Estimation of the density of regression errors by pointwise model selection

Sandra Plancade

ESTIMATION OF THE DENSITY OF REGRESSION ERRORS BY POINTWISE MODEL SELECTION

S. PLANCADE

ABSTRACT. This paper presents two results: a density estimator and an estimator of regression error density. We first propose a density estimator constructed by model selection, which is adaptive for the quadratic risk at a given point. Then we apply this result to estimate the error density in an homoscedastic regression framework $Y_i = b(X_i) + \epsilon_i$, from which we observe a sample (X_i, Y_i). Given an adaptive estimator $\widehat{b}$ of the regression function, we apply the density estimation procedure to the residuals $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$. We get an estimator of the density of ϵ_i whose rate of convergence for the quadratic pointwise risk is the maximum of two rates: the minimax rate we would get if the errors were directly observed and the minimax rate of convergence of $\widehat{b}$ for the quadratic integrated risk.

February 18, 2009

MSC 2000 Subject Classifications. 62G07-62G08

Keywords and phrases. density, regression error, pointwise model selection, adaptivity.

1. INTRODUCTION

Consider a sample (X_i, Y_i) from the homoscedastic regression framework:

$$Y_i = b(X_i) + \epsilon_i \tag{1}$$

where the (ϵ_i) are unobserved independent identically distributed (i.i.d.) data with common density f, with zero mean and independent of the (X_i). The main goal of this paper is to propose an estimator for the density of ϵ_i, and to provide an upper bound for the quadratic risk of this estimator at a fixed point x_0.

The main issue in regression problems is to predict Y_i by measuring only X_i. The first step in such study is the estimation of the regression function $b(x) = \mathbb{E}[Y|X = x]$. This question has already been studied at length. The second step consists in studying the variations of Y_i around its conditional mean, which are characterized by the density of the errors (ϵ_i).

The knowledge of an estimator of the error density has many applications: for example, it allows model validation and, combined with an estimator of the regression function, it provides confidence intervals for future observations Y. The reader is referred to Efromovich (2005) for practical applications. Many papers are devoted to density estimation but the difficulty in our problem is to estimate the density from a sample (ϵ_i) which is not observed. The natural approach consists in computing proxies of the (ϵ_i), i.e. quantities based on the data which estimate the true (ϵ_i), and applying to them a density estimation

Université Paris Descartes, MAP5, UMR CNRS 8145.

procedure as if they were the true error sample. Observing that $\epsilon_i = Y_i - b(X_i)$, we naturally estimate the errors by the residuals $(\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i))$, where $\widehat{b}$ is an estimator of the regression function. Efromovich applies this strategy with a thresholding density estimation procedure (see for example Efromovich (2005)). He gets an estimator of the density of the (ϵ_i) whose L^2-risk reaches the same minimax rate of convergence we would obtain if the (ϵ_i) were observed. Nevertheless, this result requires strong conditions of regularity on the regression function b, and on the density of the (X_i) and (ϵ_i). Another estimator is built in Plancade (2008) by model selection. Its L^2-risk has a rate equal to the maximum of the minimax rates of estimation of b and f if the sample (ϵ_i) was observed. Let us also mention the papers Akritas and Keilegom (2001) and Kiwitt et al. (2008) which propose estimators of the regression errors distribution functions. But to the author's knowledge, no paper studies pointwise estimation of the error density by any method.

The estimators presented in this paper are based on a pointwise model selection procedure. Model selection theory has been initiated by Birgé and Massart (see for example Birgé and Massart (1998)), and adapted to regression function estimation in Baraud (2002) in the study of integrated quadratic risks. We will use here the estimator $\widehat{b}$ of b proposed in Baraud (2002), constructed by a model selection procedure based on least square estimators. Although the principle of pointwise model selection is the same, the techniques to carry it out are different. In particular, the key tool to prove the adaptivity of classical model selection estimators is the Talagrand inequality, whereas the adaptivity of pointwise model selection estimators comes out of a simpler Bernstein inequality. The techniques developed in this paper are based on Laurent et al. (2008), in which they develop these methods in a different framework.

This paper presents two results. On the one hand, we build a density estimator which proves to be adaptive for the pointwise risk over some classical classes of regularity. Such estimators have been constructed using kernel methods in Butucea (2001), with the same adaptivity properties, along with minimax results over Sobolev classes. Nevertheless, our estimator is completely data driven, whereas the estimation procedure in Butucea (2001) brings into play upper bounds on unknown quantities. The second result proceeds from the application of the above density estimation procedure to residuals from the framework (1). We get an estimator of the error density, whose pointwise rate of convergence is the maximum of these two rates: the pointwise minimax rate of estimation of f we would get if the errors (ϵ_i) where observed and the L^2-minimax rate of estimation of b.

The paper is organized as follows. Section 2 presents the estimator of the error density and the main result. The theoretical tools used to obtain this result are described in Sections 3 and 4. More precisely, Section 3 is devoted to the construction of a density estimator by pointwise model selection, and the study of its convergence properties. In Section 4, we present an estimator of the regression function and apply the density estimation procedure described in Section 3 to the residuals. Section 5 is dedicated to numerical results. Most of the proofs are gathered in Section 6.

2. Main result

2.1. **Notations.** Let t be a function defined on an interval I of $\mathbb{R}$ and μ be a density on I. We consider different norms of t:

$$\|t\|_\infty := \sup_{x\in I} |t(x)|, \quad \|t\| := \left(\int_I t^2(x)dx\right)^{1/2}, \quad \|t\|_\mu := \left(\int_I t^2(x)\mu(x)dx\right)^{1/2}.$$

Besides, we consider the following spaces of functions over I:

$$L^2(I) := \{t : I \to \mathbb{R}, \|t\| < +\infty\}, \quad L^\infty(I) := \{t : I \to \mathbb{R}, \|t\|_\infty < +\infty\}.$$

If t is a function k times differentiable, we denote by $t^{(k)}$ its k-th derivative.
For every set A, we denote by $\mathbb{1}_A$ the indicator function of A, that is $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise.
For every function $t : \mathbb{R} \to \mathbb{R}$, we denote by t^* the Fourier transform of t:

$$t^*(u) = \int_{x\in\mathbb{R}} t(x)e^{-iux}dx, \quad \forall u \in \mathbb{R}$$

For every linear space S_m we denote by t_m the L^2-orthogonal projection of t onto S_m.

We consider the following Sobolev classes, for every $\alpha, L > 0$:

$$W(\alpha, L) = \{F \in L^2(\mathbb{R}), \frac{1}{2\pi}\int_{\mathbb{R}} |F^*(u)|^2 u^{2\alpha}du \le L^2\}.$$

The Hölder classes are defined as follows. For every $\beta, L > 0$, and r the largest integer less than β, let:

$$\mathcal{H}(\beta, L) = \{F \in L^2(\mathbb{R}), |F^{(r)}(x) - F^{(r)}(y)| \le L|x - y|^{\beta-r}, \forall x, y \in \mathbb{R}\}$$

Finally, for every $x \in \mathbb{R}$, we denote by $E(x)$ its integer part, that is $E(x) \in \mathbb{Z}$ and:

$$E(x) \le x < E(x) + 1.$$

All throughout the paper, C_i denotes a universal numerical constant, and C, C', C" denote numerical constants which only depends on the given constants of the problem and may change from one line to another.

2.2. **Assumptions.** We consider a $3n$-sample $(X_i, Y_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,2n\}}$ from the regression framework (1), where the (X_i) are i.i.d, the (ϵ_i) are i.i.d, independent of the (X_i) and $\mathbb{E}(\epsilon_1) = 0$. We suppose also that the following assumption holds.

$\mathbf{H_0}$(f) : The density f is upper bounded by $\nu := \|f\|_\infty$ and is supported on $I = \mathbb{R}$ or on a known compact set I, that we will suppose equal to $[-1, 1]$.

We define two collections of functions, one on $\mathbb{R}$ and one on $[-1, 1]$
We consider collections of functions on $\mathbb{R}$ constructed from the sine-cardinal function:

$$\phi(x) := \frac{\sin(\pi x)}{\pi x}$$

For every $m > 0$, $k \in \mathbb{Z}$, we consider $\phi_{m,k}(x) := \sqrt{m}\phi(mx - k)$ for every $x \in \mathbb{R}$, and A_m is the following model:

$$A_m = \text{vect}\{\phi_{m,k}, k \in \mathbb{Z}\} \tag{2}$$

The collection of models incorporates the models A_m for m belonging to a grid of step $1/B$, B being a fixed positive integer:

$$\mathcal{A}_n := \{A_m, m \in \frac{1}{B}\mathbb{N}, m \leq M_n\}$$

and $M_n \leq n$.

We consider also collections of functions on $[-1,1]$ constructed from the compact wavelet decomposition. We only recall here the definition of wavelet bases, the reader is referred to Meyer (1990) for more details. Let ψ be a function, called mother wavelet, supported on a compact set $[-B,B]$ of regularity r, which satisfies the following conditions:

1) $\psi, \ldots, \psi^{(r)}$ are bounded on $[-B,B]$
2) For every $0 \leq k \leq r$, and $\ell \geq 1$ there exists a constant C_ℓ such that:
$$|\psi^{(k)}(x)| \leq C_\ell(1+|x|)^{-\ell} \ , \ \forall x \in [-B,B]$$
3) $\int_{-B}^{B} x^k \psi(x)dx = 0 \ , \ \forall \ 0 \leq k \leq r,$
4) The set of functions $\{\psi_{j,k} : x \to 2^{j/2}\psi(2^{j/2}x-k), (j,k) \in \mathbb{Z}^2\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

Consider a function φ called the father wavelet of regularity r and supported on $[-B,B]$ which satisfies Assumptions 1) et 2) above, and the following assumptions:

3') $\int_{-B}^{B} \varphi(x)dx = 1$
4') The set of functions $\{\varphi_k : x \to \varphi(x-k), k \in \mathbb{Z}\} \cup \{\psi_{j,k}, j \in \mathbb{N}, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

See Meyer (1990) for examples of such functions ψ and φ. The set $\{\psi_{j,k}, j \geq 0, k \in \mathbb{Z}\} \cup \{\varphi_k, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2[-1,1]$. As ψ is supported on $[-B,B]$, the restriction of $\psi_{j,k}$ to $[-1,1]$ is identically equal to zero for all $j \in \mathbb{N}$ and $k \notin [-2^j-B, 2^j+B]$. Let us denote $\Gamma(j) := \mathbb{Z} \cap [-2^j - B, 2^j + B]$. Similarly, φ_k is identically equal to zero for all $k \notin [-B-1, B+1] = \Gamma(0)$. Finally, we consider the following models:

$$B_m := \text{vect}(\{\psi_{j,k}, j = 0, \ldots, m-1, k \in \Gamma(j)\} \cup \{\varphi_k, k \in \Gamma(0)\})$$

and the collection of models:

$$\mathcal{B}_n := \{B_m, m \in \mathbb{N}^*, 2^m \leq M_n\}$$

with $M_n \leq n$.

Proposition 2.1. *1) For every $m > 0$:*

$$\| \sum_{k \in \mathbb{Z}} \phi_{m,k}^2 \|_\infty \leq \sqrt{m}. \tag{3}$$

2) There exists a constant $K(B)$ such that, for every $m \in \mathbb{N}^*$,

$$\| \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \psi_{j,k}^2 + \sum_{k \in \Gamma(0)} \varphi_k^2 \|_\infty \leq K^2(B)\sqrt{2^m}. \tag{4}$$

From now on, we use common notations for these two collections of models. The collection $\mathcal{M}_n$ is $\mathcal{A}_n$ if f is supported on $\mathbb{R}$, and $\mathcal{B}_n$ if f is supported on $[-1,1]$. We denote by S_m the model A_m or B_m and $\mathcal{M}_n = \{S_m, m \in J_n\}$. Moreover, we denote by :

$$S_m = \text{vect}(\chi_\lambda, \lambda \in I_m\}$$

where the functions χ_λ denote the $\psi_{j,k}$ and the φ_k, or the $\phi_{m,k}$. Thus according to Inequalities (3) and (4), we have:

$$\| \sum_{\lambda \in I_m} \chi_\lambda^2 \|_\infty \leq K^2 D_m$$

with $D_m = m$ and $K = 1$ for the sine-cardinal models, and $D_m = 2^m$ and $K = K(B)$ for the wavelets models.

We make different assumptions, for the cases of f supported on $\mathbb{R}$ or on a compact set :

H$_1$ (β) : Take $I = \mathbb{R}$, we consider the collection of model $\mathcal{M}_n = \mathcal{A}_n$. We assume that there exist $\beta > 0$ and $C_0 > 0$ such that for every model $A_m \in \mathcal{A}_n$, the L^2-orthogonal projection f_m of f onto A_m satisfies:

$$\|f - f_m\|_\infty \leq C_0 D_m^{-\beta}$$

Moreover, we suppose that f is Lipschitz, i.e. there exists a constant $L > 0$ such that for every $x, y \in I$, $|f(x) - f(y)| \leq L|x - y|$.

The following Proposition gives conditions ensuring that Assumption H_1 holds. The proof is given in Section 7.

Proposition 2.2. *If $f \in W(\alpha, L)$ with $\alpha > 1/2$, then $\|f - f_m\|_\infty \leq D_m^{\alpha - 1/2}$. Moreover, if $\alpha > 3/2$ then f is Lipschitz.*

H$_2$ (β) : Take $I = [-1, 1]$, we consider the collection of models $\mathcal{M}_n = \mathcal{B}_n$ with regularity r, and we suppose that $f \in \mathcal{H}(\beta, L)$ for some $1 \leq \beta \leq r$ and $L > 0$.

2.3. **Construction of the estimator and main result.** In this subsection, we give the definition of the estimator of f, the heuristical motivation concerning its construction being developed in the following sections. Let x_0 be a fixed point in I. We split the sample $(X_i, Y_i)_{i \in \{-n, \ldots, -1\} \cup \{1, \ldots, n\}}$ into three independent samples:

$$Z^- := (X_i, Y_i)_{\in \{-n, \ldots, -1\}}, \quad Z_0^+ := (X_i, Y_i)_{i \in \{1, \ldots, 2n\}}, \quad Z_1^+ := (X_i, Y_i)_{i \in \{n+1, \ldots, 2n\}} \tag{5}$$

Let $\widehat{b}$ be any estimator of b built out of the first sample Z^-. An example of such an estimator is given in Section 4. Consider the residuals from the second sample:

$$\widehat{\epsilon}_i := Y_i - \widehat{b}(X_i) \ , \ i \in \{1, \ldots, 2n\}$$

Given Z^-, the $(\widehat{\epsilon}_i)$ are i.i.d. with common density denoted by f^-.

