\beta Z_k^2/\hat{\sigma}^2 + b)]\,\mathrm{d}q} \qquad \text{for } j = 1, \ldots, p.$$

This formula can be derived from the Appendix of Leung and Barron [18].

Remark 2. When the variance is known, we can give a formula similar to (8) for the estimator of Leung and Barron [18]. Let us consider the same setting, with $p \leq n$. Then, when the distribution $\{\pi_m,\ m \in \mathcal{M}\}$ is given by (7), the estimator (3) with weights w_m given by (4) takes the form (8) with $c_j = \mathrm{e}^{\beta Z_j^2/\sigma^2}/(p^\alpha \mathrm{e}^{2\beta} + \mathrm{e}^{\beta Z_j^2/\sigma^2})$.

3. The performance

3.1. A general risk bound

The next result gives an upper bound on the L^2-risk of the estimation procedure. We remind the reader that $n \geq 3$ and set

$$\phi(x) = \tfrac{1}{2}(x - 1 - \log x), \tag{9}$$

which is decreasing on $]0, 1[$.

Theorem 1. *Assume that β and N_* fulfill the condition*

$$\beta < 1/4 \quad \textit{and} \quad N_* \geq 2 + \frac{\log n}{\phi(4\beta)}, \tag{10}$$

with ϕ defined by (9). Assume also that $L_m \geq \dim(\mathcal{S}_m)/2$ for all $m \in \mathcal{M}$. Then, we have the following upper bounds on the L^2-risk of the estimator $\hat{\mu}$:

$$\begin{aligned}&\mathbb{E}(\|\mu - \hat{\mu}\|^2)\\ &\quad \leq -(1+\varepsilon_n)\frac{\bar{\sigma}^2}{\beta}\log\left[\sum_{m\in\mathcal{M}} \pi_m \mathrm{e}^{-\beta[\|\mu - \Pi_{\mathcal{S}_m}\mu\|^2 - \dim(\mathcal{S}_m)\sigma^2]/\bar{\sigma}^2 - L_m}\right] + \frac{\sigma^2}{2\log n}\end{aligned} \tag{11}$$

$$\leq (1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar\sigma^2}{\beta}(L_m-\log\pi_m)\right\}+\frac{\sigma^2}{2\log n}, \tag{12}$$

where $\varepsilon_n=(2n\log n)^{-1}$ *and* $\bar\sigma^2=\sigma^2+\|\mu-\Pi_{\mathcal{S}_*}\mu\|^2/N_*$.

The proof Theorem 1 is delayed to Section 6.3. Let us comment on this result.

We would like to compare the bounds of Theorem 1 with the minimum over $\mathcal{M}$ of the risks given by (2). Roughly, the bound (12) states that the estimator $\hat\mu$ achieves the best trade-off between the bias $\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2/\sigma^2$ and the complexity term $C_m=L_m-\log\pi_m$. More precisely, we derive from (12) the (cruder) bound

$$\mathbb{E}(\|\mu-\hat\mu\|^2)\leq(1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{1}{\beta}C_m\sigma^2\right\}+R_n^*\sigma^2, \tag{13}$$

with

$$\varepsilon_n=\frac{1}{2n\log n}\quad\text{and}\quad R_n^*=\frac{1}{2\log n}+\frac{\|\mu-\Pi_{\mathcal{S}_*}\mu\|^2}{\beta N^*\sigma^2}\sup_{m\in\mathcal{M}}C_m.$$

In particular, if C_m is of order $\dim(\mathcal{S}_m)$, then (13) allows to compare the risk of $\hat\mu$ and the infimum of the risks (2). We discuss this point in the following example.

Example. Assume that the family $\mathcal{M}$ has an index of complexity (K,a), as defined in [2], which means that $|\{m\in\mathcal{M},\dim(\mathcal{S}_m)=d\}|\leq K\mathrm{e}^{ad}$ for all $d\geq 1$. If we choose

$$\pi_m=\frac{\mathrm{e}^{-(a+1/2)\dim(\mathcal{S}_m)}}{\sum_{m'\in\mathcal{M}}\mathrm{e}^{-(a+1/2)\dim(\mathcal{S}_{m'})}}\quad\text{and}\quad L_m=\dim(\mathcal{S}_m)/2, \tag{14}$$

then we have $C_m\leq(a+1)\dim(\mathcal{S}_m)+\log(3K)$. Therefore, when β is given by (15) and $p\leq\kappa n$ for some $\kappa<1$, we have

$$\mathbb{E}(\|\mu-\hat\mu\|^2)\leq(1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{a+1}{\beta}\dim(\mathcal{S}_m)\sigma^2\right\}+R_n'\sigma^2,$$

with

$$R_n'=\frac{\log(3K)}{\beta}+\frac{1}{2\log n}+\frac{\|\mu-\Pi_{\mathcal{S}_*}\mu\|^2}{\sigma^2}\times\frac{(a+1)\kappa+n^{-1}\log(3K)}{\beta(1-\kappa)}.$$

In particular, for a given index of complexity (K,a) and a given κ, the previous bound gives an oracle inequality.

Let us discuss the bound (11). It may look somewhat cumberstone, but it improves (12) when there are several good models to estimate μ. For example, we can derive from (11) the bound

$$\mathbb{E}(\|\mu-\hat\mu\|^2)\leq(1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar\sigma^2}{\beta}(L_m-\log\pi_m)\right\}$$

$$+\inf_{\delta\geq 0}\left\{\delta-\frac{\bar{\sigma}^2}{\beta}\log|\mathcal{M}_\delta|\right\}+\frac{\sigma^2}{2\log n},$$

where $\mathcal{M}_\delta$ is the set made of those m^* in $\mathcal{M}$ fulfilling

$$\|\mu-\Pi_{\mathcal{S}_{m^*}}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}(L_{m^*}-\log\pi_{m^*})\leq\delta+\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}(L_m-\log\pi_m)\right\}.$$

In the extreme case where all the quantities $\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}(L_m-\log\pi_m)$ are equal, (11) then improves (12) by a factor of $\beta^{-1}\bar{\sigma}^2\log|\mathcal{M}|$.

