ESAIM: PS July 2008, Vol. 12, p. 438–463
DOI: 10.1051/ps:2007050

ESAIM: Probability and Statistics
www.esaim-ps.org

DENSITY ESTIMATION WITH QUADRATIC LOSS: A CONFIDENCE INTERVALS METHOD

Pierre Alquier[1, 2]

Abstract. We propose a feature selection method for density estimation with quadratic loss. This method relies on the study of unidimensional approximation models and on the definition of confidence regions for the density thanks to these models. It is quite general and includes cases of interest like detection of relevant wavelets coefficients or selection of support vectors in SVM. In the general case, we prove that every selected feature actually improves the performance of the estimator. In the case where features are defined by wavelets, we prove that this method is adaptative near minimax (up to a log term) in some Besov spaces. We end the paper by simulations indicating that it must be possible to extend the adaptation result to other features.

Mathematics Subject Classification. 62G07, 62G15, 62G20, 68T05, 68Q32.

Received 10 October 2007.

1. Introduction: the density estimation setting

1.1. Notations

Let us assume that we are given a measure space $(\mathcal{X}, \mathcal{B}, \lambda)$ where λ is positive and σ-finite, and a probability measure P on $(\mathcal{X}, \mathcal{B})$ such that P has a density with respect to λ:

$$P(\mathrm{d}x) = f(x)\lambda(\mathrm{d}x).$$

We assume that we observe a realization of the canonical process $(X_1, ..., X_N)$ on $(\mathcal{X}^N, \mathcal{B}^{\otimes N}, P^{\otimes N})$. Our objective here is to estimate f on the basis of the observations $X_1, ..., X_N$.

More precisely, let $\mathcal{L}^2(\mathcal{X}, \lambda)$ denote the Hilbert space of all measurable functions from $(\mathcal{X}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B}_R)$ where $\mathcal{B}_R$ is the Borel σ-algebra on $\mathbb{R}$. We will write $\mathcal{L}^2(\mathcal{X}, \lambda) = \mathcal{L}^2$ for short. In the whole paper we will assume that $f \in \mathcal{L}^2$. Let us put, for any $(g, h) \in \left(\mathcal{L}^2\right)^2$:

$$\mathrm{d}^2(g, h) = \int_{\mathcal{X}} \Big(g(x) - h(x)\Big)^2 \lambda(\mathrm{d}x),$$

Keywords and phrases. Density estimation, support vector machines, kernel algorithms, thresholding methods, wavelets.

[1] Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, France; alquier@ensae.fr

[2] Laboratoire de Statistique, CREST 3, avenue Pierre Larousse, 92240 Malakoff, France.

Article published by EDP Sciences

and let $\|.\|$ and $\langle .,. \rangle$ denote the corresponding norm and scalar product. We are here looking for an estimator $\hat{f}$ that tries to minimize our objective:

$$\mathrm{d}^2(\hat{f}, f).$$

We assume that the statistician chooses a dictionary of functions $(f_1, ..., f_m) \in \left(\mathcal{L}^2\right)^m$, with $m \in \mathbb{N}^*$. There is no particular assumptions about this family: it is not necessarily linearly independent for example. In a first time, we assume that these functions are not data dependant, but we will see later in the paper how to include the case where $f_k(.) = K(X_k, .)$ for some kernel K for example.

1.2. Objective

Density estimation under quadratic loss is a classical problem in statistics and a lot of work has been done, we refer the reader to the general introduction by Tsybakov [23] and the references therein for example. There is a wide range of applications, among them let us mention multiclass pattern recognition (by the estimation of the density of every class and then classification of a pattern by likelihood maximization), or image segmentation, see Zhang *et al.* [25] for example.

The objective here is to provide a practical algorithm to select and aggregate the functions f_k that are relevant to perform density estimation.

In the case of a wavelet basis, such algorithms are known and are based on coefficient thresholding, see Härdle *et al.* [15] and the references therein for an introduction. Under suitable hypotheses, these estimators are able to reach the minimax rate of convergence up to a $\log N$ factor on some spaces of function with an unknown regularity β, as in Tsybakov [23] or Donoho *et al.* [13] for some particular Besov spaces. We show that in this case our algorithm produces a soft-thresholded estimator that reaches the same rate of convergence.

We focus also particularly on the case of kernel methods and support vector machines (SVM). SVM are a class of learning algorithm introduced by Boser *et al.* in the case of classification [5]. They were later generalized by Vapnik [24] to regression, and density estimation of a real-valued random variables with the Kolmogorov-Smirnov distance as a loss function:

$$\mathrm{d}_{KS}^2(g, h) = \sup_{x \in \mathcal{X}} \left| \int_{-\infty}^{x} g(t)\mathrm{d}t - \int_{-\infty}^{x} h(t)\mathrm{d}t \right|.$$

Note that a lot of variants of SVM were introduced in order to modify the set of support vectors (of basis kernel functions used in the estimation of the function). For example Tipping [22] introduced Relevance Vector Machine: in this variant of SVM, the support vectors are meant to be close to the center of clusters of data. Blanchard *et al.* [4] proposed to perform a principal component analysis on the space induced by the kernel. Here, we generalize the definition of SVM for density estimation to the quadratic loss and propose a method that justifies the use of a wide range of heuristics to select the set of support vectors. The choice of a kernel is also of interest for practitionners. For example, the Gaussian kernel is very often used, but the choice of its parameter remains a problem. Algorithms using several kernels (for example the Gaussian kernel with different values for the parameter) were proposed, see for example Ratsch *et al.* [19], often without theoretical justifications. Our method allows the use of multiple kernels.

The guarantee obtained here is that every selected feature actually improves the performance of the estimator: the quadratic distance to f decreases. Moreover the estimator is sparse, that means that often only a few of the functions f_k are actually selected. From this point of view the method can be seen as an implementation of Rissanen's MDL [20].

1.3. Organization of the paper

The method is an adaptation to the case of density estimation of the method we proposed in [2] for regression estimation. In a first time, we are going to study estimators of f in every unidimensional approximation model $\{\alpha f_k(.), \alpha \in \mathbb{R}\}$. Note that these models are too small and the obtained estimators do not have good properties in general. But they are used to obtained, by a PAC bound, confidence regions on f that have a very simple

geometry. We then propose an iterative method that selects and aggregate such estimators in order to build a suitable estimator of f (Sect. 2). For the sake of simplicity, we describe the method in this section for a family (f_k) that is not allowed to be data-dependant.

In Section 3 we make several comments on this algorithm in the case of some classical dictionaries $(f_1, ..., f_m)$.

In Section 4 we focus more particularly on the statistical point of view: we study the rate of convergence of the obtained estimator in the case of a basis of wavelets.

Section 5 is devoted to technical improvements and generalizations of the method. Improvements consists in more accurate PAC bounds leading to tighter confidence regions. Generalizations consists in including the case where the basis functions (f_k) are allowed to be data-dependant.

In Section 6 we provide some simulations in order to compare the practical performances of our estimator with the density estimators described in [13].

Finally, Section 7 is dedicated to the proofs of the various results given in the whole paper.

2. Estimation method

2.1. The dictionary of functions

We choose a family of functions $(f_1, ..., f_m) \in (\mathcal{L}^2)^m$ for some $m \in \mathbb{N}$. For the sake of simplicity we assume that the functions f_k are chosen such that $\|f_k\| = 1$.

In order to be able to justify our method, some requirements are necessary about the family $(f_k)_k$, depending on the available information on f (additional to the fact that $f \in \mathcal{L}^2$).

In the most general case, there is no additional information about f and we shall require that every f_k is a bounded function (by a constant C_k).

However, if we make some assumption about f, we can weaken the condition about the family $(f_k)_k$. We give the different cases in the following array.

Information on f	**Requirement on the dictionary**
no hypothesis	for any $k \in \{1, ..., m\}$ we have $\|f_k\|_{+\infty} \leq C_k < +\infty$
there are some known $q \in]1, +\infty[$ and $c > 0$ such that: $\left(\int_{\mathcal{X}} f^q \lambda(\mathrm{d}x)\right)^{\frac{1}{q}} \leq c$ notation: $\mathcal{H}(q)$	there are some $(c_1, ..., c_m) \in \mathbb{R}^m$ such that for any $k \in \{1, ..., m\}$ $\left(\int_{\mathcal{X}} \lvert f_k\rvert^{2p} \lambda(\mathrm{d}x)\right)^{\frac{1}{p}} \leq c_k^2 < +\infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. We put $C_k = c_k\sqrt{c}$.
there is some known $c > 0$ such that $\|f\|_{+\infty} \leq c$ notation: $\mathcal{H}(+\infty)$	no requirement necessary we put $C_k = \sqrt{c}$

2.2. Unidimensional models

Let us choose $k \in \{1, ..., m\}$ and consider the unidimensional model $\mathcal{M}_k = \{\alpha f_k(.), \alpha \in \mathbb{R}\}$. Remark that the orthogonal projection (denoted by $\Pi_{\mathcal{M}_k}$) of f on $\mathcal{M}_k$ is known, it is namely:

$$\Pi_{\mathcal{M}_k} f(.) = \overline{\alpha}_k f_k(.)$$

where:

$$\overline{\alpha}_k = \arg\min_{\alpha \in \mathbb{R}} \mathrm{d}^2(\alpha f_k, f) = \int_{\mathcal{X}} f_k(x) f(x) \lambda(\mathrm{d}x).$$

A natural estimator of this coefficient is:

$$\hat{\alpha}_k = \frac{1}{N}\sum_{i=1}^{N} f_k(X_i).$$

Actually, we can give a non-asymptotic control on the error of this estimator.

Theorem 2.1. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, \quad (\hat{\alpha}_k - \overline{\alpha}_k)^2 \leq \beta(\varepsilon, k)\right\} \geq 1-\varepsilon$$

where

$$\beta(\varepsilon, k) = \frac{4\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\left[\frac{1}{N}\sum_{i=1}^{N} f_k(X_i)^2 + C_k^2\right].$$

The proof is given in Section 7, more precisely in Section 7.1 page 456. Let us notice that this theorem does not require any condition on the true density function f.

2.3. The selection algorithm

What follows is based on the following remark:

$$(\hat{\alpha}_k - \overline{\alpha}_k)^2 = \mathrm{d}^2\left(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k\right).$$

Let us put:

$$\mathcal{CR}_{k,\varepsilon} = \left\{g \in \mathcal{L}^2, \quad \mathrm{d}^2(\hat{\alpha}_k f_k, \Pi_{\mathcal{M}_k} g) \leq \beta(\varepsilon, k)\right\}.$$

Then Theorem 2.1 implies the Corollary 2.2.

