4

From Model Selection to Adaptive Estimation

Lucien Birgé[1]
Pascal Massart[2]

4.1 Introduction

Many different model selection information criteria can be found in the literature in various contexts including regression and density estimation. There is a huge amount of literature concerning this subject and we shall, in this paper, content ourselves to cite only a few typical references in order to illustrate our presentation. Let us just mention AIC, C_p or C_L, BIC and MDL criteria proposed by Akaike (1973), Mallows (1973), Schwarz (1978), and Rissanen (1978) respectively. These methods propose to select among a given collection of parametric models that model which minimizes an empirical loss (typically squared error or minus log-likelihood) plus some penalty term which is proportional to the dimension of the model. From one criterion to another the penalty functions differ by factors of $\log n$, where n represents the number of observations.

The reasons for choosing one penalty rather than another come either from information theory or Bayesian asymptotic computations or approximate evaluations of the risk on specific families of models. Many efforts were made to understand in what circumstances these criteria allow to identify the *right* model asymptotically (see Li (1987) for instance). Much less is known about the performances of the estimators provided by these methods from a *nonparametric point of view*. Let us consider the particular context of density estimation in $\mathbb{L}^2$ for instance. By a nonparametric point of view, we mean that the unknown density does not necessarily belong to any of the given models and that the best model should approximately realize the best trade-off between the risk of estimation within the model and the distance of the unknown density to the model. When the models have good approximation properties (following Grenander (1981) these models will be called sieves), an adequate choice of the penalty can produce *adaptive* estimators in the sense that they estimate a density of unknown

[1]Université Paris VI and URA CNRS 1321
[2]Université Paris Sud and URA CNRS 743

smoothness at the rate which one would get if the degree of smoothness were known. Notable results in that direction have been obtained by Barron & Cover (1991) who use the MDL criterion when the models are chosen as ε-nets and by Polyak & Tsybakov (1990) who select the order of a Fourier expansion via Mallow's C_p for regression. One should also mention the results on penalized spline smoothing by Wahba and various coauthors (see Wahba (1990) for an extensive list of references).

This paper is meant to illustrate by a few theorems and applications, mainly directed towards adaptive estimation in Besov spaces, the power and versatility of the method of penalized minimum contrast estimation on sieves. A more general approach to the theory will be given in the companion paper Barron, Birgé & Massart (1995). We shall here content ourselves to consider linear sieves and the particular contrast which defines projection estimators for density estimation. These restrictions will allow us to make an extensive use of a recent and very powerful exponential inequality of Talagrand (1994) on the fluctuations of empirical processes which greatly simplifies the presentation and proofs. The choice of the penalty derives from the control of the risk on a fixed sieve. From that respect our approach presents some similarity with the method of structural minimization of the risk of Vapnik (1982). Minimum contrast estimators on a fixed sieve have been studied in great detail in Birgé & Massart (1994). For projection estimators their results can roughly be summarized as follows: s is an unknown density in $\mathbb{L}^2(\mu)$ to be estimated using a projection estimator acting on a linear sieve S of dimension D and the loss function is proportional to the square of the distance induced by the norm. Under reasonable conditions on the structure of the space S one gets a quadratic risk of the order of $\|s-\pi(s)\|^2 + D/n$ if one denotes by $\pi(s)$ the projection of s on S. This is essentially the classical decomposition between the square of the bias and the variance. The presence of a D/n term corresponding to a D-dimensional approximating space is not surprising for those who are familiar with Le Cam's developments about the connections between the dimension (in the metric sense) of a space and the minimax risk on this space. One should see Le Cam (1973) and (1986, Chapter 16) for further details.

Our main purpose, in this paper, is to show that if we replace the single sieve S by a collection of linear sieves S_m, $m \in \mathcal{M}_n$, with respective dimensions D_m and suitable properties, and introduce a penalty function $\mathrm{pen}(m)$ of the form $\mathcal{L}(m)D_m/n$, one gets a risk which, up to some multiplicative constant, realizes the best trade-off between $\|s - s_m\|^2$ and $\mathcal{L}(m)D_m/n$. Here s_m is the best approximant of s in S_m and $\mathcal{L}(m)$ is either uniformly bounded or possibly of order $\log n$ when too many of the sieves have the same dimension D_m. Note also that $\mathrm{pen}(m)$ will be allowed to be random. We shall show that some more or less recently introduced methods of adaptive density estimation like the unbiased cross validation (Rudemo 1982), or the hard thresholding of wavelet empirical coefficients (Donoho, John-

stone, Kerkyacharian & Picard 1993) can be viewed as special instances of penalized projection estimators. In order to emphasize the flexibility and potential of the methods of penalization we shall play with different families of sieves and penalties and propose some new adaptive estimators especially in the context of wavelet expansions in nonhomogeneous Besov spaces and piecewise polynomials with non equally spaced knots.

4.2 The statistical framework

4.2.1 The model and the estimators

We observe n i.i.d. random variables $X_1, \ldots, X_n$ with values on some measurable space $\mathcal{X}$ and common density s with respect to some measure μ. We assume that s belongs to the Hilbert space $\mathbb{L}^2(\mu)$ with norm $\|\cdot\|$ and denote by $\|\cdot\|_p$ the norm in $\mathbb{L}^p(\mu)$ for $1 \leq p \leq \infty$ and $p \neq 2$. We first consider an N_n-dimensional linear subspace $\bar{\mathcal{S}}_n$ of $\mathbb{L}^2(\mu)$, then choose a finite family $\{\bar{S}_m \,|\, m \in \mathcal{M}_n\}$ of linear subspaces of $\bar{\mathcal{S}}_n$, each $\bar{S}_m$ being a D_m-dimensional subspace of $\mathbb{L}^2(\mu)$ and finally for each $m \in \mathcal{M}_n$ we take a convex subset $S_m \subset \bar{S}_m$. In most cases, $S_m = \bar{S}_m$. The set $\mathcal{M}_n$ usually depends on n and more generally all the elements bearing a subscript (like m or m') which belong to $\mathcal{M}_n$. In order to keep the notations as simple as possible we shall systematically omit the subscript n when m is already present and also when the dependence on n is clear from the context. All real numbers that we shall introduce and which are not indexed by m or n are "fixed constants". We shall also denote by $\mathcal{S}_n$ the union of the S_m's, by s_m and $\bar{s}_n$ the projections of s onto S_m and $\bar{\mathcal{S}}_n$ respectively, by $\mathbb{P}$ the joint distribution of the observations X_i's when s obtains and by $\mathbb{E}$ the corresponding expectation. The centered empirical operator ν_n on $\mathbb{L}^2(\mu)$ is defined by

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} t(X_i) - \int_{\mathcal{X}} t(x)s(x)d\mu(x) \quad \text{for all} \quad t \in \mathbb{L}^2(\mu).$$

Let us consider on $\mathcal{X} \times \mathcal{S}_n$ the contrast function $\gamma(x,t) = -2t(x) + \|t\|^2$ where $\|\cdot\|$ denotes the norm in $\mathbb{L}^2(\mu)$. The empirical version of this contrast is $\gamma_n(t) = (1/n)\sum_{i=1}^{n}\gamma(X_i,t)$. Minimizing $\gamma_n(t)$ over $\bar{S}_m$ leads to the classical projection estimator $\hat{s}_m$ on $\bar{S}_m$ and we shall denote by $\hat{s}_n$ the projection estimator on $\bar{\mathcal{S}}_n$. If $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ is an orthonormal basis of $\bar{S}_m$ one gets:

$$\hat{s}_m = \sum_{\lambda\in\Lambda_m} \hat{\beta}_\lambda\varphi_\lambda \quad \text{with} \quad \hat{\beta}_\lambda = \frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i) \quad \text{and} \quad \gamma_n(\hat{s}_m) = -\sum_{\lambda\in\Lambda_m}\hat{\beta}_\lambda^2.$$

In order to define the penalty function, we associate to each S_m a weight $L_m \geq 1$. The use of those weights will become clear later but let us just

mention here that in the examples, either $L_m \asymp 1$ or $L_m \asymp \log n$ depending on the number of sieves with the same dimension D_m. For each $m \in \mathcal{M}_n$ the value of the penalty function $\mathrm{pen}(m)$ is defined by

$$\mathrm{pen}(m) = \tilde{K}_m(X_1, \ldots, X_n)\frac{L_m D_m}{n} \tag{1}$$

where $\tilde{K}_m$ is a positive random variable independent of the unknown s. Typically one must think of $\tilde{K}_m$ as a fixed constant (independent of m and n) or as a random variable which is, with a large probability and uniformly with respect to m and n, bounded away from zero and infinity. Then, in both cases, $\mathrm{pen}(m)$ is essentially proportional to $L_m D_m/n$.

A penalized projection estimator (PPE for short) is defined as any $\tilde{s} \in S_{\tilde{m}} \subset \mathcal{S}_n$ such that

$$\gamma_n(\tilde{s}) + \mathrm{pen}(\tilde{m}) = \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \gamma_n(t) + \mathrm{pen}(m) \right) \quad \text{if} \quad \tilde{s} \in S_{\tilde{m}}. \tag{2}$$

If such a minimizer does not exist one rather takes an approximate minimizer and chooses $\tilde{s}$ satisfying

$$\gamma_n(\tilde{s}) + \mathrm{pen}(\tilde{m}) \leq \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \gamma_n(t) + \mathrm{pen}(m) \right) + \frac{1}{n}.$$

We shall assume in the sequel that (2) holds, the modifications needed to handle the extra $1/n$ term being straightforward.

In the sequel we shall distinguish between two different situations corresponding to different structures of the family of sieves: nested and non-nested. The nested situation can be described by the following assumption

N: Nested family of sieves *We assume that the integer N_n is given satisfying $N_n \leq n\Gamma^{-2}$ for some fixed constant Γ that $m \mapsto D_m$ is a one-to-one mapping, and that one of the two equivalent sets of assumptions holds:*

(i) *$\|u\|_\infty \leq \Phi\sqrt{D_m}\|u\|$ for all $m \in \mathcal{M}_n$ and $u \in \bar{S}_m$ where Φ is a fixed constant and $D_m \leq N_n$ for all m. Moreover, $D_m < D_{m'}$ implies that $\bar{S}_m \subset \bar{S}_{m'}$ and $S_m \subset S_{m'}$;*

(ii) *$\bar{\mathcal{S}}_n$ is a finite-dimensional subspace of $\mathbb{L}^2(\mu)$ with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ and the cardinality of $\bar{\Lambda}_n$ is $|\bar{\Lambda}_n| = N_n$. A family of subsets $\{\Lambda_m\}_{m \in \mathcal{M}_n}$ of $\bar{\Lambda}_n$ with $|\Lambda_m| = D_m$ is given, $\bar{S}_m$ is the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ and $\|\sum_{\lambda \in \Lambda_m} \varphi_\lambda^2\|_\infty \leq D_m\Phi^2$. Moreover for all m and m', the inequality $D_m < D_{m'}$ implies that $S_m \subset S_{m'}$ and $\Lambda_m \subset \Lambda_{m'}$.*

The equivalence between (i) and (ii) follows from Lemma 6 of Birgé & Massart (1994). Assumption **N** will typically be satisfied when $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ is either a bounded basis or a localized basis in natural order. In this case, the

usual choices for $\mathcal{M}_n$ will be either a finite subset of $\mathbb{N}$ (and then $m \mapsto D_m$ is increasing) or a totally ordered family of sets. In some situations, the basis $\{\varphi_\lambda\}_{\lambda\in\bar{\Lambda}_n}$ is given (Fourier expansions for instance) from which one defines $\bar{S}_m$. In other cases (piecewise polynomials for instance) one starts with the family $\{\bar{S}_m\}_{m\in\mathcal{M}_n}$ which is the natural object to consider.

In the non-nested situation we shall distinguish a particular situation which is of special interest:

Case S: Non-nested subsets of a basis *Let $\{\varphi_\lambda\}_{\lambda\in\bar{\Lambda}_n}$ be an orthonormal system in $\mathbb{L}^2(\mu)$ with $|\bar{\Lambda}_n| = N_n$ and $\bar{\mathcal{S}}_n$ be the linear span of $\{\varphi_\lambda\}_{\lambda\in\bar{\Lambda}_n}$. Each $m \in \mathcal{M}_n$ is a subset of $\bar{\Lambda}_n$ with cardinality D_m and $S_m = \bar{S}_m$ is the linear span of $\{\varphi_\lambda\}_{\lambda\in m}$.*

Particular choices of $\mathcal{M}_n$ and of the penalty function lead to various classical estimators. Here are three illustrations.