3.2. How to choose the parameters β and $\{L_m, m\in\mathcal{M}\}$

The choice of the tuning parameters β and $\{L_m, m\in\mathcal{M}\}$ is important in practice. The choice $L_m=\dim(\mathcal{S}_m)/2$ seems to be the most accurate since it satisfies the conditions of Theorem 1 and minimizes the right-hand side of (11) and (12). We shall mostly use this one in the following, but there are some cases where it is easier to use some (sharp) upper bound of the dimension of $\mathcal{S}_m$ instead of $\dim(\mathcal{S}_m)$ itself; see, for example, Section 4.2.

We now turn to the choice of the parameter β. The largest parameter β fulfilling condition (10) is

$$\beta=\frac{1}{4}\phi^{-1}\left(\frac{\log n}{N_*-2}\right)<\frac{1}{4}. \tag{15}$$

The choice of this value for β seems to be advisable, since it minimizes the right-hand side of (11) and (12). Nevertheless, Bayesian arguments [16] suggest taking a larger value for β, namely $\beta=1/2$. We discuss this issue below in the example of Section 2.3.

For the sake of simplicity, we will restrict our discussion to the case where the variance is known (see [14] for the case where the variance is unknown). We consider the weights (4) proposed by Leung and Barron, with the probability distribution $\pi_m=(1+p^{-1})^{-p}p^{-|m|}$. According to Remark 2 of Section 2.3, the estimator $\hat{\mu}$ then takes the form

$$\hat{\mu}=\sum_{j=1}^{p}s_\beta(Z_j/\sigma)Z_jv_j,\qquad\text{with } Z_j=\langle Y,v_j\rangle \text{ and } s_\beta(z)=\frac{\mathrm{e}^{\beta z^2}}{p\mathrm{e}^{2\beta}+\mathrm{e}^{\beta z^2}}. \tag{16}$$

First, we note that a choice $\beta>1/2$ is not recommended. Indeed, we can compare the shrinkage coefficient $s_\beta(Z_j/\sigma)$ to a threshold at level $T=(2+\beta^{-1}\log p)\sigma^2$ since $s_\beta(Z_j/\sigma)\geq\frac{1}{2}\mathbf{1}_{\{Z_j^2\geq T\}}$. For $\mu=0$, the risk of $\hat{\mu}$ is then larger than a quarter of the risk of the threshold estimator $\hat{\mu}_T=\sum_{j=1}^{p}\mathbf{1}_{\{Z_j^2\geq T\}}Z_jv_j$, namely,

$$\mathbb{E}(\|0-\hat{\mu}\|^2)=\sum_{j=1}^{p}\mathbb{E}(s_\beta(Z_j/\sigma)^2Z_j^2)\geq\tfrac{1}{4}\sum_{j=1}^{p}\mathbb{E}\big(\mathbf{1}_{\{Z_j^2\geq T\}}Z_j^2\big)=\tfrac{1}{4}\mathbb{E}(\|0-\hat{\mu}_T\|^2).$$

Now, when the threshold T is of order $2K \log p$ with $K < 1$, the threshold estimator is known to behave poorly for $\mu = 0$; see [7] Section 7.2. Therefore, a choice $\beta > 1/2$ would give poor results, at least when $\mu = 0$.

The next proposition justifies the use of any $\beta \leq 1/2$ by a risk bound similar to (12). For $p \geq 1$ and $\beta > 0$, we introduce the numerical constants $\gamma_\beta(p) = \sqrt{2 + \beta^{-1} \log p}$ and

$$c_\beta(p) = 0.6 \vee \sup\left\{ \frac{\int_{\mathbb{R}} (x - (x+z) s_\beta(x+z))^2 \mathrm{e}^{-z^2/2} \, \mathrm{d}z}{\sqrt{2\pi}(\min(x^2, \gamma_\beta(p)^2) + \gamma_\beta(p)^2/p)};\ x \in [0, 4\gamma_\beta(p)] \right\}.$$

The constant $c_\beta(p)$ can be numerically computed. For example, $c_{1/2}(p) \leq 1$ for any $3 \leq p \leq 10^6$; see Figure 1.

Proposition 2. *For $3 \leq p \leq n$ and $\beta \in [1/4, 1/2]$, the L^2-risk of the estimator (16) is upper bounded by*

$$\mathbb{E}(\|\mu - \hat{\mu}\|^2) \leq \|\mu - \Pi_{\mathcal{S}_*}\mu\|^2 + c_\beta(p) \inf_{m \in \mathcal{M}} [\|\Pi_{\mathcal{S}_*}\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + (2 + \beta^{-1} \log p)(|m| + 1)\sigma^2].$$

In the light of this risk bound, the choice $\beta = 1/2$ seems to be a good one in this case and corresponds to the choice of Hartigan [16]. The proof of this result can be found in the technical report [14], together with further comments on the choice of the parameter β. We emphasize that the risk bound stated in Proposition 2 differs from the best trade-off between the bias and the

