The Annals of Statistics
2000, Vol. 28, No. 5, 1302–1338

ADAPTIVE ESTIMATION OF A QUADRATIC FUNCTIONAL BY MODEL SELECTION

BY B. LAURENT AND P. MASSART

Université Paris Sud

We consider the problem of estimating $\|s\|^2$ when s belongs to some separable Hilbert space and one observes the Gaussian process $Y(t) = \langle s, t\rangle + \sigma L(t)$, for all $t \in \mathbb{H}$, where L is some Gaussian isonormal process. This framework allows us in particular to consider the classical "Gaussian sequence model" for which $\mathbb{H} = l_2(\mathbb{N}^*)$ and $L(t) = \sum_{\lambda\geq 1} t_\lambda \varepsilon_\lambda$, where $(\varepsilon_\lambda)_{\lambda\geq 1}$ is a sequence of i.i.d. standard normal variables. Our approach consists in considering some at most countable families of finite-dimensional linear subspaces of $\mathbb{H}$ (the *models*) and then using model selection via some conveniently penalized least squares criterion to build new estimators of $\|s\|^2$. We prove a general *nonasymptotic* risk bound which allows us to show that such penalized estimators are adaptive on a variety of collections of sets for the parameter s, depending on the family of models from which they are built. In particular, in the context of the Gaussian sequence model, a convenient choice of the family of models allows defining estimators which are adaptive over collections of hyperrectangles, ellipsoids, l_p-bodies or Besov bodies. We take special care to describe the conditions under which the penalized estimator is efficient when the level of noise σ tends to zero. Our construction is an alternative to the one by Efroïmovich and Low for hyperrectangles and provides new results otherwise.

1. Introduction.

The framework. We consider the following extension of the standard linear Gaussian model to a possibly infinite-dimensional setting. Given some separable Hilbert space $\mathbb{H}$, one observes

$$Y(t) = \langle s, t\rangle + \sigma L(t) \quad \text{for all } t \in \mathbb{H},$$

where L is some centered Gaussian isonormal process; that is, L maps $\mathbb{H}$ isometrically onto some Gaussian subspace of $\mathbb{L}_2(\Omega)$. We shall say that Y is a Gaussian linear process with mean s and variance σ^2. Our purpose is to propose new adaptive estimators of $\|s\|^2$. The Gaussian framework that we introduce here could appear useless or at least unusual to the reader. In fact, it will turn out to be convenient for covering both the infinite-dimensional "white noise model" introduced by Ibragimov and Khasminskii for which $\mathbb{H} = \mathbb{L}_2([0, 1])$ and $L(t) = \int t(x)\, dW(x)$, where W is a standard Brownian motion, and the finite-dimensional linear model for which $\mathbb{H} = \mathbb{R}^N$ and $L(t) = \langle \zeta, t\rangle$, where ζ is a standard N-dimensional Gaussian vector. Given some Hilbertian basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ of $\mathbb{H}$, where Λ is a finite or countable set, one can equivalently

Received December 1998; revised April 2000.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20, 62J02.

Key words and phrases. Adaptive estimation, quadratic functionals, model selection, Besov bodies, l_p-bodies, Gaussian sequence model, efficient estimation.

describe the observation Y by

$$Y_\lambda = \beta_\lambda + \sigma\varepsilon_\lambda, \qquad \lambda \in \Lambda,$$

where $\{\varepsilon_\lambda\}_{\lambda\in\Lambda}$ is a family of i.i.d. standard normal random variables and $\{\beta_\lambda\}_{\lambda\in\Lambda}$ is the family of coordinates of s. When $\Lambda = \mathbb{N}^*$, this model is known as the Gaussian sequence model. The white noise model (or its equivalent discrete version, the Gaussian sequence model) has been considered by many authors since it represents in some sense an "ideal laboratory" for nonparametric inference. One can indeed hope to transpose the estimation methods developed within this framework, which is especially simple from a probabilistic point of view, to other more complicated situations such as density or regression estimation. It is exactly in this spirit that we will deal with the statistical framework described above, choosing to write the level of noise σ as $\sigma = n^{-1/2}$ in order to allow easy comparisons of the results obtained within this framework and other ones, such as density estimation on the basis of n i.i.d. observations. Let us now recall what is known about the problem of estimating $\|s\|^2$ in the white noise, Gaussian sequence or density frameworks.

Estimating $\|s\|^2$ with prior information on s. First it is important to say that, even from a purely minimax point of view when s belongs to some given set $\mathcal{S}$, there is at this time no complete answer to the following question:

(Q) Taking the usual distance on $\mathbb{R}$ as a loss function, what is the order of the minimax risk over $\mathcal{S}$ when estimating $\|s\|^2$?

This problem is really puzzling since two related questions have been solved for quite a long time. Indeed, when estimating the function itself with various loss functions, one can identify the order of the minimax risk in terms of the metric dimension of $\mathcal{S}$ [see the landmark paper by Birgé (1983) on this topic]. Concerning the problem of estimating a linear functional, the order of the risk is entirely determined by the modulus of continuity of the functional over $\mathcal{S}$ with respect to Hellinger distance [see Donoho and Liu (1991) where one will also find some results for nonlinear functionals which are not, however satisfactory for quadratic functionals]. Let us now turn to the estimation of a quadratic functional. Bickel and Ritov (1988) were the first to point out the following remarkable phenomenon for the estimation rates of $\theta = \|s\|^2$ in the density estimation context (more precisely if one observes n i.i.d. variables with common density s with respect to the Lebesgue measure on the real line). Assume that s belongs to some Hölderian ball $\mathcal{S}$ with radius R and index of smoothness α; then it is possible to construct some estimator $\hat{\theta}_n$ (depending on $\mathcal{S}$) such that, if $\alpha > 1/4$, $\hat{\theta}_n$ is an asymptotically $\sqrt{n}$-efficient estimator of θ, while it achieves the rate of convergence $n^{-4\alpha/(1+4\alpha)}$ whenever $\alpha \le 1/4$, this rate being the order of the minimax risk over $\mathcal{S}$. Corresponding results for the Gaussian sequence model have been obtained by Donoho and Nussbaum (1990), the smoothness assumptions being in this context replaced by geometric assumptions for the sequence $(\beta_\lambda)_{\lambda\ge1}$ such as $\sum_{\lambda\ge1}\lambda^{2\alpha}\beta_\lambda^2 \le R^2$, which means that $(\beta_\lambda)_{\lambda\ge1}$ belongs to some ellipsoid. It is worth noticing that

estimating $\|s\|^2$ is a key step to constructing estimators of more general integral functionals of s. The interested reader will find some details about the density framework in Laurent (1996), where simpler estimators of $\|s\|^2$ than those used by Bickel and Ritov are also introduced [see also Birgé and Massart (1995) for minimax lower bounds concerning smooth functionals of the density]. These results solve question (Q) for some particular sets $\mathcal{S}$ such as ellipsoids (similar results hold for hyperrectangles) when $\mathcal{S}$ is a set of sequences or Hölderian balls when $\mathcal{S}$ is a set of functions. This suggests that a general answer to question (Q) should take into account not only the modulus of continuity of the functional as in Donoho and Liu's theory (the quadratic functional $\|s\|^2$ is indeed Lipschitz over any of the sets $\mathcal{S}$ described above) but also the "size" of $\mathcal{S}$ in a sense that we do not know. Our feeling is that the metric dimension successfully used by Birgé for the estimation of s itself might not be appropriate as suggested by the new results established in this paper concerning l_p-bodies, for $p < 2$. Indeed, for the Gaussian sequence model, under the assumption $\sum_{\lambda \geq 1} \lambda^{p(1/2+\alpha-1/p)} |\beta_\lambda|^p \leq R^p$ with $\alpha > 1/p - 1/2$, we shall show in Section 3 the existence of a $\sqrt{n}$-convergent estimator provided that $\alpha > 1/p - 1/4$ when $p \geq 4/3$ and $\alpha > 1/2$ when $p \leq 4/3$. The striking fact here is that our result depends on p while the metric dimension of the l_p-body is known to depend only on α [see for instance Birgé and Massart (2000a)]. Unfortunately we do not know whether our result is optimal or not.