Corollary 2.2. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, \quad f \in \mathcal{CR}_{k,\varepsilon}\right\} \geq 1-\varepsilon.$$

So for any k, $\mathcal{CR}_{k,\varepsilon}$ is a confidence region at level k for f. Moreover, $\mathcal{CR}_{k,\varepsilon}$ being convex we have Corollary 2.3.

Corollary 2.3. *For any $\varepsilon > 0$ we have:*

$$P^{\otimes N}\left\{\forall k \in \{1,...,m\}, \forall g \in \mathcal{L}^2, \quad \mathrm{d}^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, f) \leq \mathrm{d}^2(g,f) - \mathrm{d}^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, g)\right\} \geq 1-\varepsilon.$$

It just means that for any g, $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$ is a better estimator than g. We indicate in Proposition 2.4 how to compute this projection.

Proposition 2.4. *We have*

$$\Pi_{\mathcal{CR}_{k,\varepsilon}} g = g + b f_k$$

where

$$b = \hat{\alpha}_k - \langle g, f_k\rangle - \mathrm{sgn}\left(\hat{\alpha}_k - \langle g, f_k\rangle\right)\sqrt{\beta(\varepsilon,k)}$$

where sgn is the sign function given by $sgn(x) = \mathbb{1}_{\mathbb{R}_+}(x) - \mathbb{1}_{\mathbb{R}_-}(x)$. *This also implies that*

$$\mathrm{d}^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g, g) = b^2.$$

The proof is given in Section 7, more precisely in Section 7.2 page 459.

So we propose the following algorithm:

- we choose $\varepsilon > 0$, a constant $\kappa > 0$ and start with $g_0 = 0$ (the values κ and ε are discussed in Rem. 2.2);
- at each step n, we choose an indice $k(n)$ using any heuristic we want, it is of course allowed to be data-dependant, then we take:

$$g_{n+1} = \Pi_{\mathcal{CR}_{k(n),\varepsilon}} g_n,$$

  the choice of $k(n)$ is discussed in Remark 2.1;
- we choose a stopping time

$$n_s = \inf\left\{n \in \mathbb{N}, \quad \mathrm{d}^2(g_n, g_{n+1}) \leq \kappa\right\}$$

  and put $\hat{f} = g_{n_s}$.

Corollary 2.3 implies that:

$$P^{\otimes N}\left\{\mathrm{d}^2(\hat{f}, f) \leq \mathrm{d}^2(0, f) - \sum_{n=0}^{n_s-1} \mathrm{d}^2(g_n, g_{n+1})\right\} \geq 1 - \varepsilon.$$

Remark 2.1. This result suggests to choose the sequence $k(n)$ in order to maximize

$$\sum_{n=0}^{n_s-1} \mathrm{d}^2(g_n, g_{n+1}).$$

However, the maximization over the whole sequence can be computationally intensive.

So we suggest, at each step, to take for $k(n)$ the direction k such that the improvement $\mathrm{d}^2(\Pi_{\mathcal{CR}_{k,\varepsilon}} g_n, g_n)$ is maximal. This is very close to the greedy algorithms already used in the context of regression estimation, see Barron *et al.* [3] for example. Note however that this is not necessarily the optimal choice.

Remark 2.2. The choice for the value of κ should not be a problem when we expect our estimator to reach a specified rate of convergence r_N (see for example Sect. 4 p. 445): we just require that $\kappa = \kappa(N) \ll r_N$.

The choice of ε is more problematic. In Section 4 we will see that a choice like $\varepsilon = N^{-3/2}$ is enough to reach the optimal rate of convergence in some classical examples in statistics. However, we remark in simulations (Sect. 6 p. 451) that the method performs badly when the confidence regions are two large, and this is the case when ε is too small. A good value for ε in practice should be a constant like 0.10 or 0.05. See Section 6 for more details.

2.4. Remarks on the intersection of the confidence regions

Actually, Corollary 2.2 (p. 441) could motivate another method. Note that:

$$\forall k \in \{1, ..., m\}, f \in \mathcal{CR}_{k,\varepsilon} \Leftrightarrow f \in \bigcap_{k=1}^{m} \mathcal{CR}_{k,\varepsilon}.$$

Definition 2.1. Let us put, for any $I \subset \{1, ..., m\}$:

$$\mathcal{CR}_{I,\varepsilon} = \bigcap_{k \in I} \mathcal{CR}_{k,\varepsilon},$$

and:

$$\hat{f}_I = \Pi_{\mathcal{CR}_{I,\varepsilon}} 0.$$

The estimator $\hat{f}_{\{1,...,m\}}$ can be reached by solving the following optimization problem:

$$\begin{cases} \min\limits_{g \in \mathcal{L}^2} \|g\|^2, \\ \quad \text{subject to:} \\ \forall k \in \{1, ..., m\} : \quad \langle g - \hat{\alpha}_k f_k, f_k \rangle - \sqrt{\beta(\varepsilon, k)} \leq 0, \\ \forall k \in \{1, ..., m\} : \quad -\langle g - \hat{\alpha}_k f_k, f_k \rangle - \sqrt{\beta(\varepsilon, k)} \leq 0. \end{cases}$$

The problem can be solved in dual form:

$$\max_{\gamma \in \mathbb{R}^m} \left[-\sum_{i=1}^m \sum_{k=1}^m \gamma_i \gamma_k \langle f_i, f_k \rangle + 2 \sum_{k=1}^m \gamma_k \hat{\alpha}_k - 2 \sum_{k=1}^m |\gamma_k| \sqrt{\beta(\varepsilon, k)} \right]$$

with solution $\gamma^* = (\gamma_1^*, ..., \gamma_m^*)$ and:

$$\hat{f}_{\{1,...,m\}} = \sum_{k=1}^m \gamma_k^* f_k.$$

From a statistical point of view, as:

$$2 \sum_{k=1}^m \gamma_k^* \hat{\alpha}_k = 2 \sum_{k=1}^m \gamma_k^* \frac{1}{N} \sum_{i=1}^N f_k(X_i) = \frac{2}{N} \sum_{i=1}^N \hat{f}_{\{1,...,m\}}(X_i)$$

the problem becomes:

$$\max_{\gamma \in \mathbb{R}^m} \left[\frac{2}{N} \sum_{i=1}^N \hat{f}_{\{1,...,m\}}(X_i) - \left\| \hat{f}_{\{1,...,m\}} \right\|^2 - 2 \sum_{k=1}^m |\gamma_k| \sqrt{\beta(\varepsilon, k)} \right].$$

We can see this as a maximization of a penalized criterion. The idea of penalizing likelihood to obtain good estimators when the model is very large is quite standard since the pioneering work of Akaike [1] or Good and Gaskins [14]. Here, a new penalty term is derived of the algorithm.

3. Examples

3.1. The histogram

Here, we just make things more explicit in a classical example in statistics, the histogram, so the reader more interested in kernel methods can skip this subsection. Let us assume that λ is a finite measure and let $A_1, ..., A_m$ be a partition of $\mathcal{X}$. We put, for any $k \in \{1, ..., m\}$:

$$f_k(.) = \frac{\mathbb{1}_{A_k}(.)}{\lambda(A_k)}.$$

Note that these functions satisfies requirement given in Section 2.1 (case with no assumption on f) with:

$$C_k = \frac{1}{\lambda\left(A_k\right)}.$$

In this context we have:

$$\begin{aligned}
\overline{\alpha}_k &= \frac{P\left(X \in A_k\right)}{\lambda\left(A_k\right)}, \\
\hat{\alpha}_k &= \frac{\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{A_k}\left(X_i\right)}{\lambda\left(A_k\right)}, \\
\beta(\varepsilon, k) &= \frac{4\left(1+\log\frac{2m}{\varepsilon}\right)\left(\frac{|\{i:X_i\in A_k\}|}{N}+1\right)}{N\lambda\left(A_k\right)}.
\end{aligned}$$

Finally, note that all the confidence regions $\mathcal{CR}_{k,\varepsilon}$ are orthogonal in this case. So the order of projection does not affect the obtained estimator here, and we can take:

$$\hat{f} = \Pi_{\mathcal{CR}_{m,\varepsilon}} ... \Pi_{\mathcal{CR}_{1,\varepsilon}} 0.$$

We have:

$$\hat{f}(x) = \sum_{k=1}^{m} \left(\hat{\alpha}_k - \sqrt{\beta(\varepsilon,k)}\right)_+ f_k(x)$$

where, for any $y \in \mathbb{R}$ we have: $(y)_+ = \max(y, 0) = y \vee 0$.

In this case Corollary 2.3 (p. 441) becomes:

$$\begin{aligned}
P^{\otimes N}\Bigg\{ \mathrm{d}^2(\hat{f}, f) &\leq \mathrm{d}^2(0, f) - \sum_{k=1}^{m}\left(\hat{\alpha}_k - \sqrt{\beta(\varepsilon,k)}\right)_+^2 \\
&= \sum_{k=1}^{m}\left[(\overline{\alpha}_k)^2 - \left(\hat{\alpha}_k - \sqrt{\beta(\varepsilon,k)}\right)_+^2\right] \Bigg\} \geq 1-\varepsilon.
\end{aligned}$$

3.2. **Kernel estimators**

Here, we assume that $\mathcal{X} = \mathbb{R}$ and that f is compactly supported, say for example $[0, 1]$. We put, for any $m \in \mathbb{N}$ and $k \in \{1, ..., m\}$:

$$f_k(x) = \frac{K\left(\frac{k}{m}, x\right)}{\sqrt{\int_{\mathcal{X}} K^2\left(\frac{k}{m}, u\right)\lambda(\mathrm{d}u)}}$$

where K is some function $\mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and we obtain some estimator that has the form of a kernel estimator:

$$\hat{f}(x) = \sum_{k=1}^{m} \tilde{\alpha}_k K\left(\frac{k}{m}, x\right).$$

Moreover, is possible to use a multiple kernel estimator. Let us choose $n \in \mathbb{N}$, $h \in \mathbb{N}$, h kernels $K_1, ..., K_h$ and put, for any $k = i + n * j \in \{1, ..., m = hn\}$:

$$f_k(x) = \frac{K_j\left(\frac{i}{n}, x\right)}{\sqrt{\int_{\mathcal{X}} K_j^2\left(\frac{i}{n}, u\right)\lambda(\mathrm{d}u)}}.$$

We obtain a multiple kernel estimator:

$$\hat{f}_{\{1,\dots,m\}}(x) = \sum_{i=1}^{n}\sum_{j=1}^{h} \tilde{\alpha}_{i+nj} K_j\left(\frac{i}{n}, x\right).$$

However, note that the use of kernel functions is more justified in large dimension where we will take as basis functions: $K_j(X_i, .)$, namely data-dependant functions. We show in Section 5 that our algorithm can be extended to this case.