Let $\widehat{\nu}_n^-$ be an estimator of $\nu^- := \|f^-\|_\infty$ built from the sample Z_1^+ such that the probability $P[\{\nu^-/2 \leq \widehat{\nu}_n^- < 2\nu^-\}^c]$ decreases exponentially in n. We give an explicit construction of $\widehat{\nu}_n^-$ in Section 3.5.

For every model $S_m = \text{vect}\{\chi_\lambda, \lambda \in I_m\}$, we consider the projection density estimator associated to the sample $(\widehat{\epsilon}_i)$:

$$\widehat{f}_m^- = \sum_{\lambda \in I_m} \left(\frac{1}{n} \sum_{i=1}^n \chi_\lambda(\widehat{\epsilon}_i) \right) \chi_\lambda. \tag{6}$$

The selected model is:

$$\widehat{m} = \arg \min_{m \in J_n} \left[\sup_{j \in J_m, j \geq m} \{(\widehat{f}_j^- - \widehat{f}_m^-)^2(x_0) - A x_{j,m} \widehat{\nu}_n^- \frac{D_m + D_j}{n}\}_+ + A K^2 x_m \widehat{\nu}_n^- \frac{D_m}{n} \right] \tag{7}$$

where A is a positive constant, and (x_m) and $(x_{j,m})$ are weights of order $\ln(D_m)$ and $\ln(D_j + D_m)$ more precisely described in (19) and (20).

Finally, we define the following numerical constants, depending on the collection of models:

$$\alpha_1 = [\frac{D_1}{1+D_1} + \ln(1+D_1)]^{-1} \tag{8}$$

where $D_1 = \min\{D_m, m \in \mathcal{M}_n\}$ and:

$$\alpha_2 = \max_{x>0} x^{1/(1+x)} \tag{9}$$

and we consider a positive number α_3 such that:

$$\alpha_3 \geq \alpha_1^{1/3}\alpha_2.$$

We have $\alpha_2 < 1.4$, and $D_1 \leq 1$ so that $\alpha_1 \leq (1/2 + \ln 2)^{-1}$. This implies that $\alpha_3 = 1.4$ works.

We can prove the following result for our estimator:

Theorem 2.1. *We suppose that Assumption* $\mathbf{H_0(f)}$ *holds. Moreover, we suppose that either* $\mathbf{H_1}(\beta)$ *or* $\mathbf{H_2}(\beta)$ *hold for some* $\beta \geq 1$*, with* $M_n = E(\alpha_3 n^{1/3}) + 1$*. Then*

$$\mathbb{E}[(\widehat{f}_{\widehat{m}} - f)^2(x_0)] \leq \kappa(\frac{n}{\ln n})^{-\frac{2\beta}{2\beta+1}} + \kappa'\mathbb{E}[\|b - \widehat{b}\|_\mu^2] \tag{10}$$

for some constant κ *and* κ' *depending on the parameters of the problem but not on* n*.*

Comments: Suppose that $f \in W(\beta, L)$ for some $\beta > 3/2$, then (10) holds.

On the one hand, Butucea (2001) proves that the minimax rate of estimation of a density over Sobolev class $W(\beta, L)$ is $n^{-\frac{2\beta-1}{2\beta}}$. She also proves that the adaptive minimax rate of convergence (which is the best rate of convergence for adaptive estimators over all classes of convergence $W(\beta, L)$) is $(n/\ln n)^{-\frac{2\beta-1}{2\beta}}$.

On the other hand, we present in Section 4 an adaptive estimator $\widehat{b}$ of b which reaches the minimax rate over Besov balls, from Baraud (2002).

Thus, the rate of convergence of our estimator is the maximum of the two following rates:

- the minimax rate of estimation of b over Besov balls.
- the minimax rate of estimation we would obtain for f if the (ϵ_i) were directly observed.

An analogous comment holds if Assumption $\mathbf{H_2}(\beta)$ holds.

3. Density estimation by pointwise model selection

In this section, we present a density estimation procedure which products adaptive estimators for the pointwise risk. This procedure is the one which is applied to the pseudo observations $\widehat{\epsilon_i}$ of ϵ_i.

The results of this section require weaker assumptions on regularity than the error density estimation, and are staten for a more general collection of models. The assumptions considered here are satisfied in particular by the collections defined in Section 2. We consider a collection of model $\mathcal{M}_n = \{S_m, m \in J_n\}$ which satisfies:

$\mathbf{H_{dens}}(\beta)$: For every $m \in J_n$ and $\{\chi_\lambda, \lambda \in I_m\}$ an orthonormal basis of S_m, there exists an positive number D_m such that:

$$\|f - f_m\|_\infty \leq C_0 D_m^{-\beta} \tag{11}$$

$$\|\sum_{\lambda \in I_m} \chi_\lambda^2\|_\infty \leq K\sqrt{D_m} \tag{12}$$

for some $K > 0$. We denote by $M_n = \max_{m \in J_n} D_m$ and we suppose that $M_n \leq n$. Moreover, we suppose that the collection $\mathcal{M}_n$ si rich enough. More precisely, we assume that there exists a constant $M \geq 1$ such that for every n, for every $\alpha \in]0,1[$ such that $n^\alpha M \leq M_n$, there exists a model m and

$$n^\alpha \leq D_m \leq M n^\alpha \tag{13}$$

Remark: The Property (13) is satisfied by the collections described in Section 2, and by most of the clasical collections.

3.1. **A preliminary risk bound.** Let $(V_1, \ldots, V_{2n})$ be a i.i.d. sample drawn from a density g, split into two samples:

$$Z_0 := (V_i)_{i \in \{1,\ldots,n\}}, \quad Z_1 := (V_i)_{i \in \{n+1\ldots,2n\}} \tag{14}$$

Let x_0 be a fixed point in I. For every model $m \in \mathcal{M}_n$, let $\widehat{g}_m$ be the projection estimator of g on S_m from the sample Z_0:

$$\widehat{g}_m = \sum_{\lambda \in I_m} (\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i))\chi_\lambda \tag{15}$$

Let g_m be the L^2-projection of g onto S_m. Observing that $\mathbb{E}[\widehat{g}_m(x_0)] = g_m(x_0)$, we get the following bias-variance decomposition for every model m:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] + (g_m - g)^2(x_0)$$

On the one hand, the variance term is replaced by a bound obtained thanks to (13) in $\mathbf{H_{dens}}(\beta)$. Indeed:

$$\mathbb{E}[(\widehat{g}_m(x_0) - g_m(x_0))^2] = Var[\sum_{\lambda \in I_m} (\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i))\chi_\lambda(x_0)] = Var[\frac{1}{n}\sum_{i=1}^n (\sum_{\lambda \in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0))]$$

As the (V_i) are i.i.d. we get:

$$\begin{aligned}
\mathbb{E}[(\widehat{g}_m(x_0) - g_m(x_0))^2] &= \frac{1}{n} Var[\sum_{\lambda \in I_m} \chi_\lambda(V_1)\chi_\lambda(x_0)] \\
&\leq \frac{1}{n}\mathbb{E}[(\sum_{\lambda \in I_m} \chi_\lambda(V_1)\chi_\lambda(x_0))^2] = \frac{1}{n}\int_{x \in \mathbb{R}} (\sum_{\lambda \in I_m} \chi_\lambda(x)\chi_\lambda(x_0))^2 g(x)dx \\
&\leq \frac{\|g\|_\infty}{n}\int_{x \in \mathbb{R}} (\sum_{\lambda \in I_m} \chi_\lambda(x)\chi_\lambda(x_0))^2 dx.
\end{aligned}$$

We develop the square in the integral:

$$\mathbb{E}[(\widehat{g}_m(x_0) - g_m(x_0))^2] \leq \frac{\|g\|_\infty}{n} \sum_{\lambda,\lambda' \in I_m} [\int_{x\in\mathbb{R}} \chi_\lambda(x)\chi_{\lambda'}(x)dx]\chi_\lambda(x_0)\chi_{\lambda'}(x_0)$$

Using that the functions (χ_λ) are orthonormal and (13) leads to:

$$\mathbb{E}[(\widehat{g}_m(x_0) - g_m(x_0))^2] \leq \frac{\|g\|_\infty}{n} \sum_{\lambda\in I_m} \chi^2_\lambda(x_0) \leq K^2\|g\|_\infty \frac{D_m}{n} = K^2\nu\frac{D_m}{n}$$

This bound is standard for a variance term. Finally, for every model $m \in J_n$ we have the following non adaptive bound for $\widehat{g}_m$:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] \leq (g - g_m)^2(x_0) + K^2\nu\frac{D_m}{n} \tag{16}$$

3.2. **Construction of the adaptive estimator.** The model selection procedure developed by Birgé and Massart relies on this idea: the best model among the collection $\mathcal{M}_n$ is the one which minimizes the squared bias-variance sum above, thus the natural idea consists in building an estimator of the right hand side in (16) and selecting the model $\widehat{m}$ which minimizes it.

The term $K^2\nu D_m/n$ is estimated by $K^2\widehat{\nu}_n D_m/n$ where $\widehat{\nu}_n$ is an estimator of ν defined in Section 3.3.

Let us consider the bias term $(g - g_m)^2(x_0)$. Contrary to the L^2-bias term $\|g - g_m\|^2$ in classical model selection procedure, the pointwise bias term $(g_m - g)^2(x_0)$ is not easy to estimate. We replace $(g_m - g)^2(x_0)$ by $\sup_{j\in J_n, D_j \geq D_m}(g_j - g_m)^2(x_0)$. Indeed, those two terms have the same order according to (11) in $\mathbf{H_{dens}}(\beta)$:

$$\begin{aligned}
\sup_{j\in J_n, D_m\leq D_j} (g_j - g_m)^2(x_0) &\leq \sup_{j\in J_n, D_m\leq D_j\leq M_n} (g_j - g)^2(x_0) + (g_m - g)^2(x_0) \\
&\leq \sup_{j\in J_n, D_m\leq D_j} C_0 D_j^{-2\beta} + C_0 D_m^{-2\beta} \\
&\leq 2C_0 D_m^{-2\beta} \qquad (17)
\end{aligned}$$

Then we define the best theoretical model as:

$$m_{opt} := \arg\min_{m\in J_n} [\sup_{j\in J_n, D_j\geq D_m} (g_j(x_0) - g_m(x_0))^2 + pen(m)] := \arg\min_{m\in J_n} [Crit(m)] \tag{18}$$

where $pen(m) := AK^2 x_m \widehat{\nu}_n \frac{D_m}{n}$, A is a positive constant and x_m a weight of order $\ln(D_m)$. More precisely

$$x_m := \max\{B_1 \ln(1 + D_m); \frac{B_2}{\widehat{\nu}_n} \ln^2(1 + D_m)\frac{D_m}{n}\} \tag{19}$$

where (B_1, B_2) are constants with $B_1 > 16/A$ and $B_2 > 128K^2/A$. Asymptotically $x_m = B_1 \ln(1 + D_m)$.

Then, the natural idea would be to replace $(g_j - g_m)^2(x_0)$ by $(\widehat{g}_j - \widehat{g}_m)^2(x_0)$, one can notice that this estimator is biased. In fact:

$$\mathbb{E}[(\widehat{g}_m - \widehat{g}_j)^2(x_0)] = (g_j - g_m)^2(x_0) + \mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2]$$

and the last term is a variance-type term. Therefore we use the following bound:

$$\begin{aligned}\mathbb{E}[((\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0))^2] &\leq 2(\mathbb{E}[(\widehat{g}_j-g_j)^2(x_0)]+\mathbb{E}[(\widehat{g}_m-g_m)^2(x_0)])\\ &\leq 2K^2\|g\|_\infty\frac{D_j+D_m}{n}\end{aligned}$$

The last inequality is established by the same upper bounds as (16). So $(g_j-g_m)^2(x_0)$ is replaced by the positive part of $(\widehat{g}_m-\widehat{g}_j)^2(x_0)-AK^2\widehat{\nu}_n x_{j,m}\frac{D_j+D_m}{n}$ where $x_{j,m}$ is a weight of order $\ln(1+D_m+D_j)$:

$$x_{j,m}:=\max\{2B_1\ln(1+D_j+D_m);\frac{B_2}{\widehat{\nu}_n}\ln^2(1+D_j+D_m)\frac{D_j+D_m}{n}\}. \tag{20}$$

Finally the selected model $\widehat{m}$ is $\widehat{m}=\arg\min_{m\in\mathcal{M}_n}\widehat{Crit}(m)$ where:

$$\widehat{Crit}(m)=\sup_{j\in J_n, D_j\geq D_m}[(\widehat{g}_j-\widehat{g}_m)^2(x_0)-AK^2\widehat{\nu}_n x_{j,m}\frac{D_j+D_m}{n}]_+ + pen(m). \tag{21}$$

Our estimator of g is $\widehat{g}_{\widehat{m}}$.

3.3. **Estimation of ν.** In this section, we propose an estimator $\widehat{\nu}_n$ of $\nu=\|g\|_\infty$ constructed from the sample Z_1. Let m_0 be a medium-size model. More precisely, let $\gamma\in]1/3,1/2[$ and $m_0=\min\{m\in J_n: D_{m_0}\geq n^\gamma\}$ and $p_0=D_{m_0}$. We define:

$$\widehat{\nu}_n:=\|\widehat{g}_{m_0}\|_\infty$$

The following results hold:

Proposition 3.1. *Suppose that Assumption* $\mathbf{H_{dens}}(\beta)$ *holds, and that for every model* $m\in\mathcal{M}_n$ *the functions* $\{\chi_\lambda\}_{\lambda\in I_m}$ *are continuous. Then for every* n *such that:*

$$(\mathbf{A_1})\quad C_0 p_0^{-\beta}<\nu/6$$

Then there exists a numerical constant C_1 *such that:*

$$P[\widehat{\nu}_n\leq\nu/2]\leq 2\exp(-\frac{C_1}{K^2}\nu\frac{n}{p_0}) \tag{22}$$

If in addition:

$$(\mathbf{A_2})\quad \frac{p_0}{\sqrt{n}}\leq\frac{\nu}{3K^2}$$

then there exist numerical constants C_2, C_3 *and* C_4 *such that:*

$$P[\widehat{\nu}_n\geq 2\nu]\leq\exp(-\frac{C_2}{K^2}\nu\frac{n}{p_0})+\exp(-\frac{C_3}{K^4}\nu^2\frac{n^{3/2}}{p_0^2})+\exp(-\frac{C_4}{K^2}\frac{n}{p_0^2}) \tag{23}$$

Comments:

1) There exists an integer N which depends on (K,β,C_0) such that for every $n\geq N$, $(\mathbf{A_1})$ and $(\mathbf{A_2})$ hold.