An analogue of Mallows' C_L

Assuming that **N** holds we define the penalty by $\mathrm{pen}(m) = K\Phi^2 D_m/n$. This gives a sequence of parametric problems with an increasing number of parameters and a penalty proportional to the number of parameters. This is an analogue in density estimation of Mallows' C_L method for the regression framework—see for instance Mallows (1973) or Li (1987).

Cross-validation

Assume again that **N** holds. A particular choice of the penalty function leads to a well-known method of selecting the order of an expansion:

Proposition 1 *Assume that we are in the nested situation described by Assumption* **N** *and that*

$$\mathrm{pen}(m) = \frac{2}{n(n+1)} \sum_{i=1}^{n} \sum_{\lambda\in\Lambda_m} \varphi_\lambda^2(X_i).$$

The resulting PPE $\tilde{s}$ is the projection estimator on $S_{\tilde{m}}$ where $\tilde{m}$ is chosen by the unbiased cross-validation method.

Proof: Let us recall that the ideal m (in view of minimizing the quadratic loss) should minimize $\|s - \hat{s}_{m'}\|^2$ or equivalently $\int \hat{s}_{m'}^2 - 2\int \hat{s}_{m'} s$ with respect to $m' \in \mathcal{M}_n$. Since this quantity involves the unknown s, it has to be estimated and the unbiased cross-validation method defines $\hat{m}$ as the minimizer with respect to $m \in \mathcal{M}_n$ of

$$\int \hat{s}_m^2 d\mu - \frac{2}{n(n-1)} \sum_{i\neq i'} \sum_{\lambda\in\Lambda_m} \varphi_\lambda(X_i)\varphi_\lambda(X_{i'}).$$

Since

$$\int \hat{s}_m^2 d\mu = \frac{1}{n^2} \sum_{i,i'} \sum_{\lambda\in\Lambda_m} \varphi_\lambda(X_i)\varphi_\lambda(X_{i'})$$

one finds $\hat{m}$ as the minimizer of

$$-\frac{n+1}{n-1}\int \hat{s}_m^2 d\mu + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2(X_i).$$

On the other hand the PPE selects $\tilde{m}$ as the minimizer of

$$\begin{aligned}\gamma_n(\hat{s}_m)+\mathrm{pen}(m) &= \int \hat{s}_m^2 d\mu - \frac{2}{n}\sum_{i=1}^{n}\hat{s}_m(X_i)+\mathrm{pen}(m)\\ &= -\int \hat{s}_m^2 d\mu + \frac{2}{n(n+1)}\sum_{i=1}^{n}\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2(X_i)\end{aligned}$$

which implies that $\hat{m}=\tilde{m}$ and the conclusion follows. □

In this case, assuming that $L_m = 1$, the estimator $\tilde{K}_m(X_1,\ldots,X_n)$ is given by

$$\frac{2}{n+1}\sum_{i=1}^{n}\frac{1}{m}\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2(X_i).$$

Threshold estimators

We now consider the situation described by Case **S**, $\mathcal{M}_n$ being the family of all (nonempty) subsets of $\bar{\Lambda}_n$ and $\mathrm{pen}(m)=\tilde{L}_n D_m/n$ where $\tilde{L}_n$ is a (possibly random) variable independent of m, we have to minimize over all possible subsets m of $\bar{\Lambda}_n$ the quantity

$$\gamma_n(\hat{s}_m)+\mathrm{pen}(m) = -\sum_{\lambda\in m}\hat{\beta}_\lambda^2 + \frac{\tilde{L}_n D_m}{n} = -\sum_{\lambda\in m}\left(\hat{\beta}_\lambda^2 - \frac{\tilde{L}_n}{n}\right).$$

The solution $\tilde{m}$ is the set of the λ's such that $\hat{\beta}_\lambda^2 > \tilde{L}_n/n$ which leads to a threshold estimator introduced, in the context of white noise models, by Donoho & Johnstone (1994)

$$\tilde{s} = \sum_{\lambda\in\bar{\Lambda}_n}\hat{\beta}_\lambda\varphi_\lambda \mathbb{I}_{\{\hat{\beta}_\lambda^2 > \tilde{L}_n/n\}}.$$

These three examples are indeed typical of the two major types of model selection for projection estimators: selecting the order of an expansion or selecting a subset of a basis. We shall later give a formal treatment of these two problems.

4.2.2 Besov spaces and examples of sieves

The target function s, in most of our illustrations, will be assumed to belong to some classical function spaces that we introduce below. We assume in this section that μ is Lebesgue measure.

Besov spaces

We shall consider here various forms of Besov spaces $B_{\alpha\, p\, \infty}(\mathcal{A})$ with $\alpha > 0$, $1 \leq p \leq \infty$, and three different types of supporting sets $\mathcal{A}$:

- Some compact interval which, without loss of generality, can be taken as $[0,1]$ and then $\mathcal{A} = [0,1]$;
- The torus $\mathbb{T}$ which we shall identify to the interval $[0, 2\pi]$, then $\mathcal{A} = [0, 2\pi]$ and we deal with periodic functions;
- Some compact interval $[-A, A]$ ($\mathcal{A} = [-A, A]$) but in this case we shall consider it as the restriction of the Besov space on the whole real line to the set of functions which have a compact support in $(-A, A)$.

Let us first recall some known facts on Besov spaces which can be found in the books by DeVore & Lorentz (1993) or Meyer (1990). Following DeVore & Lorentz (1993, page 44) we define the r-th order differences of a function t defined on $\mathcal{A}$ by

$$\Delta_h^r(t,x) = \sum_{k=0}^{r} \left(\begin{array}{c} r \\ k \end{array} \right) (-1)^{r-k} t(x+kh).$$

The Besov space $B_{\alpha\, p\, \infty}(\mathcal{A})$ will be the space of functions t on $\mathcal{X}$ such that

$$\sup_{y>0} y^{-\alpha} \omega_r(t,y)_p < +\infty \quad \text{where} \quad \omega_r(t,y)_p = \sup_{0<h\leq y} \|\Delta_h^r(t,.)\|_p$$

and $r = [\alpha] + 1$ (DeVore & Lorentz 1993, page 55). As a particular case, we get the classical Hölder spaces when $p = \infty$. One should notice that since we always work on a compact interval $\mathcal{A}$, $\mathbb{L}^p$-norms on $\mathcal{A}$ are easy to compare and $\omega_r(t,y)_p \geq C(p)\omega_r(t,y)_2$ for $p \geq 2$. This implies that $B_{\alpha\, p\, \infty}(\mathcal{A}) \subset B_{\alpha\, 2\, \infty}(\mathcal{A})$. Therefore, if $p \geq 2$ we can restrict ourselves to considering only the larger space $B_{\alpha\, 2\, \infty}(\mathcal{A})$ since we are looking for upper bounds for the risk.

Wavelet expansions

Let us consider an orthonormal wavelet basis $\{\varphi_{j,k} \,|\, j \geq 0, k \in \mathbb{Z}\}$ of $\mathbb{L}^2(\mathbb{R}, dx)$ (see Meyer (1990) for details) with the following conventions: $\varphi_{0,k}$ are translates of the father wavelet and for $j \geq 1$, the $\varphi_{j,k}$'s are affine transforms of the mother wavelet. One will also assume that these wavelets are compactly supported and have *regularity* r in the following sense: all their moments up to order r are 0. Let $t \in \mathbb{L}^2(\mathbb{R}, dx)$ be some function with compact support in $(-A, A)$. Changing the indexing of the basis if necessary we can write the expansion of t on the wavelet basis as:

$$t = \sum_{j \geq 0} \sum_{k=1}^{2^j M} \beta_{j,k} \varphi_{j,k}, \tag{3}$$

where $M \geq 1$ is a finite integer depending on A and the lengths of the wavelet's supports. For any $j \in \mathbb{N}$, we denote by $\Lambda(j)$ the set of indices $\{(j,k)\,|\,1 \leq k \leq 2^j M\}$ and if $m \subset \Lambda = \sum_{j\geq 0} \Lambda(j)$ we put $m(j) = m \cap \Lambda(j)$. Let B_0 denote the space of functions t such that $\Sigma_\infty(t) = \sum_{j\geq 0} 2^{j/2} \sup_{\lambda\in\Lambda(j)} |\beta_\lambda| < +\infty$. From Bernstein's inequality (Meyer 1990, Chapter 2, Lemma 8)

$$\|t\|_\infty \leq \Phi_\infty \Sigma_\infty(t) \qquad \text{for all} \qquad t \in B_0, \tag{4}$$

where Φ_∞ only depends on the choice of the basis. We also define $\bar{V}_J$, for $J \in \mathbb{N}$, to be the linear span of $\{\varphi_\lambda\,|\,\lambda \in \Lambda(j), 0 \leq j \leq J\}$; then $2^J M \leq \mathrm{Dim}(\bar{V}_J) = N < 2^{J+1}M$ and it follows from (4) that there exists a constant Φ, namely $\Phi^2 = 2\Phi_\infty^2/M$, such that

$$\|t\|_\infty \leq \Phi\sqrt{N}\|t\| \qquad \text{for all} \qquad t \in \bar{V}_J. \tag{5}$$

Let t be given by (3) with $\alpha < r+1$; if t belongs to the Besov space $B_{\alpha\,p\,\infty}([-A,A])$ then (Kerkyacharian & Picard 1992)

$$\sup_{j\geq 0} 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})} \left(\sum_{\lambda\in\Lambda(j)} |\beta_\lambda|^p \right)^{1/p} = |\!|\!| t |\!|\!| < +\infty. \tag{6}$$

One derives from equation (6) that $\sup_{\lambda\in\Lambda(j)} |\beta_\lambda| \leq 2^{-j(\alpha+\frac{1}{2}-\frac{1}{p})} |\!|\!| t |\!|\!|$ which proves the inclusion $B_{\alpha\,p\,\infty}([-A,A]) \subset B_0$ provided that $\alpha > 1/p$.

Piecewise polynomials

Without loss of generality we shall restrict our attention to piecewise polynomial spaces on $[0,1]$. A linear space S_m of piecewise polynomials is characterized by $m = (r, \{b_0 = 0 < b_1 < \ldots < b_D = 1\})$ where r is the maximal degree of the polynomials (that we shall essentially keep fixed in the sequel) and $\{b_0 = 0 < b_1 < \ldots < b_D = 1\}$ is a nondecreasing sequence which generates a partition of $[0,1]$ into D intervals. Such a space has the dimension $D_m = D(r+1)$. We shall distinguish between regular piecewise polynomials for which all intervals have the same length and general piecewise polynomials with arbitrary intervals subject to the restriction that their lenghts are multiples of a fixed value $1/N$ with $N \in \mathbb{N}$. In this case the b_j's are of the form N_j/N where the N_j's are integers and the corresponding set of m's will be denoted by $\mathcal{P}_N^r$. The reasons for restricting the values of the b_j's to a grid are given in Birgé & Massart (1994, Section 3) and we shall not insist on that. When dealing with regular partitions, we shall restrict, in order to get a nested family of sieves, to dyadic partitions generated by the grid $\{j2^{-J_m}, 0 \leq j \leq 2^{J_m}\}$ where J_m is an integer. The corresponding set of m's for $0 \leq J_m \leq J$ will be denoted by $\bar{\mathcal{P}}_J^r$.

We shall need hereafter some properties of these spaces of polynomials. Let us first recall (see Whittaker & Watson (1927, pp. 302-305) for details) that the Legendre polynomials $Q_j, j \in \mathbb{N}$ are a family of orthogonal polynomials in $\mathbb{L}^2([-1,1], dx)$ such that Q_j has degree j and

$$|Q_j(x)| \leq 1 \quad \text{for all} \quad x \in [-1,1], \quad Q_j(1) = 1, \quad \int_{-1}^{1} Q_j^2(t)dt = \frac{2}{2j+1}.$$

As a consequence, the family of polynomials $R_j(x) = \sqrt{2j+1}Q_j(2x-1)$ is an orthonormal basis for the space of polynomials on $[0,1]$ and if H is a polynomial of degree r such that $H(x) = \sum_{j=0}^{r} a_j R_j(x)$,

$$|H(x)|^2 \leq \left(\sum_{j=0}^{r} a_j^2\right)\left(\sum_{j=0}^{r} 2j+1\right) = (r+1)^2 \sum_{j=0}^{r} a_j^2.$$

Hence $\|H\|_\infty \leq (r+1)\|H\|$. Therefore any polynomial H of degree r on an interval $[a,b]$ satisfies $\|H\|_\infty \leq (r+1)(b-a)^{-1/2}\|H\|$ from which one deduces that for $H \in S_m$

$$\|H\|_\infty \leq \frac{r+1}{\sqrt{h}}\|H\| \quad \text{where} \quad h = \inf_{1 \leq j \leq D}\{b_j - b_{j-1} \,|\, b_j > b_{j-1}\}. \tag{7}$$

Therefore, if s is a function on $[a,b]$ and H_s its projection on the space of polynomials of degree $\leq r$ on $[a,b]$, one gets

$$\|H_s\|_\infty \leq \frac{r+1}{(b-a)^{1/2}}\|H_s\| \leq \frac{r+1}{(b-a)^{1/2}}\|s\| \leq (r+1)\|s\|_\infty \tag{8}$$

and this inequality remains true for the projections on spaces of piecewise polynomials since it only depends on the degree and not on the support.