4. Optimality of the procedure in a classical example in statistics

In this whole section, we assume that $\mathcal{X} = [0, 1]$ and that λ is the Lebesgue measure.

4.1. General remarks when $(f_k)_k$ is an orthonormal family and condition $\mathcal{H}(+\infty)$ is satisfied

In Sections 4.1 and 4.2, we study the rate of convergence of our estimator in the special case where $(f_k)_{k\in\mathbb{N}^*}$ is an orthonormal basis of $\mathcal{L}^2$. Note that in this case all the order of application of the projections $\Pi_{\mathcal{CR}_{k,\varepsilon}}$ does not matter because these projections works on orthogonal directions. So we can define, once m is chosen:

$$\hat{f} = \Pi_{\mathcal{CR}_{m,\varepsilon}} \dots \Pi_{\mathcal{CR}_{1,\varepsilon}} 0 = \Pi_{\mathcal{CR}_{\{1,\dots,m\},\varepsilon}} 0.$$

Note that:

$$\hat{f}(x) = \sum_{k=1}^{m} \operatorname{sgn}(\hat{\alpha}_k)\left(|\hat{\alpha}_k| - \sqrt{\beta(\varepsilon, k)}\right)_+ f_k(x),$$

and so $\hat{f}$ is a soft-thresholded estimator.

Let us also make the following remark. If we assume that $\|f\|_\infty \le c$ (that is condition $\mathcal{H}(+\infty)$ is satisfied, see Sect. 2.1 p. 440), we have:

$$\mathrm{d}^2(f, 0) = \int_0^1 f^2(x)\mathrm{d}x \le c\int_0^1 f(x)\mathrm{d}x = c.$$

So the region:

$$\mathcal{B} = \left\{ g \in \mathcal{L}^2 : \forall k \in \mathbb{N}^*, \int_{\mathcal{X}} g(x) f_k(x) \lambda(\mathrm{d}x) \le \sqrt{c} \right\}$$

is convex, and contains f. So the projection on $\mathcal{B}$, $\Pi_{\mathcal{B}}$ can only improve $\hat{f}$. We put:

$$\tilde{f} = \Pi_{\mathcal{B}} \hat{f}. \tag{4.1}$$

Note that this transformation is needed to obtain Theorem 4.1, but does not have practical incidence in general. Actually:

$$\tilde{f}(x) = \sum_{k=1}^{m} \operatorname{sgn}(\hat{\alpha}_k)\left\{\left(|\hat{\alpha}_k| - \sqrt{\beta(\varepsilon, k)}\right)_+ \wedge \sqrt{c}\right\} f_k(x),$$

where we let $a \wedge b$ denote $\min(a, b)$ for any $(a, b) \in \mathbb{R}^2$.

4.2. **Rate of convergence in Besov spaces**

Definition 4.1. Let $\tilde{\phi}(.)$ and $(\tilde{\psi}_{j,k})_{(j,k)\in\mathbb{N}\times\mathbb{Z}}$ be a Daubechies wavelets basis with a given regularity R (see Daubechies [11]). We define the periodized wavelets:

$$\forall x\in\mathbb{R},\quad \phi(x)=\sum_{\ell\in\mathbb{Z}}\tilde{\phi}(x-\ell), \forall (j,k)\in\mathbb{N}\times\mathbb{Z}, \forall x\in\mathbb{R},\quad \psi_{j,k}(x)=\sum_{\ell\in\mathbb{Z}}\tilde{\psi}_{j,k}(x-\ell).$$

The idea of periodizing compactly supported wavelets is due to Daubechies [11], and very simply described in some papers by Cai, see [6] for example. It allows to give a simple definition for Besov spaces of functions on the interval.

Definition 4.2. We define the ball in a periodized Besov space, for $0<s\leq R+1$ and $p,q\geq 1$ and $D\in\mathbb{R}_+$:

$$B_{s,p,q}(D)=\left\{g:[0,1]\to\mathbb{R},\quad g(.)=\alpha\phi(.)+\sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k}(.), \sum_{j=0}^{\infty}2^{jq\left(s-\frac{1}{2}-\frac{1}{p}\right)}\left[\sum_{k=1}^{2^j}|\beta_{j,k}|^p\right]^{\frac{q}{p}}=\|g\|_{s,p,q}^q\leq D\right\},$$

with obvious changes for $p=+\infty$ or $q=+\infty$ (see for example Donoho *et al.* [13]). We also define the ball in a weak periodized Besov space, for $\rho,\pi>0$, $D'\in\mathbb{R}_+$ with $\frac{1}{2}\left(\frac{\pi}{\rho}-1\right)\leq R+1$:

$$W_{\rho,\pi}(D')=\left\{g:[0,1]\to\mathbb{R},\quad g(.)=\alpha\phi(.)+\sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k}(.),\quad \sup_{\lambda>0}\lambda^{\rho}\sum_{j=0}^{\infty}2^{j\left(\frac{\pi}{2}-1\right)}\sum_{k=1}^{2^j}\mathbb{1}_{\{|\beta_{j,k}|>\lambda\}}\leq D'\right\},$$

as done by Cohen [10].

Let us remark that $B_{s,p,q}(D)$ is a set of functions with regularity s while $W_{\rho,\pi}(D')$ is a set of functions with regularity:

$$s'=\frac{1}{2}\left(\frac{\pi}{\rho}-1\right).$$

Theorem 4.1. *Let us assume that $\mathcal{H}(\infty)$ is satisfied: $\|f\|_\infty<c$. Let us assume that*

$$f\in B_{s,p,q}(D)$$

with $R+1\geq s>\frac{1}{p}$, $1\leq q\leq\infty$, $2\leq p\leq+\infty$, or that

$$f\in B_{s,p,q}(D)\cap W_{\frac{2}{2s+1},2}(D')$$

with $R+1\geq s>\frac{1}{p}$, $1\leq p\leq+\infty$, with unknown constants s, p and q. Let us choose:

$$\{f_1,...,f_m\}=\{\phi\}\cup\{\psi_{j,k},j=1,...,2^{\lfloor\frac{\log N}{\log 2}\rfloor},k=1,...,2^j\}$$

(so $\frac{N}{2}\leq m\leq N$) and $\varepsilon=N^{-3/2}$ in the definition of $\tilde{f}$. Then we obtain:

$$P^{\otimes N}\left[\mathrm{d}^2(\tilde{f},f)\right]\leq\mathcal{C}\left(\frac{\log N}{N}\right)^{\frac{2s}{2s+1}},$$

where $\tilde{f}$ is the estimator defined by Equation 4.1 (page 445), and where:

$$\mathcal{C}=\mathcal{C}(s,p,q,D,D',c)$$

does not depend on N nor on other characteristics of the functions f.

The proof of this theorem is also given in Section 7.3 page 459. Let us remark that Theorem 4.1 is adaptative near minimax (up to a $\log N$ term), see Härdle *et al.* [15], or Donoho, Johnstone *et al.* [13].

5. Improvements and generalization of Theorem 2.1

It appears in simulations that the bound on $\mathrm{d}^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k) = (\hat{\alpha}_k - \overline{\alpha}_k)^2$, as given by Theorem 2.1, has to be very sharp if we want to obtain a good estimator. Actually, as pointed out by Catoni [9], the symmetrization technique used in the proof of Theorem 2.1 causes the loss of a factor 2 in the bound because we upper bound the variance of two samples instead of one. So it is possible to obtain sharper bounds. In this section, we try to use this remark to improve our bound, using techniques already used by Catoni [7].

We then remark that a technique due to Seeger [21] allows to include the case of data-dependant basis functions (f_k) and to deal with SVM in particular.

First, remark that the estimation technique described in Section 2 does not necessarily require a bound on $(\hat{\alpha}_k - \overline{\alpha}_k)^2$. Actually, a simple confidence interval on $\overline{\alpha}_k$ is sufficient.

5.1. An improvement of Theorem 2.1

Theorem 5.1. *We assume that for every k, $\|f_k\| \leq C_k < +\infty$. For any $\varepsilon > 0$, for any $\beta_{k,1}, \beta_{k,2}$ such that:*

$$0 < \beta_{k,j} < \frac{N}{C_k}, \quad j \in \{1,2\},$$

with $P^{\otimes N}$-probability at least $1-\varepsilon$ we have:

$$\forall k \in \{1,...,m\}, \quad \alpha_k^{\inf}(\varepsilon, \beta_{k,1}) \leq \overline{\alpha}_k \leq \alpha_k^{\sup}(\varepsilon, \beta_{k,2})$$

with:

$$\alpha_k^{\sup}(\varepsilon, \beta_{k,2}) = \frac{N - N \exp\left[\frac{1}{N}\sum_{i=1}^N \log\left(1 - \frac{\beta_{k,2}}{N} f_k(X_i)\right) - \frac{\log \frac{2m}{\varepsilon}}{N}\right]}{\beta_{k,2}}$$

and:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) = \frac{N \exp\left[\frac{1}{N}\sum_{i=1}^N \log\left(1 + \frac{\beta_{k,1}}{N} f_k(X_i)\right) - \frac{\log \frac{2m}{\varepsilon}}{N}\right] - N}{\beta_{k,1}}.$$

The proof is given in Section 7.4 page 461.

First, let us see why this theorem really improves Theorem 2.1, at least when N is large. Let us define:

$$V_k = P\left\{\left[f_k(X) - P\left(f_k(X)\right)\right]^2\right\}$$

and let us choose:

$$\beta_{k,1} = \beta_{k,2} = \sqrt{\frac{N \log \frac{2m}{\varepsilon}}{V_k}}.$$

Then we obtain:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) = \hat{\alpha}_k - \sqrt{\frac{2V_k \log \frac{2m}{\varepsilon}}{N}} + \mathcal{O}_P\left(\frac{\log \frac{2m}{\varepsilon}}{N}\right)$$

and:

$$\alpha_k^{\sup}(\varepsilon, \beta_{k,2}) = \hat{\alpha}_k + \sqrt{\frac{2V_k \log \frac{2m}{\varepsilon}}{N}} + \mathcal{O}_P\left(\frac{\log \frac{2m}{\varepsilon}}{N}\right).$$

So, the first order term for $\mathrm{d}^2(\hat{\alpha}_k f_k, \overline{\alpha}_k f_k)$ is:

$$\frac{2V_k \log \frac{2m}{\varepsilon}}{N},$$

there is an improvement by a factor 4 when we compare this bound to Theorem 2.1.