2) The condition of continuity of the (χ_λ) prohibits the piecewise continuous bases, like for example the histograms. Nevertheless, similar upper bounds can be obtained with localised bases, included the histograms. Besides, the collections of model in which $\widehat{\nu}_n$ and $\widehat{g}_{\widehat{m}}$ are computed can be different.

3.4. Upper bound for the pointwise risk of $\widehat{g}_{\widehat{m}}$.

Theorem 3.1. *Assume that* $\mathbf{H_{dens}}(\beta)$, $(\mathbf{A_1})$ *and* $(\mathbf{A_2})$ *hold, there exists a constant* κ *which depends on* (α, B_1, B_2) *such that the following inequality holds:*

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \leq \kappa(1+\nu)(\frac{n}{\ln n})^{-2\beta/(2\beta+1)} + \mathcal{R}_n$$

with: $\mathcal{R}_n = (\nu + K^2 M_n)^2 \exp(-\frac{C_1 \nu}{K^2}\frac{n}{p_0}) + (\frac{n}{\ln n})^{2\beta/(2\beta+1)} K^2 p_0 [\exp(-\frac{C_2}{K^2}\nu\frac{n}{p_0}) + \exp(-\frac{C_3}{K^4}\nu^2\frac{n^{3/2}}{p_0^2}) + \exp(-\frac{C_4}{K^2}\frac{n}{p_0^2})]$

Comments:
1) According to (13), p_0 is of order $O(n^\gamma)$ with $\gamma \in]1/3, 1/2[$. Thus we get immediately that $\mathcal{R}_n \leq C/n$ for some constant C depending on (ν, K, p_0).

2) If $g \in \mathcal{H}(\beta, L)$ then g satisfies $\mathbf{H_{dens}}(\beta)$. Besides, Stone (1980) proves that the minimax rate of convergence over the set of k times continuously differentiable functions in density estimation is $n^{-2k/(2k+1)}$. Tsybakov generalized this result to Hölder classes of functions for every $\beta > 0$ (see Tsybakov (2004)). Moreover Lepski and Spokoiny (1997) show that the adaptive minimax rate of convergence over Hölder classes for the white noise model is $(n/\ln n)^{-2\beta/(2\beta+1)}$. This allows to believe that the adaptive rate of convergence over Hölder classes for density is also $(n/\ln n)^{-2\beta/(2\beta+1)}$. So our estimator seems to be adaptive over Hölder classes.

According to the comments about Theorem 2.1, our estimator is also adaptive over Sobolev classes.

3.5. Application to the estimation of the error density.

Now we go back to the initial issue, the estimation of the error density, and clarify the estimator defined in Section 2. Let us recall that our goal is to build an estimator of the error density f out of a sample $(X_i, Y_i)_{\{i=-n,\dots,-1\}\cup\{1,\dots 2n\}}$ from regression framework (1). The sample is split into three independent samples Z^-, Z_0^+ and Z_1^+ defined in (5). Let $\widehat{b}$ be an estimator of b computed from the sample Z^- and $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$ for $i \in \{1, \dots, 2n\}$ the residuals from the two other samples. Given Z^-, $\widehat{b}$ is fixed and the $(\widehat{\epsilon}_i)$ have a density f^-. Let us give f^- explicitly. Let F be any function, then:

$$\begin{aligned}
\mathbb{E}[F(\widehat{\epsilon}_1)|Z^-] &= \mathbb{E}[F(\epsilon_1 + (b - \widehat{b})(X_1))|Z^-] \\
&= \int_{t\in\mathbb{R}} \int_{x=0}^{1} F(t + (b-\widehat{b})(x)\mu(x) f(t) dx dt \\
&= \int_{u\in\mathbb{R}} F(u) \int_{x=0}^{1} f(u - (b-\widehat{b})(x))\mu(x) dx du.
\end{aligned}$$

Hence:

$$f^-(t) = \int_0^1 f(t - (b-\widehat{b})(x)\mu(x) dx.$$

We can easily deduce from this expression that f^- is upper bounded by $\nu^- := \|f^-\|_\infty \leq \nu$ for every Z^-.

Now, we apply the density estimation procedure presented in Section 3 to the sample $(\widehat{\epsilon}_i)$. For every model m, let $\widehat{f}_m^-$ be the projection estimator defined in (6). Let $m_0 = \min\{m : D_m \geq n^\gamma\}$, and:

$$\widehat{\nu}_n^- := \|\widehat{f}_{m_0}^-\|_\infty.$$

Then the density estimation procedure is applied to the residuals and provides the estimator $\widehat{f}_{\widehat{m}}^-$, where $\widehat{m}$ is the selected model (7).

Let us explain the basic guidelines of this result. The risk of the estimator $\widehat{f}_{\widehat{m}}^-$ comes from two consecutive approximations of different nature: the first one consists in replacing the errors (ϵ_i) by the residuals, and the second one is a density estimation error. These two approximations appear in the following inequality:

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f)^2(x_0)] \leq 2\{\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}. \tag{24}$$

On the one hand, for a fixed sample Z^-, we prove (see Lemma 6.3) that f^- satisfies the Assumption $\mathbf{H_{dens}}(\beta)$ so

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)|Z^-] \leq \kappa(\frac{n}{\ln n})n^{-2\beta/(2\beta+1)} + \frac{C}{n}$$

By taking the expectation over Z^-, we get the first term in Theorem 2.1. Actually, the constant C depends on f^- and so on Z^- and we need to study it more carefully to obtain this result (see Section 5).

On the other hand f^- is the density of $\widehat{\epsilon}_i = \epsilon_i + (b - \widehat{b})(X_i)$, so the difference between f and f^- can be expressed in function of $(b - \widehat{b})$. More precisely, we will prove that:

$$\mathbb{E}[(f - f^-)^2(x_0)] \leq C\mathbb{E}[\|\widehat{b} - b\|^2].$$

4. An adaptive estimator of the regression function

In this section, we briefly exhibit an estimator $\widehat{b}$ of b which suits to our setting. This is the estimator which is implemented in the simulations. The regression function estimator presented here results from Baraud's works (see Baraud (2002) and Baraud (2000)), gathered in Plancade (2008). Consider the following assumption:

$\mathbf{H_3}$: The density μ of X_1 is supported on a compact J, and is lower bounded by a $m_0 > 0$ and upper bounded by $m_1 < +\infty$.

Let us consider a collection of finite dimensional models Σ_n which satisfies the following assumptions:

$\mathbf{H_b}$: Σ_n is included in a global model S_n with dimension smaller than $n^{1/2-d}$ for some $d > 0$. Furthermore, there exists some nonnegative constants Γ and R such that

$$|\{m \in \mathcal{M}_n(resp.\Sigma_n) : D_m = n\} \leq \Gamma D^R$$

for every $D \in \mathbb{N}^*$. Finally, there exists a constant K such that:

$$\|t\|_\infty \leq K\sqrt{N_n}\|t\|, \ \forall t \in S_n.$$

For every model $m \in \Sigma_n$, let $\widehat{b}_m$ be the least squares estimator of b:

$$\widehat{b}_m := \arg\min_{t \in S_m} \gamma_n(t) \quad \text{where} \quad \gamma_n(t) := \frac{1}{n}\sum_{i=-n}^{-1}(Y_i - t(X_i))^2,$$

and the selected model is $\widehat{m} = \arg\min_{m\in\Sigma_n}[\gamma_n(\widehat{b}_m) + \widehat{\sigma}_n^2 \frac{D_m}{n}]$ where $\widehat{\sigma}_n^2$ is an estimator of the variance of ϵ_1: let V_n be a space of dimension $E(n/2)$ which includes the global model S_n, then:

$$\widehat{\sigma}_n^2 = \frac{1}{n - E(n/2)} \inf_{t\in S_n} (Y_i - t(X_i))^2$$

Let us define $\widehat{b} = \widehat{b}_{\widehat{m}}$ if $\|\widehat{b}_{\widehat{m}}\| \leq n$ and $\widehat{b} = 0$ otherwise then:

$$\mathbb{E}[\|b - \widehat{b}\|_\mu^2] \leq C \inf_{m\in\Sigma_n} [\|b - b_m\|^2 + \sigma^2 \frac{D_m}{n}].$$

Finally, classical results about approximation theory in Besov spaces lead to the following statement: if b belong to the Besov space $\mathcal{B}_2^{\alpha,\infty}$, then $\mathbb{E}[\|\widehat{b} - b\|_\mu^2] \leq Cn^{-2\alpha/(2\alpha+1)}$. This entails the following Corollary:

Corollary 4.1. *Suppose that Assumptions* $\mathbf{H_0(f)}$, $\mathbf{H_b}$ *and* $\mathbf{H_1}(\beta)$ *or* $\mathbf{H_2}(\beta)$ *hold, and suppose that* $b \in \mathcal{B}_2^{\alpha,\infty}$ *with* $\alpha \geq \beta - 1/2$ *then:*

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f)^2(x_0)] \leq \kappa(\frac{n}{\ln n})^{-\frac{2\beta}{2\beta+1}}$$

for some constant κ *independent from* n.

In other words, if b is smoother than f, the rate of convergence of $\widehat{f}_{\widehat{m}}^-$ is the optimal rate we would get if the (ϵ_i) were directly observed.

5. Simulations

5.1. **Density estimation.** This section illustrates the density estimation procedure presented in Section 3, with the sine-cardinal collection of models $\mathcal{A}_n$ described in (2). We choose $B = 10$ and $M_n = \sqrt{n}$. We draw 50 samples $(V_1, \ldots, V_n)$ of size $n = 200$, 500, 2000 of i.i.d. variables with gaussian distribution (denoted by $\mathcal{N}(0,1)$) and with Laplace density $g(x) = \frac{1}{2}\exp(-|x|)$ (denoted by $\mathcal{L}(1)$). Let J be the set of 150 regularly spaced points on $[-5,5]$. For each sample and for every point $x \in J$ we compute an estimator $\widehat{g}_{\widehat{m}}(x)$ as follows, assuming that the maximum of the density ν is known:

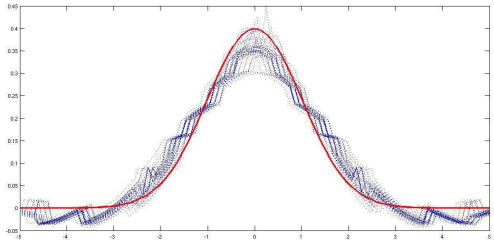
- First we compute the projection density estimators $(\widehat{g}_m(x))$ for every $m \in \frac{1}{10}\mathbb{N}$, $m \leq M_n$ and every $x \in J$ (cf (15)).
- Then for every $x \in J$, we select the best model as:

$$\widehat{m} = \arg\min\{\sup_{j\geq m}[(\widehat{g}_j - \widehat{g}_m)^2(x) - \alpha\nu\ln(1+j+m)\frac{j+m}{n}]_+ + \beta\nu\ln(1+m)\frac{m}{n}$$

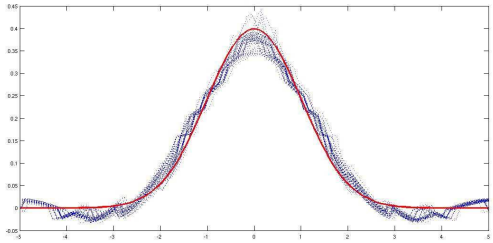
  with $\alpha = 5$ and $\beta = 15$.
- We plot the set of points $\{(x, \widehat{g}_{\widehat{m}}(x)), x \in J\}$

In Figure 1, each graph presents the 50 curves of $\widehat{g}_{\widehat{m}}$ for a given density g_i and a given n.

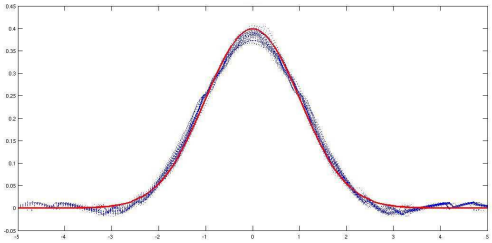
$\mathbf{V_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$



n=200

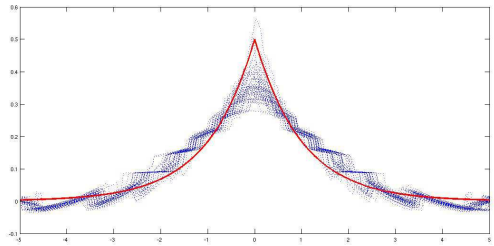


n=500

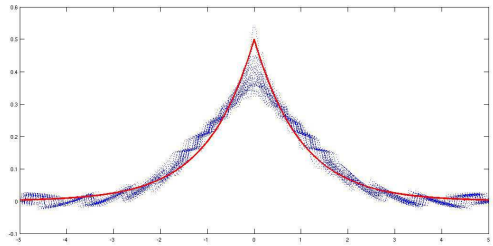


n=2000

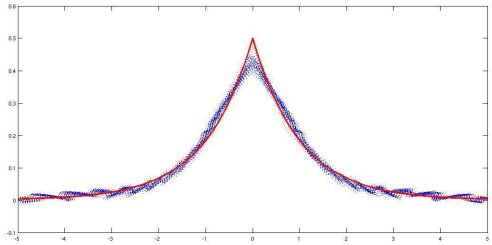
$\mathbf{V_i} \sim \mathcal{L}(\mathbf{1})$



n=200



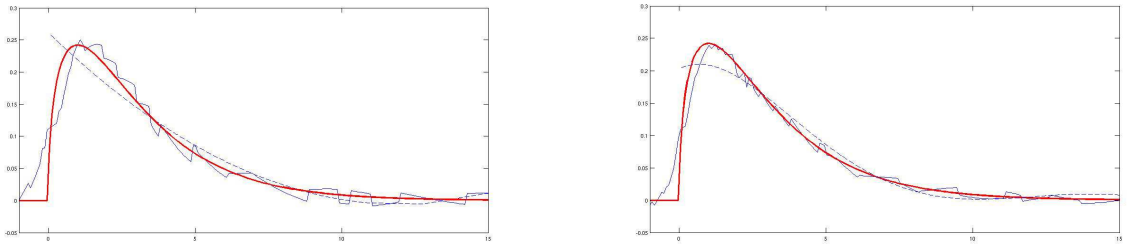
n=500



n=2000

FIGURE 1. Beam of 50 density estimators curves (blue dotted lines) built from i.i.d. samples of size n=200, 500 and 2000 of density $\mathcal{N}(0,1)$ and $\mathcal{L}(1)$ (red thick line), in sine-cardinal bases.