4.3 Presentation of some of the results

From now on, we shall have to introduce various constants to set up the assumptions, describe the penalty function, state the results and produce the proofs. In order to clarify the situation we shall stick to some fixed conventions and give to the letters κ, C (or c) and K, with various sub- or supscripts, a special meaning. The constants used to set up the assumptions will be denoted by various letters but the three letters above will be reserved. $\kappa_1, \ldots$ denote universal (numerical) constants which are kept fixed throughout the paper. $K, K', \ldots$ are constants to be chosen by the statistician or to be used as generic constants in some assumptions. Finally $C, c, C', \ldots$ denote constants which come out from the computations and proofs and depend on the other constants given in the assumptions. One shall also use $C(\cdot, \cdot, \cdots)$ to indicate more precisely the dependence on

various quantities and especially those which are related to the unknown s. The value of K or C is fixed throughout a proof but, in order to keep the notations simple, we shall use the same notation for different constants when one goes from one proof or example to another.

Before giving the formal results let us describe a few typical and illustrative examples (more will be given later) of applications of these results together with a sketch of proof in order to make them more appealing. We shall distinguish between the two situations described above: nested and non-nested.

4.3.1 Nested models

We assume that **N** holds and $S_m = \bar{S}_m$ and we choose either a deterministic or a random penalty function of the form

$$\text{pen}(m) = K\Phi^2 \frac{D_m}{n} \qquad \text{or} \qquad \text{pen}(m) = \frac{K}{n(n+1)} \sum_{i=1}^{n} \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i)$$

where K is a suitably chosen constant. We recall from Section 4.2.1 that the choice $K = 2$ corresponds to Mallows' C_L or cross-validated estimators. We shall prove below that under such assumptions, one gets, as expected

$$\mathbb{E}[\|\tilde{s} - s\|^2] \leq C \inf_{m \in \mathcal{M}_n} [\|s_m - s\|^2 + D_m/n].$$

Assuming that the true s belongs to some unknown Besov space $B_{\alpha\, 2\, \infty}$ with $\alpha > 0$ and choosing a convenient basis with good approximation properties with respect to such spaces (wavelet basis, dyadic splines or Fourier basis), we shall get the usual and optimal $n^{-\alpha/(2\alpha+1)}$ rate of convergence for our penalized estimator (see Example 1 below).

Remarks: The constant $\sqrt{K}$ should be larger than some universal constant involved in some suitable exponential inequality. A reasonable conjecture (by analogy with the gaussian case) is that a lower bound for K is one which means that our results should hold for the classical cross-validated estimators.

4.3.2 Selecting a subset of a wavelet basis

We consider the wavelet basis of regularity r and the notations introduced in Section 4.2.2 and assume that Case **S** obtains with $\bar{\Lambda}_n = \sum_{0 \leq j \leq J_n} \Lambda(j)$ where J_n is given by $2^{J_n} \asymp n/(\log^2 n)$. Then the dimension N_n of $\bar{\mathcal{S}}_n$ satisfies $2^{J_n} M < N_n < 2^{J_n+1} M$ and $D_m = |m|$.

Thresholding

$\mathcal{M}_n$ is taken to be *all* the subsets of $\bar{\Lambda}_n$ and the penalty function is given by $K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K')(\log n)|m|/n$. As mentioned in Section 4.2.1, the PPE

is then a threshold estimator. We recall that $\hat{s}_n$ is the projection estimator on the largest sieve $\bar{\mathcal{S}}_n$. It comes from (4) that $\Phi_\infty \Sigma_\infty(\hat{s}_n)$ is an estimator of an upper bound of $\|s\|_\infty$ provided that s belongs to B_0. In this case we shall prove that

Proposition 2 *Let $\tilde{s}$ be the threshold estimator given by*

$$\tilde{s} = \sum_{\lambda \in \bar{\Lambda}_n} \hat{\beta}_\lambda \varphi_\lambda \mathbb{I}_{\{\hat{\beta}_\lambda^2 > \tilde{T}\}}, \qquad \text{with} \quad \tilde{T} = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K') \log n/n$$

where K has to be larger than a universal constant and K' is an arbitrary positive number. Provided that s belongs to B_0, the following upper bound holds for any $q \geq 1$,

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C \inf_{m \in \mathcal{M}_n} \left[\|s_m - s\|^2 + \log n \frac{D_m}{n} \right]^{q/2} \tag{9}$$

as soon as

$$\Phi_\infty \left[\Sigma_\infty(s) - \Sigma_\infty(\bar{s}_n) \right] \leq K'. \tag{10}$$

Either one knows an upper bound for $\Phi_\infty \Sigma_\infty(s)$ and one should choose K' to be this upper bound or (10) will hold only for n large enough. Assuming that s belongs to some unknown Besov space $B_{\alpha\, p\, \infty}$, with $r + 1 > \alpha > 1/p$ (and therefore $s \in B_0$) the resulting rate of convergence is $(\log n/n)^{\alpha/(2\alpha+1)}$. There is an extra power of $\log n$ in the rate but it should be noticed that (9) holds for a set of densities which is larger than the Besov spaces. With a different thresholding strategy, the same rates have been obtained by Donoho et al. (1993).

Special strategy for Besov spaces

We introduce a smaller family of sieves which has the same approximation properties *in the Besov spaces* than the previous one. It can be described as follows. Let us first introduce an auxiliary positive and decreasing function l defined on $[1, +\infty)$ with $l(1) < 1$. For each pair of integers J, j' with $0 \leq j' \leq J$, let $\mathcal{M}_J^{j'}$ be the collection of subsets m of Λ such that $m(j) = \Lambda(j)$ for $0 \leq j \leq j'$ and $|m(j)| = [|\Lambda(j)| l(j - j')]$ for $j' < j \leq J$, where $[x]$ denotes the integer part of x. We define $\mathcal{M}_n = \sum_{0 \leq j' \leq J_n} \mathcal{M}_{J_n}^{j'}$ and $\text{pen}(m) = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K') L|m|/n$ where K will be larger than a universal constant, K' is an arbitrary positive number and L is a fixed weight depending on l. The resulting penalized estimator $\tilde{s}$ will then satisfy

Proposition 3 *Let s belong to B_0 and (10) be satisfied. If the function l is such that $2^{-j'} \log |\mathcal{M}_{J_n}^{j'}|$ is bounded by a fixed constant then, for any $q \geq 1$,*

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C \inf_{m \in \mathcal{M}_n} \left[\|s_m - s\|^2 + \frac{D_m}{n} \right]^{q/2}.$$

We shall see in Example 4 that one can choose l satisfying the required conditions and such that the resulting bias leads to the right rate of convergence $n^{-\alpha/(2\alpha+1)}$ for all Besov spaces $B_{\alpha\,p\,\infty}$, with $r+1>\alpha>1/p$ simultaneously.

4.3.3 Variable weights and piecewise polynomials

Up to know we only considered situations where the weights L_m did not depend on m. The following example is meant to illustrate the advantage of letting the weights vary with the models in some cases. It deals with piecewise polynomials. Let us fix some maximal degree r for our polynomials and take $2^{J_n} \asymp n/\log^2 n$. We consider the family of sieves $\{S_m\}_{m\in\mathcal{M}_n}$ where $\mathcal{M}_n = \mathcal{P}^r_{2^{J_n}}$. The S_m's are the corresponding piecewise polynomials of degree $\leq r$ described in Section 4.2.2. One should notice that since $\bar{\mathcal{P}}^r_{J_n} \subset \mathcal{M}_n$, this family of sieves includes in particular piecewise polynomials based on regular dyadic partitions. Let us define $L_m = 1$ when $m \in \bar{\mathcal{P}}^r_{J_n}$ and $L_m = \log n$ otherwise. In this situation, it is wiser to choose $\bar{\mathcal{S}}_n$ as the space of piecewise polynomials based on the finest possible partition generated by the sequence $\{j2^{-J_n}\}_{0\leq j\leq 2^{J_n}}$ and with degree $2r$ instead of r; then $N_n = (2r+1)2^{J_n}$. With such a choice the squares of the elements of all the sieves will belong to $\bar{\mathcal{S}}_n$.

Proposition 4 *Let us choose* $\mathrm{pen}(m) = K(\|\hat{s}_n\|_\infty + K')L_m D_m/n$ *and assume that* s *is bounded, then the PPE* $\tilde{s}$ *satisfies, for any* $q \geq 1$,

$$\mathbb{E}[\|\tilde{s}-s\|^q] \leq C \inf_{m\in\mathcal{M}_n}\left[\|s-s_m\|^2 + \frac{L_m D_m}{n}\right]^{q/2}.$$

For an arbitrary s, the method hunts for a partition which provides, up to a $\log n$ factor, the best trade-off between the dimension of the partition and the bias. But if s belongs to some Besov space $B_{\alpha\,2\,\infty}$ with $\alpha < r+1$, then the estimator achieves the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$.

4.3.4 Sketch of the proofs

In order to prove results of the form

$$\mathbb{E}[\|s-\tilde{s}\|^q] \leq C_0 \inf_{m\in\mathcal{M}_n}\left[\|s-s_m\|^2 + L_m D_m/n\right]^{q/2}$$

we always follow the same basic scheme (with various additional technicalities). m is defined as the minimizer with respect to $m' \in \mathcal{M}_n$ of $\|s_{m'}-s\|^2 + L_{m'}D_{m'}/n$. Using a powerful result of Talagrand (1994), we begin to prove that with probability larger than $1-p_{m'}\exp(-c\sqrt{\xi})$, for any $m' \in \mathcal{M}_n$ and uniformly for $t \in S_{m'}$

$$\nu_n(t-s_m) \leq \frac{1}{4}\left(\|t-s\|^2 + \|s_m-s\|^2\right) + \frac{\xi}{n} + C\left[\frac{L_{m'}D_{m'}}{n} + \frac{L_m D_m}{n}\right]. \quad (11)$$

By assumption, the $L_{m'}$'s are chosen in such a way that $\sum_{m' \in \mathcal{M}_n} p_{m'} \leq C_1$ which implies that the control (11) holds for all m' simultaneously with probability larger than $1 - C_1 \exp(-c\sqrt{\xi})$. In particular (11) holds with $t = \tilde{s}$ and $m' = \tilde{m}$. We then use the following simple lemma:

Lemma 1 *Let $\tilde{s} = \hat{s}_{\tilde{m}}$ be the PPE associated with the penalty function* $\mathrm{pen}(\cdot)$, *m a given element of $\mathcal{M}_n$ and s_m the projection of the true underlying density s onto S_m. The following inequality holds:*

$$\|s - \tilde{s}\|^2 \leq \|s - s_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\tilde{m}) + 2\nu_n(\tilde{s} - s_m). \qquad (12)$$

Proof: The conclusion follows from the fact that $\gamma_n(\tilde{s}) + \mathrm{pen}(\tilde{m}) \leq \gamma_n(s_m) + \mathrm{pen}(m)$ and the following inequalities:

$$\gamma_n(t) = \|t\|^2 - 2\nu_n(t) - 2\int t s d\mu \quad \text{for all} \quad t \in \mathcal{S}_n;$$

$$\begin{aligned} \|s - t\|^2 - \|s - s_m\|^2 &= \|t\|^2 - \|s_m\|^2 + 2\int (s_m - t)s \, d\mu \\ &= \mathbb{E}\left(\gamma_n(t) - \gamma_n(s_m)\right). \qquad \square \end{aligned}$$

Using (11) and (12) simultaneously we get

$$\|s - \tilde{s}\|^2 \leq 3\|s - s_m\|^2 + 2[\mathrm{pen}(m) - \mathrm{pen}(\tilde{m})] + \frac{4\xi}{n} + 4C\left[\frac{L_{\tilde{m}} D_{\tilde{m}}}{n} + \frac{L_m D_m}{n}\right].$$

If $\mathrm{pen}(m')$ is defined in such a way that for all $m' \in \mathcal{M}_n$,

$$2\,C\frac{L_{m'} D_{m'}}{n} \leq \mathrm{pen}(m') \leq K\frac{L_{m'} D_{m'}}{n},$$

one gets with probability larger than $1 - C_1 \exp(-c\sqrt{\xi})$

$$\|s - \tilde{s}\|^2 \leq 3\|s - s_m\|^2 + (4C + 2K)\frac{L_m D_m}{n} + \frac{4\xi}{n}.$$

One concludes using the following elementary lemma since $L_m D_m \geq 1$.