Remark that this particular choice for $\beta_{k,1}$ and $\beta_{k,2}$ is valid as soon as:

$$\sqrt{\frac{N \log \frac{2m}{\varepsilon}}{V_k}} < \frac{N}{C_k}$$

or equivalently as soon as N is greater than

$$\frac{C_k^2 \log \frac{2m}{\varepsilon}}{V_k}.$$

In practice, however, this particular $\beta_{k,1}$ and $\beta_{k,2}$ are unknown. We can use the following procedure (see Catoni [9]). We choose a value $a > 1$ and:

$$B = \left\{ a^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{C_k}}{\log a} \right\rfloor - 1 \right\}.$$

By taking a union bound over all possibles values of B, with:

$$|B| \leq \frac{\log \frac{N}{C_k}}{\log a}$$

we obtain Corollary 5.2.

Corollary 5.2. *Under condition $\mathcal{H}(+\infty)$, for any $a > 1$, for any $\varepsilon > 0$, with $P^{\otimes N}$-probability at least $1 - \varepsilon$ we have, for any $k \in \{1, ..., m\}$:*

$$\sup_{\beta \in B} \alpha_k^{\inf}\left(\frac{\varepsilon \log a}{\log N - \log C_k}, \beta\right) \leq \overline{\alpha}_k \leq \inf_{\beta \in B} \alpha_k^{\sup}\left(\frac{\varepsilon \log a}{\log N - \log C_k}, \beta\right),$$

with:

$$B = \left\{ a^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{C_k}}{\log a} \right\rfloor - 1 \right\}.$$

Note that the price to pay for the optimization with respect to $\beta_{k,1}$ and $\beta_{k,2}$ was just a $\log \log N$ factor.

5.2. **The histogram example continued**

We apply here the improved bounds in the case of the histogram introduced in Section 3.1 page 443.
In this case:

$$\begin{aligned}
\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) &= \frac{N}{\beta_{k,1}} \left\{ \left[\left(1 + \frac{\beta_{k,1}}{\lambda(A_k)N}\right)^{|\{i:X_i \in A_k\}|} \frac{\varepsilon}{2m} \right]^{\frac{1}{N}} - 1 \right\} \\
&= \frac{N}{\beta_{k,1}} \left[\left(1 + \frac{\beta_{k,1}}{\lambda(A_k)N}\right)^{\hat{\alpha}_k} \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} - 1 \right].
\end{aligned}$$

Remember that, for any $x \geq 0$:

$$(1+x)^{\gamma} \geq 1 + \gamma x + \frac{\gamma(\gamma-1)}{2}x^2$$

and so we have:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} \left[1 - \frac{\beta_{k,1}(1-\hat{\alpha}_k)}{2N}\right] - \frac{N}{\beta_{k,1}} \left[1 - \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}}\right].$$

Now, we take the grid:

$$B = \left\{2^l, 0 \leq l \leq \left\lfloor \frac{\log \frac{N}{C_k}}{\log 2} \right\rfloor - 1\right\}.$$

Remark that, for any β in:

$$\left[1, \frac{N}{2C_k}\right]$$

there is some $b \in B$ such that $\beta \leq b \leq 2\beta$, and so:

$$\alpha_k^{\inf}(\varepsilon, b) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} \left[1 - \frac{\beta_{k,1}(1-\hat{\alpha}_k)}{2N}\right] - \frac{N}{2\beta_{k,1}} \left[1 - \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}}\right].$$

This allows us to choose whatever value for $\beta_{k,1}$ in

$$\left[1, \frac{N}{2C_k}\right].$$

Let us choose:

$$\beta_{k,1} = \sqrt{\frac{N^2\left[\left(\frac{\varepsilon}{2m}\right)^{\frac{-1}{N}} - 1\right]}{\hat{\alpha}_k(1-\hat{\alpha}_k)}}$$

that is allowed for N large enough. So we have:

$$\alpha_k^{\inf}(\varepsilon, \beta_{k,1}) \geq \hat{\alpha}_k \left(\frac{\varepsilon}{2m}\right)^{\frac{1}{N}} - \sqrt{\hat{\alpha}_k(1-\hat{\alpha}_k)\left[\left(\frac{\varepsilon}{2m}\right)^{\frac{-1}{N}} - 1\right]}.$$

With the union bound term (over the grid B) we obtain:

$$\begin{aligned}
\alpha_k^{\inf}\left(\frac{\varepsilon \log 2}{\log \frac{N}{C_k}}, \beta_{k,1}\right) &\geq \hat{\alpha}_k \left(\frac{\varepsilon \log 2}{2m \log \frac{N}{C_k}}\right)^{\frac{1}{N}} - \sqrt{\hat{\alpha}_k(1-\hat{\alpha}_k)\left[\left(\frac{\varepsilon \log 2}{2m \log \frac{N}{C_k}}\right)^{\frac{-1}{N}} - 1\right]} \\
&= \hat{\alpha}_k - \sqrt{\frac{\hat{\alpha}_k(1-\hat{\alpha}_k)\log \frac{2m \log \frac{N}{C_k}}{\varepsilon \log 2}}{N}} + \mathcal{O}\left(\frac{\log \frac{m \log N}{\varepsilon}}{N}\right),
\end{aligned}$$

remark that we have this time the "real" variance term of $\mathbb{1}_{A_k}(X)$:

$$\hat{\alpha}_k(1-\hat{\alpha}_k) = \frac{|\{i : X_i \in A_k\}|}{N}\left(1 - \frac{|\{i : X_i \in A_k\}|}{N}\right).$$

5.3. **A generalization to data-dependent basis functions**

We now extend the previous method to the case where the family $(f_1, ..., f_m)$ is allowed to be data-dependant, in a particular sense. This subsection requires some modifications of the notations of Section 2.

Definition 5.1. For any $m' \in \mathbb{N}^*$ we define a function $\Theta_{m'} : \mathcal{X} \to \left(\mathcal{L}^2\right)^{m'}$. For any $i \in \{1, ..., N\}$ we put:

$$\Theta_{m'}(X_i) = (f_{i,1}, ..., f_{i,m'}),$$

and assume that $\Theta_{m'}$ is such that:

$$\|f_{i,k}\| = 1$$

for any $(i,k) \in \{1, ..., N\} \times \{1, ..., m'\}$. Finally, consider the family of functions:

$$(f_1, ..., f_m) = (f_{1,1}, ..., f_{1,m'}, ..., f_{N,1}, ..., f_{N,m'}).$$

So we have $m = m'N$ (of course, m' is allowed to depend on N). Let us take, for any $i \in \{1, ..., N\}$:

$$P_i(.) = P^{\otimes N}(.|X_i).$$

We assume that we have known finite constants $C_{i,k}$ such that $\|f_{i,k}(x)\|_{+\infty} \leq C_{i,k}$ for any i and k. Finally, we put:

$$\overline{\alpha}_{i,k} = \arg\min_{\alpha \in \mathbb{R}} \mathrm{d}^2(\alpha f_{i,k}, f).$$

Theorem 5.3. *For any $\varepsilon > 0$, for any $\beta_{i,k,1}, \beta_{i,k,2}$ such that:*

$$0 < \beta_{i,k,j} < \frac{N-1}{C_{i,k}}, \quad j \in \{1, 2\},$$

with $P^{\otimes N}$-probability at least $1 - \varepsilon$, for any $i \in \{1, ..., N\}$ and $k \in \{1, ..., m\}$ we have:

$$\tilde{\alpha}_k^{\inf}(\varepsilon, \beta_{i,k,1}) \leq \overline{\alpha}_k \leq \tilde{\alpha}_k^{\sup}(\varepsilon, \beta_{i,k,2})$$

with:

$$\tilde{\alpha}_k^{\sup}(\varepsilon, \beta_{i,k,2}) = \frac{N - 1 - (N-1)\exp\left[\frac{1}{N-1}\sum_{j \neq i} \log\left(1 - \frac{\beta_{i,k,2}}{N-1} f_{i,k}(X_j)\right) - \frac{\log\frac{2m'N}{\varepsilon}}{N-1}\right]}{\beta_{i,k,2}}$$

and:

$$\tilde{\alpha}_k^{\inf}(\varepsilon, \beta_{i,k,1}) = \frac{(N-1)\exp\left[\frac{1}{N-1}\sum_{j \neq i} \log\left(1 + \frac{\beta_{i,k,1}}{N-1} f_{i,k}(X_j)\right) - \frac{\log\frac{2m'N}{\varepsilon}}{N-1}\right] - N + 1}{\beta_{i,k,1}}.$$

The proof of this theorem is also given in Section 7 (Sect. 7.4 p. 461).

Example 5.1 (multiple kernel SVM)**.** This example is the continuation of the kernel estimator example, studied in Section 3.2 page 444.

We propose the following choice:

$$\Theta_{m'}(X_i) = \left\{ \frac{K_1(X_i, .)}{\int_{\mathcal{X}} K_1^2(X_i, .)\lambda(\mathrm{d}x)}, ..., \frac{K_{m'}(X_i, .)}{\int_{\mathcal{X}} K_{m'}^2(X_i, .)\lambda(\mathrm{d}x)} \right\}$$

for some family of functions $\mathcal{X}^2 \to \mathbb{R}$: $(K_1, ..., K_{m'})$. Note that we have $m = m'N$. In this case, the estimator is under the form:

$$\forall x \in \mathcal{X}, \quad \hat{f}(x) = \sum_{j=1}^{m'} \sum_{i=1}^{N} \tilde{\alpha}_{i,j} K_j(X_i, x),$$

and the number of $\tilde{\alpha}_{i,j} \neq 0$ is expected to be small. This estimator has the form of a SVM (if $h = 1$ and K_1 is a Mercer's kernel). However, if we take the general form of the algorithm, we can see that it is possible to use whatever heuristic to choose the next pair (i, j), and so we can use a wide range of methods to choose the set of support vectors.

For example, if N is large, we can use only the following method:

- use a clustering algorithms on the data to obtain c clusters;
- at each step, try to use only one vector from each cluster, for example the one that is the closest to the mean point of the cluster.

In this case have only to try c projection instead of N. This proposition is suggested by Tipping's Relevance Vector Machine [22].

One of the most used kernels is the Gaussian kernel. If $\mathcal{X}$ is a metric space, with a distance $\delta(.,.)$, we choose $\gamma \in \mathbb{R}_+^*$ and we put:

$$K(x, x') = \exp\left[-\gamma\delta^2(x, x')\right].$$

In practice, the choice of γ is problematic. Here, we can choose a grid of values $(\gamma_1, ..., \gamma_h) \in (\mathbb{R}_+^*)^h$ and take:

$$K_j(x, x') = \exp\left[-\gamma_j\delta^2(x, x')\right]$$

and let the algorithm selects the relevant values of γ_j.

6. Simulations

The whole simulations and estimations are performed with the software R [18].