Figure 2 presents a comparison between our pointwise model selection estimator, and a global model selection estimator, computed following the procedure developped by Massart (2007), Section 7, for sample of size $n = 500$, 2000 with common density $\chi^2(3)$. The global model selection estimator (dotted blue line) is computed in a mixed piecewise polynomial an trionometric polynomial basis using matlab programs available on Yves Rozenholc's web page (http://www.math-info.univ-paris5.fr/ rozen/). The pointwise model selection estimator (solid blue line) is built following the procedure described above, on the set J of 150 regularly spaced points on $[-1, 15]$. We observe that the pointwise model selection estimator (in solid blue line) fits the true density (in red thick line) for a smaller sample size than the global model selection estimator.



n=500 n=2000

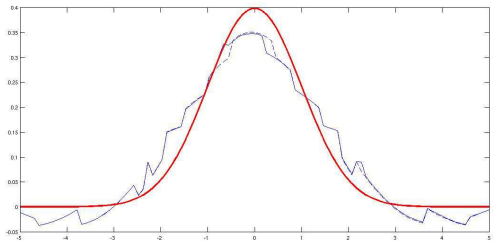
FIGURE 2. Pointwise model selection estimator (solid blue line) and global model selection estimator (dotted blue line) for a sample of size n=500, 2000 of density $\chi^2(3)$ (red thick line)

5.2. **Error density estimation.** This section proposes illustrations of the error density estimator described in Section 2, with the following procedure:

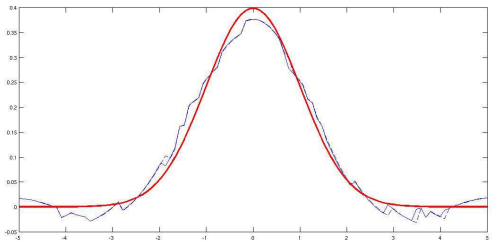
- We draw a sample $(X_1, \ldots, X_{2n})$ with common density f_X uniform on $[0, 1]$ and $\chi^2(3)$. We draw also a sample $(\epsilon_1, \ldots, \epsilon_{2n})$ with common density f from a distribution $\mathcal{N}(0, 1)$ and $\mathcal{L}(1)$. We choose a regression function $b(x) = x^3 + 5x$ and $b(x) = \exp(-|x|)$ and compute the sample $(Y_1, \ldots, Y_{2n})$ where $Y_i = b(X_i) + \epsilon_i$.
- From the sample $\{(X_i, Y_i)\}_{i=1\ldots n}$, we compute an estimator $\widehat{b}$ of b following the procedure described in Section 4, using mixed piecewise polynomial and trigonometric polynomial basis (see Comte et al. (2008)).
- We compute the residuals from the second sample $(\widehat{\epsilon_i})_{i=n+1,\ldots,2n}$, where $\widehat{\epsilon_i} = Y_i - \widehat{b}(X_i)$.
- Let J be a set of 150 regularly spaced points on $[-5, 5]$ and apply the density estimation procedure described in Section 5.1 to the residuals $(\widehat{\epsilon_i})_{i=n+1,\ldots,2n}$.

Figure 3 presents the error density estimator (dotted blue line) and the theoretical estimator we get by applying the density estimation procedure of Section 5.1 directly to the sample $(\epsilon_i)_{i=n+1,\ldots,2n}$. The thick line is the true density of ϵ_1.

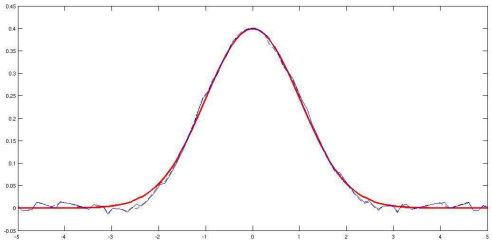
$$\mathbf{X_i} \sim \mathcal{U}[\mathbf{0},\mathbf{1}],\ \epsilon_\mathbf{i} \sim \mathcal{N}(\mathbf{0},\mathbf{1}),\ \mathbf{b(x)} = \mathbf{x^3} + \mathbf{5x}$$



n=200

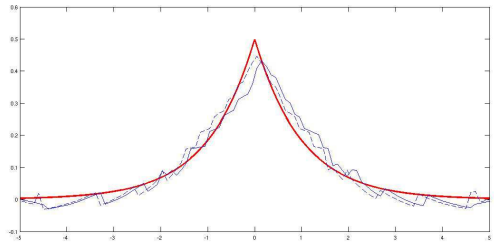


n=500

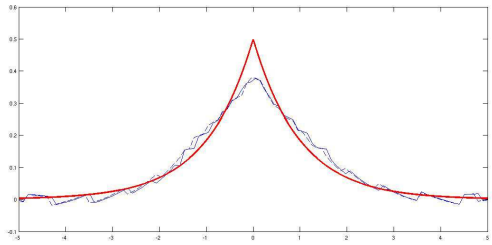


n=2000

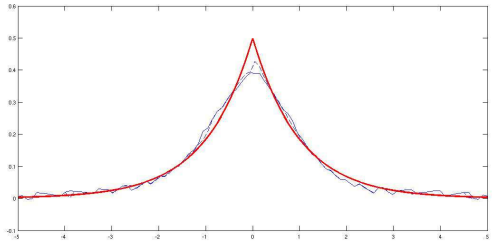
$$\mathbf{X_i} \sim \chi^\mathbf{2}(\mathbf{3}),\ \epsilon_\mathbf{i} \sim \mathcal{L}(\mathbf{1}),\ \mathbf{b(x)} = \exp(-|\mathbf{x}|)$$



n=200



n=500



n=2000

FIGURE 3. Error density estimator (solid blue line), theoretical estimator we would get if the errors were observed (dotted blue line) and true density (thick red line).

We have also checked that the error density estimator hardly depends on the designs' distribution.

6. Proofs

6.1. **Proof of Theorem** 3.1. Let Z_1 be fixed. Let us denote by $\mathbb{E}_1[.]$ the conditional expectation $\mathbb{E}[.|Z_1]$ and $P_1[.]$ the conditional probability $P[.|Z_1]$. We first prove the following Claim:

Claim 1. *If Assumption* $\mathbf{H_{dens}}(\beta)$ *holds, there exist constants* κ *and* κ' *which depend on* (C_0, B_1, B_2, K) *such that the following inequality holds:*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times 1\!\!1_{\{\widehat{\nu}_n \geq \nu/2\}} \leq \kappa[Crit(m_{opt}) + (g_{m_{opt}}(x_0) - g(x_0))^2] + \frac{\kappa'(1+\nu)}{n} \quad (25)$$

Proof of Claim 1.
For every $j, m \in J_n$, we denote by:

$$H(j,m) := AK^2 x_{j,m} \widehat{\nu}_n \frac{D_m + D_j}{n}.$$

The basic idea of the proof is to upper bound $\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]1\!\!1_{\{\widehat{\nu}_n \geq \nu/2\}}$ by the sum of two terms:

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)1\!\!1_{\{\widehat{\nu}_n \geq \nu/2\}}] \leq 2(\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+]1\!\!1_{\{\widehat{\nu}_n \geq \nu/2\}} + \mathbb{E}_1[\mathcal{U}_{opt}]) \quad (26)$$

where $\mathbb{E}[\mathcal{U}_{opt}]$ is a quantity with same order as $Crit(m_{opt})$. Besides:

$$\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] \leq \int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x]dx \quad (27)$$

and the quantity $\mathcal{U}_{opt}$ will be chosen such that the probability under the integral decreases exponentially in n. Let us consider a first result:

Lemma 6.1. *For every* $\delta > 0$, $x > 0$ *and for every model* m*:*

$$P_1[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x] \leq \sum_{j \in J_n, D_j \geq D_m} \exp[-C(x,j,m)]$$

where $C(x,j,m) = \min\{\frac{u}{\nu K^2}\frac{nx}{D_j + D_m} + Aux_{j,m}\frac{\widehat{\nu}_n}{\nu}; \frac{u'n\sqrt{x}}{K^2(D_j+D_m)} + \frac{u'\sqrt{A}}{K}\sqrt{x_{j,m}\widehat{\nu}_n\frac{n}{D_j+D_m}}\}$

and $u = 1/(8(1+\frac{1}{\delta}))$ *and* $u' = 1/(4\sqrt{2}\sqrt{1+\frac{1}{\delta}})$.

Proof of Lemma 6.1**:** The empirical criterion $\widehat{Crit}(m)$ (defined in (21)) is built from $Crit(m)$ (defined in (18)) by replacing the unknown (g_j) by their empirical means $(\widehat{g}_j)$, so the deviation between $\widehat{Crit}(m)$ and $Crit(m)$ is upper bounded with Bernstein Inequality (see Appendix). More precisely:

$$\begin{aligned} & P_1[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x] \\ \leq\ & P_1[\sup_{j \in J_n, D_j \geq D_m}((\widehat{g}_j - \widehat{g}_m)^2(x_0) - H(j,m))_+ \geq (1+\delta)\sup_{j \in J_n, D_j \geq D_m}(g_j - g_m)^2(x_0) + x] \end{aligned}$$

As $\sup_{j\in J_n, D_j\geq D_m}(g_j-g_m)^2(x_0)+x$ is positive, we omit the positive part $(.)_+$:

$$\begin{aligned}
& P_1[\widehat{Crit}(m)\geq (1+\delta)Crit(m)+x] \\
\leq\ & P_1[\sup_{j\in J_n, D_j\geq D_m}((\widehat{g}_j-\widehat{g}_m)^2(x_0)-H(j,m))\geq (1+\delta)\sup_{j\in J_n, D_j\geq D_m}(g_j-g_m)^2(x_0)+x] \\
\leq\ & \sum_{j\in J_n, D_j\geq D_m} P_1[(\widehat{g}_j-\widehat{g}_m)^2(x_0)\geq (1+\delta)(g_j-g_m)^2(x_0)+x+H(j,m)] \\
:=\ & \sum_{j\in J_m, D_j\geq D_m} P_{j,m}
\end{aligned}$$

and for every (j,m):

$$P_{j,m}=P_1[(\widehat{g}_j-\widehat{g}_m)^2(x_0)\geq (1+\delta)(g_j-g_m)^2(x_0)+(1+\frac{1}{\delta})(\sqrt{\frac{x+H(j,m)}{(1+\frac{1}{\delta})}})^2]$$

It follows from the inequality $(x+y)^2\leq x^2(1+1/a)+y^2(1+a)$, $\forall\ x,y\in\mathbb{R}, a>0$ that:

$$\begin{aligned}
P_{j,m} &\leq P_1[(\widehat{g}_j-\widehat{g}_m)^2(x_0)\geq (|(g_j-g_m)(x_0)|+\sqrt{\frac{x+H(j,m)}{1+\frac{1}{\delta}}})^2] \\
&= P_1[|(\widehat{g}_j-\widehat{g}_m)(x_0)|\geq |(g_j-g_m)(x_0)|+\sqrt{\frac{x+H(j,m)}{1+\frac{1}{\delta}}}] \\
&\leq P_1[|(\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0)|+|(g_j-g_m)(x_0)|\geq |(g_j-g_m)(x_0)| \\
&\qquad\qquad +\sqrt{\frac{x+H(j,m)}{1+\frac{1}{\delta}}}] \\
&= P_1[|\frac{1}{n}\sum_{i=1}^n (U_i-\mathbb{E}(U_i))|\geq \sqrt{\frac{x+H(j,m)}{1+\frac{1}{\delta}}}]
\end{aligned}$$

where $U_i=\sum_{\lambda\in I_j}\chi_\lambda(V_i)\chi_\lambda(x_0)-\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)$ and $\mathbb{E}(U_i)=(g_j-g_m)(x_0)$. Let us compute the terms v and c involved in Bernstein Inequality (Theorem 7.1).

By the same methods as in (16) we get:

$$\mathbb{E}_1(U_1^2)\leq 2\mathbb{E}_1[\widehat{g}_j^2(x_0)+\widehat{g}_m^2(x_0)]\leq 2\nu K^2(D_j+D_m):=v$$

Let ℓ be an integer greater than 2, then:

$$\begin{aligned}
\mathbb{E}_1[(U_1)_+^\ell] &\leq \mathbb{E}_1[U_1^2]\times\|U_1\|_\infty^{\ell-2} \\
&\leq v^2[\|\sum_{\lambda\in I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\|_\infty+\|\sum_{\lambda\in I_j}\chi_\lambda(V_1)\chi_\lambda(x_0)\|_\infty]^{\ell-2} \\
&\leq v^2[\|\sqrt{\sum_{\lambda\in I_m}\chi_\lambda^2(V_1)}\sqrt{\sum_{\lambda\in I_m}\chi_\lambda^2(x_0)}\|_\infty+\|\sqrt{\sum_{\lambda\in I_j}\chi_\lambda^2(V_1)}\sqrt{\sum_{\lambda\in I_j}\chi_\lambda^2(x_0)}\|_\infty]^{\ell-2}
\end{aligned}$$

Finally, Assumption $\mathbf{H_{con}}$ leads to:

$$\mathbb{E}_1[(U_1)_+^l]\leq v^2[K^2(D_j+D_m)]^{l-2} \tag{28}$$

and $c = K^2(D_j + D_m)$. Let us denote:

$$\epsilon = \sqrt{\frac{x + H(j,m)}{1+\frac{1}{\delta}}} \geq \frac{1}{\sqrt{2(1+\frac{1}{\delta})}}(\sqrt{x} + \sqrt{H(j,m)})$$

Then Bernstein Inequality provides the following upper bound for $P_{j,m}$

$$P_{j,m} \leq 2\exp[-\min(\frac{n\epsilon^2}{4v}; \frac{n\epsilon}{4c})]$$

Moreover:

$$\begin{aligned}\frac{n\epsilon^2}{4v} &= \frac{nx}{8(1+\frac{1}{\delta})\nu K^2(D_j+D_m)} + \frac{AK^2 x_{j,m}\widehat{\nu}_n(D_j+D_m)}{8(1+\frac{1}{\delta})\nu K^2(D_j+D_m)} \\ &= \frac{u}{\nu K^2}\frac{nx}{D_j+D_m} + Aux_{j,m}\frac{\widehat{\nu}_n}{\nu}\end{aligned}$$

$$\begin{aligned}\frac{n\epsilon}{4c} &\geq \frac{n\sqrt{x}}{4\sqrt{2}\sqrt{1+\frac{1}{\delta}}K^2(D_j+D_m)} + n\sqrt{Ax_{j,m}\widehat{\nu}_n\frac{D_j+D_m}{n}}\frac{1}{4\sqrt{2}\sqrt{1+\frac{1}{\delta}}K^2(D_j+D_m)} \\ &= \frac{u'n\sqrt{x}}{K^2(D_j+D_m)} + \frac{u'\sqrt{A}}{K}\sqrt{x_{j,m}\widehat{\nu}_n\frac{n}{D_j+D_m}}\end{aligned}$$

which provides the upper bound of Lemma 6.1. □

We deduce from Lemma 6.1 an upper bound for a quantity of the kind $P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x]$:

Lemma 6.2. *Let δ and x some positive numbers, then for every Z_1:*

1) $P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)(\sup_{j\in J_n, D_j \geq D_{m_{opt}}}(g_j - g)^2(x_0) + Crit(m_{opt})) + x\} \cap \{\widehat{m} > m_{opt}\}]$

$$\leq \sum_{j\in J_n, D_j \geq D_{m_{opt}}} \exp[-C(x,j,m_{opt})] + 2\sum_{m\in J_n}\exp[-C(x,m)]$$

where $C(x,m) = \min\{\frac{2u}{K^2}\frac{nx}{\nu D_m} + 2Au\frac{x_m\widehat{\nu}_n}{\nu}; \frac{u'}{K}\frac{n\sqrt{x}}{D_m} + \frac{u'\sqrt{A}}{K}\sqrt{\frac{nx_m\widehat{\nu}_n}{D_m}}\}$.