Lemma 2 *Let X be a nonnegative random variable satisfying $X^2 \leq a + K_1 t/n$ with probability larger than $1 - K_2 \exp(-K_3\sqrt{t})$ for all $t > 0$. Then for any number $q \geq 1$*

$$\mathbb{E}[X^q] \leq 2^{(q/2-1)^+}\left[a^{q/2} + K_2\Gamma(q+1)\left(\frac{K_1}{nK_3}\right)^{q/2}\right].$$

The case of a random penalty requires extra arguments but the basic ideas are the same.

4.4 The theorems

4.4.1 Talagrand's Theorem

All our results rely upon an important theorem of Talagrand (1994) which can be considered, if stated in a proper form, as an analogue of an inequality for Gaussian processes by Cirel'son, Ibragimov & Sudakov (1976). Let us first recall this inequality in the case of a real-valued non-centered process in order to emphasize the similarity between the two results.

Theorem 1 *Let $X_t, t \in T$ be a real valued gaussian process with bounded sample paths and $v = \sup_t \mathrm{Var}(X_t)$. Then for $\xi > 0$*

$$\mathbb{P}\left[\sup_t (X_t - \mathbb{E}[X_t]) \geq \mathbb{E}\left[\sup_t (X_t - \mathbb{E}[X_t])\right] + \xi\right] \leq \frac{2}{\sqrt{2\pi v}} \int_\xi^{+\infty} e^{-x^2/(2v)} dx \leq \exp\left[-\frac{1}{2}\frac{\xi^2}{v}\right].$$

Although Talagrand did not state his theorem[3] in such a form one can actually write it as follows:

Theorem 2 *Let $X_1, \ldots, X_n$ be n i.i.d. random variables, $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. random signs ($+1$ or -1 with probability $1/2$) independent of the $X_i's$ and $\{f_t, t \in T\}$ a family of functions that are uniformly bounded by some constant b. Let $v = \sup_{t\in T} \mathrm{Var}(f_t(X_1))$. There exists universal constants $\kappa_2 \geq 1$ and κ_1 such that for any positive ξ*

$$\mathbb{P}\left[\sup_t \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n f_t(X_i) - \mathbb{E}[f_t(X_i)]\right) \geq \kappa_2 \mathbb{E}\left[\sup_t \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f_t(X_i)\right|\right] + \xi\right] \leq \exp\left[-\kappa_1\left(\frac{\xi^2}{v} \wedge \frac{\xi\sqrt{n}}{b}\right)\right]. \quad (13)$$

In the sequel κ_1 and κ_2 will always denote the two constants appearing in (13).

Talagrand's Theorem has many useful statistical consequences, such as the following extension of a result from Mason & van Zwet (1987): the control of χ^2-type statistics $\mathcal{K}_{n,D}^2 = \sum_{j=1}^D (X_j - np_j)^2/np_j$ with $(X_1, \ldots, X_k)$ a multinomial random vector with distribution $\mathcal{M}(n, p_1, \ldots, p_k)$ and $D \leq k$. If $\delta = \inf_{1\leq j\leq D} p_j$, $x > 0$ and $\varepsilon > 0$, the following inequality holds:

$$\mathbb{P}\left[\mathcal{K}_{n,D}^2 \geq (1+\varepsilon)\kappa_2^2 D + x\right] \leq 2\exp\left[\frac{-\kappa_1 \varepsilon x}{1+\varepsilon}\left(1 \wedge \sqrt{\frac{n\delta}{x}}\right)\right]. \quad (14)$$

[3] During the final revision of our article we became aware of improvements by Talagrand (1995) and Ledoux (1995) of Theorem 2 that might lead to more explicit lower bounds for our penalty functions.

This inequality implies Mason and van Zwet's inequality (Mason & van Zwet 1987, Lemma 3) when $x \leq n\delta$ and provides more information on the tail distribution of $\mathcal{K}^2_{n,D}$ since it holds without any restriction on x. The proof is given in Section 4.6.

4.4.2 Selecting the order of an expansion

In this section, we shall restrict ourselves to the nested case. The first result deals with the analogue of Mallows' C_L.

Theorem 3 *Assume that* **N** *holds, choose some positive* θ *and define the penalty function by* $\mathrm{pen}(m) = (\kappa_2^2\Phi^2 + \theta)D_m/n$. *Then for any* $q \geq 1$

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C(q, \|s\|, \Phi, \Gamma, \theta) \inf_{m \in \mathcal{M}_n} \left[\frac{D_m}{n} + \|s_m - s\|^2\right]^{q/2}. \qquad (15)$$

In view of Proposition 1, the following theorem applies to cross-validated projection estimators provided that the conjecture $\kappa_2 = 1$ is true, but cross-validation would also make sense with different values of the constant K_n in (16) below.

Theorem 4 *Assume that* **N** *is satisfied, that* S_m *is the linear span of* $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ *for all* m*'s, and that*

$$\inf_{m \in \mathcal{M}_n} \frac{1}{D_m} \int \left(\sum_{\lambda \in \Lambda_m} \varphi_\lambda^2\right) s d\mu = a > 0;$$

$$\mathrm{pen}(m) = \frac{K_n}{n(n+1)} \sum_{i=1}^{n} \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i) \quad \textit{with} \quad K_n = \frac{n+1}{n}(\kappa_2^2 + \theta) \qquad (16)$$

for some positive θ. *Then for any* $q \geq 1$

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C(q, \|s\|, \Phi, \Gamma, a, \theta) \inf_{m \in \mathcal{M}_n} \left[\frac{D_m}{n} + \|s_m - s\|^2\right]^{q/2}. \qquad (17)$$

4.4.3 Extension

In order to analyze some particular types of estimators which were first proposed by Efroimovich (1985) (see Example 2 below), it is useful to have some more general version of Theorem 3. $\bar{\mathcal{S}}_n$ is a finite-dimensional space with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ and $|\bar{\Lambda}_n| = N_n$. Let $\bar{\mathcal{M}}_n$ be a finite collection of sets (not necessarily totally ordered by inclusion but having a maximal element $\bar{m}_n$) and $m' \mapsto \Lambda_{m'}$ be an increasing mapping from $\bar{\mathcal{M}}_n$ into the family of nonvoid subsets of $\bar{\Lambda}_n$. We define $S_{m'}$ as the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$; then $D_{m'} = |\Lambda_{m'}|$. Let $\mathcal{M}_n$ be a totally ordered subfamily of $\bar{\mathcal{M}}_n$ containing $\bar{m}_n$ and for which Assumption **N** holds. One defines for

each $m' \in \bar{\mathcal{M}}_n$ an associate $\tau(m') \in \mathcal{M}_n$ which is the smallest m such that $m \supset m'$. Assuming that the penalty function satisfies the inequality

$$(\kappa_2^2\Phi^2 + \theta)D_{\tau(m')}/n \leq \text{pen}(m') \leq KD_{\tau(m')}/n,$$

it is easily seen that the bound (15) still holds for the PPE $\tilde{s}$ based on the larger family of sieves $\{S_m\}_{m\in\bar{\mathcal{M}}_n}$, where s_m is defined as before, since the proof only involves the larger spaces $S_{\tau(m')}$ and the values of the penalty function. If we assume that for each $m' \in \bar{\mathcal{M}}_n$

$$D_{m'} \geq \delta D_{\tau(m')} \tag{18}$$

for some fixed positive constant δ, the penalty $\text{pen}(m') = \delta^{-1}(\kappa_2^2\Phi^2 + \theta)D_{m'}/n$ will satisfy the above inequality and since $\|s_{m'} - s\| \geq \|s_{\tau(m')} - s\|$ the following bound remains valid:

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq \delta^{-q/2}C(q, \|s\|, \Phi, \Gamma, \theta) \inf_{m'\in\bar{\mathcal{M}}_n} \left[\frac{D_{m'}}{n} + \|s_{m'} - s\|^2\right]^{q/2}.$$

4.4.4 Selecting a subset of a basis

Let us now study the more general situation of a rich and possibly non-nested family of sieves. We shall use the assumption

B: *$R_n(s) = \sup_{t\in\bar{\mathcal{S}}_n} \|t\|^{-2} \int t^2 sd\mu$ is finite and there exists a family of weights $L_m \geq 1$, $m \in \mathcal{M}_n$ and a fixed constant Δ such that*

$$\sum_{m\in\mathcal{M}_n} \exp(-L_m D_m) \leq \Delta. \tag{19}$$

Our first theorem deals with a bounded situation where $S_m \neq \bar{S}_m$.

Theorem 5 *Assume that $\|t\|_\infty \leq B_n$ for all $t \in \mathcal{S}_n$ and that* **B** *holds with $R_n(s) \leq B_n$, $\text{pen}(m)$ being possibly random. Then, for any $q \geq 1$,*

$$\begin{aligned}&\mathbb{E}[\|\tilde{s} - s\|^q \mathbb{I}_{\tilde{\Omega}}] \\ &\quad\leq\ C(q)\left[\inf_{m\in\mathcal{M}_n}\left[\|s - s_m\|^q + \mathbb{E}[(\text{pen}(m))^{q/2}\mathbb{I}_{\tilde{\Omega}}]\right] + \Delta(B_n/n)^{q/2}\right]\end{aligned}$$

if $\tilde{\Omega}$ is defined by

$$\tilde{\Omega} = \{\text{pen}(m) \geq \kappa_1^{-1}(3 + 5\sqrt{2}\kappa_2 + 4\kappa_2^2)B_n L_m D_m/n \quad \textit{for all} \quad m \in \mathcal{M}_n\}.$$

The boundedness restrictions on $\mathcal{S}_n$ and $R_n(s)$ being rather unpleasant we would like to be dispensed with them. A more general situation can be handled in the following way. We recall that $\bar{s}_n$ is the projection of s on $\bar{\mathcal{S}}_n$, $\hat{s}_n$ the projection estimator defined on $\bar{\mathcal{S}}_n$ and t_m the projection of t on S_m.

Theorem 6 *Assume that* **B** *holds, μ is a finite measure and there exists some orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\bar\Lambda_n}$ of $\bar{\mathcal{S}}_n$ which satisfies $\|\varphi_\lambda\|_\infty \le \Phi\sqrt{N_n}$ for all $\lambda\in\bar\Lambda_n$ with $N_n = |\bar\Lambda_n| = n/(\theta_n \log n)$ where Φ is a fixed constant and $\{\theta_k\}_{k\ge1}$ a sequence converging to infinity. Suppose that a real function ψ is given on $\bar{\mathcal{S}}_n$ such that for all $t\in\bar{\mathcal{S}}_n$ and $m\in\mathcal{M}_n$, $\|t_m\|_\infty \le \psi(t)$ and*

$$|\psi(\bar s_n) - \psi(\hat s_n)| \le \Phi'\sqrt{N_n}\sup_{\lambda\in\bar\Lambda_n}|\nu_n(\varphi_\lambda)|. \tag{20}$$

Let us define the penalty function by

$$\mathrm{pen}(m) = K\frac{L_m D_m}{n}(\psi(\hat s_n)+K'), \quad \textit{with} \quad K \ge \frac{2}{\kappa_1}(3+5\sqrt{2}\kappa_2+4\kappa_2^2) \tag{21}$$

where K' is a positive constant to be chosen by the statistician. Then, for any $q\ge1$,

$$\begin{aligned}\mathbb{E}[\|\tilde s - s\|^q] \;\le\; & C(q,K,K',\Delta,\Psi(s))\inf_{m\in\mathcal{M}_n}\left[\|s-s_m\| + \frac{L_m D_m}{n}\right]^{q/2} \\ & + n^{-q/2}C'(q,K,K',\Phi,\Phi',\{\theta_k\},\Psi(s),\|s\|)\end{aligned}$$

provided that the following conditions are satisfied:

$$R_n(s) \le \psi(\bar s_n) + K' \qquad \textit{and} \qquad \Psi(s) = \sup_n \psi(\bar s_n) < +\infty. \tag{22}$$

4.5 Examples

4.5.1 Nested models

Example 1 We assume here that the true s belongs to some unknown Besov space $B_{\alpha\,2\,\infty}(\mathcal{A})$ and that $\mathcal{M}_n = \{0,\dots,J_n\}$. If $\mathcal{A} = [-A,A]$, let $J_n = [\log n]$, $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ be a wavelet basis of regularity r and $\Lambda_m = \sum_{0\le j\le m}\Lambda(j)$, then S_m is the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$.If $\mathcal{A} = [0,1]$, S_m is the space of piecewise polynomials of degree $\le r$ based on the dyadic partition generated by the grid $\{j2^{-m}, 0\le j\le 2^m\}$ and $J_n = [\log n]$. If $\mathcal{A} = \mathbb{T}$, S_m is the set of trigonometric polynomials of degree $\le m$ and $J_n = n$. Provided that $\alpha < r+1$, the approximation properties of s by s_m which are collected in Section 4.7.1 lead, for each of our three cases, to a bias control of the form $\|s-s_m\| \le C(s)D_m^{-\alpha}$. Assumption N (ii) is satified by the Fourier basis and N (i) by the piecewise polynomials because of (7) and by the wavelets by (5). Therefore Theorem 3 applies and choosing m in such a way that $D_m \asymp n^{1/(1+2\alpha)}$ we get a rate of convergence of order $n^{-\alpha/(2\alpha+1)}$ provided that $\alpha < r+1$ except for the trigonometric polynomials which allow to deal with all values of α simultaneously.