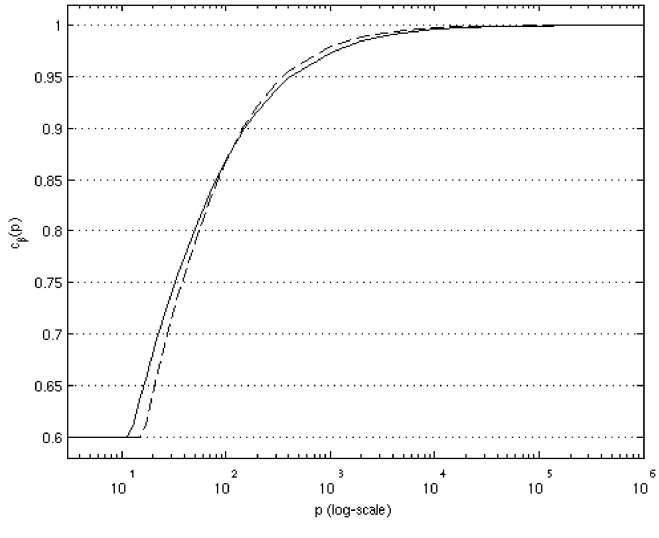

Figure 1. Plots of $p \mapsto c_{1/2}(p)$ (solid line) and $p \mapsto c_{1/4}(p)$ (dashed line).

variance term by a factor of $\log p$. This is unavoidable from a minimax point of view, as observed by Donoho and Johnstone [13].

4. Choice of the models and the weights in two different settings

4.1. Estimation of BV functions

We consider here the functional setting

$$\mu_i = f(x_i), \qquad i = 1, \dots, n, \tag{17}$$

where $f : [0, 1] \to \mathbb{R}$ is an unknown function and $x_1, \dots, x_n$ are n deterministic points of $[0, 1]$. We assume, for simplicity, that $0 = x_1 < x_2 < \cdots < x_n < x_{n+1} = 1$ and $n = 2^{J_n} \geq 8$. We set $J^* = J_n - 1$ and $\Lambda^* = \bigcup_{j=0}^{J^*} \Lambda(j)$ with $\Lambda(0) = \{(0, 0)\}$ and $\Lambda(j) = \{j\} \times \{0, \dots, 2^{j-1} - 1\}$ for $j \geq 1$. For $(j, k) \in \Lambda^*$, we define $v_{j,k} \in \mathbb{R}^n$ by

$$[v_{j,k}]_i = 2^{(j-1)/2}\big(\mathbf{1}_{I^+_{j,k}}(i) - \mathbf{1}_{I^-_{j,k}}(i)\big), \qquad i = 1, \dots, n,$$

with $I^+_{j,k} = \{1 + (2k + 1)2^{-j}n, \dots, (2k + 2)2^{-j}n\}$ and $I^-_{j,k} = \{1 + 2k2^{-j}n, \dots, (2k + 1)2^{-j}n\}$. The family $\{v_{j,k},\ (j, k) \in \Lambda^*\}$ corresponds to the image of the points $x_1, \dots, x_n$ by a Haar basis (see Section 6.4) and it is orthonormal for the scalar product

$$\langle x, y\rangle_n = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

We use the collection of models $\mathcal{S}_m = \operatorname{span}\{v_{j,k}, (j, k) \in m\}$ indexed by $\mathcal{M} = \mathcal{P}(\Lambda^*)$ and thereby adopt the setting of Section 2.3. We choose the distribution π given by (7) with $p = n/2$ and $\alpha = 1$. We also set $b = 1$ and take some β satisfying $\beta \leq \phi^{-1}(2\log(n/2)/n)/2$. According to Proposition 1, the estimator (6) then takes the form

$$\hat{\mu} = \sum_{j=0}^{J^*} \sum_{k\in\Lambda(j)} \left(\frac{Z_{j,k}\exp(n\beta Z^2_{j,k}/\hat{\sigma}^2)}{\mathrm{e}n/2 + \exp(n\beta Z^2_{j,k}/\hat{\sigma}^2)}\right) v_{j,k}, \tag{18}$$

with $Z_{j,k} = \langle Y, v_{j,k}\rangle_n$ and $\hat{\sigma}^2 = 2(\langle Y, Y\rangle^2_n - \sum_{j=0}^{J^*}\sum_{k\in\Lambda(j)} Z^2_{j,k})$.

The next corollary gives the rate of convergence of this estimator when f has bounded variation, in terms of the norm $\|\cdot\|_n$ induced by the scalar product $\langle\cdot,\cdot\rangle_n$.

Corollary 1. *In the setting described above, there exists a numerical constant C such that for any function f with bounded variation $V(f)$,*

$$\mathbb{E}(\|\mu - \hat{\mu}\|^2_n) \leq C \max\left\{\left(\frac{V(f)\sigma^2\log n}{n}\right)^{2/3}, \frac{V(f)^2}{n}, \frac{\sigma^2\log n}{n}\right\}.$$

The proof is delayed to Section 6.4. The minimax rate in this setting is $(V(f)\sigma^2/n)^{2/3}$. So, the rate of convergence of the estimator differs from the minimax rate by a factor of $(\log n)^{2/3}$. We can actually obtain a rate-minimax estimator by using a smaller collection of models similar to the one in the next section, but then we do not have the nice formula (18).

4.2. Regression on Besov space $\mathcal{B}^{\alpha}_{p,\infty}[0,1]$

We again consider the setting (17) with $f:[0,1]\to\mathbb{R}$ and introduce an $L^2([0,1],\mathrm{d}x)$-orthonormal family $\{\phi_{j,k}, j\geq 0, k=1,\dots,2^j\}$ of compactly support wavelets with regularity r. We will use models generated by finite subsets of wavelets. If we want our estimator to share some good adaptive properties on Besov spaces, we shall introduce a family of models induced by the compression algorithm of Birgé and Massart [7]. This collection turns out to be slightly more intricate than the family used in the previous section. We start with some $\kappa<1$ and set $J_*=\lfloor\log(\kappa n/2)/\log 2\rfloor$. The largest approximation space we will consider is $\mathcal{F}_*=\mathrm{span}\{\phi_{j,k}, j=0,\dots,J_*, k=1,\dots,2^j\}$, whose dimension is bounded by κn. For $1\leq J\leq J_*$, we define

$$\mathcal{M}_J=\left\{m=\bigcup_{j=0}^{J_*}\{j\}\times A_j,\ \text{with } A_j\in\Lambda_{j,J}\right\},$$

where $\Lambda_{j,J}=\{\{1,\dots,2^j\}\}$ when $j\leq J-1$ and

$$\Lambda_{j,J}=\left\{A\subset\{1,\dots,2^j\}:|A|=\lfloor 2^J/(j-J+1)^3\rfloor\right\}\qquad\text{when } J\leq j\leq J_*.$$