2) $P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq 2(1+\delta)Crit(m_{opt}) + 2\sup_{j\in J_n, D_j \leq D_{m_{opt}}} H(m_{opt}, j) + 2(\widehat{g}_{m_{opt}} - g)^2(x_0)$

$$+2x\} \cap \{\widehat{m} \leq m_{opt}\}] \leq \sum_{j\in J_n, D_j \geq D_{m_{opt}}} \exp[-C(x,j,m)]$$

Proof of Lemma 6.2**:**

• Let us prove inequality 1).

$$P_1[\{(\widehat{g}_{\widehat{m}}-g)^2(x_0) \geq (1+\delta)(\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)+Crit(m_{opt}))+x\}\cap\{\widehat{m}>m_{opt}\}]$$

$$\leq P_1[\{(\widehat{g}_{\widehat{m}}-g)^2(x_0) \geq (1+\delta)\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)+\widehat{Crit}(\widehat{m})+x\}\cap\{\widehat{m}>m_{opt}\}]$$

$$+P_1[\widehat{Crit}(\widehat{m})\geq(1+\delta)Crit(m_{opt})]$$

By definition of $\widehat{m}$, $\widehat{Crit}(\widehat{m})=\inf_{m\in J_n}\widehat{Crit}(m)\leq\widehat{Crit}(m_{opt})$ thus we get from Lemma 6.1:

$$\begin{aligned} P_1[\widehat{Crit}(\widehat{m})\geq(1+\delta)Crit(m_{opt})] &\leq P[\widehat{Crit}(m_{opt})\geq(1+\delta)Crit(m_{opt}) \\ &\leq \sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp[-C(x,j,m_{opt})] \end{aligned}$$

Besides for every model m, $Crit(m)\geq pen(m)$ according to the definition of $Crit(m)$, and if $\widehat{m}\leq m_{opt}$, $\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)\geq(g_{\widehat{m}}-g)^2(x_0)$, thus:

$$P_1[\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)+\widehat{Crit}(\widehat{m})+x\}\cap\{\widehat{m}>m_{opt}\}]$$

$$\begin{aligned} &\leq P_1[(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)(g_{\widehat{m}}-g)^2(x_0)+pen(\widehat{m})+x] \\ &\leq \sum_{m\in J_n}P_1[(\widehat{g}_m-g)^2(x_0)\geq(1+\delta)(g_m-g)^2(x_0)+pen(m)+x]:=\sum_{m\in J_n}P_m \end{aligned}$$

The quantities P_m are upper bounded in the same way as $P_{j,m}$ in the proof of Lemma 6.1, so we only give the outline of the proof. First of all we have for every model m:

$$\begin{aligned} P_m &\leq P_1[|(\widehat{g}_m-g_m)(x_0)|\geq\frac{1}{1+\frac{1}{\delta}}\sqrt{pen(m)+x}] \\ &= P_1[|\frac{1}{n}\sum_{i=1}^n U_i-\mathbb{E}(U_i)|\geq\frac{1}{1+\frac{1}{\delta}}\sqrt{pen(m)+x}] \end{aligned}$$

where $U_i=\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)$. We apply Bernstein Inequality with the following quantities v and c:

$$\mathbb{E}[U_1^2]\leq\nu K^2D_m:=v^2$$

For every integer $l\geq 2$, similarly to inequality (28) we have:

$$\mathbb{E}[(U_1)_+^l]\leq v^2(K^2D_m)^{l-2}$$

thus $c=K^2D_m$. Then Bernstein Inequality provides inequality 1), exactly like in the proof of Lemma 6.1.

• Let us prove now inequality 2) in Lemma 6.2. If $\widehat{m}\leq m_{opt}$ given that $pen(m)$ is always positive:

$$\widehat{Crit}(\widehat{m})\geq\sup_{j\in J_n, D_j\geq D_{\widehat{m}}}[(\widehat{g}_j-\widehat{g}_{\widehat{m}})^2(x_0)-H(j,\widehat{m})]\geq(\widehat{g}_{m_{opt}}-\widehat{g}_{\widehat{m}})^2(x_0)-H(m_{opt},\widehat{m})$$

Moreover $(\widehat{g}_{\widehat{m}}-g)^2(x_0)\leq 2[(\widehat{g}_{\widehat{m}}-\widehat{g}_{m_{opt}})^2(x_0)+(\widehat{g}_{m_{opt}}-g)^2(x_0)]$, thus:

$$\begin{aligned}\widehat{Crit}(\widehat{m}) &\geq \frac{1}{2}(\widehat{g}_{\widehat{m}}-g)^2(x_0)-(\widehat{g}_{m_{opt}}-g)^2(x_0)-H(m_{opt},\widehat{m})\\ &\geq \frac{1}{2}(\widehat{g}_{\widehat{m}}-g)^2(x_0)-(\widehat{g}_{m_{opt}}-g)^2(x_0)-\sup_{j\in J_n, D_j\leq D_{m_{opt}}} H(m_{opt},j)\end{aligned}$$

Hence:

$$\begin{aligned}&P_1[\{\frac{1}{2}(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)Crit(m_{opt})+\sup_{j\in J_n, D_j\leq D_{m_{opt}}} H(m_{opt},j)\\ &\qquad +(\widehat{g}_{m_{opt}}-g)^2(x_0)+x\}\cap\{\widehat{m}\leq m_{opt}\}]\\ &\leq P[\widehat{Crit}(\widehat{m})\geq(1+\delta)Crit(m_{opt})+x]\leq \sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp[-C(x,j,m_{opt})]\end{aligned}$$

which concludes the proof of Lemma 6.2. □

Let us define:

$$\mathcal{U}_{opt}:=2(\widehat{g}_{\widehat{m}}-g)^2(x_0)+2(1+\delta)Crit(m_{opt})+2\sup_{j\in J_n, D_j\leq D_{m_{opt}}} H(m_{opt},j)+\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)$$

for some constant $\delta>0$ defined later. According to inequalities 1) and 2) in Lemma 6.2, we have:

$$P_1[(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq\mathcal{U}_{opt}+x]\leq 2\sum_{m\in J_n}\exp(-C(x,m))+2\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-C(x,j,m))$$

Hence:

$$\begin{aligned}&\mathbb{E}_1[((\widehat{g}_{\widehat{m}}-g)^2(x_0)-\mathcal{U}_{opt})_+]\leq\int_0^{+\infty}P_1[(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq\mathcal{U}_{opt}+x]dx\\ &\leq 2\int_0^{+\infty}[\sum_{m\in J_n}\exp(-C(x,m))+\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-C(x,j,m))]dx \qquad (29)\end{aligned}$$

Besides, for every constant C:

$$\int_0^{+\infty}\exp(-Cx)dx=\frac{1}{C},\quad \int_0^{+\infty}\exp(-C\sqrt{x})dx=\frac{2}{C^2}.$$

Thus

$$\begin{aligned}\int_0^{+\infty}\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-C(x,j,m_{opt}))dx\leq\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\Bigg[&\exp(-Aux_{j,m_{opt}}\frac{\widehat{\nu}_n}{\nu})\frac{K^2\nu}{u}\frac{D_j+D_{m_{opt}}}{n}\\ &+\exp(-\frac{u'\sqrt{A}}{K}\sqrt{x_{j,m_{opt}}\widehat{\nu}_n\frac{n}{D_j+D_{m_{opt}}}})\frac{K^4}{u'^2}\frac{(D_j+D_{m_{opt}})^2}{n^2}\Bigg]\end{aligned}$$

Moreover, for every $j\in J_n$, $(D_j+D_{m_{opt}})/n\leq 2$, and if $\widehat{\nu}_n/\nu\geq 1/2$ then:

$$\int_0^{+\infty}\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-C(x,j,m_{opt}))dx\leq\frac{2}{n}\left[\frac{K^2\nu}{u}\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-\frac{1}{2}Aux_{j,m_{opt}})(D_j+D_{m_{opt}})\right.$$

$$+\frac{2K^4}{u'^2}\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-\frac{u'\sqrt{A}}{K}\sqrt{\frac{x_{j,m_{opt}}n\widehat{\nu}_n}{D_j+D_{m_{opt}}}})(D_j+D_{m_{opt}})\Bigg]$$

The term $\int_0^{+\infty}\sum_{j\in J_n, j\geq m_{opt}}\exp(-C(x,j,m_{opt}))dx$ has order $1/n$ as soon as:

$$\begin{cases}\exp(-\frac{1}{2}Aux_{j,m_{opt}})\leq(1+D_j+D_{m_{opt}})^{-(2+a)}\\ \exp(-\frac{u'\sqrt{A}}{K}\sqrt{\frac{x_{j,m_{opt}}n\widehat{\nu}_n}{D_j+D_{m_{opt}}}})\leq(1+D_j+D_{m_{opt}})^{-(2+a)}\end{cases}$$

for some $a>0$ which is equivalent to:

$$\begin{cases}x_{j,m_{opt}}\geq\frac{2(2+a)}{Au}\ln(1+D_j+D_m)\\ x_{j,m_{opt}}\geq\frac{(2+a)^2K^2}{u'^2A\widehat{\nu}_n}\times\frac{D_j+D_{m_{opt}}}{n}\ln^2(1+D_j+D_{m_{opt}})\end{cases}$$

This is guaranteed if:

$$x_{j,m}\geq\max\{\frac{16}{A}(2+a)(1+\frac{1}{\delta})^2\ln(1+D_j+D_m);$$
$$\frac{32K^2}{A\widehat{\nu}_n}(1+\frac{1}{\delta})^2(2+a)^2\ln^2(1+D_j+D_m)\frac{D_j+D_m}{n}\} \tag{30}$$

Let $B_1>32/A$ and $B_2>128K^2/A$ be the constants involved in the definition (20) of the $(x_{j,m})$ and let consider $\delta>0$ and $a>0$ such that $2B_1\geq\frac{32}{A}(2+a)(1+\frac{1}{\delta})^2$ and $B_2\geq 32K^2A(1+\frac{1}{\delta})^2(2+a)^2$. Then $x_{j,m_{opt}}$ satisfies inequality (30), and there exists a constant C which depends on (A,B_1,B_2,K) such that:

$$\int_0^{+\infty}\left(\sum_{j\in J_n, D_j\geq D_{m_{opt}}}\exp(-C(x,j,m_{opt}))\right)dx\leq(1+\nu)\frac{C}{n}$$

The same type of computation yields:

$$\int_0^{+\infty}\left(\sum_{m\in J_n}\exp(-C(x,m))\right)dx\leq(1+\nu)\frac{C}{n}$$

Then inequality (29) leads to:

$$\mathbb{E}_1[((\widehat{g}_{\widehat{m}}-g)^2(x_0)-\mathcal{U}_{opt})_+]\leq(1+\nu)\frac{C}{n} \tag{31}$$

Besides, for every $D_j\leq D_{m_{opt}}$, $H(m_{opt},j)\leq 2pen(m_{opt})$. Moreover:

$$\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)\leq 2[\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g_{m_{opt}})^2(x_0)+(g_{m_{opt}}-g)^2(x_0)]$$

Hence:

$$\begin{aligned}\mathbb{E}_1[\mathcal{U}_{opt}] &\leq 3(g_{m_{opt}}-g)^2(x_0)+2(1+\delta)Crit(m_{opt})+4pen(m_{opt})\\ &\quad+\sup_{j\in J_n, D_j\geq D_{m_{opt}}}(g_j-g_{m_{opt}})^2(x_0)\\ &\leq C'Crit(m_{opt})+3(g_{m_{opt}}-g)^2(x_0) \qquad (32)\end{aligned}$$

By gathering inequalities (26), (31) and (32), we get inequality (25).