One can also use the cross-validated estimator defined in Theorem 5 if we assume that (16) is satisfied.

Example 2 $\bar{\mathcal{S}}_n$ is a finite-dimensional space with an orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\bar{\Lambda}_n}$ and $|\bar{\Lambda}_n| = N_n$. Let $\bar{\mathcal{M}}_n$ be the collection of all subsets of $\{1,\ldots,J_n\}$ and $\{\Lambda(j)\}_{1\le j\le J_n}$ be a partition of $\bar{\Lambda}_n$. For any $m'\in\bar{\mathcal{M}}_n$ we define $\Lambda_{m'} = \sum_{j\in m'}\Lambda(j)$ and $S_{m'}$ to be the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_{m'}}$. Let $\mathcal{M}_n$ be the collection of the sets $\{1,\ldots,j\}$ with $1\le j\le J_n$ and assume that **N** holds for $\mathcal{M}_n$. Let us choose the penalty function $\mathrm{pen}(m') = K|\Lambda_{m'}|/n$. Then the corresponding PPE will minimize

$$\sum_{j\in m'}\left[-\sum_{\lambda\in\Lambda(j)}\hat{\beta}_\lambda^2 + |\Lambda(j)|\frac{K}{n}\right]$$

with respect to m' and the solution is clearly to keep only those indices j for which $\sum_{\lambda\in\Lambda(j)}\hat{\beta}_\lambda^2 \ge |\Lambda(j)|K/n$. This is the level thresholding estimator introduced by Efroimovich (1985) for trigonometric expansions and more recently in Kerkyacharian, Picard & Tribouley (1994) with wavelets expansions. We can deal with this example using the extension of Theorem 3 given in Section 4.4.3 provided that the dimension of $S_{m'}$ has the required property (18). This property will be clearly satisfied if $|\Lambda(j)| \ge (1+\rho)|\Lambda(j-1)|$ for all j's and some $\rho > 0$. Comparing these results with those of Example 1 we notice that this method performs exactly as the methods described in Example 1. This means that in such a case one cannot do better with the larger family $\{S_{m'}\}_{m'\in\bar{\mathcal{M}}_n}$ than with the simpler one $\{S_m\}_{m\in\mathcal{M}_n}$.

4.5.2 Selecting a subset of a wavelet basis

We shall now provide some details and proofs about the results announced in Section 4.3.2. We follow hereafter the notations and assumptions of that section. We recall that it follows from (5) that $\|\varphi_\lambda\|_\infty \le \Phi\sqrt{N_n}$ for all $\lambda\in\bar{\Lambda}_n$ and that N_n has been chosen of order $n/(\log^2 n)$ so that all the structural assumptions concerning the basis which are required to apply Theorem 6 are satisfied.

Example 3 (Thresholding) Following the set-up given in Section 4.3.2 we want to apply Theorem 6, $\mathcal{M}_n$ being the family of *all* the subsets of $\bar{\Lambda}_n$ and the penalty function being given by $\mathrm{pen}(m) = K(\Phi_\infty\Sigma_\infty(\hat{s}_n) + K')\log nD_m/n$.

Proof of Proposition 2: Note first that $R_n(s)\le\|s\|_\infty < +\infty$ since $s\in B_0$. The number of models m with a given cardinality $|m| = D$ is bounded by $\begin{pmatrix} N_n \\ D\end{pmatrix} < (eN_n/D)^D$. Hence Assumption **B** holds with our choice $L_m = \log n$. Following (4), let us choose $\psi(t) = \Phi_\infty\Sigma_\infty(t)$ for all $t\in\bar{\mathcal{S}}_n$. Then

$$|\psi(\bar{s}_n) - \psi(\hat{s}_n)| \le \Phi_\infty\Sigma_\infty(\bar{s}_n - \hat{s}_n) \le \Phi_\infty \sup_\lambda|\nu_n(\varphi_\lambda)|\sum_{j=0}^{J_n}2^{j/2}$$

which implies (20). It remains to check (22) which is immediate from (10) and $R_n(s) \leq \|s\|_\infty \leq \Phi_\infty \Sigma_\infty(s)$. □

When applied to Besov spaces, Proposition 2 gives

Corollary 1 *Assume that s belongs to some Besov space $B_{\alpha,p,\infty}$ with $r+1 > \alpha > 1/p$. Then the threshold estimator described above satisfies, for any $q \geq 1$,*

$$\mathbb{E}[\|\tilde{s} - s\|^q] = \mathcal{O}\Big((\log n/n)^{q\alpha/(1+2\alpha)}\Big)$$

provided that (10) holds.

Proof: Since s belongs to B_0, Proposition 2 applies and provides an upper bound for the risk of the form $C[\|s_m - s\|^2 + \log n D_m/n]^{q/2}$ for any subset m of $\bar{\Lambda}_n$. Although Proposition 6 of Section 4.7.2 has been designed for the smaller family of sieves to be considered in the next example, we can a fortiori apply it with the larger collection of sieves that we are handling here. The choices $J = J_n$ and $2^{j'} \asymp (n/\log n)^{1/(1+2\alpha)}$ ensure the existence of some m such that

$$D_m = \mathcal{O}\left(\left(\frac{n}{\log n}\right)^{1/(1+2\alpha)}\right) \quad \text{and} \quad \|s - s_m\|^2 = \mathcal{O}\left(\left(\frac{n}{\log n}\right)^{-2\alpha/(1+2\alpha)}\right)$$

which leads to the expected rate. □

Example 4 (Special strategy for Besov spaces) We follow the set-up given in Section 4.3.2. Let us first give more information about the computation of the estimator. Since the penalty has the form $\tilde{K}L|m|/n$, the estimator will take a rather simple form, despite the apparent complexity of the family of sieves. We have to minimize over the values of m in $\mathcal{M}_n = \sum_{0 \leq j' \leq J_n} \mathcal{M}_{J_n}^{j'}$ the quantity $\tilde{K}L|m|/n - \sum_{\lambda \in m} \hat{\beta}_\lambda^2$. This optimization can be carried out in two steps, first with respect to $m \in \mathcal{M}_{J_n}^{j'}$ for fixed j' and then with respect to j'. The first step amounts to minimize

$$\sum_{j' < j \leq J_n} \left[\frac{\tilde{K}L}{n}|m(j)| - \sum_{\lambda \in m(j)} \hat{\beta}_\lambda^2\right].$$

Since for a given j, $|m(j)|$ is fixed, the operation amounts to selecting the set $\hat{m}^{j'}(j)$ corresponding to the largest $[|\Lambda(j)|l(j-j')]$ coefficients $|\hat{\beta}_\lambda|$ for each $j > j'$. This is analogous but different from a thresholding procedure. Instead of selecting the coefficients which are larger than some threshold, one merely fixes the number of coefficients one wants to keep equal to $|m(j)|$ and takes the largest ones. For each j' the minimization of the criterion leads to the element $\hat{m}^{j'}$ of $\mathcal{M}_{J_n}^{j'}$. One should notice that all the elements of

$\mathcal{M}_{J_n}^{j'}$ have the same cardinality $2^{j'}Q(j')$. Therefore one selects j' in order to minimize

$$-\sum_{\lambda\in\hat{m}^{j'}}\hat{\beta}_\lambda^2+\tilde{K}L2^{j'}Q(j')/n$$

which only requires a few comparisons because the number of j''s is of the order of $\log n$. We now want to apply Theorem 6 to the family $\mathcal{M}_n$ with $\mathrm{pen}(m)=K(\Phi_\infty\Sigma_\infty(\hat{s}_n)+K')LD_m/n$.

Proof of Proposition 3: The proof follows exactly the lines of the preceding proof with the same function ψ, the only difference being the new choice of the weight L which is now independent of n. Since $|m|\geq M2^{j'}$ for all $m\in\mathcal{M}_{J_n}^{j'}$, we get

$$\sum_{m\in\mathcal{M}_n}\exp(-\kappa_1 L|m|)\leq\sum_{j'}\exp(-\kappa_1 LM2^{j'}+\log|\mathcal{M}_{J_n}^{j'}|).$$

(19) follows if we choose $L\geq 2M^{-1}\sup_{j'}(2^{-j'}\log|\mathcal{M}_{J_n}^{j'}|)$ which achieves the proof. □

Let us now conclude with an evaluation of the risk of $\tilde{s}$ when the target function s belongs to some Besov space. From now on, let l be the function $l(x)=x^{-3}2^{-x}$.

Corollary 2 *Assume that s belongs to some Besov space $B_{\alpha,p,\infty}$ with $r+1>\alpha>1/p$. Then the estimator $\tilde{s}$ satisfies, for any $q\geq 1$,*

$$\mathbb{E}[\|\tilde{s}-s\|^q]=\mathcal{O}\Big(n^{-q\alpha/(1+2\alpha)}\Big)$$

provided that (10) holds.

Proof: It follows the lines of the proof of Corollary 1 by applying again Proposition 6 which is exactly tuned for our needs. One chooses $2^{j'}\asymp n^{1/(1+2\alpha)}$ and one concludes by Proposition 3. □

4.5.3 Variable weights and piecewise polynomials

Example 5 We exactly follow the definition of the family of sieves given in Section 4.3.3. and try to apply Theorem 6.

Proof of Proposition 4: $R_n(s)$ is clearly bounded by $\|s\|_\infty$. Since there is at most one sieve per dimension when $m\in\bar{\mathcal{P}}_{J_n}^r$ and since the number of different partitions including D nonvoid intervals (and therefore corresponding to a sieve of dimension $D(r+1)$) is bounded by $(e2^{J_n}/D)^D$, (19) is satisfied and Assumption **B** holds. Recalling that whatever m and $t\in S_m$, $t^2\in\bar{\mathcal{S}}_n$, we conclude that $R_n(s)\leq\|\bar{s}_n\|_\infty$. Let $\psi(t)=\|t\|_\infty$. Since by (8) $\psi(\bar{s}_n)\leq(2r+1)\|s\|_\infty$, (22) is satisfied. It remains to find a basis of

$\bar{\mathcal{S}}_n$ with the required properties. We take the basis which is the union of the Legendre polynomials on each elementary intervals. Due to the properties of these polynomials mentioned in Section 4.2.2, the required bound on $\|\varphi_\lambda\|_\infty$ holds. Finally, denoting by I_j the interval $[(j-1)2^{-J_n}, j2^{-J_n})$ we get by (7)

$$\begin{aligned}\|\psi(\hat{s}_n) - \psi(\bar{s}_n)\|_\infty &= \sup_{1\le j\le 2^{J_n}} \|(\hat{s}_n - \bar{s}_n)\mathbb{I}_{I_j}\|_\infty \\ &\le \sup_{1\le j\le 2^{J_n}} (2r+1)2^{J_n/2}\|(\hat{s}_n - \bar{s}_n)\mathbb{I}_{I_j}\| \\ &\le (2r+1)2^{J_n/2}\sqrt{2r+1}\sup_{\lambda\in\bar{\Lambda}_n} \nu_n(\varphi_\lambda)\end{aligned}$$

which gives (20) and Theorem 6 applies. □

Following the arguments of Example 1, one can conclude that the estimator will reach the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ for all Besov spaces $B_{\alpha\,2\,\infty}([0,1])$ with $\alpha < r+1$ since in this case the best choice of m corresponds to a regular partition and therefore $L_m = 1$. For other densities, the risk comes within a $\log n$ factor to the risk obtained by the estimator build on the best partition if s were known.