To m in $\mathcal{M}=\bigcup_{J=1}^{J_*}\mathcal{M}_J$, we associate $\mathcal{F}_m=\mathrm{span}\{\phi_{j,k},(j,k)\in m\}$ and define the model $\mathcal{S}_m$ by

$$\mathcal{S}_m=\{(f(x_1),\dots,f(x_n)), f\in\mathcal{F}_m\}\subset\mathcal{S}_*=\{(f(x_1),\dots,f(x_n)), f\in\mathcal{F}_*\}.$$

When $m\in\mathcal{M}_J$, the dimension of $\mathcal{S}_m$ is bounded from above by

$$\dim(\mathcal{S}_m)\leq\sum_{j=0}^{J-1}2^j+\sum_{j=J}^{J_*}\frac{2^J}{(j-J+1)^3}\leq 2^J\left[1+\sum_{k=1}^{J_*-J+1}k^{-3}\right]\leq 2.2\cdot 2^J \tag{19}$$

and $\dim(\mathcal{S}_*)\leq\kappa n$. Also, note that the cardinality of $\mathcal{M}_J$ is $|M_J|=\prod_{j=J}^{J_*}\binom{2^j}{\lfloor 2^J/(j-J+1)^3\rfloor}$. To estimate μ, we use the estimator $\hat{\mu}$ given by (6) with β given by (15) and

$$L_m=1.1\cdot 2^J$$

and

$$\pi_m=\left[2^J(1-2^{J_*})\prod_{j=J}^{J_*}\binom{2^j}{\lfloor 2^J/(j-J+1)^3\rfloor}\right]^{-1}\qquad\text{for } m\in\mathcal{M}_J.$$

The next corollary gives the rate of convergence of the estimator $\hat{\mu}$ when f belongs to some Besov ball $\mathcal{B}^{\alpha}_{p,\infty}(R)$ with $1/p < \alpha < r$ (we refer to De Vore and Lorentz [12] for a precise definition of Besov spaces). As is usual in this setting, we express the result in terms of the norm $\|\cdot\|_n^2 = \|\cdot\|^2/n$ on $\mathbb{R}^n$.

Corollary 2. *For any $p, R > 0$ and $\alpha \in]1/p, r[$, there exists some constant C not depending on n and σ^2 such that the estimator $\hat{\mu}$ defined above fulfills*

$$\mathbb{E}(\|\mu - \hat{\mu}\|_n^2) \le C \max\left\{ \left(\frac{\sigma^2}{n}\right)^{2\alpha/(2\alpha+1)}, \frac{1}{n^{2(\alpha-1/p)}}, \frac{\sigma^2}{n} \right\}$$

for any μ given by (17) with $f \in \mathcal{B}^{\alpha}_{p,\infty}(R)$.

The proof is delayed to Section 6.5. We recall that the rate $1/n^{2\alpha/(2\alpha+1)}$ is minimax in this framework; see Yang and Barron [24]. Therefore, the above procedure is rate-minimax over all of the Besov balls $\mathcal{B}^{\alpha}_{p,\infty}(R)$ with parameters $p > 1/r$ and $\alpha \ge \alpha_p = (1+\sqrt{1+2p})/2p$. We note that $\alpha_p \in]1/p, 1/p + 1/2[$.

5. Conclusion

In this paper, we adapt the mixing procedure of Leung and Barron [18] to handle regression when the variance of the noise is unknown. The resulting estimator does not require the data to be split into two (or more) sets and can thus handle frameworks like fixed-design regression. It also reduces, in some cases, to a simple shrinkage estimator and can be computed efficiently. Our main result exploits the Gibbs form of the weight w_m to provide a new risk bound for $\beta < 1/4$. Another nice feature of this bound is that it underlines the interest of mixing (compared to model selection) when there are several good models to approximate μ. We emphasize that we only need a small number of 'degrees of freedom' to estimate the variance and obtain an oracle bound with a $1 + \varepsilon_n$ constant in front of the bias term. In practice, the choice of the tuning parameter β is crucial: the choice $\beta > 1/2$ is not recommended in the case of Section 3.2, but any choice of $\beta \le 1/2$ is suitable. From a more theoretical point of view, we give new approximation bounds for BV functions in Proposition 3 of Section 6.4.

6. Proofs

To save space, we give here only the main lines of the proofs. The details can be found in the technical report [14].

6.1. Proof of Proposition 1

We express the weights w_m in terms of the Z_j's and $\tau = \alpha \log p + b$:

$$w_m = \frac{\exp(\sum_{k\in m}(\beta Z_k^2/\hat\sigma^2 - \tau))}{\sum_{m'\in\mathcal{M}}\exp(\sum_{k\in m'}(\beta Z_k^2/\hat\sigma^2 - \tau))}.$$

Since $c_j = \sum_{m\in\mathcal{M}} \mathbb{1}_{j\in m} w_m$, we obtain (8) by factoring out j for any m including j as $\{j\}\cup m'$ for some subset m' not including j.

6.2. A preliminary lemma

The next lemma gives a control on the deviations of $\hat\sigma^2$.