6.1. Description of the example

We simulate X_i for $i \in \{1, ..., N\}$ with $N = 2^{10} = 1024$, where the variables $X_i \in [0, 1] \subset \mathbb{R}$ are i.i.d. from a distribution with a given density f with respect to the Lebesgue measure.

When then try to estimate f on the basis on the sample, using the method described in this paper as well as classical non-parametric estimators. Here, we will use five methods. The first estimation method will be a multiple kernel estimator obtained by our method and the second is the classical Parzen-Rosenblatt estimate as implemented in R (so a kernel estimator using one single kernel). The third one is a wavelet estimator obtained by our algorithm while the fourth and the fifth are two wavelet estimators (using the same wavelets basis) build using asymptotic considerations as done by Donoho *et al.* [12]: one is soft-thresholded and the other is hard-thresholded.

6.2. The estimators

6.2.1. *Wavelet estimators based on asymptotic considerations*

We choose $\tilde{\phi}$ and $(\tilde{\psi}_{j,k})_{j,k}$ as the Daubechies wavelets with regularity $R = 7$, and define ϕ and $(\psi_{j,k})_{j,k}$ as the periodized version of these wavelets (see Def. 4.1 p. 446). For convenience let us put $\psi_{-1,0} = \phi$ and $S_j = \{0, 1, ..., 2^j - 1\}$ if $j \geq 0$, and $S_{-1} = \{0\}$.

We take:

$$\hat{\alpha}_{j,k} = \frac{1}{N}\sum_{j=1}^{N}\psi_{j,k}(X_i).$$

For a given $\kappa \geq 0$ and $J \in \mathbb{N}$, we define the hard-thresholded estimator:

$$\tilde{f}_{HT}(.) = \sum_{j=-1}^{J}\sum_{k\in S_j}\hat{\alpha}_{j,k}\mathbb{1}(|\hat{\alpha}_{j,k}| \geq \kappa t_{j,N})\psi_{j,k}(.)$$

and the soft-thresholded estimator

$$\tilde{f}_{ST}(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} sgn\left(\hat{\alpha}_{j,k}\right) \left(|\hat{\alpha}_{j,k}| - \kappa' t_{j,N}\right)_+ \psi_{j,k}(.)$$

where:

$$t_{j,N} = \sqrt{\frac{j}{N}}.$$

Actually, according to the asymptotic theory given by Donoho *et al.* [12] we must choose J in such a way that:

$$2^J \sim t_{J,N}^{-1}.$$

Here, we choose $\kappa = 2$ and $\kappa' = 1$ (experimental results led to the choice of a different threshold in the soft and in the hard case) and $J = 6$.

6.2.2. *Wavelet estimators obtained by confidence intervals*

We also use the same family of functions, and we apply our thresholding method, described in Section 4. Note that we simply have $m = 2^{J+1}$.

We use an asymptotic version of our confidence intervals inspired by our theoretical confidence intervals:

$$\overline{\alpha}_{j,k} \in \left[\hat{\alpha}_{j,k} \pm \sqrt{2\frac{\log \frac{2m}{\varepsilon} V_{j,k}}{N}}\right]$$

where $V_{j,k}$ is the estimated variance of $\psi_{j,k}(X)$:

$$V_{j,k} = \frac{1}{N}\sum_{i=1}^{N}\left[\psi_{j,k}(X_i) - \frac{1}{N}\sum_{h=1}^{N}\psi_{j,k}(X_h)\right]^2.$$

Let us remark that the union bound are always "pessimistic", and that we use a union bound argument over all the m models despite only a few of them are effectively used in the estimator. So, we propose to actually use the individual confidence interval for each model, replacing: the $\log \frac{2m}{\varepsilon}$ by $\log \frac{1}{\varepsilon}$.

So the estimator is the following (note the similarity with the soft-thresholded estimator, but with a non-constant threshold value):

$$\hat{f}_{WAV}(.) = \sum_{j=-1}^{J} \sum_{k \in S_j} \operatorname{sgn}\left(\hat{\alpha}_{j,k}\right) \left(|\hat{\alpha}_{j,k}| - \sqrt{2\frac{\log \frac{2}{\varepsilon} V_{j,k}}{N}}\right)_+ \psi_{j,k}(.).$$

We choose $\varepsilon = 0.1$.

6.2.3. *Multiple kernel estimator obtained by confidence intervals*

Finally, we use the kernel estimator described in example 5.1 page 450, with function K:

$$K_j(u,v) = \exp\left[-2^{2j}(u-v)^2\right]$$

and $j \in \{1, ..., h = 6\}$. We add the constant function 1 to the family, so $m = 1 + hN$.

Here again we use the individuals confidence intervals, and the asymptotic version of this intervals.

We let $\hat{f}_{KER}(.)$ denote the obtained estimator, with $\kappa = 0.0001$ and $\varepsilon = 0.1$.

6.2.4. *Parzen-Rosenblatt estimator*

We also use the Parzen-Rosenblatt estimator with the default settings (gaussian kernel) in R as a benchmark. As this estimator uses a single kernel, we cannot hope to obtain adaptation with it, and so its performances are expected to be the worst among the estimator studied here.

6.3. **The density functions**

We use six density functions in the experiences, given in three groups with increasing difficulty. The first group contains densities from well-known parametric families. The second one contains densities taken from the paper by Marron and Wand [16], that are difficult to estimate because of multimodality. The third one contains mixtures of densities from the previous groups with uniform densities, leading to non-continuous densities.

6.3.1. *Densities from simple parametric families*

We take a density from the Gaussian family and a density from the Laplace family, restricted to $[0,1]$:

$$f_{GAUSS}(x) \propto \exp\Big[-20(0.5-x)^2\Big]\mathbb{1}_{[0,1]}(x)$$

$$f_{LAPLACE}(x) \propto \exp\Big[-5|0.5-x|\Big]\mathbb{1}_{[0,1]}(x).$$

6.3.2. *Marron and Wand's densities*

We take two densities taken from the paper by Marron and Wand [16]. Theses densities are mixture of Gaussian that are quite difficult to estimate because of the presence of components with very different variance and multimodality. Here, we translate these densities in order to have all the modes in the unit interval (this is the meaning of the $6x-3$ transform) and we restrict them to this interval. So we put:

$$f_{CLAW}(x) \propto g_{CLAW}(x) = \mathbb{1}_{[0,1]}(x) \times \left\{ \exp\Big[-\big(1-(6x-3)\big)^2\Big] + \sum_{\ell=0}^{4} \exp\left[-50\left(\left(\frac{\ell}{2}-1\right)-(6x-3)\right)^2\right]\right\}$$

$$f_{SMOOTHCOMB}(x) \propto \mathbb{1}_{[0,1]}(x) \times \left\{ \sum_{\ell=0}^{5} \exp\left[-2^{2\ell-1}\left(\frac{63}{32}\right)^2\left(\frac{65-96(0.5)^\ell}{21}-(6x-3)\right)^2\right]\right\}.$$

6.3.3. *Non-continuous densities*

Here, we define mixtures of the previous distributions with uniform distributions in order to obtain non-continuous densities:

$$f_{GAUSS-UNIF}(x) \propto \exp\left[-20(0.5-x)^2\right]\mathbb{1}_{[0,1]}(x) + \frac{1}{2}\mathbb{1}_{[0.6,0.8]}(x)$$

$$f_{CLAW-UNIF}(x) \propto g_{CLAW}(x) + \frac{1}{2} \times \mathbb{1}_{[0.6,0.8]}(x).$$

6.4. **Experiments and results**

For each density function, the whole experiment (including simulations and computation of each estimator) is repeated 20 times. The results are reported in Table 1. For each estimator, we give:

- the mean distance of the estimated density to the true density;
- the standard deviation of this distance.

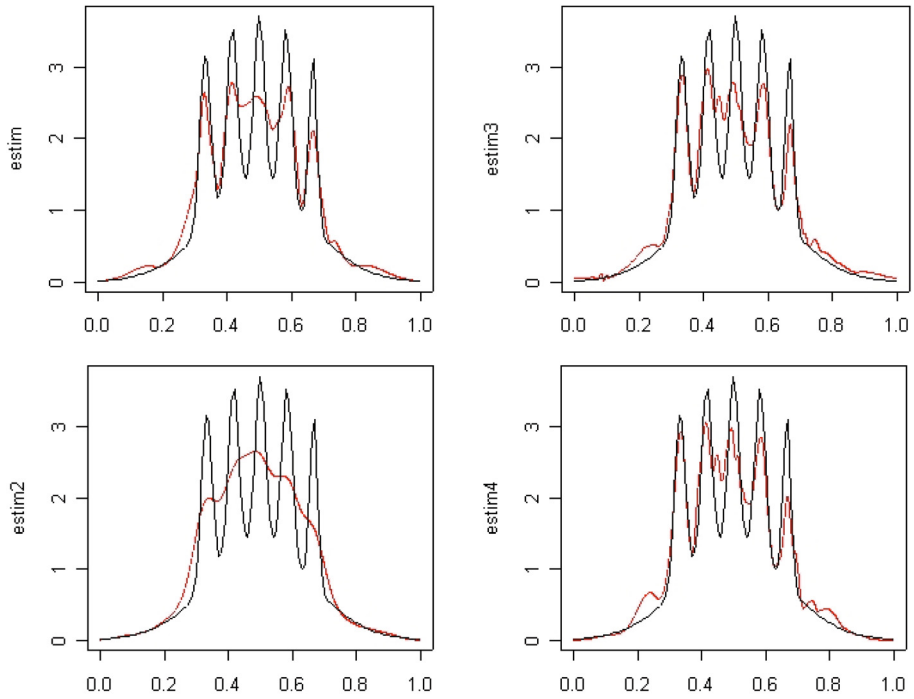

FIGURE 1. Estimators for f_{CLAW} (the true density is represented in black on every graphic). Top-left (labelled "estim"): multi-kernel estimate. Down-left (labelled "estim2"): R standard Parzen-Rosenblatt estimate. Top-right (labelled "estim3"): wavelet estimate based on confidence intervals. Down-right (labelled "estim4"): soft-thresholded wavelet estimate.

Note that our estimators built with confidence intervals (CI) are based on the quadratic loss. However, the results are also reported using the absolute loss and the supremum loss in order to check the performances of our estimator with different losses.

We also represent the different estimators together with the true densities for the $CLAW$ and $SMOOTHCOMB$ case (Figs. 1 and 2).

6.5. **Comments**

First of all, note that the simplest estimator, the R Parzen-Rosenblatt estimator, achieves the best results for the densities taken from quite simple parametric families ($GAUSS$, $LAPLACE$ and even $GAUSS - UNIF$"). However, for the three others densities, its results are the worst off all. Figure 1 shows actually that this estimator fails to capture the multiscale aspect of these three densities: of course, wavelets and multi-kernel estimates where introduced to take this aspect into account.