Remarks: A similar strategy of introducing variable weights could be applied in the same way to deal with the situation described in Example 3. It would lead to similar results and give the right rate of convergence in Besov spaces $B_{\alpha\,2\,\infty}([0,1])$ when $1/2 < \alpha < r+1$. But the resulting estimator would not be a thresholding estimator anymore since the penalty would not be proportional to the dimension of the sieve.

4.6 Proofs

4.6.1 Inequalities for χ^2 statistics

Let $\|a\|$ denote the euclidean norm in $\mathbb{R}^D$. Inequality (13) implicitly contains the following bound on χ^2-type statistics which is of independent interest.

Proposition 5 *Let $X_1, \cdots, X_n$ be i.i.d. random variables and $Z_n = \sqrt{n}\nu_n$ the corresponding normalized empirical operator. Let $\varphi_1, \cdots, \varphi_D$ be a finite set of real functions. Let $v = \sup_{\|a\|\le 1} \mathbb{E}[(\sum_{j=1}^D a_j\varphi_j(X_1))^2]$ and $b^2 = \|\sum_{j=1}^D \varphi_j^2\|_\infty$. The following inequality holds for all positive t and ε:*

$$\mathbb{P}\left[\sum_{j=1}^D Z_n^2(\varphi_j) \ge (1+\varepsilon)\kappa_2^2((Dv)\wedge b^2) + x\right] \le 2\exp\left[\frac{-\kappa_1\varepsilon}{1+\varepsilon}\left(\frac{x}{v}\wedge\frac{\sqrt{nx}}{b}\right)\right].$$

Proof: We denote by Z'_n the symmetrized empirical process defined by $Z'_n(f) = n^{-1/2}\sum_{i=1}^n \varepsilon_i f(X_i)$, where the ε_i's are independent Rademacher random variables independent from the X_i's. Let

$$Y = \sup_{\|a\|\le 1} \left| Z_n\left(\sum_{j=1}^D a_j\varphi_j\right)\right| \qquad \text{and} \qquad Y' = \sup_{\|a\|\le 1} \left| Z'_n\left(\sum_{j=1}^D a_j\varphi_j\right)\right|.$$

From the well known duality formula $\sup_{\|a\|\le 1} |\sum_{j=1}^D a_j b_j| = \|b\|$ and the linearity of Z_n and Z'_n we derive that

$$\text{(i)} \quad Y = \left[\sum_{j=1}^D Z_n^2(\varphi_j)\right]^{1/2} \qquad \text{and} \qquad \text{(i')} \quad Y' = \left[\sum_{j=1}^D Z_n^{'2}(\varphi_j)\right]^{1/2}.$$

In order to apply Theorem 2 we first control Y'. It comes from (i') and Jensen's inequality that

$$\mathbb{E}(Y') \le \left[\sum_{j=1}^D \mathbb{E}(Z_n^{'2}(\varphi_j))\right]^{1/2} \le \left[\sum_{j=1}^D \mathbb{E}(\varphi_j^2(X_1))\right]^{1/2} \le \sqrt{Dv} \wedge b$$

and therefore Theorem 2 yields

$$\mathbb{P}\left[Y \ge \kappa_2(\sqrt{Dv}\wedge b) + \xi\right] \le 2\exp\left[-\kappa_1\left(\frac{\xi^2}{v} \wedge \frac{\xi\sqrt{n}}{b}\right)\right].$$

Since for $\varepsilon > 0$, $(\alpha+\beta)^2 \le \alpha^2(1+\varepsilon)+\beta^2(1+\varepsilon^{-1})$, we get for $x = (1+\varepsilon^{-1})\xi^2$

$$\mathbb{P}[Y^2 \ge (1+\varepsilon)\kappa_2^2((Dv)\wedge b^2) + x] \le 2\exp\left[-\frac{\kappa_1}{1+\varepsilon^{-1}}\left(\frac{x}{v}\wedge\frac{\sqrt{nx}}{b}\right)\right]$$

and the result follows from (i). □

In order to enlight the power of this bound, let us see what it gives for the standard χ^2 statistics. We want to prove (14). Considering a partition of $[0,1]$ by intervals $(I_j)_{1\le j\le D}$ such that the length of each I_j is equal to p_j and applying Proposition 5 with X_i uniformly distributed on $[0,1]$, $\varphi_j = (1/\sqrt{p_j})\mathbb{I}_{I_j}$, $v = 1$ and $b^2 = 1/\delta \ge D$ we get the required bound (14) for the χ^2 statistics $\mathcal{K}^2_{n,D}$ which has the same distribution as $\sum_{j=1}^D Z_n^2(\varphi_j)$.

4.6.2 Proof of Theorems 3 and 4

Without loss of generality we can assume that the index set $\mathcal{M}_n$ is chosen in such a way that $m = D_m$ and that $\bar{\mathcal{S}}_n$ is the largest of the S_m's, which we shall assume throughout the proof. Let m be some element of $\mathcal{M}_n$ which minimizes the sum $m/n + \|s - s_m\|^2, m'$ an arbitrary element in $\mathcal{M}_n$ and

$t \in S_{m'}$. We define $w(t) = \|s-t\| + \|s-s_m\|$ and apply Talagrand's Theorem to the family of functions $f_t = (t-s_m)/w(t)$ for $t \in S_{m'}$. It will be convenient to use the following form of Theorem 2 which is more appropriate for our needs:

$$\mathbb{P}\left[\sup_t \nu_n(f_t) \geq \kappa_2 \mathbb{E} + \xi\right] \leq \exp\left[-n\kappa_1\left(\frac{\xi^2}{v} \wedge \frac{\xi}{b}\right)\right] \tag{23}$$

with

$$b = \sup_t \|f_t\|_\infty; \quad v = \sup_t \mathrm{Var}(f_t(X)); \quad \mathbb{E} = \mathbb{E}\left[\sup_t \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_t(X_i)\right|\right].$$

To control $\mathbb{E}$, we shall distinguish between two cases:
(a) If $m \leq m'$ then $s_m \in S_{m'}$. Let $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$ be an orthonormal basis of $\bar{S}_{m'}$. Then $t - s_m = \sum_{\lambda \in \Lambda_{m'}} \beta_\lambda \varphi_\lambda$ and

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i (t-s_m)(X_i)\right)^2\right] &\leq \|t-s_m\|^2 \sum_{\lambda \in \Lambda_{m'}} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \varphi_\lambda(X_i)\right)^2\right] \\ &= \frac{1}{n}\sum_{\lambda \in \Lambda_{m'}} \mathbb{E}[\varphi_\lambda^2(X_1)]\|t-s_m\|^2.\end{aligned}$$

Since $w(t) \geq \|t-s_m\|$ and Assumption **N** holds we get

$$\mathbb{E}^2 \leq \frac{1}{n}\int \Psi_{m'}^2 s d\mu \qquad \text{where} \qquad \Psi_{m'}^2 = \sum_{\lambda \in \Lambda_{m'}} \varphi_\lambda^2 \leq \Phi^2 m'. \tag{24}$$

(b) If $m > m'$, one uses the decomposition $t - s_m = (t - s_{m'}) + (s_{m'} - s_m)$ and the inequalities $w(t) \geq \|s-t\| \geq \|s - s_{m'}\| \geq \|s_m - s_{m'}\|$ and $\|s-t\| \geq \|s_{m'} - t\|$ to get by similar arguments

$$\begin{aligned}\mathbb{E} &\leq \mathbb{E}\left[\sup_t \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{(t-s_{m'})(X_i)}{w(t)}\right|\right] + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{(s_m - s_{m'})(X_i)}{\inf_t w(t)}\right|\right] \\ &\leq \left(\frac{1}{n}\int \Psi_{m'}^2 s d\mu\right)^{1/2} + \left(\frac{1}{n}\frac{\int (s_m - s_{m'})^2 s d\mu}{\|s_m - s_{m'}\|^2}\right)^{1/2}.\end{aligned} \tag{25}$$

One concludes in both cases that $\mathbb{E} \leq E_{m'}$ with

$$E_{m'} = \left(\frac{1}{n}\int \Psi_{m'}^2 s d\mu\right)^{1/2} + \mathbb{I}_{\{m' < m\}} \Phi \sqrt{\frac{m}{n}}. \tag{26}$$

Let us now fix $\eta > 0$, $\bar{m} = m \vee m'$ and for $\xi > 0$ define $x_{m'} = x_{m'}(\xi)$ by $n x_{m'}^2 = \xi^2 + \eta \bar{m}$. Notice that for any $u \in S_m$ and $t \in S_{m'}$, $\|t-u\|_\infty \leq \|t-u\|\Phi\sqrt{\bar{m}}$ from which we get

$$\|f_t\|_\infty \leq \frac{\|t-s_m\|_\infty}{\|t-s_m\|} \leq \Phi\sqrt{\bar{m}}; \;\; \mathrm{Var}(f_t(X)) \leq \frac{\int (s_m - t)^2 s d\mu}{\|t-s_m\|^2} \leq \|s\|\Phi\sqrt{\bar{m}}.$$

Then (23) implies that

$$\mathbb{P} = \mathbb{P}\left[\sup_{t\in S_{m'}} \frac{\nu_n(t-s_m)}{w(t)} \geq \kappa_2 E_{m'} + x_{m'}\right] \leq \exp\left[\frac{-n\kappa_1}{\Phi\sqrt{\bar{m}}}\left(\frac{x_{m'}^2}{\|s\|} \wedge x_{m'}\right)\right].$$

Since $nx_{m'}^2 \geq \sqrt{\eta\bar{m}/2}(\xi+\sqrt{\eta\bar{m}})$; $x_{m'}\sqrt{2n} \geq \xi+\sqrt{\eta\bar{m}}$ and $m' \leq \bar{m} \leq n\Gamma^{-2}$, one gets

$$\begin{aligned}\mathbb{P} &\leq \exp\left[-\frac{\kappa_1}{\Phi}\left(\xi+\sqrt{\eta\bar{m}}\right)\left(\frac{\sqrt{\eta/2}}{\|s\|} \wedge \frac{\sqrt{n}}{\sqrt{2\bar{m}}}\right)\right] \\ &\leq \exp\left[-\frac{\kappa_1}{\Phi\sqrt{2}}\left(\xi+\sqrt{\eta m'}\right)\left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma\right)\right].\end{aligned}$$

Denoting by Ω_ξ the following event

$$\Omega_\xi = \left\{\sup_{t\in S_{m'}} \frac{\nu_n(t-s_m)}{w(t)} \leq \kappa_2 E_{m'} + x_{m'}(\xi) \quad \text{for all} \quad m' \in \mathcal{M}_n\right\} \tag{27}$$

we see that since the m''s are all different positive integers

$$1-\mathbb{P}[\Omega_\xi] \leq \exp\left[-\xi\frac{\kappa_1}{\Phi\sqrt{2}}\left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma\right)\right]\sum_{j=1}^{\infty}\exp\left[-\sqrt{j}\frac{\kappa_1\sqrt{\eta}}{\Phi\sqrt{2}}\left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma\right)\right].$$

If Ω_ξ is true, Lemma 1 implies that

$$\begin{aligned}\|s-\tilde{s}\|^2 &\leq \|s-s_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\tilde{m}) \\ &\quad + 2(\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi))(\|s-\tilde{s}\| + \|s-s_m\|).\end{aligned} \tag{28}$$

We shall again distinguish between two cases and repetedly use the inequality $2ab \leq \alpha^2 a^2 + \alpha^{-2}b^2$.
(a) If $\tilde{m} < m$ one applies (26) and (24) to get

$$\left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right]^2 \leq 8\kappa_2^2\Phi^2\frac{m}{n} + \frac{2}{n}\left[\xi^2+\eta m\right] = \frac{2}{n}\left[m(4\Phi^2\kappa_2^2+\eta)+\xi^2\right],$$

from which (28) becomes

$$\begin{aligned}\|s-\tilde{s}\|^2 &\leq \|s-s_m\|^2 + \mathrm{pen}(m) + \frac{1}{2}(\|s-\tilde{s}\|^2 + \|s-s_m\|^2) \\ &\quad + \frac{8}{n}\left[m(4\Phi^2\kappa_2^2+\eta)+\xi^2\right]\end{aligned}$$

and finally

$$\|s-\tilde{s}\|^2 \leq \frac{16}{n}\left[m(4\Phi^2\kappa_2^2+\eta)+\xi^2\right] + 3\|s-s_m\|^2 + 2\mathrm{pen}(m). \tag{29}$$