Adaptive estimation of $\|s\|^2$. All the estimators of $\|s\|^2$ that one can find in the literature cited above suffer from the same drawback: they depend on the a priori knowledge that s belongs to some set $\mathcal{S}$ (such as some given Hölderian ball or some given ellipsoid for instance). From this point of view, Efroïmovitch and Low (1996) have obtained an important improvement of the previous results. In the context of the Gaussian sequence model, by using a procedure which is close to Lepskii's method [as introduced in Lepskii (1990, 1992)] they propose an estimator $\hat{\theta}_n$ of θ with the following *adaptive* properties. For any positive R and α, provided that the sequence $(\beta_\lambda)_{\lambda \geq 1}$ satisfies the condition $\beta_\lambda^2 \lambda^{2\alpha+1} \leq R^2$ for all λ [which means that $(\beta_\lambda)_{\lambda \geq 1}$ belongs to some hyperrectangle] one has:

1. $\hat{\theta}_n$ is asymptotically efficient if $\alpha > 1/4$;
2. $\mathbb{E}\big[(\hat{\theta}_n - \theta)^2\big] \leq b_n/n$, where b_n tends to infinity when n goes to infinity as slowly as desired, if $\alpha = 1/4$.

   Note that they also consider an estimator $\hat{\theta}_n$ which is $\sqrt{n}$-consistent in the whole range $\alpha \geq 1/4$. For both estimates one has

$$\mathbb{E}\big[(\hat{\theta}_n - \theta)^2\big] \leq C(R, \alpha)(n^{-2} \log(n))^{4\alpha/(1+4\alpha)} \quad \text{if } \alpha < 1/4.$$

Since the minimax quadratic risk for estimating θ on a given hyperrectangle is of order $n^{-8\alpha/(1+4\alpha)}$ whenever $\alpha < 1/4$, the estimator $\hat{\theta}_n$ misses the optimal rate within the factor $(\log(n))^{4\alpha/(1+4\alpha)}$. Efroïmovitch and Low (1996) actually show that this is really the price to pay for adaptation, which means that this logarithmic factor is unavoidable if you do not know in advance to what hyperrectangle s belongs. The reader should take note of the fact that the index

α that we use here is different from the one used in Efroïmovitch and Low (1996). This choice will turnout to be convenient when connecting smoothness assumptions on functions with geometrical constraints on the coefficients of functions in a proper basis.

Description of our method and results. Our approach to building adaptive estimators is based on model selection via penalization. This method has been successfully developed to estimate adaptively a function s in various contexts [see Birgé and Massart (1997), Barron, Birgé and Massart (1999), Baraud (1997) or Birge and Massart (2000b)]. Although we shall deal with a general Gaussian framework, we are presenting our approach in the context of the Gaussian sequence model by sake of simplicity. We consider some collection $\mathscr{M}$ of subsets of $\mathbb{N}^*$ and a *penalty function* pen: $\mathscr{M} \to \mathbb{R}^+$. Our penalized estimator of $\theta = \sum_{\lambda \geq 1} \beta_\lambda^2$ is then simply defined by

$$\hat{\theta} = \sup_{m \in \mathscr{M}} \left[\sum_{\lambda \in m} Y_\lambda^2 - \mathrm{pen}(m) \right]. \tag{1.1}$$

Our main theorem (see Theorem 1 in Section 2 below) provides a *nonasymptotic* bound for $\mathbb{E}\left[\left(\hat{\theta} - \theta - \left(2/\sqrt{n}\right) \sum_{\lambda \geq 1} \beta_\lambda \varepsilon_\lambda\right)^2\right]$ when the penalty function is conveniently chosen (an explicit expression for the penalty function is given in the statement of Theorem 1). Such a bound can also be used for asymptotic purposes and is especially useful for specifying under which condition on the sequence $(\beta_\lambda)_{\lambda \geq 1}$, $\hat{\theta}$ is asymptotically efficient. The choice of the penalty function is very important and influences the order of magnitude of the risk bound. The penalty pen(m) depends, of course, on the cardinality of m but also on the complexity of the whole collection $\mathscr{M}$. It should be noticed that such a dependency also appears in Birgé and Massart (2000b) in the same context of the Gaussian sequence model, but with a very different expression for the penalty. This means that an appropriate penalty function to estimate s is not necessarily convenient to estimate $\|s\|^2$ and vice versa.

Our general risk bound can be used to analyze the adaptivity property of the penalized estimator over various families of sets of parameters $(S_a)_{a \in A}$. Of course the geometric nature of the sets S_a, $a \in A$ will heavily depend on the collection $\mathscr{M}$ and more precisely of its approximation properties. For instance, taking first $\mathscr{M}$ as the nested family $\mathscr{M}_{\mathrm{nest}}$ of sets $\{1, \ldots, D\}$, $D \in \mathbb{N}^*$, let us define the penalized estimator as

$$\hat{\theta} = \sup_{D \in \mathbb{N}^*} \left[\sum_{\lambda \leq D} Y_\lambda^2 - \frac{1}{n} \left(D + 1 + 2\sqrt{(D+1)x_D} + 2x_D \right) \right], \tag{1.2}$$

where $x_D = C \log D$ for any positive integer D, for some given constant $C > 2$. Then $\hat{\theta}$ will have the same adaptivity properties over the set of hyperrectangles as the estimator of Efroïmovitch and Low. Note that we even get some modest but objective gain with respect to Efroïmovitch and Low's result since our estimator is $\sqrt{n}$-efficient instead of $\sqrt{n}$-convergent when the index α of the hyperrectangle is not too small, that is, $\alpha > 1/4$. It also has

analogous adaptivity properties with respect to the collection of l_p-bodies $\sum_{\lambda\in\mathbb{N}^*}\lambda^{p(\alpha-1/p+1/2)}|\beta_\lambda|^p \le R^p$ for $p>2$.

We can, moreover, profit by the flexibility of the model selection via penalization method and produce other estimators with new adaptivity properties by simply enlarging the collection $\mathscr{M}_{\text{nest}}$. If, for instance, we take $\mathscr{M}$ as the collection $\mathscr{M}_{\text{all}}$ of all the finite subsets of $\mathbb{N}^*$ and choose the penalty adequately, we shall show that the penalized estimator still has the adaptivity properties of the preceding one with respect to the collection of hyperrectangles or ellipsoids but furthermore has new adaptivity properties with respect to l_p-bodies for $p<2$. In particular we shall prove that under the assumption $\sum_{\lambda\ge1}\lambda^{p(1/2+\alpha-1/p)}|\beta_\lambda|^p \le R^p$ with $\alpha>1/p-1/2$, the estimator is $\sqrt{n}$-efficient provided that $\alpha>1/p-1/4$ when $p\ge 4/3$ and $\alpha>1/2$ when $p\le 4/3$. Otherwise some nonparametric rates of convergence arise but, as previously mentioned, we do not know if our results are optimal or not, simply because of the lack of lower bounds for the minimax risk on a given l_p-body when $p<2$. We shall also construct penalized estimators with improved adaptive convergence properties (the gain is a logarithmic factor), by considering some specially designed collection of sets $\mathscr{M}$ such that $\mathscr{M}_{\text{nest}}\subset\mathscr{M}\subset\mathscr{M}_{\text{all}}$.

In each example that we shall consider, we shall indicate an easy way to compute the corresponding penalized estimator. Indeed, since definition (1.1) involves some optimization over the collection $\mathscr{M}$, one could have legitimate doubts about the computability of the penalized estimator, especially in situations where $\mathscr{M}$ is taken to be a very large collection of sets like $\mathscr{M}_{\text{all}}$. Fortunately, we shall show that for all the examples that we shall consider, the computation reduces to some optimization over the set of integers, exactly as in (1.2), which this time can be expected to be performed with the help of a computer.

2. Estimation via model selection.

2.1. *Description of the framework.* Assume that one observes a Gaussian linear process Y with mean s and variance $1/n$ on some Hilbert space $\mathbb{H}$, endowed with the scalar product $\langle\cdot,\cdot\rangle$. We recall that this means

$$Y(t)=\langle s,t\rangle+\frac{1}{\sqrt{n}}L(t),\qquad t\in\mathbb{H}, \tag{2.1}$$

where $s\in\mathbb{H}$ is unknown, L is some isonormal Gaussian process on $\mathbb{H}$ [see Dudley (1973)], and L is a linear isometry from $\mathbb{H}$ to some Gaussian subspace of $\mathbb{L}_2(\Omega,\mathbb{P})$. In particular, the covariance of the process is defined by $\operatorname{Cov}(L(t),L(t'))=\langle t,t'\rangle$.