Lemma 1. *Consider an integer N larger than 2 and a random variable X such that NX is distributed as a χ^2 of dimension N. Then, for any $0 < a < 1$,*

$$\mathbb{E}[(a - X)_+] \le \mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] \le \frac{2}{(1-a)(N-2)}\exp(-N\phi(a)), \tag{20}$$

with $\phi(a) = \frac{1}{2}(a - 1 - \log a) > \frac{1}{4}(1-a)^2$.

Proof. The first inequality follows directly from the opposite monotonicity of x and $(a/x - 1)_+$. For the second inequality, we start with the Markov bound

$$\mathbb{P}\left(X \le \frac{1}{t}\right) \le \mathrm{e}^{\lambda/t}\left(1 + \frac{2\lambda}{N}\right)^{-N/2} \qquad \text{for any } \lambda \ge 0.$$

Choosing $\lambda = N(t-1)/2$, we obtain

$$\mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] = a\int_{1/a}^{+\infty} \mathbb{P}\left(X \le \frac{1}{t}\right)\mathrm{d}t \le a\int_{1/a}^{+\infty} \exp\left(\frac{N}{2}(1 - 1/t)\right)\frac{\mathrm{d}t}{t^{N/2}}.$$

Iterating integrations by parts finally gives, for $0 < a < 1$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] &\le \exp\left(\frac{N}{2}(1-a)\right)\frac{2a^{N/2}}{N-2}\sum_{k\ge 0}\frac{a^k(N/2)^k}{\prod_{i=0}^{k-1}(N/2+i)} \\ &\le \frac{2}{(1-a)(N-2)}\exp(-N\phi(a)).\end{aligned}$$

□

6.3. Proof of Theorem 1

To keep formulas short, we write d_m for the dimension of $\mathcal{S}_m$ and use the following notation for the various projections:

$$\hat{\mu}_* = \Pi_{\mathcal{S}_*} Y, \qquad \mu_* = \Pi_{\mathcal{S}_*}\mu \quad \text{and} \quad \mu_m = \Pi_{\mathcal{S}_m}\mu, \qquad m \in \mathcal{M}.$$

We derive from Theorem 1 of Leung and Barron [18] that

$$S(\hat{\mu}) = \sum_{m\in\mathcal{M}} w_m \left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\sigma^2} + 2d_m + \left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)\frac{\|\hat{\mu}_m - \hat{\mu}\|^2}{\hat{\sigma}^2}\right] - p$$

is an unbiased estimate of $\sigma^{-2}\mathbb{E}(\|\mu_* - \hat{\mu}\|^2)$ (note that the β here is half of the β in [18]). We control the last term with the bound

$$\sum_{m\in\mathcal{M}} w_m \|\hat{\mu}_m - \hat{\mu}\|^2 = \sum_{m\in\mathcal{M}} w_m \|\hat{\mu}_* - \hat{\mu}_m\|^2 - \|\hat{\mu} - \hat{\mu}_*\|^2 \le \sum_{m\in\mathcal{M}} w_m \|\hat{\mu}_* - \hat{\mu}_m\|^2$$

and get

$$\begin{aligned} S(\hat{\mu}) \le{}& \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right] \sum_{m\in\mathcal{M}} w_m \left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] - p \\ &+ \sum_{m\in\mathcal{M}} w_m \left(2d_m - \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\frac{L_m}{\beta}\right), \end{aligned}$$

where $(x)_+ = \max(0, x)$. First, note that when $L_m \ge d_m/2$ we have

$$\begin{aligned} 2d_m - \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\frac{L_m}{\beta} &\le \left[2 - \frac{\hat{\sigma}^2}{2\beta\sigma^2} - \frac{1}{2\beta}\left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right] d_m \\ &\le \min\left(0, 2 - \frac{\hat{\sigma}^2}{2\beta\sigma^2}\right) d_m \le 0. \end{aligned}$$

Therefore, setting $\hat{\delta}_\beta = (4\beta\sigma^2/\hat{\sigma}^2 - 1)_+$, we get

$$S(\hat{\mu}) \le \frac{\hat{\sigma}^2}{\sigma^2}(1 + \hat{\delta}_\beta) \sum_{m\in\mathcal{M}} w_m \left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] - p.$$

Let us introduce the Kullback divergence between two probability distributions $\{\alpha_m, m \in \mathcal{M}\}$ and $\{\pi_m, m \in \mathcal{M}\}$ on $\mathcal{M}$,

$$\mathcal{D}(\alpha|\pi) = \sum_{m\in\mathcal{M}} \alpha_m \log\frac{\alpha_m}{\pi_m} \ge 0,$$

and the function

$$\mathcal{E}_\beta^\pi(\alpha) = \sum_{m\in\mathcal{M}} \alpha_m \left[\frac{\|\hat\mu_* - \hat\mu_m\|^2}{\hat\sigma^2} + \frac{L_m}{\beta} \right] + \frac{1}{\beta}\mathcal{D}(\alpha|\pi).$$

The latter function is convex on the simplex $S_{\mathcal{M}}^+ = \{\alpha \in [0,1]^{|\mathcal{M}|}, \sum_{m\in\mathcal{M}} \alpha_m = 1\}$ and can be interpreted as a free energy function. It is minimal for the Gibbs measure $\{w_m, m \in \mathcal{M}\}$, so for any $\alpha \in S_{\mathcal{M}}^+$,

$$S(\hat\mu) \le (1+\hat\delta_\beta)\left[\sum_{m\in\mathcal{M}} \alpha_m \left[\frac{\|\hat\mu_* - \hat\mu_m\|^2}{\sigma^2} + \frac{\hat\sigma^2}{\beta\sigma^2} L_m \right] + \frac{\hat\sigma^2}{\beta\sigma^2}\mathcal{D}(\alpha|\pi) \right] - p.$$