For $CLAW$ and $SMOOTHCOMB$, our wavelet estimators (built with confidence intervals) reaches the best values, while for $CLAW - UNIF$ it is our multiple kernel estimator. However, notice that the results of the three wavelets estimators and of the multi-kernel estimator are quite close in every experiences. Is is remarquable that our estimators, introduced thanks to considerations of the quadratic loss, performs as well as the other estimators even for the absolute loss and for the supremum loss.

TABLE 1. Results of the experiments. For each density, for each estimator and for each loss we give the mean value of the loss (top of the case, in bold) and the standard deviation (down). "Loss 2" is the quadratic distance $\mathrm{d}^2(\hat{f}, f)$, while "loss 1" is the absolute distance $\int_0^1 |\hat{f}(x) - f(x)|\mathrm{d}x$ and "loss ∞" is the uniform distance $\sup_{x\in[0,1]} |\hat{f}(x) - f(x)|$.

DENSITY	Loss	Multiple Kernel (CI)	Parzen-Rosenblatt (R default)	Wavelets (CI)	Wavelets (soft-th)	Wavelets (hard-th)
GAUSS	2	**0.0093** 0.0025	**0.0062** 0.0043	**0.0200** 0.0059	**0.0154** 0.0057	**0.0189** 0.0105
	1	**0.0785** 0.0107	**0.0560** 0.0174	**0.1098** 0.0154	**0.0957** 0.0170	**0.1089** 0.0230
	∞	**0.2187** 0.0775	**0.1891** 0.0687	**0.5111** 0.1490	**0.4289** 0.1608	**0.3551** 0.3296
LAPLACE	2	**0.0196** 0.0056	**0.0110** 0.0040	**0.0319** 0.0061	**0.0234** 0.0048	**0.0260** 0.0103
	1	**0.1091** 0.0168	**0.0726** 0.0124	**0.1271** 0.0106	**0.1071** 0.0107	**0.1170** 0.0155
	∞	**0.4852** 0.1105	**0.4254** 0.1135	**0.7023** 0.1078	**0.6132** 0.1387	**0.5863** 0.3331
CLAW	2	**0.1041** 0.0256	**0.2544** 0.0105	**0.0807** 0.0192	**0.0811** 0.0184	**0.0858** 0.0197
	1	**0.2053** 0.0256	**0.3147** 0.0097	**0.1867** 0.0216	**0.1851** 0.0196	**0.1961** 0.0213
	∞	**1.2031** 0.2411	**1.5552** 0.1350	**1.1292** 0.2364	**1.1614** 0.2806	**1.2335** 0.2619
SMOOTHCOMB	2	**0.0970** 0.0117	**0.2932** 0.0103	**0.0710** 0.0097	**0.0872** 0.0098	**0.0811** 0.0171
	1	**0.2163** 0.0169	**0.4202** 0.0166	**0.2027** 0.0167	**0.2151** 0.0142	**0.2060** 0.0178
	∞	**1.5362** 0.1503	**1.7919** 0.0403	**0.9920** 0.1628	**1.3889** 0.1603	**1.5111** 0.1788
GAUSS-UNIF	2	**0.0285** 0.0052	**0.0245** 0.0049	**0.0377** 0.0054	**0.0356** 0.0060	**0.0423** 0.0115
	1	**0.1137** 0.0106	**0.0994** 0.0121	**0.1419** 0.0100	**0.1357** 0.0102	**0.1423** 0.0203
	∞	**0.6243** 0.0952	**0.6432** 0.0732	**0.7373** 0.1025	**0.7160** 0.1078	**0.8036** 0.2022
CLAW-UNIF	2	**0.0798** 0.0144	**0.1648** 0.0037	**0.0850** 0.0124	**0.0847** 0.0128	**0.1057** 0.0240
	1	**0.1917** 0.0158	**0.2673** 0.0040	**0.2005** 0.0149	**0.1989** 0.0158	**0.2165** 0.0178
	∞	**0.9654** 0.1351	**1.3057** 0.1064	**1.1000** 0.2031	**1.1061** 0.2174	**1.2305** 0.3085

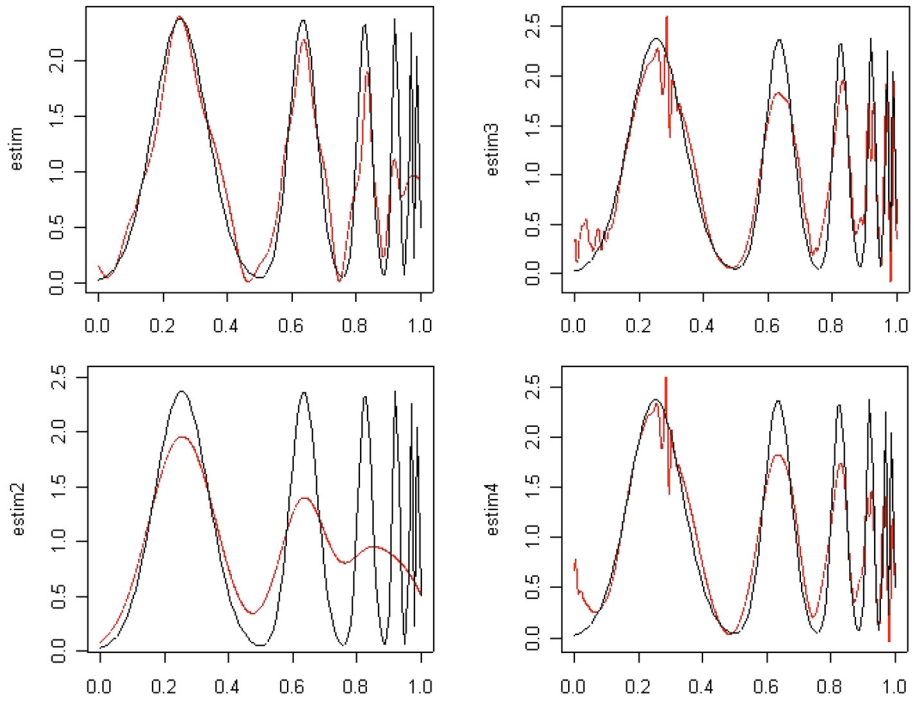

FIGURE 2. Estimators for $f_{SMOOTHCOMB}$.

7. Proofs

7.1. Proof of Theorem 2.1 of Section 2

Before we give the proof, let us state two lemmas. The first one is a variant of a lemma by Catoni [9], the second one is due to Panchenko [17].

Lemma 7.1. *Let $(T_1, ..., T_{2N})$ be a random vector taking values in $\mathbb{R}^{2N}$ distributed according to a distribution $\mathcal{P}^{\otimes 2N}$. For any $\eta \in \mathbb{R}$, for any measurable function $\lambda : \mathbb{R}^{2N} \to \mathbb{R}_+^*$ that is exchangeable with respect to its $2 \times 2N$ arguments, we have:*

$$\mathcal{P}^{\otimes 2N} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \left\{ T_{i+N} - T_i \right\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right) \leq \exp\left(-\eta\right)$$

and the reverse inequality:

$$\mathcal{P}^{\otimes 2N} \exp\left(\frac{\lambda}{N} \sum_{i=1}^{N} \left\{ T_i - T_{i+N} \right\} - \frac{\lambda^2}{N^2} \sum_{i=1}^{2N} T_i^2 - \eta \right) \leq \exp\left(-\eta\right),$$

where we write:

$$\eta = \eta\left(T_1, ..., T_{2N}\right)$$
$$\lambda = \lambda\left(T_1, ..., T_{2N}\right)$$

for short.

Proof of Lemma 7.1. In order to prove the first inequality, we write:

$$\mathcal{P}^{\otimes 2N}\exp\left(\frac{\lambda}{N}\sum_{i=1}^{N}\Big\{T_{i+N}-T_i\Big\}-\frac{\lambda^2}{N^2}\sum_{i=1}^{2N}T_i^2-\eta\right)=$$
$$\mathcal{P}^{\otimes 2N}\exp\left(\sum_{i=1}^{N}\log\cosh\left\{\frac{\lambda}{N}\left(T_{i+N}-T_i\right)\right\}-\frac{\lambda^2}{N^2}\sum_{i=1}^{2N}T_i^2-\eta\right).$$

We now use the inequality:

$$\forall x\in\mathbb{R},\log\cosh x\leq\frac{x^2}{2}\cdot$$

We obtain:

$$\log\cosh\left\{\frac{\lambda}{N}\left(T_{i+N}-T_i\right)\right\}\leq\frac{\lambda^2}{2N^2}\left(T_{i+N}-T_i\right)^2\leq\frac{\lambda^2}{N^2}\left(T_{i+N}^2+T_i^2\right).$$

The proof for the reverse inequality is exactly the same. □

Lemma 7.2 (Panchenko [17], Cor. 1)**.** *Let us assume that we have i.i.d. variables $T_1,...,T_N$ (with distribution $\mathcal{P}$ and values in $\mathbb{R}$) and an independent copy $T'=(T_{N+1},...,T_{2N})$ of $T=(T_1,...,T_N)$. Let $\xi_j(T,T')$ for $j\in\{1,2,3\}$ be three measurable functions taking values in $\mathbb{R}$, and $\xi_3\geq 0$. Let us assume that we know two constants $A\geq 1$ and $a>0$ such that, for any $u>0$:*

$$P^{\otimes 2N}\left[\xi_1(T,T')\geq\xi_2(T,T')+\sqrt{\xi_3(T,T')u}\right]\leq A\exp(-au).$$

Then, for any $u>0$:

$$P^{\otimes 2N}\left\{P^{\otimes 2N}\left[\xi_1(T,T')|T\right]\geq P^{\otimes 2N}\left[\xi_2(T,T')|T\right]+\sqrt{P^{\otimes 2N}\left[\xi_3(T,T')|T\right]u}\right\}\leq A\exp(1-au).$$

The proof of this lemma can be found in Panchenko's paper, [17]. We can now give the proof of Theorem 2.1.