The following frameworks are easily seen to be of type (2.1).

Finite-dimensional Gaussian regression. One observes

$$Y_i=s_i+\varepsilon_i,\qquad i=1,\dots,n, \tag{2.2}$$

where $(\varepsilon_1,\dots,\varepsilon_n)$ are independent standard normal variables. We consider $\mathbb{H}=\mathbb{R}^n$ endowed with the scalar product $\langle x,y\rangle=(1/n)\sum_{i=1}^n x_iy_i$ and set

$s = (s_1, \ldots, s_n)$. Model (2.1) is obtained by setting, for all $t = (t_1, \ldots, t_n) \in \mathbb{R}^n$, $Y(t) = (1/n)\sum_{i=1}^n t_i Y_i$ and $L(t) = (1/\sqrt{n})\sum_{i=1}^n t_i \varepsilon_i$.

Conversely, if model (2.1) is observed, then we recover the Gaussian regression model with fixed design by considering an orthonormal basis of $\mathbb{R}$, say $(e_1, \ldots, e_n)$, and by setting $Y_i = Y(ne_i)$, $s_i = n\langle s, e_i\rangle$ and $\varepsilon_i = \sqrt{n}L(e_i)$.

The Gaussian sequence model. In the Gaussian sequence model, one observes

$$Y_\lambda = \beta_\lambda + \frac{1}{\sqrt{n}}\varepsilon_\lambda, \qquad \lambda \in \mathbb{N}^*, \tag{2.3}$$

where $(\varepsilon_\lambda)_{\lambda\in\mathbb{N}^*}$ is a sequence of independent standard normal variables.

Setting $\mathbb{H} = l_2(\mathbb{N}^*)$ endowed with the usual scalar product $\langle \beta, \gamma\rangle = \sum_{\lambda\in\mathbb{N}^*}\beta_\lambda\gamma_\lambda$ and $s = (\beta_\lambda)_{\lambda\in\mathbb{N}^*}$, we define for any $t = (\alpha_\lambda)_{\lambda\in\mathbb{N}^*} \in \mathbb{H}$, $Y(t) = \sum_{\lambda\in\mathbb{N}^*}\alpha_\lambda Y_\lambda$ and $L(t) = \sum_{\lambda\in\mathbb{N}^*}\alpha_\lambda\varepsilon_\lambda$ and we see that (2.3) implies (2.1).

Conversely, if one observes $\{Y(t), t \in l_2(\mathbb{N}^*)\}$ according to model (2.1), then we recover the Gaussian sequence model by setting for all $\lambda \in \mathbb{N}^*$, $Y_\lambda = Y(\phi_\lambda)$, $\beta_\lambda = \langle s, \phi_\lambda\rangle$ and $\varepsilon_\lambda = L(\phi_\lambda)$ where $(\phi_\lambda)_{\lambda\in\mathbb{N}^*}$ is the canonical basis of $l_2(\mathbb{N}^*)$.

The multivariate white noise model. One observes

$$Z(x) = \int_{[0,1]^d} \mathbb{1}_{[0,x_1]\times\cdots\times[0,x_d]}(u)s(u)\,du + \frac{1}{\sqrt{n}}W(x)$$

for all $x = (x_1, \ldots, x_d) \in [0,1]^d$, where W is the standard Wiener process on $[0,1]^d$. We consider $\mathbb{H} = \mathbb{L}_2([0,1]^d)$ endowed with its usual scalar product. We set $Y(t) = \int_{[0,1]^d} t(u)\,dZ(u)$ and $L(t) = \int_{[0,1]^d} t(u)\,dW(u)$.

Conversely, if one observes $\{Y(t), t \in \mathbb{L}_2([0,1]^d)\}$, according to model (2.1) then one a fortiori observes $Z(x) = Y(\mathbb{1}_{[0,x_1]\times\cdots\times[0,x_d]})$ for all $x = (x_1, \ldots, x_d) \in [0,1]^d$. Since $W(x) = L(\mathbb{1}_{[0,x_1]\times\cdots\times[0,x_d]})$ is a standard Wiener process, Z is indeed defined from a white noise model.

2.2. *The estimation procedure.* Our aim is to estimate $\|s\|^2 = \langle s, s\rangle$ from observation (2.1). We want to present an adaptive estimation method based on model selection. To better understand its interest and the way it works, it is useful to recall first the minimax approach for which one can use an estimator defined from a single finite-dimensional linear model.

The minimax approach. Let us take some D-dimensional linear subspace S of $\mathbb{H}$. Given some orthonormal basis $(\phi_\lambda, \lambda \in \Lambda)$ of S, since the orthogonal projection of s on S can be written as $\sum_{\lambda\in\Lambda}\langle s, \phi_\lambda\rangle\phi_\lambda$, it is natural to consider the projection estimator $\hat{s} = \sum_{\lambda\in\Lambda} Y(\phi_\lambda)\phi_\lambda$. It is easy to verify that

$$\hat{s} = \arg\min_{v\in S}(\|v\|^2 - 2Y(v)),$$

which shows that $\hat{s}$ does not depend on the particular choice of the basis $(\phi_\lambda, \lambda \in \Lambda)$. It is instructive to study the behavior of the statistics $\|\hat{s}\|^2$. Since

$$\hat{s} = \sum_{\lambda\in\Lambda}\langle s, \phi_\lambda\rangle\phi_\lambda + \frac{1}{\sqrt{n}}\sum_{\lambda\in\Lambda} L(\phi_\lambda)\phi_\lambda,$$

we obtain

$$\|\hat{s}\|^2 = \sum_{\lambda\in\Lambda}\langle s,\phi_\lambda\rangle^2 + \frac{2}{\sqrt{n}}\sum_{\lambda\in\Lambda}\langle s,\phi_\lambda\rangle L(\phi_\lambda) + \frac{1}{n}\sum_{\lambda\in\Lambda}L^2(\phi_\lambda).$$

From this identity, we derive that $\hat{\theta} = \|\hat{s}\|^2 - D/n$ is an unbiased estimator of $\|\pi_S(s)\|^2$ where $\pi_S(s) = \sum_{\lambda\in\Lambda}\langle s,\phi_\lambda\rangle\phi_\lambda$ denotes the orthogonal projection of s onto S. We can easily compute the quadratic risk of $\tilde{\theta}$ as an estimator of $\theta = \|s\|^2$. Indeed, we notice that

$$\tilde{\theta} - \theta - \frac{2L(s)}{\sqrt{n}} = -\|s-\pi_S(s)\|^2 + \frac{2}{\sqrt{n}}L(\pi_S(s)-s) + \frac{1}{n}\sum_{\lambda\in\Lambda}(L^2(\phi_\lambda)-1).$$

Since the variables $L(\pi_S(s)-s)$ and $\sum_{\lambda\in\Lambda}L^2(\phi_\lambda)$ are independent with respective distributions $\mathcal{N}(0,\|s-\pi_S(s)\|^2)$ and $\chi^2(D)$, we derive that

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{\theta}-\theta-\frac{2L(s)}{\sqrt{n}}\right)^2\right] &= \|s-\pi_S(s)\|^4 + \frac{4}{n}\|s-\pi_S(s)\|^2 + \frac{2D}{n^2}\\ &\le 3\|s-\pi_S(s)\|^4 + \frac{2(D+1)}{n^2}.\end{aligned} \tag{2.4}$$

From this inequality, we see that an ideal choice of S would be to make the trade-off between the squared bias term $\|s-\pi_S(s)\|^4$ and the variance term D/n^2. To be more concrete, let us take $\mathbb{H} = \mathbb{L}_2([0,1])$. Then, some prior smoothness assumption on s such as s belongs to the class of Hölderian functions

$$\mathcal{H}_\alpha(L) = \{t\in\mathbb{L}_2([0,1]), |t(x)-t(y)|\le L|x-y|^\alpha, \forall x,y\in[0,1]\},$$

leads to the existence of some subspace S (such as histograms with D regular pieces) such that

$$\sup_{s\in\mathcal{H}_\alpha(L)}\|s-\pi_S(s)\|^2 \le CL^2D^{-2\alpha}.$$

Therefore, choosing D in a way that $D/n^2 \sim L^4D^{-4\alpha}$ ensures that, for some universal constant C',

$$\sup_{s\in\mathcal{H}_\alpha(L)}\mathbb{E}\left[\left(\tilde{\theta}-\theta-\frac{2L(s)}{\sqrt{n}}\right)^2\right] \le C'L^{4/(1+4\alpha)}n^{-8\alpha/(1+4\alpha)}.$$

In that case, $\tilde{\theta}$ is an asymptotically efficient estimator of θ with asymptotic variance $4\|s\|^2$ whenever $\alpha > 1/4$.