We fix a probability distribution $\alpha \in S_{\mathcal{M}}^+$ and take the expectation in this last inequality to get

$$\begin{aligned}\mathbb{E}[S(\hat\mu)] \le \big(1+\mathbb{E}(\hat\delta_\beta)\big) \sum_{m\in\mathcal{M}} \alpha_m \mathbb{E}\left[\frac{\|\hat\mu_* - \hat\mu_m\|^2}{\sigma^2} \right] \\ + \mathbb{E}\left[\frac{\hat\sigma^2}{\beta\sigma^2}(1+\hat\delta_\beta) \right] \left[\sum_{m\in\mathcal{M}} \alpha_m L_m + \mathcal{D}(\alpha|\pi) \right] - p.\end{aligned}$$

Since $\hat\sigma^2/\sigma^2$ is stochastically larger than a random variable X with $\chi^2(N_*)/N_*$ distribution, Lemma 1 ensures that

$$\mathbb{E}\left[\frac{\hat\sigma^2}{\sigma^2}\hat\delta_\beta \right] \le \mathbb{E}[\hat\delta_\beta] \le \frac{2}{(1-4\beta)(N_*-2)} \exp(-N_*\phi(4\beta)),$$

with $\phi(x) = (x - 1 - \log(x))/2$. Furthermore, the condition $N_* \ge 2 + (\log n)/\phi(4\beta)$ guarantees that

$$\frac{2}{(1-4\beta)(N_*-2)} \exp(-N_*\phi(4\beta)) \le \frac{2\phi(4\beta)\mathrm{e}^{-2\phi(4\beta)}}{(1-4\beta)n\log n} \le \frac{1}{2n\log n} = \varepsilon_n.$$

Putting the pieces together, we obtain

$$\begin{aligned}\frac{\mathbb{E}[\|\mu - \hat\mu\|^2]}{\sigma^2} &= \frac{\|\mu-\mu_*\|^2}{\sigma^2} + \mathbb{E}[S(\hat\mu)] \\ &\le \frac{\|\mu-\mu_*\|^2}{\sigma^2} + (1+\varepsilon_n) \sum_{m\in\mathcal{M}} \alpha_m \mathbb{E}\left[\frac{\|\hat\mu_* - \hat\mu_m\|^2}{\sigma^2} \right] \\ &\quad + \left(\frac{\bar\sigma^2}{\beta\sigma^2} + \varepsilon_n \right) \left[\sum_{m\in\mathcal{M}} \alpha_m L_m + \mathcal{D}(\alpha|\pi) \right] - p \\ &\le (1+\varepsilon_n) \left[\sum_{m\in\mathcal{M}} \alpha_m \left[\frac{\|\mu-\mu_m\|^2}{\sigma^2} - d_m + \frac{\bar\sigma^2}{\beta\sigma^2} L_m \right] + \frac{\bar\sigma^2}{\beta\sigma^2}\mathcal{D}(\alpha|\pi) \right] + \varepsilon_n p.\end{aligned}$$

This inequality holds for any non-random probability distribution $\alpha \in S^+_{\mathcal{M}}$, so it holds, in particular, for the Gibbs measure

$$\alpha_m = \frac{\pi_m}{\mathcal{Z}_\beta} \exp\left[-\frac{\beta}{\bar{\sigma}^2}(\|\mu - \mu_m\|^2 - d_m\sigma^2) - L_m \right], \qquad m \in \mathcal{M},$$

where $\mathcal{Z}_\beta$ normalizes the sum of the α_m's to one. For this choice of α_m, we obtain

$$\frac{\mathbb{E}[\|\mu - \hat{\mu}\|^2]}{\sigma^2} \le -\frac{(1+\varepsilon_n)\bar{\sigma}^2}{\beta\sigma^2} \log\left[\sum_{m\in\mathcal{M}} \pi_m \exp\left[-\frac{\beta}{\bar{\sigma}^2}(\|\mu - \mu_m\|^2 - d_m\sigma^2) - L_m \right] \right] + \varepsilon_n p,$$

which ensures (11) since $p \le n$. To get (12), simply note that

$$\sum_{m\in\mathcal{M}} \pi_m \exp\left[-\frac{\beta}{\bar{\sigma}^2}(\|\mu - \mu_m\|^2 - d_m\sigma^2) - L_m \right]$$
$$\ge \pi_{m^*} \exp\left[-\frac{\beta}{\bar{\sigma}^2}(\|\mu - \mu_{m^*}\|^2 - d_{m^*}\sigma^2) - L_{m^*} \right]$$

for any $m^* \in \mathcal{M}$.

6.4. Proof of Corollary 1

We start by proving some results on the approximation of BV functions with the Haar wavelets. We remain in the setting of Section 4.1, with $0 = x_1 < x_2 < \cdots < x_n < x_{n+1} = 1$ and $n = 2^{J_n}$. For $0 \le j \le J_n$ and $p \in \Lambda(j)$, we define $t_{j,p} = x_{p2^{-j}n+1}$. We also set $\phi_{0,0} = 1$ and

$$\phi_{j,k} = 2^{(j-1)/2}\left(\mathbf{1}_{[t_{j,2k+1}, t_{j,2k+2})} - \mathbf{1}_{[t_{j,2k}, t_{j,2k+1})}\right) \qquad \text{for } 1 \le j \le J_n \text{ and } k \in \Lambda(j).$$

This family of Haar wavelets is orthonormal for the positive semi-definite quadratic form

$$(f, g)_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i)$$

on functions mapping $[0, 1]$ into $\mathbb{R}$. For $0 \le J \le J_n$, we write f_J for the projection of f onto the linear space spanned by $\{\phi_{j,k}, 0 \le j \le J, k \in \Lambda(j)\}$ with respect to $(\cdot, \cdot)_n$, namely,

$$f_J = \sum_{j=0}^{J} \sum_{k\in\Lambda(j)} c_{j,k}\phi_{j,k}, \qquad \text{with } c_{j,k} = (f, \phi_{j,k})_n.$$