Proof of Theorem 3.1. We prove the following claim:

Claim 2. *If Assumption* $\mathbf{H_{dens}}(\beta)$ *holds, there exist constants* κ *and* κ' *which depend on* (β, B_1, B_2)*, and a universal constant* C_1 *such that*

(33)
$$\mathbb{E}[(\widehat{g}_{\widehat{m}}-g)^2(x_0)1\!\!1_{I_{\{\widehat{\nu}_n\geq\nu/2\}}}] \leq \kappa(1+2\nu)(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}+\kappa' K^2 p_0(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}P[\widehat{\nu}_n > 2\nu]$$

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)1\!\!1_{I_{\{\widehat{\nu}_n<\nu/2\}}}] \leq (\nu + K^2 M_n)^2 P[\widehat{\nu}_n < \frac{1}{2}\nu]. \tag{34}$$

Proof of Claim 2

• Let us prove inequality (33). First of all, we notice that if $\widehat{\nu}_n \geq \nu/2$ then for every model m:

$$x_m \leq \max\{B_1 \ln(1+D_m), 2B_2 \ln(1+D_m)\frac{D_m}{n} \leq B_3 \ln(1+D_m)\}$$

with $B_3 = \max(B_1, 2B_2)$. Thus:

$$\begin{aligned} Crit(m)1\!\!1_{\{\widehat{\nu}_n\geq\nu/2\}} &\leq 2[\sup_{D_j\geq D_m}(g_j-g)^2(x_0)+(g_m-g)^2(x_0)] + x_m\widehat{\nu}_n\ln(1+D_m)\frac{D_m}{n} \\ &\leq 2C_0 D_m^{-2\beta} + B_3\ln(1+D_m)\widehat{\nu}_n\frac{D_m}{n} \\ &\leq C(1+\widehat{\nu}_n)[D_m^{-2\beta}+\ln(1+D_m)\frac{D_m}{n}] \end{aligned}$$

for some constant C depending on (A, K, β, C_0). Let us denote $F(m) = D_m^{-2\beta} + \ln(1+D_m)\frac{D_m}{n}$ and $m_1 = \arg\min F(m)$. Then:

$$F(m_1) \leq F((\frac{n}{\ln(1+n)})^{1/(2\beta+1)}) \leq 2(\frac{n}{\ln n})^{-2\beta)/(2\beta+1)}$$

Remark 1. *We give here an upper bound for* D_{m_1}*, which will be useful in the proof of Theorem* 2.1*. The model* m_1 *satisfies:*

$$\frac{2\beta}{D_{m_1}^{2\beta+1}} = \frac{1}{n}[(\frac{D_{m_1}}{1+D_{m_1}} + \ln(1+D_{m_1})]$$

Besides, the function $m \to (\frac{D_m}{1+D_m} + \ln(1+D_m)$ *is increasing so:*

$$\frac{2\beta}{D_{m_1}^{2\beta+1}} \geq \frac{\alpha_1}{n} \Rightarrow D_{m_1} \leq \alpha_1^{1/(2\beta+1)}\alpha_2 n^{-1/(2\beta+1)}$$

where α_1 *and* α_2 *are defined in* (8) *and* (9)*.*

Besides:

- If $D_{m_{opt}} \leq D_{m_1}$, then:

$$(\widehat{g}_{m_{opt}} - g)^2(x_0) \leq C_0 D_{m_{opt}}^{-2\beta} \leq C_0 D_{m_1}^{-2\beta} \leq F(m_1)$$

- If $D_{m_{opt}} > D_{m_1}$:

$$\begin{aligned}(g_{m_{opt}} - g)^2(x_0) &\leq 2[(g_{m_{opt}} - g_{m_1})^2(x_0) + (g_{m_1} - g)^2(x_0)] \\ &\leq 2 \sup_{j\in J_n, j\geq m_{opt}} (g_{m_{opt}} - g_j)^2(x_0) + C(\beta, L)D_{m_1}^{-(2\beta-1)} \\ &\leq Crit(m_{opt}) + F(m_1) \\ &\leq Crit(m_1) + F(m_1) \\ &\leq C(1+\widehat{\nu}_n)F(m_1)\end{aligned}$$

Hence in these two cases:

$$Crit(m_{opt}) + (g - g_{m_{opt}})^2(x_0) \leq C(1+\widehat{\nu}_n)F(m_1).$$

Thus according to inequality (25) in Claim 1 we have:

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq \max(\kappa, \kappa')(1+\widehat{\nu}_n)(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}$$

And by integrating this result over the sample Z_1 we get:

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}}] \leq C(1+\mathbb{E}[\widehat{\nu}_n])(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}. \tag{35}$$

Moreover we have proved in (36) that $\widehat{\nu}_n = \|\widehat{g}_{m_0}\|_\infty \leq K^2 p_0$. Thus

$$\mathbb{E}[\widehat{\nu}_n] = \mathbb{E}[\widehat{\nu}_n 1_{\{\widehat{\nu}_n \leq 2\nu\}}] + \mathbb{E}[\widehat{\nu}_n 1_{\{\widehat{\nu}_n > 2\nu\}}] \leq 2\nu + K^2 p_0 P[\widehat{\nu}_n > 2\nu]$$

By reporting this result in (35), we get inequality (33).

• Let us prove inequality (34). For every model $m \in J_n$, $(\widehat{g}_m - g)^2(x_0) \leq (|\widehat{g}_m(x_0)| + \nu)^2$. Besides:

$$\begin{aligned}(\widehat{g}_m)^2(x_0) &= \sum_{\lambda\in I_m} (\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i))\chi_\lambda(x_0))^2 \leq \frac{1}{n}\sum_{i=1}^n (\sum_{\lambda\in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0))^2 \leq \|\sum_{\lambda\in I_m} \chi_\lambda^2\|_\infty^2 \\ &\leq K^4 D_m^2 \qquad (36)\end{aligned}$$

Hence:

$$P[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\widehat{\nu}_n < \nu/2}] \leq (KM_n + \nu)^2 P[\widehat{\nu}_n < \frac{1}{2}\nu]$$

and inequality (22) in Proposition 3.1 ends the proof of (34). □

Theorem 3.1 results directly from Claim 2: $P[\widehat{\nu}_n > 2\nu]$ and $P[\widehat{\nu}_n < \nu/2]$ are upper bounded by Proposition (3.1):

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}}] \leq \kappa(1+2\nu)\frac{n}{\ln n})^{-2\beta/(2\beta+1)}$$
$$+\kappa' K^2 p_0 (\frac{n}{\ln n})^{-2\beta/(2\beta+1)})[\exp(-\frac{C_2}{K^2}\nu\frac{n}{p_0}) + \exp(-\frac{C_3}{K^4}\nu^2\frac{n^{3/2}}{p_0^2}) + \exp(-\frac{C_4}{K^2}\frac{n}{p_0^2})]$$

Then the combination of inequalities (37) and (34) ends the proof of Theorem 3.1 □

6.2. **Proof of Proposition** 3.1. • We prove inequality (22). Let $x_1 \in I$ be such that $g(x_1) \geq 5\nu/6$, then by definition of $\widehat{\nu}_n$:

$$\begin{aligned} P[\widehat{\nu}_n \leq \nu/2] &\leq P[\widehat{g}_{m_0}(x_1) \leq \nu/2] \\ &= P[(\widehat{g}_{m_0} - g_{m_0})(x_1) \leq 5\nu/6 - g_{m_0}(x_1) - \nu/3] \\ &\leq P[(\widehat{g}_{m_0} - g_{m_0})(x_1) \leq (g - g_{m_0})(x_1) - \nu/3] \end{aligned}$$

According to (11), we get:

$$P[\widehat{\nu}_n \leq \nu/2] \leq P[(\widehat{g}_{m_0} - g_{m_0})(x_1) \leq C_0 p_0^{-\beta} - \nu/3]$$

and with condition ($\mathbf{A_1}$):

$$\begin{aligned} P[\widehat{\nu}_n \leq \nu/2] &\leq P[(\widehat{g}_{m_0} - g_{m_0})(x_1) \leq -\nu/6] \\ &\leq P[|(\widehat{g}_{m_0} - g_{m_0})(x_1)| \geq \nu/6] \\ &= P[|\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}(U_i)| \geq \nu/6] \end{aligned} \tag{37}$$

where $U_i = \sum_{\lambda \in I_{m_0}} \chi_\lambda(V_i)\chi_\lambda(x_1)$. This term is upper bounded with Bernstein Inequality, with the following parameters:

$$\begin{aligned} \mathbb{E}[U_1^2] &= \mathbb{E}[(\sum_{\lambda \in I_{m_0}} \chi_\lambda(V_1)\chi_\lambda(x_1))^2] = \int_I (\sum_{\lambda \in I_{m_0}} \chi_\lambda(x)\chi_\lambda(x_1))^2 g(x)dx \\ &\leq \nu \sum_{\lambda,\lambda' \in I_{m_0}} [\int_I (\chi_\lambda(x)\chi_{\lambda'}(x)dx]\chi_\lambda(x_1)\chi_{\lambda'}(x_1)) = \nu \sum_{\lambda \in I_{m_0}} \chi_\lambda^2(x_1) \end{aligned}$$

as the $\{\chi_\lambda\}$ are orthonormal. Finally, Assumption (13) in $\mathbf{H_{dens}}(\beta)$ leads to:

$$\mathbb{E}[U_1^2] \leq \nu K^2 p_0 := v^2$$

Let l be an integer greater than 2, then:

$$\begin{aligned} \mathbb{E}[(X_1)_+^l] &\leq \mathbb{E}[U_1^2] \times \|U_1\|_\infty^{l-2} \\ &\leq v^2 \| \sum_{\lambda \in I_{m_0}} \chi_\lambda(V_1)\chi_\lambda(x_0)\|_\infty^{l-2} \\ &\leq v^2 [\sqrt{\| \sum_{\lambda \in I_{m_0}} \chi_\lambda^2(V_1)\|_\infty} \sqrt{\sum_{\lambda \in I_{p_0}} \chi_\lambda^2(x_0)}]^{l-2} \\ &\leq v^2 (K^2 p_0)^{l-2} \end{aligned}$$

thus $c = K^2 p_0$. Bernstein Inequality applied to (37) provides (22).

• We prove inequality (23). Let $\widehat{x_1}$ be such that $\widehat{g}_{m_0}(\widehat{x}_1) \geq 5\widehat{\nu}_n/6$, then:

$$\begin{aligned} P[\widehat{\nu}_n \geq 2\nu] &\leq P[\frac{6}{5}\widehat{g}_{m_0}(\widehat{x}_1) \geq 2\nu] \\ &= P[\frac{6}{5}(\widehat{g}_{m_0} - g_{m_0})(\widehat{x}_1) \geq \frac{4}{5}\nu + \frac{6}{5}(\nu - g_{m_0}(\widehat{x}_1))]. \end{aligned}$$

By definition of ν,

$$P[\widehat{\nu}_n \geq 2\nu] \leq P[\frac{6}{5}(\widehat{g}_{m_0} - g_{m_0})(\widehat{x}_1) \geq \frac{4}{5}\nu + \frac{6}{5}(g - g_{m_0})(\widehat{x}_1))].$$

According to (11) of $\mathbf{H_{dens}}(\beta)$:

$$P[\widehat{\nu}_n \geq 2\nu] \leq P[(\widehat{g}_{m_0} - g_{m_0})(\widehat{x}_1) \geq \frac{2}{3}\nu - C_0 p_0^{-\beta}]$$

and Assumption $(\mathbf{A_1})$ leads to:

$$P[\widehat{\nu}_n \geq 2\nu] \quad \leq \quad P[(\widehat{g}_{m_0} - g_{m_0})(\widehat{x}_1) \geq \frac{1}{2}\nu] \leq P[\sup_{x\in I}(\widehat{g}_{m_0} - g_{m_0})(x) \geq \frac{1}{2}\nu].$$

Let:

$$\begin{aligned} Z &= \sup_{x\in I}(\widehat{g}_{m_0} - g_{m_0})(x) \\ &= \sup_{x\in I} \frac{1}{n}\sum_{i=1}^{n}\{\sum_{\lambda\in I_{m_0}} (\chi_\lambda(V_i)\chi_\lambda(x) - \mathbb{E}[\chi_\lambda(V_i)\chi_\lambda(x)])\} \end{aligned}$$

We upper bound $P[Z \geq \nu/2]$ with Talagrand Inequality (see Theorem 7.2 in Section 7), but the set of functions:

$$\mathcal{F} = \{\varphi_x : u \to \sum_{\lambda\in I_{m_0}} \chi_\lambda(x)\chi_\lambda(u) - \mathbb{E}[\chi_\lambda(x)\chi_\lambda(V_1)], x \in I\}$$

is not countable. Nevertheless, for every u the application $x \to \varphi_x(u)$ is continuous (as the (χ_λ) are continuous), so :

$$Z = \sup_{x\in I} \frac{1}{n}\sum_{i=1}^{n}\varphi_x(V_i) = \sup_{x\in I\cap\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n}\varphi_x(V_i)$$

by density of $\mathbb{Q}\cap I$ in I, and $\mathbb{Q}\cap I$ is countable, which allows us to apply Talagrand Inequality. For every $x \in I$:

$$\begin{aligned} &\sum_{\lambda\in I_{m_0}} [\frac{1}{n}\sum_{i=1}^{n}(\chi_\lambda(V_i)\chi_\lambda(x) - \mathbb{E}[\chi_\lambda(V_i)\chi_\lambda(x)])]^2 \\ \leq \{\sum_{\lambda\in I_{m_0}} \chi_\lambda^2(x)\} &\times \{\sum_{\lambda\in I_{m_0}} [\frac{1}{n}\sum_{i=1}^{n}(\chi_\lambda(V_i) - \mathbb{E}[\chi_\lambda(V_i)])]^2\} \\ &\leq K^2 p_0 \sum_{\lambda\in I_{m_0}} [\frac{1}{n}\sum_{i=1}^{n}(\chi_\lambda(V_i) - \mathbb{E}[\chi_\lambda(V_i)])]^2 \end{aligned}$$

Thus:

$$\begin{aligned}\mathbb{E}[Z] &\leq K^2 p_0 \sum_{\lambda\in I_{m_0}} \mathbb{E}[(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i) - \mathbb{E}[\chi_\lambda(V_i)])^2] \\ &= K^2 p_0 \sum_{\lambda\in I_{m_0}} Var(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i)) \\ &= \frac{K^2 p_0}{n} \sum_{\lambda\in I_{m_0}} Var(\chi_\lambda(V_1)) \\ &\leq \frac{K^2 p_0}{n} \sum_{\lambda\in I_{m_0}} \mathbb{E}[\chi_\lambda^2(V_1)]\end{aligned}$$

Hence

$$\mathbb{E}[Z] \leq \frac{K^2 p_0}{n}$$

Let us compute the variance term v. For every $x \in I$:

$$\begin{aligned}Var(\sum_{\lambda\in I_{m_0}} \chi_\lambda(V_1)\chi_\lambda(x)) &\leq \mathbb{E}[(\sum_{\lambda\in I_{m_0}} \chi_\lambda(V_1)\chi_\lambda(x))^2] \\ &= \int_I (\sum_{\lambda\in I_{m_0}} \chi_\lambda(u)\chi_\lambda(x))^2 g(u)du \\ &\leq \nu \int_I (\sum_{\lambda\in I_{m_0}} \chi_\lambda(u)\chi_\lambda(x))^2 du \\ &= \nu \sum_{\lambda,\lambda'\in I_{m_0}} [\int_I \chi_\lambda(u)\chi_{\lambda'}(u)du]\chi_\lambda(x)\chi_{\lambda'}(x)\end{aligned}$$