The main drawback of this minimax approach is that the choice of the subspace S and of its dimension D depends on the prior smoothness class $\mathcal{H}_\alpha(L)$. Our strategy to overcome this difficulty consists of considering some preliminary collection of models $(S_m)_{m\in\mathcal{M}}$ where $\mathcal{M}$ is some finite or countable set that may depend on n and defining our estimator from the corresponding collection of projection estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ via some model selection criterion. This criterion relies upon the following idea.

Heuristics of the model selection method. For any $m \in \mathcal{M}$, let S_m be some D_m-dimensional linear subspace of $\mathbb{H}$, s_m denote the orthogonal projection of s onto S_m and $\tilde{\theta}_m$ be the unbiased estimator of $\|s_m\|^2$ defined by $\tilde{\theta}_m = \|\hat{s}_m\|^2 - D_m/n$. From inequality (2.4), we derive that "the best" model from the point of view of minimizing the quadratic risk of $\tilde{\theta}_m$ as an estimator of θ should minimize $\|s - s_m\|^2 + C\sqrt{D_m}/n$ or equivalently $-\|s_m\|^2 + C\sqrt{D_m}/n$. Since $\|s_m\|^2$ is unknown, it is natural to replace it by the unbiased estimator $\tilde{\theta}_m$. This leads to the idea of minimizing $-\|\hat{s}_m\|^2 + D_m/n + C\sqrt{D_m}/n$ to get a proper data-driven model choice. Our selection criterion will indeed be close to the latter since we shall consider some penalty function pen: $\mathcal{M} \to \mathbb{R}^+$ and define

$$\hat{m} = \arg\min_{m \in \mathcal{M}} (-\|\hat{s}_m\|^2 + \text{pen}(m)).$$

The main issue is that pen(m) will be taken greater than the heuristically determined penalty term $D_m/n + C\sqrt{D_m}/n$ in order to take into account the "complexity" of the collection of models. We finally define our penalized estimator of θ as

$$\hat{\theta} = \|\hat{s}_{\hat{m}}\|^2 - \text{pen}(\hat{m}) = \sup_{m \in \mathcal{M}} (\|\hat{s}_m\|^2 - \text{pen}(m)).$$

We turn now to the main result of the paper, which we shall illustrate in the next sections.

2.3. *The main theorem.* We address the problem of constructing risk bounds for penalized estimators which depend on a proper choice of the penalty function. In the statement of Theorem 1 below, the parameter n which appears in (2.1) is fixed and our bounds involve numerical constants that do not depend on n. Hence the Hilbert space involved in (2.1) as well as the collection of models $(S_m)_{m \in \mathcal{M}}$ or the penalty function pen$(\cdot)$ are allowed to depend on n.

THEOREM 1. *Let $\mathbb{H}$ be some Hilbert space endowed with scalar product $\langle \cdot, \cdot \rangle$. One observes the Gaussian process $\{Y(t), t \in \mathbb{H}\}$, where $Y(t)$ is given by* (2.1). *Let $\mathcal{M}^*$ be some finite or countable set and for any $m \in \mathcal{M}^*$, let S_m denote some linear subspace of $\mathbb{H}$ with finite dimension $D_m > 0$. We consider $(x_m)_{m \in \mathcal{M}^*}$ to be some family of nonnegative real numbers. Let, for any $m \in \mathcal{M}^*$,* pen(m) *satisfy*

$$n\,\text{pen}(m) \geq (D_m + 1) + 2\sqrt{(D_m + 1)x_m} + 2x_m. \tag{2.5}$$

Let $\mathcal{M}$ be either $\mathcal{M}^*$ or $\mathcal{M}^* \cup \{0\}$, with $S_0 = \{0\}$ and pen$(0) = 0$. *Let $\hat{s}_m$ be the projection estimator of s over S_m.*

We consider the collection of estimators $(\hat{\theta}_m)_{m \in \mathcal{M}}$ of $\theta = \|s\|^2$, given by

$$\hat{\theta}_m = \|\hat{s}_m\|^2 - \text{pen}(m)$$

and define

$$\hat{\theta} = \sup_{m \in \mathcal{M}} \hat{\theta}_m. \tag{2.6}$$

Let r be some positive real number. Then, whenever

$$\Sigma_r = \sum_{m \in \mathcal{M}} D_m^{r/2} e^{-x_m} < +\infty, \tag{2.7}$$

$\hat{\theta}$ is almost surely finite and

$$\mathbb{E}_s\left[\left|\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}}\right|^r\right] \leq \inf_{m \in \mathcal{M}} \mathbb{E}_s\left[\left(-\hat{\theta}_m + \theta + \frac{2L(s)}{\sqrt{n}}\right)_+^r\right] + C_1(r)\frac{(\Sigma_r + 1)}{n^r}, \tag{2.8}$$

where $C_1(r)$ is some numerical constant depending only on r. Moreover, for any $m \in \mathcal{M}$,

$$\begin{aligned}\mathbb{E}_s\left[\left(-\hat{\theta}_m + \theta + \frac{2L(s)}{\sqrt{n}}\right)_+^r\right] \leq C_2(r)\Bigg[\|s - s_m\|^{2r} + \frac{(D_m^{r/2} + 1)}{n^r} \\ + \left(\mathrm{pen}(m) - \frac{D_m}{n}\right)^r\Bigg],\end{aligned} \tag{2.9}$$

where s_m is the orthogonal projection of s over S_m and $C_2(r)$ is a numerical constant depending only on r.

COMMENTS. (i) There is some ambiguity in the definition of $\{Y(t), t \in \mathbb{H}\}$ since the isonormal process $\{L(t), t \in \mathbb{H}\}$ is defined up to some negligible event that may change for each $t \in \mathbb{H}$. In other words, if $\mathbb{H}$ is infinite dimensional, one cannot guarantee that there exists a given version of $\{L(t), t \in \mathbb{H}\}$ such that $L(t)(\omega)$ is linear with respect to t for almost all $\omega \in \Omega$. Nevertheless, the definition of our estimator only involves some given countable collection of finite-dimensional linear subspaces $(S_m)_{m \in \mathcal{M}}$ and $\{Y(t), t \in \bigcup_{m \in \mathcal{M}} S_m\}$. It is easy to see that there exists some version of L which is linear on the algebraic linear span S of $\bigcup_{m \in \mathcal{M}} S_m$ and such a version is implicitly used to define a linear version of Y on S.

(ii) If $\mathcal{M}$ is finite, then any possible $\hat{m}$ which minimizes $-\|\hat{s}_m\|^2 + \mathrm{pen}(m)$ leads to the same value for $\hat{\theta}_{\hat{m}}$ which is precisely our estimator $\hat{\theta} = \sup_{m \in \mathcal{M}} \hat{\theta}_m$. If $\mathcal{M}$ is infinite, there is no guarantee that the minimum of $-\|\hat{s}_m\|^2 + \mathrm{pen}(m)$ is achieved. Nevertheless, it always makes sense to consider $\hat{\theta} = \sup_{m \in \mathcal{M}} \hat{\theta}_m$; this is the reason why we use such a definition of $\hat{\theta}$ in the statement of Theorem 1 rather than $\hat{\theta} = \hat{\theta}_{\hat{m}}$. Moreover, Theorem 1 ensures that $\hat{\theta}$ is fortunately almost surely finite.