We also consider, for $1 \le J \le J_n$, an approximation of f *à la* Birgé and Massart [6],

$$\tilde{f}_J = f_{J-1} + \sum_{j=J}^{J_n} \sum_{k\in\Lambda'_j(j)} c_{j,k}\phi_{j,k},$$

where $\Lambda'_J(j) \subset \Lambda(j)$ is the set of indices k we obtain when we select the $K_{j,J}$ largest coefficients $|c_{j,k}|$ among $\{|c_{j,k}|,\ k \in \Lambda(j)\}$, with $K_{j,J} = \lfloor (j-J+1)^{-3} 2^{J-2} \rfloor$ for $1 \le J \le j \le J_n$. Note that the number of coefficients $c_{j,k}$ in $\tilde{f}_J$ is bounded from above by

$$1 + \sum_{j=1}^{J-1} 2^{j-1} + \sum_{j \ge J} (j-J+1)^{-3} 2^{J-2} \le 2^{J-1} + 2^{J-2} \sum_{p \ge 1} p^{-3} \le 2^J.$$

The next proposition states approximation bounds for f_J and $\tilde{f}_J$ in terms of the (semi-)norm $\|f\|_n^2 = (f,f)_n$.

Proposition 3. *When f has bounded variation $V(f)$, we have*

$$\|f - f_J\|_n \le 2V(f) 2^{-J/2} \qquad \text{for } J \ge 0 \tag{21}$$

and

$$\|f - \tilde{f}_J\|_n \le cV(f) 2^{-J} \qquad \text{for } J \ge 1, \text{ with } c = \sum_{p \ge 1} p^3 2^{-p/2+1}. \tag{22}$$

Proof. (21) and (22) are based on the following fact.

Lemma 2. *When f has bounded variation $V(f)$, we have*

$$\sum_{k \in \Lambda(j)} |c_{j,k}| \le 2^{-(j+1)/2} V(f) \qquad \text{for } 1 \le j \le J_n.$$

Proof. We assume, for simplicity, that f is non-decreasing. Then, we have

$$c_{j,k} = \langle f, \phi_{j,k} \rangle_n = \frac{2^{(j-1)/2}}{n} \left[\sum_{i \in I^+_{j,k}} f(x_i) - \sum_{i \in I^-_{j,k}} f(x_i) \right],$$

with $I^+_{j,k}$ and $I^-_{j,k}$ defined in Section 4.1. Since $|I^+_{j,k}| = 2^{-j} n$ and f is non-decreasing,

$$\begin{aligned} |c_{j,k}| &\le \frac{2^{(j-1)/2}}{n} |I^+_{j,k}| \left[f\left(x_{(2k+2)2^{-j}n}\right) - f\left(x_{(2k)2^{-j}n}\right) \right] \\ &\le 2^{-(j+1)/2} \left[f\left(x_{(2k+2)2^{-j}n}\right) - f\left(x_{(2k)2^{-j}n}\right) \right], \end{aligned}$$

and Lemma 2 follows. □

We first prove (21). Since the $\{\phi_{j,k}, k \in \lambda(j)\}$ have disjoint supports, we have, for $0 \le J \le J_n$,

$$\|f - f_J\|_n \le \sum_{j>J} \left\| \sum_{k \in \Lambda(j)} c_{j,k} \phi_{j,k} \right\|_n \le \sum_{j>J} \left[\sum_{k \in \Lambda(j)} |c_{j,k}|^2 \underbrace{\|\phi_{j,k}\|_n^2}_{=1} \right]^{1/2} \le \sum_{j>J} \sum_{k \in \Lambda(j)} |c_{j,k}|.$$

Formula (21) then follows from Lemma 2.

To prove (22), we introduce the set $\Lambda''_J(j) = \Lambda(j) \setminus \Lambda'_J(j)$. Then, for $1 \le J \le J_n$, we have

$$\|f - \tilde{f}_J\|_n \le \sum_{j=J}^{J_n} \left[\sum_{k \in \Lambda''_J(j)} |c_{j,k}|^2 \underbrace{\|\phi_{j,k}\|_n^2}_{=1} \right]^{1/2} \le \sum_{j=J}^{J_n} \left[\max_{k \in \Lambda''_J(j)} |c_{j,k}| \sum_{k \in \Lambda(j)} |c_{j,k}| \right]^{1/2}.$$

The choice of $\Lambda'_J(j)$ guarantees the inequalities

$$(1 + K_{j,J}) \max_{k \in \Lambda''_J(j)} |c_{j,k}| \le \sum_{k \in \Lambda''_J(j)} |c_{j,k}| + \sum_{k \in \Lambda'_J(j)} |c_{j,k}| \le \sum_{k \in \Lambda(j)} |c_{j,k}|.$$

To complete the proof of Proposition 3, we combine this bound with Lemma 2:

$$\begin{aligned} \|f - \tilde{f}_J\|_n &\le \sum_{j \ge J} 2^{-(j+1)/2} V(f) (1 + K_{j,J})^{-1/2} \\ &\le \sum_{j \ge J} 2^{-(j+1)/2} V(f) 2^{-(J-2)/2} (j - J + 1)^3. \\ &\le V(f) 2^{-J} \sum_{p \ge 1} p^3 2^{-p/2+1}. \end{aligned}$$