(b) If $\tilde{m} \geq m$, one chooses two real numbers α and $\beta \in (0,1)$ and applies the following inequalities

$$2\|s - s_m\| \left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right] \leq \alpha^2 \left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right]^2 + \alpha^{-2}\|s - s_m\|^2;$$

$$2\|s - \tilde{s}\| \left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right] \leq \beta^2 \left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right]^2 + \beta^{-2}\|s - \tilde{s}\|^2;$$

$$\left[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)\right]^2 \leq (1+\alpha^2)\left((\kappa_2 E_{\tilde{m}})^2 + \alpha^{-2} x^2_{\tilde{m}}(\xi)\right)$$

together with (28) to derive

$$\begin{aligned}(1-\beta^{-2})\|s-\tilde{s}\|^2 \quad &\leq \quad (\alpha^2+\beta^2)(1+\alpha^2)\left[(\kappa_2 E_{\tilde{m}})^2 + \alpha^{-2} x^2_{\tilde{m}}(\xi)\right] \\ &\quad + (1+\alpha^{-2})\|s-s_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\tilde{m}).\end{aligned}$$

Since by (26), $nE^2_{\tilde{m}} = \int \Psi^2_{\tilde{m}} s d\mu$ and $nx^2_{\tilde{m}} = \xi^2 + \eta\tilde{m}$ when $\tilde{m} \geq m$, we get

$$(1-\beta^{-2})\|s-\tilde{s}\|^2 \leq (1+\alpha^{-2})\left(\|s-s_m\|^2 + (\alpha^2+\beta^2)\frac{\xi^2}{n}\right) + \mathrm{pen}(m) \quad (30)$$

provided that the penalty satisfies for all $m' \in \mathcal{M}_n$

$$\mathrm{pen}(m') \geq \frac{(\alpha^2+\beta^2)(1+\alpha^2)}{n}\left[\kappa_2^2 \int \Psi^2_{m'} s d\mu + \frac{\eta m'}{\alpha^2}\right]. \quad (31)$$

Under the assumptions of Theorem 3 we can apply (24) and (31) will hold provided that α, η and $1-\beta$ are small enough depending on θ. One then derives from (29) and (30) that in both cases with probability greater than $1 - \mathbb{P}[\Omega_\xi]$

$$\|s-\tilde{s}\|^2 \leq C_1\|s-s_m\|^2 + C_2\frac{m}{n} + C_3\frac{\xi^2}{n}.$$

Theorem 3 then follows from Lemma 2.

To prove Theorem 4 we first apply (24) and Hoeffding's inequality to get

$$\mathbb{P}\left[\left|\nu_n\left(\Psi^2_{m'}\right)\right| > \varepsilon m'\right] \leq 2\exp\left[\frac{-2n\varepsilon^2 m'^2}{4\Phi^2 m'^2}\right]$$

for any positive ε, which implies that $\mathbb{P}(\Omega_n^c) \leq 2n\Gamma^{-2}\exp[-(n\varepsilon^2)/(2\Phi^2)]$ if we denote by Ω_n the event

$$\left\{\left|\frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_{m'}} \varphi^2_\lambda(X_i) - \int \Psi^2_{m'} s d\mu\right| \leq \varepsilon m' \quad \text{for all} \quad m' \in \mathcal{M}_n\right\}.$$

If Ω_n is true

$$\begin{aligned}\mathrm{pen}(m') \quad &\geq \quad \frac{K_n}{n+1}\left(\int \Psi^2_{m'} s d\mu - \varepsilon m'\right) \\ &\geq \quad \frac{1}{n}\left(\kappa_2^2 + \frac{\theta}{3}\right)\int \Psi^2_{m'} s d\mu + \frac{\theta a m'}{3n} + \frac{\theta m'}{3n}\left(a - \varepsilon\frac{3nK_n}{\theta(n+1)}\right)\end{aligned}$$

for all $m' \in \mathcal{M}_n$ and (31) will then be satisfied provided that we choose $\alpha, \eta, \varepsilon$ and $1-\beta$ small enough, depending only on κ_2, θ and a. In order to conclude we notice that on Ω_n^c

$$\|s-\tilde{s}\|^2 \leq 2\left(\|s\|^2 + \sum_{\lambda\in\bar{\Lambda}_n} \hat{\beta}_j^2\right) \leq 2\left(\|s\|^2 + \left(\frac{n\Phi}{\Gamma^2}\right)^2\right)$$

since $\bar{\mathcal{S}}_n$ is one of the S_m's and therefore $|\hat{\beta}_\lambda| \leq \|\varphi_\lambda\|_\infty \leq \Phi\sqrt{N_n}$. Hence

$$\mathbb{E}\left[\|s-\tilde{s}\|^q \mathbb{I}_{\Omega_n^c}\right] \leq \left[2(\|s\|^2 + n^2\Phi^2\Gamma^{-4})\right]^{q/2} \frac{2n}{\Gamma^2} \exp\left[-\frac{n\varepsilon^2}{2\Phi^2}\right].$$

On the other hand on Ω_n (31) is satisfied and $\mathrm{pen}(m)$ is bounded by

$$\mathrm{pen}(m) \leq \frac{K_n}{n+1}\left(\int \Psi_m^2 s d\mu + \varepsilon m\right) \leq \frac{1}{n}(\kappa_2^2+\theta)(\Phi^2 m + \varepsilon m)$$

and finally by (30)

$$\|s-\tilde{s}\|^2 \mathbb{I}_{\Omega_n} \leq C_1\|s-s_m\|^2 + C_2\frac{m}{n} + C_3\frac{\xi^2}{n}$$

which allows us to conclude by Lemma 2.

4.6.3 Proof of Theorems 5 and 6

Let m be some fixed element of $\mathcal{M}_n$, m' an arbitrary element in $\mathcal{M}_n$ and $t \in S_{m'}$. Once again, we want to apply Theorem 2 to the family of functions $f_t = (t-s_m)/w(t)$ where $w(t) = (\|s-t\| + \|s-s_m\|) \vee 2x_{m'}$ with $t \in S_{m'}$, $x_{m'}^2 = x_{m'}^2(\xi) = B_n(\xi^2 + \kappa_1^{-1} D_{m'} L_{m'})/n$ and $\xi \geq 1$. We get

$$\|f_t\|_\infty \leq \frac{\|t-s_m\|_\infty}{2x_{m'}} \leq \frac{2B_n}{2x_{m'}}; \quad \mathrm{Var}(f_t(X)) \leq \frac{\int (s_m-t)^2 s d\mu}{\|t-s_m\|^2} \leq B_n;$$

$$\mathbb{E} \leq \sqrt{B_n D_{m'}/n} + \sqrt{B_n/n} \leq E_{m'} = \sqrt{2}x_{m'}$$

by the analogues of (24) and (25) since $R_n(s) \leq B_n$ and now $D_{m'} \neq m'$. Theorem 2 then implies that

$$\mathbb{P}\left[\sup_{t\in S_{m'}} \frac{\nu_n(t-s_m)}{w(t)} > \kappa_2 E_{m'} + x_{m'}(\xi)\right] \leq \exp\left[-\frac{n\kappa_1 x_{m'}^2}{B_n}\right]$$

and we use Assumption **B** with Ω_ξ defined by (27) to get

$$1 - \mathbb{P}[\Omega_\xi] \leq \sum_{m'\in\mathcal{M}_n} \exp[-\kappa_1\xi^2 + D_{m'}L_{m'}] \leq \Delta \exp(-\kappa_1\xi^2). \tag{32}$$

Let $2(\kappa_2 E_{m'} + x_{m'}) = \kappa x_{m'}$. Lemma 1 implies that, if Ω_ξ is true

$$\|s - \tilde{s}\|^2 \leq \|s - s_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\tilde{m}) + \kappa x_{\tilde{m}}(\xi) w(\tilde{s}). \tag{33}$$

Then either $w(\tilde{s}) = 2x_{\tilde{m}}(\xi)$ and $x_{\tilde{m}}(\xi)w(\tilde{s}) = 2x^2_{\tilde{m}}(\xi)$ or

$$\begin{aligned} 2x_{\tilde{m}}(\xi)w(\tilde{s}) &= 2x_{\tilde{m}}(\xi)\|s - \tilde{s}\| + 2x_{\tilde{m}}(\xi)\|s - s_m\| \\ &\leq x^2_{\tilde{m}}(\xi) + \|s - s_m\|^2 + \kappa x^2_{\tilde{m}}(\xi) + \|s - \tilde{s}\|^2/\kappa. \end{aligned}$$

In both cases since $\kappa = 2(1 + \kappa_2\sqrt{2}) > 3$

$$2\kappa x_{\tilde{m}}(\xi) w(\tilde{s}) \leq \kappa \|s - s_m\|^2 + \kappa(1+\kappa) x^2_{\tilde{m}}(\xi) + \|s - \tilde{s}\|^2$$

and (33) becomes with $\kappa(1+\kappa) = 2(3 + 5\sqrt{2}\kappa_2 + 4\kappa_2^2)$

$$\|s - \tilde{s}\|^2 \leq (2+\kappa)\|s - s_m\|^2 + 2[\mathrm{pen}(m) - \mathrm{pen}(\tilde{m})] + \kappa(1+\kappa)x^2_{\tilde{m}}(\xi).$$

Therefore on the set $\Omega_\xi \cap \tilde{\Omega}$,

$$\|s - \tilde{s}\|^2 \leq 2\left[(2 + \kappa_2\sqrt{2})\|s - s_m\|^2 + \mathrm{pen}(m)\right] + \kappa(1+\kappa)B_n\xi^2/n$$

and Theorem 5 follows from (32) and an analogue of Lemma 2.

Let us now turn to Theorem 6. Let $R = \psi(\bar{s}_n) + K' \geq R_n(s)$, $\varepsilon = R/(3\Phi'\sqrt{N_n})$ and $\tilde{\Omega}$ be the event $\{\sup_\lambda |\nu_n(\varphi_\lambda)| \leq \varepsilon\}$. From our assumptions and Bernstein's inequality we get

$$\mathbb{P}[\tilde{\Omega}^c] \leq 2N_n \exp\left[\frac{-n\varepsilon^2}{2R_n(s) + \frac{2}{3}\Phi\sqrt{N_n}\varepsilon}\right] \leq 2N_n \exp\left[\frac{-K'\theta_n \log n}{2\Phi'(9\Phi' + \Phi)}\right] \tag{34}$$

since $\|\varphi_\lambda\|_\infty \leq \Phi\sqrt{N_n}$ and $\mathrm{Var}(\varphi_\lambda(X_1)) \leq \int \varphi^2_\lambda s d\mu \leq R_n(s)$. Let $B_n = 2(\psi(\bar{s}_n) - R/3 + K')$ and assume that $\tilde{\Omega}$ is true; then by (20) $\psi(\bar{s}_n) - R/3 \leq \psi(\hat{s}_n) \leq \psi(\bar{s}_n) + R/3$. Since s_m and $\hat{s}_m$ are the projections of $\bar{s}_n$ and $\hat{s}_n$ respectively on S_m, one derives that

$$R \leq B_n; \quad \sup_{m \in \mathcal{M}_n} \|\hat{s}_m\|_\infty \leq \psi(\hat{s}_n) \leq B_n; \quad \sup_{m \in \mathcal{M}_n} \|s_m\|_\infty \leq \psi(\bar{s}_n) \leq B_n$$

and for all $m \in \mathcal{M}_n$ simultaneously since $K \geq \kappa(1+\kappa)/\kappa_1$

$$\kappa(1+\kappa)B_n \frac{L_m D_m}{2n\kappa_1} \leq \mathrm{pen}(m) \leq \frac{4KL_mD_m}{3n}(\psi(\bar{s}_n) + K').$$

If $\tilde{s}'$ is the penalized projection estimator defined on the family of sieves $S'_m, m \in \mathcal{M}_n$ with a penalty given by (21) and $S'_m = \{t \in S_m \mid \|t\|_\infty \leq B_n\}$, it follows from the above inequalities that $\tilde{s} = \tilde{s}'$ and that Theorem 5 applies to $\tilde{s}'$. One can then conclude since $L_m D_m \geq 1$ that

$$\begin{aligned} \mathbb{E}[\|\tilde{s} - s\|^q \mathbb{I}_{\tilde{\Omega}}] &\leq C(q)\left[\|s - s_m\|^q + \mathbb{E}[(\mathrm{pen}(m))^{q/2}\mathbb{I}_{\tilde{\Omega}}] + \Delta(B_n/n)^{q/2}\right] \\ &\leq C(q)\left[\|s - s_m\|^q + C'(q, \Delta, \Psi(s), K, K')\left(\frac{L_m D_m}{n}\right)^{q/2}\right]. \end{aligned}$$

On the other hand, if $\tilde{\Omega}$ does not hold one uses the crude estimate

$$\|\tilde{s}\|_\infty \leq \|\psi(\hat{s}_n)\|_\infty \leq \Psi(s) + \Phi'\sqrt{N_n}\sup_j \|\nu_n(\varphi_\lambda)\|_\infty \leq \Psi(s) + \Phi\Phi' N_n$$

from which one deduces by (34) since $C'^2 = \int d\mu < +\infty$ that

$$\mathbb{E}[\|s-\tilde{s}\|^q \mathbb{I}_{\bar{\Omega}^c}] \leq 2N_n[C'(\Psi(s)+\Phi\Phi' N_n) + \|s\|]^q \exp\left[\frac{-K'\theta_n \log n}{2\Phi'(9\Phi'+\Phi)}\right],$$

which is bounded by $Cn^{-q/2}$ as required if C is large enough.