(iii) When applying Theorem 1, we shall generally take $\mathrm{pen}(m)$ as small as permitted; that is,

$$\mathrm{pen}(m) = \frac{(D_m + 1)}{n} + 2\frac{\sqrt{(D_m + 1)x_m}}{n} + 2\frac{x_m}{n}.$$

The role of the weights $(x_m)_{m \in \mathcal{M}}$ is therefore essential but might seem mysterious at a first glance. We have indeed several possible choices for $(x_m)_{m \in \mathcal{M}}$. One possibility is to choose $(x_m)_{m \in \mathcal{M}}$ in such a way that

$$\sum_{m \in \mathcal{M}} D_m^{r/2} e^{-x_m} \leq C'(r), \tag{2.10}$$

where $C'(r)$ is some numerical constant depending only on r. Combining (2.8) and (2.9) leads, under this assumption, to

$$\mathbb{E}_s\left[\left|\hat{\theta}-\theta-\frac{2L(s)}{\sqrt{n}}\right|^r\right] \le C''(r)\inf_{m\in\mathscr{M}}\left[\|s-s_m\|^{2r}+\frac{(D_m^{r/2}+1)}{n^r}+\left(\mathrm{pen}(m)-\frac{D_m}{n}\right)^r\right]. \tag{2.11}$$

(iv) One can derive from (2.11) two kinds of information about the behavior of $\hat{\theta}$. One possibility is to analyze the risk of $\hat{\theta}$. We readily get from (2.11),

$$\mathbb{E}_s\left[\left|\hat{\theta}-\theta\right|^r\right] \le 2^{(r-1)+}\left(C''(r)\inf_{m\in\mathscr{M}}\left[\|s-s_m\|^{2r}+\frac{(D_m^{r/2}+1)}{n^r}+\left(\mathrm{pen}(m)-\frac{D_m}{n}\right)^r\right]+\frac{2^r\|s\|^r}{n^{r/2}}\mathbb{E}(|\xi|^r)\right) \tag{2.12}$$

where ξ is a standard normal variable. We shall study in the next section several examples for which (2.12) leads to upper bounds for the maximal risk of $\hat{\theta}$ over various sets of parameters.

Another possibility is to use (2.11) for asymptotic analysis, which means that n goes to infinity. Taking $r=1$, if we have chosen the weights $(x_m)_{m\in\mathscr{M}}$ such that (2.10) holds, then inequality (2.11) shows that whenever

$$\inf_{m\in\mathscr{M}}\left[\|s-s_m\|^2+\frac{D_m^{1/2}}{n}+\left(\mathrm{pen}(m)-\frac{D_m}{n}\right)\right]=o(1/\sqrt{n}),$$

then $\sqrt{n}(\hat{\theta}-\theta)-2L(s)$ converges towards 0 in probability. Recalling that $L(s)$ is a centered Gaussian variable with variance θ, we see that, if the Hilbert space $\mathbb{H}$ does not depend on n, $\theta=\|s\|^2$ is also independent of n, and therefore $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotically centered normal with variance 4θ.

We intend to apply Theorem 1 to show that our estimator is adaptive in various classes of parameter sets. Since we have in view to prove that in many situations our estimator is asymptotically efficient, it is convenient to deal from now on with the case where $\mathbb{H}$ is a given infinite-dimensional Hilbert space, although our theorem clearly also applies when $\mathbb{H}$ is finite-dimensional with dimension depending on n, as in the example of the fixed design regression model. We recall below the correspondence between classes of functions in $\mathbb{L}_2([0,1])$ like Hölderian, Sobolev or Besov balls and classes of sequences in $l_2(\mathbb{N}^*)$ like hyperrectangles, ellipsoids, l_p or Besov bodies via some proper choice of a basis. This will motivate the study of the properties of our penalized estimator within the framework of the Gaussian sequence model. This study will be performed in Section 3 where the adaptive properties of the penalized estimator over various bodies in $l_2(\mathbb{N}^*)$ will be exhibited.

2.4. *Smoothness classes and bodies in* $l_2(\mathbb{N}^*)$. We want to make precise the correspondence between classes of functions included in $\mathbb{L}_2([0,1])$ and sets of

coefficients. Many classes of functions can indeed be described by the properties of their expansions on a suitable basis. For the sake of simplicity, we shall content ourselves with dealing with the Haar basis and control the variations of a function with the help of moduli of continuity. However, more general wavelet expansions and moduli of smoothness could be considered as well [we refer to Donoho and Johnstone (1998) for more details].

Following DeVore and Lorentz (1993), the $\mathbb{L}_p$-modulus of continuity $\omega(s, y)_p$ is defined by

$$(\omega(s, y)_p)^p = \sup_{0<h\le y} \int_0^{1-h} |s(x+h) - s(x)|^p \, dx \quad \text{for } 0 < y, \text{ if } 0 < p < \infty,$$

and for $p = \infty$,

$$\omega(s, y)_\infty = \sup_{0<h\le y} \sup_{x\in[0,\,1-h]} |s(x+h) - s(x)|.$$

Let $0 < \alpha < 1$, $0 < p, q \le \infty$, the function s belongs to the Besov space $\mathscr{B}^\alpha_{p,\,q}([0, 1])$, if and only if $s \in \mathbb{L}_p([0, 1])$ and

$$|s|^q_{\alpha,\,p,\,q} = \sum_{j\ge 0} 2^{j\alpha q} \omega^q(s, 2^{-j})_p < +\infty \quad \text{when } 0 < q < +\infty,$$

$$|s|_{\alpha,\,p,\,\infty} = \sup_{j\ge 0} 2^{j\alpha} \omega(s, 2^{-j})_p < +\infty \quad \text{when } q = +\infty.$$

We recall that $\alpha > (1/p - 1/2)_+$ warrants that $\mathscr{B}^\alpha_{p,\,q}([0, 1]) \subset \mathbb{L}_2([0, 1])$.

We turn now to the correspondence between Besov balls and bodies in a sequence space via Haar expansions. Let $\psi = \mathbb{1}_{[0,\,1/2[} - \mathbb{1}_{[1/2,\,1[}$, and for any integers j and k, $\psi_{j,\,k}(\cdot) = 2^{j/2}\psi(2^j \cdot - k)$. Any function $s \in \mathbb{L}_2([0, 1])$ can be expanded as

$$s = \int_0^1 s(x)\, dx + \sum_{j\ge 0} \sum_{k=1}^{2^j} \beta_{j,\,k} \psi_{j,\,k},$$

where $\beta_{j,\,k} = \int_0^1 s(x)\psi_{j,\,k}(x)\, dx$. Let $\Lambda = \{(j, k) \in \mathbb{N}^2,\ k \in \{1, \dots, 2^j\}\}$, for any integer j. We set $|\beta|_{j,\,p} = \big(\sum_{k=1}^{2^j} |\beta_{j,\,k}|^p\big)^{1/p}$ if $p < \infty$ and $|\beta|_{j,\,\infty} = \sup_{k\in\{1,\dots,2^j\}} |\beta_{j,\,k}|$. The size of the coefficients of s depends on the modulus of continuity. This can be seen by using the following classical inequality [see Devore, Jawerth and Popov (1992)], for all $j \ge 0$ and $p \ge 1$:

$$2^{j(1/2-1/p)} |\beta|_{j,\,p} \le C_p\, \omega(s, 2^{-j})_p, \tag{2.13}$$

where C_p is a constant depending only on p.

Assume first that $p \ge 1$. It follows from (2.13) that if s belongs to some Besov ball with respect to the seminorm $|\cdot|_{\alpha,\,p,\,q}$, that is,

$$\sum_{j\ge 0} 2^{j\alpha q} \omega^q(s, 2^{-j})_p \le Q^q$$

for some $Q > 0$, then

$$\sum_{j\geq 0} 2^{qj(1/2+(\alpha-1/p))}|\beta|_{j,\,p}^{q} \leq R^{q}, \tag{2.14}$$

where $R = C_{p}Q$. Similarly, if s belongs to some Besov ball with respect to the seminorm $|\cdot|_{\alpha,\,p,\,\infty}$, that is,

$$\sup_{j\geq 0} 2^{j\alpha}\omega(s, 2^{-j})_{p} \leq Q$$

for some $Q > 0$, then

$$\forall\ j \in \mathbb{N}, \qquad |\beta|_{j,\,p} \leq R2^{-j(1/2+(\alpha-1/p))}. \tag{2.15}$$