Proof of Theorem 2.1. Let $(X_{N+1},...,X_{2N})$ be an independent copy of our sample $(X_1,...,X_N)$. Let us choose $k\in\{1,...,m\}$. Let us apply Lemma 7.1 with $\mathcal{P}=P$ and, for any $i\in\{1,...,2N\}$:

$$T_i=f_k(X_i).$$

We obtain, for any measurable function $\eta_k\in\mathbb{R}$, for any measurable function $\lambda_k:\mathbb{R}^{2N}\to\mathbb{R}_+^*$ that is exchangeable with respect to its $2\times 2N$ arguments:

$$P^{\otimes 2N}\exp\left(\frac{\lambda_k}{N}\sum_{i=1}^{N}\Big\{f_k(X_{i+N})-f_k(X_i)\Big\}-\frac{\lambda_k^2}{N^2}\sum_{i=1}^{2N}f_k(X_i)^2-\eta_k\right)\leq\exp\left(-\eta_k\right)$$

and the reverse inequality:

$$P^{\otimes 2N}\exp\left(\frac{\lambda_k}{N}\sum_{i=1}^{N}\Big\{f_k(X_i)-f_k(X_{i+N})\Big\}-\frac{\lambda_k^2}{N^2}\sum_{i=1}^{2N}f_k(X_i)^2-\eta_k\right)\leq\exp\left(-\eta_k\right)$$

as well. This implies that:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i)-f_k(X_{i+N})\Big\}\leq\frac{\lambda_k}{N^2}\sum_{i=1}^{2N}f_k(X_i)^2+\frac{\eta_k}{\lambda_k}\right]\leq\exp\left(-\eta_k\right)$$

and:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_{i+N})-f_k(X_i)\Big\}\leq\frac{\lambda_k}{N^2}\sum_{i=1}^{2N}f_k(X_i)^2+\frac{\eta_k}{\lambda_k}\right]\leq\exp\left(-\eta_k\right).$$

Let us choose:

$$\lambda_k=\sqrt{\frac{N^2\eta_k}{\sum_{i=1}^{2N}f_k(X_i)^2}}$$

in both inequalities, we obtain for the first one:

$$P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i)-f_k(X_{i+N})\Big\}\geq 2\sqrt{\frac{\eta_k\sum_{i=1}^{2N}f_k(X_i)^2}{N^2}}\right]\leq\exp\left(-\eta_k\right).$$

We now apply Lemma 7.2 with the same $T_i=f_k(X_i)$, $\eta_k=u$, $A=1$, $a=1$, $\xi_2=0$,

$$\xi_1=\frac{1}{N}\sum_{i=1}^{N}\Big\{f_k(X_i)-f_k(X_{i+N})\Big\}\quad\text{and}$$

$$\xi_3=\frac{4\sum_{i=1}^{2N}f_k(X_i)^2}{N^2}.$$

We obtain:

$$\begin{aligned}P^{\otimes N}\left[\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)-P\left[f_k(X)\right]\right.&\geq\left.2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)^2+P\left[f_k(X)^2\right]\right\}}{N}}\right]\\&=P^{\otimes 2N}\left[\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)-P\left[f_k(X)\right]\geq 2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)^2+P\left[f_k(X)^2\right]\right\}}{N}}\right]\\&\leq\exp\left(1-\eta_k\right).\end{aligned}\tag{7.1}$$

Remember that if there are no assumption on f, the f_k are bounded by C_k and so we have $P\left[f_k(X)^2\right]<C_k^2$. Now, if condition $\mathcal{H}(q)$ is satisfied for $1<q<+\infty$(see Sect. 2.1 p. 440), using Hölder's inequality we have:

$$P\left[f_k(X)^2\right]\leq\left(\int_{\mathcal{X}}\left|f_k(x)\right|^{2p}\lambda(\mathrm{d}x)\right)^{\frac{1}{p}}\left(\int_{\mathcal{X}}f(x)^q\lambda(\mathrm{d}x)\right)^{\frac{1}{q}}\leq c_k^2c=C_k^2.$$

If condition $\mathcal{H}(+\infty)$ is satisfied we have:

$$P\left[f_k(X)^2\right]=\inf_{\mathcal{X}}f_k^2(x)f(x)\lambda(\mathrm{d}x)\leq c\int_{\mathcal{X}}f_k^2(x)\lambda(\mathrm{d}x)=c=C_k^2.$$

So in any case, $P\left[f_k(X)^2\right]<C_k^2$. Now, let us combine this with Inequality 7.1 and with the reverse one by a union bound argument, we have:

$$P^{\otimes N}\left[\quad\left|\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)-P\left[f_k(X)\right]\right|\geq 2\sqrt{\frac{\eta_k\left\{\frac{1}{N}\sum_{i=1}^{N}f_k(X_i)^2+C_k^2\right\}}{N}}\quad\right]\leq 2\exp\left(1-\eta_k\right).$$

We now make a union bound on $k \in \{1, ..., m\}$ and put:

$$\eta_k = 1 + \log \frac{2m}{\varepsilon}.$$

We obtain:

$$P^{\otimes N}\left[\forall k \in \{1, ..., m\}, \quad |\hat{\alpha}_k - \overline{\alpha}_k| \leq \sqrt{\beta(\varepsilon, k)}\right] \geq 1 - \varepsilon.$$

This ends the proof. □

7.2. Proof of Proposition 2.4

Proof of Proposition 2.4. Given g and k, the computation of $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$, is quite easy. First, note that $\Pi_{\mathcal{CR}_{k,\varepsilon}} g = g + bf_k$ so we just have to compute the coefficient b. Moreover, the conditions $\Pi_{\mathcal{CR}_{k,\varepsilon}} g \in \mathcal{CR}_{k,\varepsilon}$ gives:

$$\|\langle g + bf_k, f_k\rangle f_k - \hat{\alpha}_k f_k\|^2 \leq \beta(\varepsilon, k)$$

or:

$$(\langle g, f_k\rangle + b - \hat{\alpha}_k)^2 \leq \beta(\varepsilon, k).$$

There are two possibilities. If this condition is satisfied for $b = 0$, this means that $g \in \mathcal{CR}_{k,\varepsilon}$ and so $\Pi_{\mathcal{CR}_{k,\varepsilon}} g = g$. Otherwise, $\Pi_{\mathcal{CR}_{k,\varepsilon}} g$ will lie on the boundary of $\mathcal{CR}_{k,\varepsilon}$, this means that b will satisfy:

$$\langle g, f_k\rangle + b - \hat{\alpha}_k \in \left\{\pm\sqrt{\beta(\varepsilon, k)}\right\}.$$

Finally, note that $|b|$ should be minimal. This leads to the following formula:

$$b = \hat{\alpha}_k - \langle g, f_k\rangle - sgn\left(\hat{\alpha}_k - \langle g, f_k\rangle\right)\sqrt{\beta(\varepsilon, k)}.$$

□

7.3. Proof of Theorem 4.1

Proof of Theorem 4.1. First of all, let us remind the following assertion for the weak Besov spaces:

$$W_{\rho,\pi}(D') = \left\{g : [0,1] \to \mathbb{R}, \;\; g(.) = \alpha\phi(.) + \sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k}(.), \;\; \sup_{\lambda>0}\lambda^{\rho}\sum_{j=0}^{\infty}2^{j\left(\frac{\pi}{2}-1\right)}\sum_{k=1}^{2^j}\mathbb{1}_{\{|\beta_{j,k}|>\lambda\}} < D'\right\}$$

$$= \left\{g : [0,1] \to \mathbb{R}, \;\; g(.) = \alpha\phi(.) + \sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k}(.), \;\; \sup_{\lambda>0}\lambda^{\pi-\rho}\sum_{j=0}^{\infty}2^{j\left(\frac{\pi}{2}-1\right)}\sum_{k=1}^{2^j}|\beta_{j,k}|^{\pi}\mathbb{1}_{\{|\beta_{j,k}|\leq\lambda\}} < D''\right\},$$

for some $D'' \in \mathbb{R}_+$, see Cohen [10] for a proof.

Let C be a generic constant in the whole proof. We let $\mathcal{E}(\varepsilon)$ denote the following event:

$$\left\{\forall k \in \{1, ..., m\}, \quad (\hat{\alpha}_k - \overline{\alpha}_k)^2 \leq \beta(\varepsilon, k)\right\}$$

(so, Th. 2.1 ensures that $P^{\otimes N}[\mathcal{E}(\varepsilon)] \geq 1 - \varepsilon$). We have:

$$P^{\otimes N}\mathrm{d}^2(\tilde{f}, f) = P^{\otimes N}\left[\mathbb{1}_{\mathcal{E}(\varepsilon)}\mathrm{d}^2(\tilde{f}, f)\right] + P^{\otimes N}\left[\left(1 - \mathbb{1}_{\mathcal{E}(\varepsilon)}\right)\mathrm{d}^2(\tilde{f}, f)\right].$$

For the first term we still have:

$$\mathrm{d}^2(\tilde{f}, f) \leq 2(m+1)c.$$

For the second term, let us write the development of f into our wavelet basis:

$$f = \alpha\phi + \sum_{j=0}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}\psi_{j,k},$$

and:

$$\hat{f}(x) = \tilde{\alpha}\phi + \sum_{j=0}^{J}\sum_{k=1}^{2^j}\tilde{\beta}_{j,k}\psi_{j,k}$$

the estimator $\hat{f}$. Let us put:

$$J = \left\lfloor \frac{\log N}{\log 2} \right\rfloor.$$

For any $J' \leq J$ we have:

$$\begin{aligned}
\mathrm{d}^2(\tilde{f}, f) &= \mathrm{d}^2(\Pi_{\mathcal{B}}\Pi_{\mathcal{CR}_{m,\varepsilon}}...\Pi_{\mathcal{CR}_{1,\varepsilon}}0, f) \leq \mathrm{d}^2(\Pi_{\mathcal{CR}_{m,\varepsilon}}...\Pi_{\mathcal{CR}_{1,\varepsilon}}0, f) \\
&= (\tilde{\alpha}-\alpha)^2 + \sum_{j=0}^{J}\sum_{k=1}^{2^j}(\tilde{\beta}_{j,k}-\beta_{j,k})^2 + \sum_{j=J+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2 \\
&\leq (\tilde{\alpha}-\alpha)^2 + \sum_{j=0}^{J'}\sum_{k=1}^{2^j}(\tilde{\beta}_{j,k}-\beta_{j,k})^2\mathbb{1}(|\beta_{j,k}|\geq\kappa) + \sum_{j=0}^{J'}\sum_{k=1}^{2^j}\beta_{j,k}^2\mathbb{1}(|\beta_{j,k}|<\kappa) \\
&\quad + \sum_{j=J'+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2
\end{aligned}$$

for any $\kappa \geq 0$, as soon as $\mathcal{E}(\varepsilon)$ is satisfied (here again we applied Th. 2.1). In the case where $p \geq 2$ we can take:

$$J' = \left\lfloor \frac{\log N^{\frac{1}{1+2s}}}{\log 2} \right\rfloor$$

and $\kappa = 0$ to obtain:

$$\sum_{j=J'+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2 \leq \sum_{j=J'+1}^{\infty}\left(\sum_{k=1}^{2^j}\beta_{j,k}^p\right)^{\frac{2}{p}} 2^{j\left(1-\frac{2}{p}\right)}.$$