□

6.4.1. *Proof of Corollary 1*

First, note that $v_{j,k} = (\phi_{j,k}(x_1), \ldots, \phi_{j,k}(x_n))'$ for $(j,k) \in \Lambda^*$. Then, according to (21) and (22), there exists, for any $0 \le J \le J^*$, a model $m \in \mathcal{M}$ fufilling $|m| \le 2^J$ and

$$\begin{aligned} \|\mu - \Pi_{\mathcal{S}_m} \mu\|_n^2 &= \|\mu - \Pi_{\mathcal{S}_*} \mu\|_n^2 + \|\Pi_{\mathcal{S}_*} \mu - \Pi_{\mathcal{S}_m} \mu\|_n^2 \\ &\le 2c^2 V(f)^2 (2^{-J^*} \vee 2^{-2J}), \end{aligned}$$

with $c = \sum_{p \ge 1} p^3 2^{-p/2+1}$. Putting together this approximation result with Theorem 1 gives

$$\mathbb{E}(\|\mu - \hat{\mu}\|_n^2) \le C \inf_{0 \le J \le J^*} \left[V(f)^2 (2^{-J^*} \vee 2^{-2J}) + \frac{2^J \log n}{n} \sigma^2 \right]$$

for some numerical constant C, when $4\beta \le \phi^{-1}(\log n/(n/2-2))$. We refer to [14] for the case $2\beta \le \phi^{-1}(2\log(n/2)/n)$. To conclude the proof of Corollary 1, we apply the previous bound with J given by the minimum between J^* and the smallest integer such that

$$2^J \ge V(f)^{2/3} \left(\frac{n}{\sigma^2 \log n} \right)^{1/3}.$$

6.5. Proof of Corollary 2

First, according to the inequality $\binom{n}{k} \le (en/k)^k$, we have the bound for $m \in \mathcal{M}_J$,

$$-\log \pi_m \le \log 2^J + \sum_{j=J}^{J_*} \frac{2^J}{(j-J+1)^3} \log\bigl(e 2^{j-J+1}(j-J+1)^3\bigr)$$
$$\le 2^J \left(1 + \sum_{k\ge 1} k^{-3}(1 + 3\log k + k\log 2)\right) \le 4 \cdot 2^J. \tag{23}$$

Second, when f belongs to some Besov ball $\mathcal{B}^\alpha_{p,\infty}(R)$ with $1/p < \alpha < r$, Birgé and Massart [6] give the following approximation results. There exists a constant $C > 0$ such that for any $J \le J_*$ and $f \in \mathcal{B}^\alpha_{p,\infty}(R)$, there exists $m \in \mathcal{M}_J$ fulfilling

$$\|f - \bar\Pi_{\mathcal{F}_m} f\|_\infty \le \|f - \bar\Pi_{\mathcal{F}_*} f\|_\infty + \|\bar\Pi_{\mathcal{F}_*} f - \bar\Pi_{\mathcal{F}_m} f\|_\infty \le C \max\bigl(2^{-J_*(\alpha-1/p)}, 2^{-\alpha J}\bigr),$$

where $\bar\Pi_{\mathcal{F}}$ denotes the orthogonal projector onto $\mathcal{F}$ in $L^2([0,1])$. In particular, under the previous assumptions, we have

$$\|\mu - \Pi_{\mathcal{S}_m}\mu\|_n^2 \le \frac{1}{n}\sum_{i=1}^n [f(x_i) - \bar\Pi_{\mathcal{F}_m} f(x_i)]^2$$
$$\le \|f - \bar\Pi_{\mathcal{F}_m} f\|_\infty^2 \le C^2 \max\bigl(2^{-2\alpha J}, 2^{-2J_*(\alpha-1/p)}\bigr). \tag{24}$$

To conclude the proof of Corollary 2, we combine Theorem 1 with (19), (23) and (24) to obtain

$$J = \min\left(J_*, \left\lfloor \frac{\log[\max(n/\sigma^2, 1)]}{(2\alpha+1)\log 2} \right\rfloor + 1\right).$$

Acknowledgements

We would like to thank Yannick Baraud and an anonymous referee for their constructive comments on early drafts of this paper.

References

[1] Akaike, H. (1969). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217. MR0286233

[2] Baraud, Y., Giraud, C. and Huet, S. (2008). Gaussian model selection with unknown variance. *Ann. Statist.* To appear. arXiv:math/0701250v1.

[3] Barron, A. (1987). *Are Bayesian Rules Consistent in Information?* New York: Springer.

[4] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[5] Barron, A. and Cover, T. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054. MR1111806
[6] Birgé, L. and Massart, P. (2000). An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16** 1–36. MR1848840
[7] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946
[8] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73.
[9] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101
[10] Catoni, O. (1997). Mixture approach to universal model selection. Preprint 30, Laboratoire de l'Ecole Normale Supérieure, Paris.
[11] Catoni, O. (1999). Universal aggregation rules with exact bias bounds. Preprint 510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris.
[12] Devore, R. and Lorentz, G. (1993). *Constructive Approximation*. New York: Springer. MR1261635
[13] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089
[14] Giraud, C. (2007). Mixing least-squares estimators when the variance is unknown. Technical report. arXiv:0711.0372v1.
[15] Hall, P., Kay, J. and Titterington, D.M. (1990). Asymptotically optimal differencebased estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842
[16] Hartigan, J.A. (2002). Bayesian regression using Akaike priors. Preprint, Yale Univ., New Haven.
[17] Lenth, R. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31** 469–473. MR1041567
[18] Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. MR2242356
[19] Mallows, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.
[20] Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). On difference based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc. Ser. B* **67** 19–41. MR2136637
[21] Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12** 1215–1230. MR0760684
[22] Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92** 821–830. MR2234188
[23] Tsybakov, A. (2003). Optimal rates of aggregation. *COLT-2003. Lecture Notes in Artificial Intelligence* **2777** 303–313. Heidelberg: Springer.
[24] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500
[25] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** 135–161. MR1790617
[26] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28** 75–87. MR1762904
[27] Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory* **20** 176–222. MR2028357
[28] Wang, L., Brown, L., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810

Received February 2007 and revised March 2008