We can more generally consider the class of functions s satisfying

$$\sum_{j\geq 0} \frac{\omega^{p}(s, 2^{-j})_{p}}{w^{p}(2^{-j})} < +\infty$$

if $p < \infty$ where w is a given positive function on [0,1]. Note that the Besov space $\mathscr{B}_{p,\,p}^{\alpha}([0, 1])$ corresponds to the situation where $w(x) = x^{\alpha}$. Assume that

$$\sum_{j\geq 0} \frac{\omega^{p}(s, 2^{-j})_{p}}{w^{p}(2^{-j})} < +\infty.$$

It follows from inequality (2.13) that $\beta \in l_{2}(\Lambda)$ provided that $x \mapsto w(x)x^{1/2-1/p}$ is nondecreasing (and therefore bounded) if $p \leq 2$ and provided that $\sum_{j\geq 0}(w(2^{-j}))^{(1/2-1/p)^{-1}} < +\infty$ if $p > 2$. Moreover if

$$\sum_{j\geq 0} \frac{\omega^{p}(s, 2^{-j})_{p}}{w^{p}(2^{-j})} \leq Q^{p},$$

then

$$\sum_{j\geq 0} \frac{|\beta|_{j,\,p}^{p}}{R^{p}2^{pj(1/p-1/2)}w^{p}(2^{-j})} \leq 1. \tag{2.16}$$

Similarly, if $\sup_{j\geq 0}(\omega(s, 2^{-j})_{\infty}/w(2^{-j})) \leq Q$, then

$$\sup_{j\geq 0} \frac{|\beta|_{j,\,\infty}}{R2^{-1/2}w(2^{-j})} \leq 1. \tag{2.17}$$

The case $0 < p < 1$ is more involved since in this case (2.13) is not available. However, it is still true that the Besov ball with respect to the seminorm $|\cdot|_{\alpha,\,p,\,q}$ is included in a Besov body defined by (2.14) if $q < \infty$ or (2.15) if $q = \infty$, for an appropriate value of R. See Devore, Kyriasis, Leviatan and Tikhomirov (1993).

If we order the countable set Λ with the lexicographical ordering, we can identify Λ with $\mathbb{N}^{*}$. As shown by the computations above, conditions on the moduli of continuity of s can be transferred to conditions on the sequence β of

coefficients of s. This motivates the following formal definitions of bodies in $l_2(\mathbb{N}^*)$, that we shall use below. We begin with l_p-bodies.

DEFINITION 1. Let $0 < p \leq \infty$ and c be some positive and nonincreasing sequence. We define the l_p-body $\Theta_{p,c}$ as

$$\Theta_{p,c} = \left\{ \beta \in l_p(\mathbb{N}^*),\ \sum_{\lambda \in \mathbb{N}^*} \left| \frac{\beta_\lambda}{c_\lambda} \right|^p \leq 1 \right\} \quad \text{if } p < \infty,$$

$$\Theta_{\infty,c} = \left\{ \beta \in l_\infty(\mathbb{N}^*),\ \sup_{\lambda \in \mathbb{N}^*} \left| \frac{\beta_\lambda}{c_\lambda} \right| \leq 1 \right\} \quad \text{if } p = \infty.$$

Note that an l_p-body is always included in $l_2(\mathbb{N}^*)$ for $p \leq 2$. If $p > 2$, Hölder's inequality warrants that $\Theta_{p,c} \subset l_2(\mathbb{N}^*)$ whenever

$$\sum_{\lambda \in \mathbb{N}^*} c_\lambda^{(1/2-1/p)^{-1}} < \infty. \tag{2.18}$$

Moreover, $\Theta_{p,c}$ is an ellipsoid when $p = 2$ and an hyperrectangle when $p = \infty$.

We shall also deal with the scale of Besov bodies as introduced in Donoho and Johnstone (1998).

DEFINITION 2. Let $0 < p, q \leq \infty$, $\alpha > 0$, and $R > 0$. Assume that $\alpha' = 1/2 + \alpha - 1/p > 0$. Given the partition of $\mathbb{N}^*$, $\mathbb{N}^* = \sum_{j \geq 0} \Lambda(j)$, where $\Lambda(j) = \{2^j, \ldots, 2^{j+1} - 1\}$, we define the Besov body $\mathscr{B}_{\alpha,p,q}(R)$ as

$$\mathscr{B}_{\alpha,p,q}(R) = \left\{ \beta \in l_2(\mathbb{N}^*),\ \sum_{j \geq 0} |\beta|_{j,p}^q 2^{qj\alpha'} \leq R^q \right\} \quad \text{if } q < \infty,$$

$$\mathscr{B}_{\alpha,p,\infty}(R) = \left\{ \beta \in l_2(\mathbb{N}^*),\ \sup_{j \geq 0} |\beta|_{j,p} 2^{j\alpha'} \leq R \right\} \quad \text{if } q = \infty,$$

where $|\beta|_{j,p}^p = \sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^p$ if $p < \infty$ and $|\beta|_{j,\infty} = \sup_{\lambda \in \Lambda(j)} |\beta_\lambda|$.

$\mathscr{B}_{\alpha,p,p}(R)$ is essentially an l_p-body and does not bring anything new. This is not the case for $\mathscr{B}_{\alpha,p,\infty}(R)$ which contains $\mathscr{B}_{\alpha,p,p}(R)$ and that we shall use in the sequel.

All these bodies play a role when expressing smoothness constraints on the function s through constraints on its sequence of coefficients β. Indeed, inequalities (2.14) and (2.15) ensure that whenever s belongs to some Besov ball with respect to the seminorm $|\cdot|_{\alpha,p,q}$ then β belongs to some Besov body $\mathscr{B}_{\alpha,p,q}(R)$, while (2.16) and (2.17) express that more general conditions on the modulus of continuity of s imply that β belongs to some adequate l_p-body.

3. The Gaussian sequence model. A reasonable strategy to estimate $\|s\|^2$ when s belongs to some infinite-dimensional separable Hilbert space and one observes (2.1) can be described as follows. Let $(\phi_\lambda)_{\lambda \in \Lambda}$ be some orthonormal basis of $\mathbb{H}$. One can always assume that $\Lambda = \Lambda_0 \cup \mathbb{N}^*$ where Λ_0 is a finite subset

of $\mathbb{Z}^-$ which does not depend on n. Think here that this is typically what one gets when considering the Haar basis in $\mathbb{L}^2([0,1])$, taking $\phi_0 = \mathbb{1}_{[0,1]}$, and $(\phi_\lambda)_{\lambda\in\mathbb{N}^*}$ as the ordered $(\psi_{j,k})_{j\geq 0,\, 1\leq k\leq 2^j}$. Then

$$\|s\|^2 = \|s_0\|^2 + \sum_{\lambda\in\mathbb{N}^*} \beta_\lambda^2,$$

where s_0 is the orthogonal projection of s onto the linear span of $(\phi_\lambda)_{\lambda\in\Lambda_0}$. One can estimate $\|s_0\|^2$ by $\|\hat{s}_0\|^2 - |\Lambda_0|/n$, where $\hat{s}_0$ stands for the projection estimator on the linear span of $(\phi_\lambda)_{\lambda\in\Lambda_0}$. This estimator has a quadratic risk of order $1/n$ and is moreover efficient which means that

$$\sqrt{n}\left(\|\hat{s}_0\|^2 - \frac{|\Lambda_0|}{n} - \|s_0\|^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 4\|s_0\|^2).$$

The problem of estimating properly $\|s\|^2$ reduces to that of estimating $|\beta|^2 = \sum_{\lambda\in\mathbb{N}^*}\beta_\lambda^2$. This can be done on the basis of the observation of the Gaussian sequence model (2.3) where the errors ε_λ are defined by $\varepsilon_\lambda = L(\phi_\lambda)$. Any estimator T_n of $|\beta|^2$ built from the sequence $(Y_\lambda)_{\lambda\in\mathbb{N}^*}$ leads to the definition of an estimator of $\|s\|^2$ by taking $T_n' = \|\hat{s}_0\|^2 - |\Lambda_0|/n + T_n$. Since $|\Lambda_0|$ does not depend on n, the quadratic risk of T_n' will stay of the same order as that of T_n and, moreover, if T_n is efficient which means that

$$\sqrt{n}(T_n - |\beta|^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 4|\beta|^2),$$

then since $\hat{s}_0$ is independent of T_n, T_n' is also efficient; that is,

$$\sqrt{n}(T_n - |\beta|^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 4\|s\|^2).$$

Hence, all through this section, we shall focus on the problem of estimating $|\beta|^2$ when one observes the Gaussian sequence model (2.3), and produce estimators which are adaptive on a variety of bodies of $l_2(\mathbb{N}^*)$. To do so, we shall consider several examples of collections of subsets $\{\Lambda_m, m\in\mathcal{M}\}$ of $\mathbb{N}^*$ and the corresponding collection of models $\{S_m, m\in\mathcal{M}\}$, where for any $m\in\mathcal{M}$,

$$S_m = \{\beta\in l_2(\mathbb{N}^*), \beta_\lambda = 0 \quad \forall\lambda\notin\Lambda_m\}.$$

Then we shall define appropriate penalty functions and consider the corresponding penalized estimators of $|\beta|^2$ as given by (2.6).