As $f \in B_{s,p,q} \subset B_{s,p,\infty}$ we have:

$$\left(\sum_{k=1}^{2^j}\beta_{j,k}^p\right)^{\frac{2}{p}} \leq \mathrm{d}^2 2^{-2j\left(s+\frac{1}{2}-\frac{1}{p}\right)}$$

and so:

$$\sum_{j=J+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2 \leq \mathrm{d}^2 2^{-2J's} \leq \mathrm{d}^2 N^{\frac{-2s}{1+2s}},$$

and:

$$\sum_{j=0}^{J}\sum_{k=1}^{2^j}(\tilde{\beta}_{j,k}-\beta_{j,k})^2\mathbb{1}(|\beta_{j,k}|\geq\kappa)\leq\frac{8c\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\sum_{j=0}^{J}\sum_{k=1}^{2^j}1\leq\frac{8c\left[1+\log\frac{2m}{\varepsilon}\right]}{N}2^{J'+1}\leq\frac{8c\left[1+\log\frac{2m}{\varepsilon}\right]}{N}4N^{\frac{1}{1+2s}}.$$

So we obtain the desired rate of convergence. In the case where $p<2$ we let $J'=J$ and proceed as follows.

$$\begin{aligned}\sum_{j=0}^{J}\sum_{k=1}^{2^j}(\tilde{\beta}_{j,k}-\beta_{j,k})^2\mathbb{1}(|\beta_{j,k}|\geq\kappa)&\leq\frac{8c\left[1+\log\frac{2m}{\varepsilon}\right]}{N}\sum_{j=0}^{J}\sum_{k=1}^{2^j}\mathbb{1}(|\beta_{j,k}|\geq\kappa)\\&\leq\frac{8c\left[1+\log\frac{2m}{\varepsilon}\right]}{N}D'\kappa^{-\frac{2}{2s+1}}\end{aligned}$$

because f is also assumed to be in the weak Besov space. We also have:

$$\sum_{j=0}^{J}\sum_{k=1}^{2^j}\beta_{j,k}^2\mathbb{1}(|\beta_{j,k}|<\kappa)\leq D''\kappa^{2-\frac{2}{1+2s}}.$$

For the remainder term we use (see [13,15]):

$$B_{s,p,q}\subset B_{s-\frac{1}{p}+\frac{1}{2},2,q}$$

to obtain:

$$\sum_{j=J+1}^{\infty}\sum_{k=1}^{2^j}\beta_{j,k}^2\leq\mathrm{d}^2 2^{-2J\left(s+\frac{1}{2}-\frac{1}{p}\right)}\leq\mathrm{d}^2 2^{-J}$$

as $s>\frac{1}{p}$. Let us remember that:

$$\frac{N}{2}\leq m=2^J\leq N$$

and that $\varepsilon=N^{-3/2}$, and take:

$$\kappa=\sqrt{\frac{\log N}{N}}$$

to obtain the desired rate of convergence. □

7.4. **Proof of the theorems of Section 5**

Proof of Theorem 5.1. The technique used in the proof is due to Catoni [8]. Let us choose $k\in\{1,...,m\}$, and:

$$\beta\in\left(0,\frac{N}{C_k}\right).$$

We have, for any $\eta\in\mathbb{R}$:

$$P^{\otimes N}\exp\left\{\sum_{i=1}^{N}\log\left(1-\frac{\beta}{N}f_k(X_i)\right)-\eta\right\}\leq\exp\left\{N\log\Big(1-\frac{\beta}{N}P\left[f_k(X)\right]\Big)-\eta\right\}.$$

Let us choose:

$$\eta=\log\frac{2m}{\varepsilon}+N\log\Big(1-\frac{\beta}{N}P\left[f_k(X)\right]\Big).$$

We obtain:

$$P^{\otimes N}\exp\left\{\sum_{i=1}^{N}\log\left(1-\frac{\beta}{N}f_k(X_i)\right)-\log\frac{2m}{\varepsilon}-N\log\Big(1-\frac{\beta}{N}P\left[f_k(X)\right]\Big)\right\}\leq\frac{\varepsilon}{2m},$$

and so:

$$P^{\otimes N}\left\{\sum_{i=1}^{N}\log\left(1-\frac{\beta}{N}f_k(X_i)\right)\geq\log\frac{2m}{\varepsilon}+N\log\Big(1-\frac{\beta}{N}P\left[f_k(X)\right]\Big)\right\}\leq\frac{\varepsilon}{2m},$$

that becomes:

$$P^{\otimes N}\left\{P\left[f_k(X)\right]\geq\frac{N}{\beta}\left[1-\exp\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(1-\frac{\beta}{N}f_k(X_i)\right)-\frac{\log\frac{2m}{\varepsilon}}{N}\right]\right)\right\}\leq\frac{\varepsilon}{2m}.$$

We apply the same technique to:

$$P^{\otimes N}\exp\left\{\sum_{i=1}^{N}\log\left(1+\frac{\beta'}{N}f_k(X_i)\right)-\eta\right\}\leq\exp\left\{N\log\Big(1+\frac{\beta'}{N}P\left[f_k(X)\right]\Big)-\eta\right\}$$

to obtain the upper bound. We combine both result by a union bound argument. □

Proof of Theorem 5.3. Let us choose $(i,k)\in\{1,...,N\}\times\{1,...,m'\}$. Using Seeger's idea, we follow the preceding proof, replacing $P^{\otimes N}$ by P_i, and using the $N-1$ random variables:

$$\Big(f_{i,k}(X_j)\Big)_{\substack{j\in\{1,\dots,N\}\\ j\neq i}}$$

with

$$\eta=\log\frac{2m'N}{\varepsilon}+(N-1)\log\left(1-\frac{\beta}{N-1}P\left[f_{i,k}(X)\right]\right)$$

and we obtain:

$$P_i\exp\left\{\sum_{j\neq i}\log\left(1-\frac{\beta}{N-1}f_{i,k}(X_j)\right)-\log\frac{2m'N}{\varepsilon}-(N-1)\log\Big(1-\frac{\beta}{N-1}P\left[f_{i,k}(X)\right]\Big)\right\}\leq\frac{\varepsilon}{2m'N}.$$

Note that for any random variable H that is a function of the X_i:

$$P^{\otimes N}P_iH=P^{\otimes N}H.$$

So we conclude exactly in the same way as in the proof of the previous theorem and we obtain the claimed result. □

8. Conclusion

We gave a new algorithm for function selection in density estimation with quadratic loss that is able to deal with almost any classical dictionary (wavelets, kernels...), and we provided a proof that it is able to reach the minimax rate of convergence in the case of wavelets.

Simulations clearly shows that estimators obtained with this algorithm performs as well as the best known adaptative estimators, for various losses and different dictionaries. However, the theoretical optimality of our method in this more general setting remains an open problem.

Acknowledgements. I would like to thank my PhD advisor, Professor Olivier Catoni, for his kind and constant help, and Professors Patrice Bertail, Emmanuelle Gautherat and Hugo Harari-Kermadec for their remark that Panchenko's lemma could improve Theorem 2.1. Finally I would like to thank the anonymous referee for useful comments and corrections.

References

[1] H. Akaike, A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19** (1974) 716–723.
[2] P. Alquier, Iterative Feature Selection In Least Square Regression Estimation. *Ann. Inst. H. Poincaré B: Probab. Statist.* **44** (2008) 47–88.
[3] A. Barron, A. Cohen, W. Dahmen and R. DeVore, Adaptative Approximation and Learning by Greedy Algorithms, preprint (2006).
[4] G. Blanchard, P. Massart, R. Vert and L. Zwald, Kernel Projection Machine: A New Tool for Pattern Recognition. *Proceedings of NIPS* (2004).
[5] B.E. Boser, I.M. Guyon and V.N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, D. Haussler (ed.), ACM Press (1992) 144–152.
[6] T.T. Cai and L.D. Brown, Wavelet Estimation for Samples with Random Uniform Design. *Stat. Probab. Lett.* **42** (1999) 313–321.
[7] O. Catoni, Statistical learning theory and stochastic optimization, *Lecture Notes, Saint-Flour Summer School on Probability Theory (2001)*, Springer.
[8] O. Catoni, PAC-Bayesian Inductive and Transductive Learning, manuscript (2006).
[9] O. Catoni, A PAC-Bayesian approach to adaptative classification, *preprint Laboratoire de Probabilités et Modèles Aléatoires* (2003).
[10] A. Cohen, Wavelet methods in numerical analysis, in *Handbook of numerical analysis*, Vol. VII, North-Holland, Amsterdam (2000) 417–711.
[11] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, Philadelphia (1992).
[12] D.L. Donoho and I.M. Johnstone, Ideal Spatial Adaptation by Wavelets. *Biometrika* **81** (1994) 425–455.
[13] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, Density Estimation by Wavelet Thresholding. *Ann. Statist.* **24** (1996) 508–539.
[14] I.J. Good and R.A. Gaskins, Nonparametric roughness penalties for probability densities. *Biometrika* **58** (1971) 255–277.
[15] W. Härdle, G. Kerkyacharian, D. Picard and A.B. Tsybakov, *Wavelets, Approximations and Statistical Applications*. Lecture Notes in Statistics, Springer (1998).
[16] J.S. Marron and S.P. Wand, Exact Mean Integrated Square Error. *Ann. Statist.* **20** (1992) 712–736.
[17] D. Panchenko, Symmetrization Approach to Concentration Inequalities for Empirical Processes. *Ann. Probab.* **31** (2003) 2068–2081.
[18] R Development Core Team, R: A Language And Environment For Statistical Computing, *R Foundation For Statistical Computing*, Vienna, Austria, 2004. URL `http://www.R-project.org.`
[19] G. Ratsch, C. Schafer, B. Scholkopf and S. Sonnenburg, Large Scale Multiple Kernel Learning. *J. Machine Learning Research* **7** (2006) 1531–1565.
[20] J. Rissanen, Modeling by shortest data description. *Automatica* **14** (1978) 465–471.
[21] M. Seeger, PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification. *J. Machine Learning Res.* **3** (2002) 233–269.
[22] M. Tipping, The Relevance Vector Machine, in *Advances in Neural Information Processing Systems*, San Mateo, CA (2000). Morgan Kaufmann.
[23] A.B. Tsybakov, *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications, Springer (2004).
[24] V.N. Vapnik, *The nature of statistical learning theory*. Springer Verlag (1998).
[25] Zhao Zhang, Su Zhang, Chen-xi Zhang and Ya-zhu Chen, SVM for density estimation and application to medical image segmentation. *J. Zhejiang Univ. Sci. B* **7** (2006).