As the family $\{\chi_\lambda, \lambda \in I_{m_0}\}$ is orthonormal:

$$Var(\sum_{\lambda\in I_{m_0}} \chi_\lambda(V_1)\chi_\lambda(x)) \leq \nu \sum_{\lambda\in I_{m_0}} \chi_\lambda^2(x) \leq \nu K^2 p_0 := v$$

Finally, for every $x \in I$:

$$\|\sum_{\lambda\in I_{m_0}} \chi_\lambda(x)\chi_\lambda\|_\infty \leq \sqrt{\sum_{\lambda\in I_{m_0}} \chi_\lambda^2(x)} \times \|\sqrt{\sum_{\lambda\in I_{m_0}} \chi_\lambda^2}\|_\infty \leq K^2 p_0 := b$$

Besides, by Assumption $(\mathbf{A_2})$ we have:

$$P[Z \geq \frac{\nu}{2}] = P[Z \geq \mathbb{H} + (\frac{\nu}{2} - \frac{K^2 p_0}{\sqrt{n}})] \leq P[Z \geq \mathbb{H} + \frac{\nu}{6}],$$

and Talagrand Inequality provides the following upper bound:

$$\begin{aligned} P[Z \geq \frac{\nu}{2}] &\leq \exp[-C\frac{n\nu^2}{\nu K^2 p_0 + K^4 p_0^2/\sqrt{n} + K^2 p_0^2 \nu^2}] \\ &\leq \exp(-\frac{C_2}{K^2}\nu\frac{n}{p_0}) + \exp(-\frac{C_3}{K^4}\nu^2\frac{n^{3/2}}{p_0^2}) + \exp(-\frac{C_4}{K^2}\frac{n}{p_0^2}) \quad \square \end{aligned}$$

6.3. **Proof of Theorem** 2.1**.** Let consider the decomposition (24). First of all, we study the first term $\mathbb{E}[(f-f^-)^2(x_0)]$ in the right hand side of (24). If Assumption $\mathbf{H_1}(\beta)$ holds, f is Lipschitz. It is easy to check that f is Lipschitz as well if Assumption $\mathbf{H_2}(\beta)$ holds for some $\beta \geq 1$. Let us denote L the Lipschitz constant of f.

$$\begin{aligned} (f-f^-)^2(x_0) &= (\int_0^1 [f(x_0) - f(x_0 - (b-\widehat{b})(x))]\mu(x)dx)^2 \\ &\leq \int_0^1 [f(x_0) - f(x_0 - (b-\widehat{b})(x))]^2\mu(x)dx \\ &\leq L\int_0^1 [(b-\widehat{b})(x)]^{2\beta-1}\mu(x)dx \\ &\leq L\int_0^1 [(b-\widehat{b})^2(x)\mu(x)dx \\ &\leq L\|\|b-\widehat{b}\|_\mu^2 \end{aligned}$$

So

$$\mathbb{E}[(f^- - f)^2(x_0)] \leq L\mathbb{E}[\|b-\widehat{b}\|_\mu^2] \tag{38}$$

To study the second term we need the following preliminary lemma.

Lemma 6.3. *1) Let Z^- be fixed. If f satisfies Assumption $\mathbf{H_1}(\beta)$ (resp. $\mathbf{H_2}(\beta)$), so does f^-.*

2) $\nu^- \leq \nu$ almost everywhere (a.e.).

3) For every $m \in \{1,\dots,M_n\}$:

$$(\widehat{f_m^-}(x_0) - f^-(x_0))^2 \leq (m+\nu)^2 \leq (M_n+\nu)^2$$

4) Let us consider a sequence (α_n) of positive number such that $\alpha_n = o(1/\sqrt{\ln n})$. Then for every $n \in \mathbb{N}$ such that:

$$\frac{2}{\sqrt{\ln n}} + \sigma^2\alpha_n^2 \ln n \leq \frac{1}{2}$$

where $\sigma^2 = \mathbb{E}[\epsilon_1^2]$, we have:

$$P[\nu^- \leq \alpha_n] \leq 2\ln n\alpha_n^2\mathbb{E}[\|\widehat{b}-b\|_\mu^2]$$

The proof of Lemma 6.3 is given at the end of Section 6.3.

According to 1) in Lemma 6.3, for every Z^-, f^- satisfies Assumption $\mathbf{H_{dens}}(\beta)$ for some $\beta \geq 1$. Thus, according to Remark 1 in the proof of Claim 2, the result of Claim 2 remains

true if we restrict ourselves to a maximum size of model $M_n = E(\alpha_3 n^{1/3})$. Indeed, let's go back to the proof of Claim 2. The maximum size M_n is involved when we state that:

$$Crit(m_{opt}) \leq Crit(m_1) \text{ where } m_{opt} = \arg\min_{m \in J_n} Crit(m)$$

And this holds as soon as $m_1 \in J_n$. Besides, we have proved that

$$D_{m_1} \leq \alpha_1^{1/(1\beta+1)} \alpha_2 n^{1/(2\beta+1)} \leq \alpha_1^{1/3} \alpha_2 n^{1/2} \leq M_n$$

So Claim 2 provides the following upper bound:

$$\mathbb{E}_1[(\widehat{f}^- - f^-)^2(x_0)] \leq \kappa(1+\nu)(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}$$

$$+\kappa' p_0 (\frac{n}{\ln n})^{-2\beta/(2\beta+1)} P_1[\widehat{\nu}_n^- > 2\nu^-] + (\nu^- + M_n)^2 P_1[\widehat{\nu}_n^- < \frac{1}{2}\nu^-] \tag{39}$$

Let us define the sets:

$$A_1^- := \{C_0 p_0^{-\beta} < \frac{1}{6}\nu^-\} \qquad A_2^- := \{\frac{p_0}{\sqrt{n}} \leq \frac{\nu^-}{3K^2}\}$$

Then:

$$\begin{aligned}
\mathbb{E}[P_1(\widehat{\nu}_n^- > 2\nu^-)] &\leq \mathbb{E}[P_1(\widehat{\nu}_n^- > 2\nu^-) 1_{A_1^- \cap A_2^-}] + P[(A_1^-)^c] + P[(A_2^-)^c] \\
&\leq \mathbb{E}[\exp(-\frac{C_2}{K^2}\nu^- \frac{n}{p_0} + \exp(-\frac{C_3}{K^2}(\nu^-)^2 \frac{n^{3/2}}{p_0^2})] + \exp(-\frac{C_4}{K^2}\frac{n}{p_0^2}) \\
&\quad + P[(A_1^-)^c] + P[(A_2^-)^c]
\end{aligned}$$

And:

$$\begin{aligned}
\mathbb{E}[P_1(\widehat{\nu}_n^- < \nu^-/2)] &\leq \mathbb{E}[P_1(\widehat{\nu}_n^- < nu^-/2) 1_{A_1^-}] + P[(A_1^-)^c] \\
&\leq 2\mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0})] + P[(A_1^-)^c]
\end{aligned}$$

Thus inequality (39) leads to:

$$\mathbb{E}_1[(\widehat{f}^- - f^-)^2(x_0)] \leq C(1+\nu)[(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}$$

$$+p_0(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}(\mathbb{E}[\exp(-\frac{C_2}{K^2}\nu^- \frac{n}{p_0}) + \exp(-\frac{C_3}{K^2}(\nu^-)^2\frac{n^{3/2}}{p_0^2})] + \tag{40}$$

$$\exp(-\frac{C_4}{K^2}\frac{n}{p_0^2}) + P[(A_1^-)^c] + P[(A_2^-)^c]) + M_n^2(P[(A_1^-)^c] + \mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0})])]$$

Now we upper bound each term in the right side of the above expression.

$$\begin{aligned}
\mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0})] &= \mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0}) 1_{\{\nu^- \geq 2\ln n \frac{K^2 p_0}{C_2 n}\}}] + \mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0}) 1_{\{\nu^- < 2\ln n \frac{K^2 p_0}{C_2 n}\}}] \\
&\leq \mathbb{E}[\frac{1}{n^2} 1_{\{\nu^- \geq 2\ln n \frac{K^2 p_0}{C_2 n}\}}] + \mathbb{E}[1_{\{\nu^- < 2\ln n \frac{K^2 p_0}{C_2 n}\}}] \\
&\leq \frac{1}{n^2} + P[\nu^- < 2\ln n \frac{K^2 p_0}{C_2 n}]
\end{aligned}$$

Besides, as $p_0 = n^\gamma$ with $\gamma \in]1/3, 1/2[$, there exists N_1 which depends on γ, K^2 and σ^2 such that for every $n \geq N_1$, the following inequality holds:

$$\frac{2}{\sqrt{\ln n}} + 2\sigma^2 \ln n \frac{K^2 p_0}{C_2 n} \leq \frac{1}{2}$$

Then, according to 4) from Lemma 6.3:

$$P[\nu^- < 2\ln n \frac{K^2 p_0}{C_2 n}] \leq 4\ln^3 n (\frac{K^2 p_0}{C_2 n})^2 \mathbb{E}[\|\widehat{b} - b\|_\mu^2]$$

Thus:

$$\mathbb{E}[\exp(-\frac{C_1}{K^2}\nu^- \frac{n}{p_0})] \leq \frac{1}{n^2} + C\ln^3 n \frac{p_0^2}{n^2}\mathbb{E}[\|\widehat{b} - b\|_\mu^2] \tag{41}$$

Similarly, there exists an integer N_3 which depends on γ, K^2 and σ^2 such that, for every n greater than N_3:

$$\mathbb{E}[\exp(-\frac{C_3}{K^2}(\nu^-)^2 \frac{n^{3/2}}{p_0^2})] \leq \frac{1}{n^2} + C\ln^2 n \frac{p_0^2}{n^{3/2}}\mathbb{E}[\|\widehat{b} - b\|_\mu^2] \tag{42}$$

And:

$$P[(A_1^-)^c] \leq C\ln n \frac{1}{p_0^{2\beta}}\mathbb{E}[\|\widehat{b} - b\|_\mu^2]$$

$$P[(A_2^-)^c] \leq C\ln n \frac{p_0^2}{n}\mathbb{E}[\|\widehat{b} - b\|_\mu^2] \tag{43}$$

The combination of inequalities (40), (41), (42) and (43) leads to:

$$\begin{aligned}\mathbb{E}[(\widehat{f^-} - f^-)^2(x_0)] \quad &\leq \quad C(1+\nu)[(\frac{n}{\ln n})^{-2\beta/(2\beta+1)} \\ &+\mathbb{E}[\|\widehat{b} - b\|_\mu^2]\{p_0(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}[\ln^3 n\frac{p_0^2}{n^2} + \ln^2 n\frac{p_0^2}{n^{3/2}} + \ln n\frac{1}{p_0^{2\beta}} + \ln n\frac{p_0^2}{n}] \\ &\qquad +M_n^2[\ln^3 n\frac{p_0^2}{n^2} + \ln n\frac{1}{p_0^{2\beta}}]\} + p_0(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}\exp(-\frac{C_4}{K^2}\frac{n}{p_0^2})\end{aligned}$$

Besides, we suppose that $M_n \leq \alpha_3 n^{1/3}$ and $\beta \geq 1$ which entails that

$$(\frac{n}{\ln n})^{-2\beta/(2\beta+1)} \leq C(\frac{\ln n}{n})^{2/3}$$

Moreover, $n/p_0^2 = n^{1-2\gamma}$ and $1 - 2\gamma > 0$ then:

$$p_0(\frac{n}{\ln n})^{-2\beta/(2\beta+1)}\exp(-\frac{C_4}{K^2}\frac{n}{p_0^2}) \leq \frac{C'}{n}$$

Hence

$$\begin{aligned}\mathbb{E}[(\widehat{f^-} - f^-)^2(x_0)] \quad &\leq \quad C(1+\nu)[(\frac{n}{\ln n})^{-2\beta/(2\beta+1)} \\ &+\mathbb{E}[\|\widehat{b} - b\|_\mu^2] \times \ln^4 n\{\frac{p_0^3}{n^{2+2/3}} + \frac{p_0^3}{n^{3/2+2/3}} + \frac{1}{p_0 n^{2/3}} + \frac{p_0^3}{n^{1+2/3}} + \frac{p_0^2}{n^{2-2/3}} + \frac{n^{2/3}}{p_0^2}\}\end{aligned}$$

We have chosen p_0 such that $n^\gamma \leq p_0 \leq 2n^\gamma$ with γ in $]1/3, 1/2[$, which entails that the quantity:

$$\ln^4 n\{\frac{p_0^3}{n^{2+2/3}}+\frac{p_0^3}{n^{3/2+2/3}}+\frac{1}{p_0 n^{2/3}}+\frac{p_0^3}{n^{1+2/3}}+\frac{p_0^2}{n^{2-2/3}}+\frac{n^{2/3}}{p_0^2}\}$$

is upper bounded by a constant. So:

$$\mathbb{E}[(\widehat{f}^- - f^-)^2(x_0)] \leq C(1+\nu)(\frac{n}{\ln n})^{-2\beta/(2\beta+1)} + C'(1+\nu)\mathbb{E}[\|\widehat{b}-b\|_\mu^2] \tag{44}$$

Inequalities (38) and (44) conclude the proof of Theorem 2.1 □

Proof of Lemma 6.3**:**

1) • Suppose that $\mathbf{H_1}(\beta)$ holds. Let $u \in \mathbb{R}$ and $m \in \mathbb{N}^*$.

$$\begin{aligned}
(f^- - (f^-)_m)(u) &= \int_{x=0}^1 f(u-(b-\widehat{b})(x)\mu(x)dx - \sum_{k\in\mathbb{Z}}\langle f^-,\phi_{m,k}\rangle\phi_{m,k}(u) \\
&= \int_{x=0}^1 f(u-(b-\widehat{b})(x)\mu(x)dx \\
&\quad -\sum_{k\in\mathbb{Z}}[\int_{t\in\mathbb{R}}(\int_{x=0}^1 f(t-(b-\widehat{b})(x)\mu(x)dx)\phi_{m,k}(t)dt]\phi_{m,k}(u) \\
&= \int_{x=0}^1 [f(u-(b-\widehat{b})(x)) \\
&\quad -\sum_{k\in\mathbb{Z}}\int_{t\in\mathbb{R}} f(t-(b-\widehat{b})(x))\phi_{m,k}(t)dt\phi_{m,k}(u)]\mu(x)dx
\end{aligned}$$