3.1. *l_p-bodies for $p\geq 2$.* The definition of the penalized estimator that we shall consider throughout this section is as follows.

DEFINITION 3. Let $\mathcal{M} = \mathbb{N}^*$ and K be some given real number, $K > 1$. For all $m\in\mathcal{M}$, we set $x_m = K\log(m+1)$ and

$$n\,\mathrm{pen}(m) = m + 1 + 2\sqrt{(m+1)x_m} + 2x_m.$$

We define $\hat{\theta}$ by

$$\hat{\theta} = \sup_{m\in\mathcal{M}}\left(\sum_{\lambda=1}^{m} Y_\lambda^2 - \mathrm{pen}(m)\right).$$

We introduce a new body which will turn out to be convenient since it contains in some sense l_p-bodies for $p \geq 2$. Let $\gamma = (\gamma_m)_{m \in \mathbb{N}^*}$ be some nonincreasing and nonnegative sequence and $\mathscr{S}_\gamma$ be the subset of $l_2(\mathbb{N}^*)$ defined by

$$\mathscr{S}_\gamma = \left\{ \beta \in l_2(\mathbb{N}^*),\ \forall\, m \in \mathbb{N}^*,\ \sum_{\lambda > m} \beta_\lambda^2 \leq \gamma_m^2 \right\}. \tag{3.1}$$

The following theorem gives a uniform risk bound for the penalized estimator of Definition 3 over the set $\mathscr{S}_\gamma$.

THEOREM 2. *Assume that one observes* $(Y_\lambda)_{\lambda \in \mathbb{N}^*}$ *given by the Gaussian sequence model* (2.3) *and set* $\beta = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$, $\theta = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda^2$, *and* $L(\beta) = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda \varepsilon_\lambda$. *Let* K *be some constant such that* $K > 1$ *and* $\hat{\theta}$ *be the corresponding penalized estimator given by Definition* 3. *Let* $\mathscr{S}_\gamma$ *be defined by* (3.1). *For any* r *such that* $r < 2(K-1)$, *the following inequality holds*:

$$\sup_{\beta \in \mathscr{S}_\gamma} \mathbb{E}_\beta \left[\left| \hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right|^r \right] \leq C(r) \inf_{m \in \mathbb{N}^*} \left[\gamma_m^{2r} + \left(\frac{m \log(m+1)}{n^2} \right)^{r/2} \right],$$

where $C(r)$ *is some constant depending only on* r.

COMMENTS. (i) Although the above result is not asymptotic, we can use it to derive asymptotic properties for our estimator: if the term

$$\inf_{m \in \mathbb{N}^*} \left[\gamma_m^{2r} + \left(\frac{m \log(m+1)}{n^2} \right)^{r/2} \right]$$

is negligible as compared to $n^{-r/2}$, then $\hat{\theta}$ is an efficient estimator of θ. This will depend on the structure of the sequence $\gamma = (\gamma_m)_{m \in \mathbb{N}^*}$.

(ii) An ellipsoid $\Theta_{2,c}$ defined by Definition 1 is included in the set $\mathscr{S}_\gamma$ if we set $\gamma_m = c_m\, \forall\, m \in \mathbb{N}^*$. This is also the case for a hyperrectangle $\Theta_{\infty,c}$ defined by Definition 1 if we set $\gamma_m^2 = \sum_{\lambda > m} c_\lambda^2$ and for an l_p-body $\Theta_{p,c}$ with $p > 2$ and $\sum_{\lambda \in \mathbb{N}^*} c_\lambda^{(1/2-1/p)^{-1}} < \infty$ if we set $\gamma_m = (\sum_{\lambda > m} c_\lambda^{(1/2-1/p)^{-1}})^{1/2-1/p}$. This means that Theorem 2 can be used to analyze the behavior of our estimator on some arbitrary l_p-body with $p \geq 2$.

It is interesting to look at the particular situation where $\gamma_m = Rm^{-\alpha}$. Let us denote by $\mathscr{S}_\alpha(R)$ the corresponding set $\mathscr{S}_\gamma$. We derive from Comment (ii) above that the following l_p-bodies are all included in $\mathscr{S}_\alpha(R)$:

$$\mathscr{E}_{p,\alpha'}(R') = \left\{ \gamma \in l_2(\mathbb{N}^*),\ \sum_{\lambda \in \mathbb{N}^*} \lambda^{p\alpha'} |\gamma_\lambda|^p \leq (R')^p \right\} \quad \text{if } 2 \leq p < \infty, \tag{3.2}$$

$$\mathscr{E}_{\infty,\alpha'}(R') = \left\{ \gamma \in l_2(\mathbb{N}^*),\ \forall\, \lambda \in \mathbb{N}^*, |\gamma_\lambda| \leq R' \lambda^{-\alpha'} \right\} \quad \text{if } p = \infty, \tag{3.3}$$

where $\alpha' = 1/2 + \alpha - 1/p$, $R' = R$ if $p = 2$ and $R' = R(\alpha/(1/2 - 1/p))^{1/2-1/p}$ otherwise.

It should be noticed that $\mathscr{S}_\alpha(R)$ is included in $\mathscr{B}_{\alpha,2,\infty}(R)$ which is easily seen to be included in $\mathscr{S}_\alpha(R2^{2\alpha}/\sqrt{2^\alpha-1})$. This allows us to derive the following corollary of Theorem 2 which provides uniform risk bounds for the penalized estimator given by Definition 3. Given $\alpha > 0, R > 0$, these risk bounds are uniform over the Besov body $\mathscr{B}_{\alpha,2,\infty}(R)$, and therefore over the l_p-body $\mathscr{E}_{p,\alpha'}(R')$ for $p \geq 2$ and $\alpha' = 1/2 + (\alpha - 1/p)$, where $R' = R$ if $p = 2$, and $R' = R(\alpha/(1/2 - 1/p))^{1/2-1/p}$ otherwise.

COROLLARY 1. *Assume that one observes* $(Y_\lambda)_{\lambda\in\mathbb{N}^*}$ *given by the Gaussian sequence model* (2.3). *Let* $\beta = (\beta_\lambda)_{\lambda\in\mathbb{N}^*}$ *and* $\theta = \sum_{\lambda\in\mathbb{N}^*}\beta_\lambda^2$. *Let* K *be some constant such that* $K > 1$ *and* $\hat\theta$ *be the corresponding penalized estimator given by Definition* 3. *Assume that* r *is some positive real number which satisfies* $r < 2(K-1)$. *For any* $R > 0$ *and* $\alpha > 0$, *let the Besov body* $\mathscr{B}_{\alpha,2,\infty}(R)$ *be defined by Definition* 2.

Assume that $nR^2 \geq 1$, *then*

$$\sup_{s\in\mathscr{B}_{\alpha,2,\infty}(R)} \mathbb{E}_s\left[\left|\hat\theta - \theta - \frac{2L(s)}{\sqrt{n}}\right|^r\right] \leq C(r,\alpha)\left[R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)}\right],$$

where $C(r,\alpha)$ *depends only on* r *and* α. *This leads to*:

(i) *If* $\alpha \leq 1/4$,

$$\sup_{\beta\in\mathscr{B}_{\alpha,2,\infty}(R)} \mathbb{E}_\beta\left[|\hat\theta - \theta|^r\right] \leq C'(r,\alpha)\left[R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)}\right]; \tag{3.4}$$

(ii) *if* $\alpha > 1/4$,

$$\sup_{\beta\in\mathscr{B}_{\alpha,2,\infty}(R)} \mathbb{E}_\beta\left[|\hat\theta - \theta|^r\right] \leq C'(r,\alpha)\frac{R^r}{n^{r/2}}, \tag{3.5}$$

where $C'(r,\alpha)$ *depends only on* r *and* α.