4.7 Some results in approximation theory for Besov spaces

4.7.1 Linear approximations

We shall collect here some known results of approximation of Besov spaces $B_{\alpha\, p\, \infty}(\mathcal{A})$ defined in Section 4.2.2 by classical finite-dimensional linear spaces. We first assume that $p = 2$ and consider the following approximation spaces:

- If $\mathcal{A} = [0,1]$ let S be the space of piecewise polynomials of degree bounded by r with $r > \alpha - 1$ based on the partition generated by the grid $\{j/D, 0 \leq j \leq D\}$;
- If $\mathcal{A} = \mathbb{T}$ let S be the space of trigonometric polynomials of degree $\leq D$;
- If $\mathcal{A} = [-A, A]$ let S be the space $\bar{V}_J$ generated by a wavelet basis of regularity $r > \alpha - 1$ defined in Section 4.2.2 with $D = 2^J$.

Let $\pi(s)$ be the projection of s onto the approximating space S. Then in each of the three situations, with different constants $C(s)$ in each case, we get

$$\|s - \pi(s)\| \leq C(s) D^{-\alpha}. \tag{35}$$

The proof of (35) comes from DeVore & Lorentz (1993) page 359 for piecewise polynomials and page 205 for trigonometric polynomials. For the wavelet expansion we shall prove a more general result than (35) which holds when $p \leq 2$ and $\alpha > 1/p - 1/2$. From the classical inequality

$$\sum_{\lambda\in\Lambda(j)} |\beta_\lambda|^2 \leq (2^j M)^{(1-2/p)^+} \left[\sum_{\lambda\in\Lambda(j)} |\beta_\lambda|^p\right]^{2/p}$$

and (6) we derive that

$$\|s - \pi(s)\|^2 = \sum_{j>J} \sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^2 \leq |\!|\!| s |\!|\!|^2 \sum_{j>J} 2^{-2j(\alpha+\frac{1}{2}-\frac{1}{p})} (2^j M)^{(1-2/p)^+}.$$

This implies for $p = 2$

$$\|s - \pi(s)\|^2 \leq |\!|\!| s |\!|\!|^2 \sum_{j>J} 2^{-2j\alpha} = \frac{|\!|\!| s |\!|\!|^2}{4^\alpha - 1} 2^{-2J\alpha},$$

which gives (35) since $D = 2^J$. Moreover for $p < 2$ and $\alpha > 1/p - 1/2$,

$$\|s - \pi(s)\|^2 \leq |\!|\!| s |\!|\!|^2 \sum_{j>J} 2^{-2j(\alpha+\frac{1}{2}-\frac{1}{p})} = |\!|\!| s |\!|\!|^2 \frac{2^{-2J(\alpha+\frac{1}{2}-\frac{1}{p})}}{4^{\alpha+\frac{1}{2}-\frac{1}{p}} - 1}. \tag{36}$$

4.7.2 Nonlinear approximations

Starting with a wavelet basis $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ as described in Section 4.2.2, we follow the framework stated for Proposition 3 of Section 4.3.2 with $l(x) = x^{-3}2^{-x}$. We want to study the approximation properties of the union of linear spaces $S_m = \mathrm{Span}(\{\varphi_\lambda \,|\, \lambda \in m\})$ when m belongs to $\mathcal{M}_J^{j'}$.

Proposition 6 *All elements of $\mathcal{M}_J^{j'}$ have the same cardinality bounded by $\kappa' M 2^{j'}$ and $\log |\mathcal{M}_J^{j'}|$ is bounded by $\kappa'' M 2^{j'}$. Moreover, if $M 2^{j'} \geq J^3$ and s belongs to $B_{\alpha\, p\, \infty}([-A, A])$ with $p \leq 2$ and $\alpha > 1/p - 1/2$ there exists $m \in \mathcal{M}_J^{j'}$ such that*

$$\|s - s_m\|^2 \leq C |\!|\!| s |\!|\!|^2 \left(2^{-2\alpha j'} + 2^{-2J(\alpha+\frac{1}{2}-\frac{1}{p})} \right). \tag{37}$$

Remarks

- From the bounds on the cardinalities of both $\mathcal{M}_J^{j'}$ and the elements of $\mathcal{M}_J^{j'}$ one can derive that $\cup_{m \in \mathcal{M}_J^{j'}} S_m$ is a metric space with finite metric dimension (in the sense of Le Cam) bounded by $D = C' 2^{j'}$ whatever J.
- Choosing $J \asymp \alpha j'/(\alpha + 1/2 - 1/p)$, this finite dimensional nonlinear metric space approximates $B_{\alpha\, p\, \infty}([-A, A])$ within $\mathcal{O}(D^{-\alpha})$. Hence (37) provides an analogue of (35), the difference being that a nonlinear finite-dimensional space instead of a linear vector space (both with dimensions of order D) is needed to get the $D^{-\alpha}$-rate of approximation for $p < 2$.
- As a consequence, the ε-entropy of $\{s \in B_{\alpha\, p\, \infty}([-A, A]) \,|\, |\!|\!| s |\!|\!| \leq 1\}$ is of order $\varepsilon^{-1/\alpha}$ when $\alpha > 1/p - 1/2$.

Proof of Proposition 6: The bound on $|m|$ derives from

$$|m| = M\left(\sum_{j=0}^{j'} 2^j + 2^{j'}\sum_{k=1}^{J-j'} 2^k l(k)\right) < \kappa' M 2^{j'}.$$

The control of $|\mathcal{M}_J^{j'}|$ is clear for $j' = J$. Otherwise

$$|\mathcal{M}_J^{j'}| = \prod_{j=j'+1}^{J} \begin{pmatrix} M2^j \\ [M2^j l(j-j')] \end{pmatrix} = \prod_{j=1}^{J-j'} \begin{pmatrix} M2^{j'+j} \\ [M2^{j+j'} l(j)] \end{pmatrix}$$

and from the inequality $\log \binom{n}{[nx]} < nx(\log(1/x)+1)$ which holds for $0 < x \leq 1$ one gets

$$\log|\mathcal{M}_J^{j'}| \leq M2^{j'}\sum_{j=1}^{\infty} 2^j l(j)\left(\log\left(\frac{1}{l(j)}\right)+1\right)$$

and the series converges from our choice of l. If $p = 2$, the conclusion of Proposition 6 follows from (35). If $p < 2$ the bias can always be written as

$$\|s - s_m\|^2 = \sum_{j>J}\sum_{\lambda\in\Lambda(j)} \beta_\lambda^2 + \sum_{j=j'+1}^{J}\sum_{\lambda\in\Lambda(j)\setminus m(j)} \beta_\lambda^2$$

where the second term is 0 when $j' = J$. We can bound the first term by (36). In order to control the second term we shall need the following

Lemma 3 *Assume that we are given n nonnegative numbers $0 \leq b_1 \leq \ldots \leq b_n$ with $\sum_{i=1}^n b_i = B$. For any number $r > 1$ and any integer k with $1 \leq k \leq n-1$ one has*

$$\sum_{i=1}^{n-k} b_i^r \leq B^r \frac{k^{1-r}}{r-1}(1-r^{-1})^r.$$

Proof of the lemma: One can assume without loss of generality that $B = 1$ and that for $i > n-k$ the b_i's are equal to some $b \leq 1/k$. Then $\sum_{i=1}^{n-k} b_i + kb = 1$ which implies that

$$\sum_{i=1}^{n-k} b_i^r \leq b^{r-1}\sum_{i=1}^{n-k} b_i = b^{r-1}(1-kb),$$

and the conclusion follows from a maximization with respect to b. □

We choose m such that for $j > j'$, $m(j)$ corresponds to the $[M2^j l(j-j')]$ largest values of the $|\beta_\lambda|$ for $\lambda \in \Lambda(j)$. Since by assumption $M2^{j'} \geq (J-j')^3$

and $l < 1$, $1 \leq |m(j)| < |\Lambda(j)|$ and we may apply Lemma 3 with $r = 2/p$ and $k = |m(j)|$ to get

$$\begin{aligned}\sum_{\lambda \in \Lambda(j) \setminus m(j)} \beta_\lambda^2 &\leq C(p) \left(\sum_{\lambda \in \Lambda(j)} \beta_\lambda^p \right)^{2/p} (|m(j)|)^{1-2/p} \\ &\leq C(p,M) \|s\|^2 2^{-2j(\alpha+\frac{1}{2}-\frac{1}{p})} \left(2^j l(j-j')\right)^{1-2/p}\end{aligned}$$

from which we deduce that

$$\sum_{j=j'+1}^{J} \sum_{\lambda \in \Lambda(j) \setminus m(j)} \beta_\lambda^2 \leq C(p,M) \|s\|^2 2^{-2\alpha j'} \sum_{j=1}^{\infty} 2^{-2j(\alpha+\frac{1}{2}-\frac{1}{p})} \left(2^j l(j)\right)^{1-2/p}$$

and the series converges since $2^j l(j) = j^{-3}$ and $\alpha + 1/2 - 1/p > 0$. □

4.8 References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* P. N. Petrov & F. Csaki, eds, 'Proceedings 2nd International Symposium on Information Theory', Akademia Kiado, Budapest, pp. 267–281.

Barron, A. R. & Cover, T. M. (1991), 'Minimum complexity density estimation', *IEEE Transactions on Information Theory* **37**, 1034–1054.

Barron, A. R., Birgé, L. & Massart, P. (1995), Model selection via penalization, Technical Report 95.54, Université Paris-Sud.

Birgé, L. & Massart, P. (1994), Minimum contrast estimation on sieves, Technical Report 94.34, Université Paris-Sud.

Cirel'son, B. S., Ibragimov, I. A. & Sudakov, V. N. (1976), Norm of gaussian sample function, *in* 'Proceedings of the 3rd Japan-USSR Symposium on Probability Theory', Springer-Verlag, New York, pp. 20–41. Springer Lecture Notes in Mathematics 550.

DeVore, R. A. & Lorentz, G. G. (1993), *Constructive Approximation*, Springer-Verlag, Berlin.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1993), Density estimation by wavelet thresholding, Technical Report 426, Department of Statistics, Stanford University.

Efroimovich, S. Y. (1985), 'Nonparametric estimation of a density of unknown smoothness', *Theory of Probability and Its Applications* **30**, 557–568.

Grenander, U. (1981), *Abstract Inference*, Wiley, New-York.

Kerkyacharian, G. & Picard, D. (1992), 'Estimation de densité par méthode de noyau et d'ondelettes: les lieus entre la géometrie du noyau et les contraintes de régularité', *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math* **315**, 79–84.

Kerkyacharian, G., Picard, D. & Tribouley, K. (1994), L^p adaptive density estimation, Technical report, Université Paris VII.

Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* **19**, 633–667.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Ledoux, M. (1995). Private communication.

Li, K. C. (1987), 'Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set', *Annals of Statistics* **15**, 958–975.

Mallows, C. L. (1973), 'Some comments on C_p', *Technometrics* **15**, 661–675.

Mason, D. M. & van Zwet, W. R. (1987), 'A refinement of the KMT inequality for the uniform empirical process', *Annals of Probability* **15**, 871–884.

Meyer, Y. (1990), *Ondelettes et Opérateurs I*, Hermann, Paris.

Polyak, B. T. & Tsybakov, A. B. (1990), 'Asymptotic optimality of the C_p-criteria in regression projective estimation', *Theory of Probability and Its Applications* **35**, 293–306.

Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**, 465–471.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators', *Scandinavian Journal of Statistics* **9**, 65–78.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Talagrand, M. (1994), 'Sharper bounds for Gaussian and empirical processes', *Annals of Probability* **22**, 28–76.

Talagrand, M. (1995), New concentration inequalities in product spaces, Technical report, Ohio State University.

Vapnik, V. (1982), *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

Whittaker, E. T. & Watson, G. N. (1927), *A Course of Modern Analysis*, Cambridge University Press, London.