Let x be fixed in $[0,1]$ and $f^x(u) := f(u-(b-\widehat{b})(x))$, then:

$$f(u-(b-\widehat{b})(x)) - \sum_{k\in\mathbb{Z}}\int_{t\in\mathbb{R}} f(t-(b-\widehat{b})(x))\phi_{m,k}(t)dt\phi_{m,k}(u) = f^x(u) - (f^x)_m(u)$$

Besides, according to 5) in Proposition 7.1:

$$\begin{aligned}
(f^x - (f^x)_m)(u) &= \frac{1}{2\pi}\int_{|\theta|>\pi m}(f^x)^*(\theta)e^{i\theta u}d\theta \\
&= \frac{1}{2\pi}\int_{|\theta|>\pi m} f^*(\theta)e^{i\theta(b-\widehat{b}(x))}e^{i\theta u}d\theta \\
&= (f-f_m)(u+(b-\widehat{b})(x))
\end{aligned}$$

So for every $u \in \mathbb{R}$ $|(f^x - (f^x)_m)(u)| \leq \|f-f_m\|_\infty$, hence:

$$|(f^- - (f^-)_m)(u)| \leq \int_{x=0}^1 \|f-f_m\|_\infty \mu(x)dx \leq \|f-f_m\|_\infty \leq C_0 D_m^{-\beta}$$

• Let $f \in \mathcal{H}(\beta, L)$ and r the greater integer smaller than β. f is r times differentiable and its r-th derivative is upper bounded, then with classical analysis results we get:

$$\begin{aligned}(f^-)^{(r)}(t) &= \frac{\partial^r}{\partial t}\int_0^1 f(t-(b-\widehat{b})(x))\mu(x)dx \\ &= \int_0^1 \frac{\partial^r}{\partial t} f(t-(b-\widehat{b})(x))\mu(x)dx \\ &= \int_0^1 f^{(r)}(t-(b-\widehat{b})(x))\mu(x)dx\end{aligned}$$

For every $(t,u) \in [-1,1]^2$:

$$\begin{aligned}|(f^-)^{(r)}(t) - (f^-)^{(r)}(u)| &= |\int_0^1 [f^{(r)}(t-(b-\widehat{b})(x)) - f^{(r)}(u-(b-\widehat{b})(x))]\mu(x)dx \\ &\leq \int_0^1 |f^{(r)}(t-(b-\widehat{b})(x)) - f^{(r)}(u-(b-\widehat{b})(x))|\mu(x)dx \\ &\leq \int_0^1 L|t-u|^{\beta-r}\mu(x)dx \\ &= L|t-u|^{\beta-r}\end{aligned}$$

so $f^- \in \mathcal{H}(\beta, L)$.

2) Let $t \in \mathbb{R}$, according to the expression of f^-:

$$|f^-(t)| \leq \int_0^1 |f(t-(b-\widehat{b})(x))|\mu(x)dx \leq \nu$$

thus $\nu^- := \|f^-\|_\infty \leq \nu$.

3) A calculus similar to (36) implies that $|f_m^-(x_0)| \leq m$, which proves the third point.

4) Given Z^-, ϵ_1 and $(b-\widehat{b})(X_1)$ are independent, which entails:

$$\mathbb{E}[\widehat{\epsilon}_1^2|Z^-] = \mathbb{E}[\epsilon_1^2|Z^-] + \mathbb{E}[(b-\widehat{b})^2(X_1)|Z^-] + 2\mathbb{E}[\epsilon_1(b-\widehat{b})(X_1)|Z^-]$$

Moreover, $\mathbb{E}[\epsilon_1|Z^-] = 0$ hence:

$$\mathbb{E}[\widehat{\epsilon}_1^2|Z^-] = \sigma^2 + \|b-\widehat{b}\|_\mu^2$$

Thus for every $A_n > 0$:

$$\int_{|y|>A_n} f^-(y)dy \leq \frac{1}{A_n^2}\int_{|y|>A_n} y^2 f^-(y)dy \leq \frac{1}{A_n^2}(\sigma^2 + \|b-\widehat{b}\|_\mu^2)$$

which entails:

$$\int_{|y|\leq A_n} f^-(y)dy \geq 1 - \frac{\sigma^2 + \|b-\widehat{b}\|_\mu^2}{A_n^2}$$

On the other hand, $\int_{|y|\leq A_n} f^-(y)dy \leq 2\nu^- A_n$, by definition of ν^-. Hence:

$$\nu^- \geq \frac{1}{2A_n}(1 - \frac{\sigma^2 + \|b-\widehat{b}\|_\mu^2}{A_n^2})$$

for every $A_n > 0$. Thus:

$$\begin{aligned} P[\nu^- \leq \alpha_n] &\leq P[1 - \frac{\sigma^2 + \|b - \widehat{b}\|_\mu^2}{A_n^2} \leq 2A_n\alpha_n] \\ &= P[1 - (2A_n\alpha_n + \frac{\sigma^2}{A_n^2}) \leq \frac{\|b - \widehat{b}\|_\mu^2}{A_n^2}] \end{aligned}$$

Let us consider $A_n = 1/(\alpha_n\sqrt{\ln n})$, then condition (C) gives:

$$\begin{aligned} P[\nu^- \leq \alpha_n] &\leq P[1 - (\frac{2}{\sqrt{\ln n}} + \sigma^2\alpha_n^2 \ln n) \leq \|b - \widehat{b}\|_\mu^2 \ln n\alpha_n^2] \\ &\leq P[\frac{1}{2} \leq \|b - \widehat{b}\|_\mu^2 \ln n\alpha_n^2] \\ &\leq 2 \ln n\alpha_n^2 \mathbb{E}[\|b - \widehat{b}\|_\mu^2] \end{aligned}$$

□

7. Appendix

7.1. **Deviation inequalities for empirical processes.**

Theorem 7.1. *Let $(X_1, \ldots, X_n)$ be independent random variables. Let us suppose that:*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v, \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i)_+^l] \leq \frac{l!}{2} \times v \times c^{l-2}$$

for every $l \geq 2$. Let $S = \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i]$.

1) For every $\epsilon > 0$:

$$P[S \geq \sqrt{2vx} + cx] \leq \exp(-nx), \;\; P[|S| \geq \sqrt{2vx} + cx] \leq 2\exp(-nx).$$

2) Similarly, for every $\epsilon > 0$:

$$P[S \geq \epsilon] \leq \exp(-\frac{n\epsilon^2}{2(v^2 + c\epsilon)}), \quad P[|S| \geq \epsilon] \leq 2\exp(-\frac{n\epsilon^2}{2(v^2 + c\epsilon)}).$$

Theorem 7.2. *Let $(X_1, \ldots, X_n)$ be i.i.d., $\mathcal{F}$ a class of function and:*

$$Z = \sup_{t\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (t(X_i) - \mathbb{E}[t(X_i)])$$

Let us consider $\mathbb{H}$, v and b such that:

$$\mathbb{E}[|Z|] \leq \mathbb{H}, \;\; \sup_{t\in\mathcal{F}} Var(t(X_i)) \leq v, \;\; \sup_{t\in\mathcal{F}} \|t\|_\infty \leq b.$$

Then for every $\lambda > 0$:

$$P[|Z| > \mathbb{H} + \lambda] \leq \exp(-\frac{n\lambda^2}{2(v + 4\mathbb{H}b + 3b\lambda)})$$

This Theorem results directly from Theorem 1.1 in T. Klein (2005).

7.2. **Some results about projection on sine-cardinal bases.** Let m be a positive number, A_m the set of functions generated by $\{\phi_{m,k}, k \in \mathbb{Z}\}$ described in Section 2.2. The following Proposition holds:

Proposition 7.1. *1) For every $m > 0$, $k \in \mathbb{Z}$, the Fourier transform of $\phi_{m,k}$ is $\phi^*_{m,k}(t) = (1/\sqrt{m})e^{-ikt/m}1_{[-\pi m,\pi m]}(t)$*

2) The family $\{\phi_{m,k}, k \in \mathbb{Z}\}$ is orthonormal for the L^2-norm.

3) For every $m > 0$, $\| \sum_{k\in\mathbb{Z}} \phi^2_{m,k}\|_\infty \leq m$

4) $A_m = \{t \in L^2(\mathbb{R}), Supp(t^*) \subset [-\pi m, \pi m]\} = \text{span}(\phi_{m,k}, k \in \mathbb{Z})$.

5) For every $h \in L^2(\mathbb{R})$, the Fourier transform of the projection h_m of h on A_m is $h^*_m(t) = h^*(t)1_{[-\pi m,\pi m]}(t)$.

A simple calculus proves that the Fourier transform of $1_{[-\pi,\pi]}$ is $2\pi\phi$, then $\phi^* = 1_{[-\pi,\pi]}$ and 1) follows by a change of variable. Next, for every $k, l \in \mathbb{Z}$, according to the Parseval formula, we have:

$$\langle \phi_{m,k}, \phi_{m,l}\rangle = \frac{1}{2\pi}\langle \phi^*_{m,k}, \phi_{m,l}*\rangle$$

and 2) follows easily from 1). With inverse Fourier formula,

$$\phi_{m,k}(x) = (\sqrt{m}/2\pi) \int_{-\pi}^{\pi} e^{-iku}e^{iuxm}du,$$

so that $\sum_{k\in\mathbb{Z}} \phi^2_{m,k}(x) = (m/2\pi)\int_{-\pi}^{\pi} |e^{iuxm}|^2du = m$. This gives 3). Assertion 4) follows from Meyer (1990, p.22), and 5) is an immediate consequence of 4). Indeed, for every $h \in L^2(\mathbb{R})$:

$$h_m = \arg\min_{t\in A_m} \|h - t\|^2 = \arg \min_{Supp(t^*)\subset[-\pi m,\pi m]} \|h^* - t^*\| = h^*1_{[-\pi m,\pi m]}$$

Proof of Proposition 2.2

• Let $f \in W(\alpha, L)$, and $x \in \mathbb{R}$:

$$\begin{aligned}
(f - f_m)^2(x) &= [\int_{\mathbb{R}} (f^* - f^*_m)(t)e^{itx}dt]^2 = [\int_{|t|>\pi m} f^*(t)e^{itx}dt]^2 \\
&\leq \int_{|t|>\pi m} |f^*(t)|^2t^{2\alpha}dt \times \int_{|t|>\pi m} \frac{1}{t^{2\alpha}}dt \\
&\leq L^2 \times \frac{1}{(2\alpha - 1)(\pi m)^{2\alpha-1}} = \frac{C_0(L,\alpha)}{m^{2\alpha-1}} \quad \square
\end{aligned}$$

• Suppose that $\alpha > 3/2$, then for every $x, y \in \mathbb{R}$:

$$\begin{aligned}
|f(x) - f(y)| &= |\frac{1}{2\pi}\int_{\mathbb{R}} f^*(t)(e^{itx} - e^{ity})dt| \\
&= |\frac{1}{2\pi}\int_{\mathbb{R}} f^*(t)e^{it(x+y)/2}2i\sin(\frac{t(x-y)}{2})dt| \\
&\leq \frac{1}{2\pi}\int_{\mathbb{R}} |f^*(t)|2|\sin(\frac{t(x-y)}{2})|dt \\
&\leq \frac{1}{2\pi}\int_{\mathbb{R}} |f^*(t)||t(x-y)|dt
\end{aligned}$$

f is a density so for every $t \in \mathbb{R}$, $|f^*(t)| \leq 1$, thus:

$$\int_{-1}^{1} |f^*(t)||t|dt \leq 2$$

Besides, with Schwarz Inequality, we have:

$$\begin{aligned} \int_{|t|>1} |f^*(t)||t|dt &\leq \int_{|t|>1} |f^*(t)||t|^{2\alpha}dt \times \int_{|t|>1} |t|^{2(1-\alpha)}dt \\ &\leq L^2 C \end{aligned}$$

where C is a constant depending on α. Thus:

$$|f(x) - f(y)| \leq |x - y| \frac{2 + L^2 C}{2\pi}$$

which proves that f is Lipschitz. □

Acknowledgments. The author is grateful to Fabienne Comte for her helpful and constructive advices, and to Yves Rozenholc for his help with matlab programming.

References

M. G. Akritas and I. Van Keilegom. Non parametric estimation of the residuals distribution. *Scan. J. Statis*, 28(3):549–567, 2001.

Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Field*, 117(6):467–493, 2000.

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

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

C. Butucea. Exact adaptive pointwise estimation on sobolev classes of densities. *ESAIM Probab. Statist*, 5:1–31, 2001.

F. Comte, J. Dedecker, and M.L. Taupin. Adaptive density deconvolution with dependent inputs. *Math. Methods Statist.*, 17(2):87–112, 2008.

S. Efromovich. Estimation of the density of regression errors, 2005.

S. Kiwitt, E-R. Nagel, and N. Neumeyer. Empirical likelihood for the error distribution in nonparametric regression models. *Math. Methods Statist.*, 34:511–534, 2008.

C. Laurent, C. Ludena, and C. Prieur. Adaptive estimation of linear functionals by model selection. *Electron. J. Stat.*, 2:993–1020, 2008.

O.V. Lepski and V.G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Satist.*, 25(6):2512–2546, 1997.

P. Massart. *Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23*, chapter Concentration inequalities and model selection. Lecture Notes in Mathematics. Springer, Berlin, 2007.

Y. Meyer. *Ondelettes et operateurs*. Hermann, Paris, 1990.

S. Plancade. Non parametric estimation of the density of the regression noise. *C. R. Acad. Sci. Paris*, 346(7-8):461–466, 2008.

C.J. Stone. *Optimal rates of convergence for nonparametric estimators*, volume 8. Springer-Verlag, Berlin, 1980.

E. Rio T. Klein. Concentration around the mean of empirical processes. *Ann. Probab.*, 33 (3):1060–1077, 2005.

A. B. Tsybakov. *Introduction l'estimation non paramtrique*. Springer-Verlag, Berlin, 2004.

S. Plancade, MAP 5 UMR 8145
Universit Paris Descartes, 45 rue des Saints Pères 75006 Paris, France.
email: sandra.plancade@math-info.univ-paris5.fr