If the sequence $\beta = (\beta_\lambda)_{\lambda\in\mathbb{N}^*}$ *belongs to the Besov body* $\mathscr{B}_{\alpha,2,\infty}(R)$ *for some* $\alpha > 1/4$, *then*

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{\mathscr{D}} \mathscr{N}(0, 4\theta) \quad as\ n \to \infty, \tag{3.6}$$

$$n^{r/2}\mathbb{E}_\beta\left[|\hat\theta - \theta|^r\right] \to 2^r\theta^{r/2}\mathbb{E}[|\xi|^r] \quad as\ n \to \infty\ if\ r \geq 1, \tag{3.7}$$

where ξ *is a standard normal variable*.

COMMENTS. (i) When R does not depend on n, the minimax rate of convergence of $\hat\theta$ is $(\log(n)/n^2)^{2\alpha/(1+4\alpha)}$ if $\alpha \leq 1/4$ while if $\alpha > 1/4$, (3.6) ensures that $\hat\theta$ is an efficient estimator of θ. Efroïmovich and Low (1996) have proved that the logarithmic factor which appears in the rate of convergence for $\alpha < 1/4$ cannot be avoided. Using Lepskii's method for adaptation, Efroïmovich and Low (1996) have also built an estimator which is adaptive on the class of hyperrectangles $\mathscr{E}_{\infty,\alpha'}(R)$ with $\alpha' = \alpha + 1/2$ in the sense that it achieves the optimal rate $(\log(n)/n^2)^{2\alpha/(1+4\alpha)}$ for $\alpha < 1/4$ and it is $\sqrt{n}$ consistent whenever

$\alpha \geq 1/4$. Our estimator presents the theoretical advantage that it is, moreover, efficient whenever $\alpha > 1/4$ and that our risk bounds are valid for all l_p-bodies $\mathscr{E}_{p,\alpha'}(R)$ simultaneously and not only for hyperrectangles. The estimator given by Definition 3 is furthermore easily computable. Note also that results (3.4) and (3.5) are non-asymptotic and allow R to depend on n.

(ii) If we modify the definition of the penalty function and take $x_n = 1$ instead of $x_n = K\log(n+1)$ in Definition 3, it is easy to see that the resulting penalized estimator $\hat{\theta}$ achieves the rate $1/\sqrt{n}$ instead of $\log(n)/\sqrt{n}$ when $\alpha = 1/4$. Nevertheless, there is a price to pay for this: $\hat{\theta}$ is no longer efficient for $\alpha > 1/4$ since the remainder term $R_n = (1/n^r)\sum_{m\in\mathscr{M}} D_m^{r/2} e^{-x_m}$ is then of order $n^{-r/2}$.

3.2. *Arbitrary l_p-bodies.* In this section, we shall propose an estimator of θ with adaptivity properties over the set of l_p-bodies,

$$\Theta_{p,c} = \left\{\gamma \in l_p(\mathbb{N}^*),\ \sum_{\lambda\in\mathbb{N}^*} \left|\frac{\gamma_\lambda}{c_\lambda}\right|^p \leq 1\right\},$$

where $(c_\lambda)_{\lambda\in\mathbb{N}^*}$ is some positive and nonincreasing unknown sequence satisfying (2.18) if $p > 2$. If $p \leq 2$, $\Theta_{p,c}$ is included in $\Theta_{2,c}$, which is itself included in the set $\mathscr{S}_c$ defined by (3.1). Hence, one could think of considering the estimator defined by Definition 3, which is furthermore known to be adaptive on the l_p-bodies for $p \geq 2$, as shown in the previous section. So, let us consider, for any $m \in \mathbb{N}^*$, $x_m = 3\log(m+1)$ and

$$n\,\mathrm{pen}(m) = m + 1 + 2\sqrt{(m+1)x_m} + 2x_m.$$

We define

$$\hat{\theta}^{(1)} = \sup_{m\in\mathbb{N}^*}\left[\sum_{\lambda=1}^{m} Y_\lambda^2 - \mathrm{pen}(m)\right]. \tag{3.8}$$

It follows from Theorem 1 that for any $p \leq 2$,

$$\sup_{\beta\in\Theta_{p,c}} \mathbb{E}_\beta\left[\left|\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}}\right|^r\right] \leq C(r)\inf_{m\in\mathbb{N}^*}\left[c_m^{2r} + \left(\frac{m\log(m+1)}{n^2}\right)^{r/2}\right],$$

where $C(r)$ is some constant depending only on r. It turns out that this result is too crude and that one can take advantage of the fact that, when $p < 2$, nonlinear approximations perform better than linear approximations. This invites us to consider collections of models where different models may have the same dimension. A typical strategy of this kind can be described as follows.

We set, for any $(N, D) \in (\mathbb{N}^*)^2$,

$$x_{N,D} = 3D\left(1 + \log\left(\frac{N}{D}\right)\right), \tag{3.9}$$

and define

$$nw(N, D) = D + 1 + 2\sqrt{(D+1)x_{N,D}} + 2x_{N,D}. \tag{3.10}$$

Let $\widehat{\Lambda}_{N,D}$ be a set of indices corresponding to the D largest elements of the set $\{|Y_\lambda|, \lambda = 1, \ldots, N\}$. We define

$$\hat{\theta}^{(2)} = \sup_{N\in\mathbb{N}^*} \sup_{1\le D\le N} \left[\sum_{\lambda\in\widehat{\Lambda}_{N,D}} Y_\lambda^2 - w(N,D) \right]. \tag{3.11}$$

Here $\hat{\theta}^{(2)}$ can indeed be interpreted as a conveniently penalized estimator over the collection of models defined from the collection of all finite subsets of $\mathbb{N}^*$ (see the proof of Theorem 3). Since this collection involves an infinite number of models with the same dimension, the penalty function must be taken much larger than for the definition of $\hat{\theta}^{(1)}$. This means that we have gained something for the control of the bias term in the risk bound of Theorem 1 but that simultaneously we have lost something in the variance term. The idea is therefore to combine the two estimators and consider $\hat{\theta} = \hat{\theta}^{(1)} \vee \hat{\theta}^{(2)}$ which turns out to perform as well as $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$.

THEOREM 3. *Assume that one observes* $(Y_\lambda)_{\lambda\in\mathbb{N}^*}$ *given by the Gaussian sequence model* (2.3). *We set* $\beta = (\beta_\lambda)_{\lambda\in\mathbb{N}^*}$, $\theta = \sum_{\lambda\in\mathbb{N}^*} \beta_\lambda^2$ *and* $L(\beta) = \sum_{\lambda\in\mathbb{N}^*} \beta_\lambda \varepsilon_\lambda$. *Let* $\hat{\theta}^{(1)}$ *and* $\hat{\theta}^{(2)}$ *be defined by* (3.8) *and* (3.11), *respectively. We define* $\hat{\theta}$ *by*

$$\hat{\theta} = \hat{\theta}^{(1)} \vee \hat{\theta}^{(2)}.$$

Let $0 < p \le \infty$. *We consider some nonincreasing and nonnegative sequence* $c = (c_\lambda)_{\lambda\in\mathbb{N}^*}$. *There exists some absolute constant* C *such that the following inequalities hold*:

(i) *If* $p < 2$,

$$\sup_{\beta\in\Theta_{p,c}} \mathbb{E}_\beta\left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}}\right)^2\right] \le C \inf\left\{ \inf_{D\in\mathbb{N}^*}\left[c_D^4 + \frac{D\log(D+1)}{n^2}\right], \inf_{N\in\mathbb{N}^*}\left\{ \inf_{1\le D\le N}\left[\left(D^{1-2/p}c_D^2\right)^2 + \left(\frac{D(1+\log(N/D))}{n}\right)^2\right] + c_N^4\right\}\right\};$$

(ii) *if* γ *is a nonincreasing sequence, then*

$$\sup_{\beta\in\mathcal{L}_\gamma} \mathbb{E}_\beta\left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}}\right)^2\right] \le C \inf_{D\in\mathbb{N}^*}\left[\gamma_D^4 + \frac{D\log(D+1)}{n^2}\right]. \tag{3.12}$$

Moreover, $\Theta_{2,c} \subseteq \mathscr{S}_c$ *and if* $p > 2$, *assuming that condition* (2.18) *holds*, $\Theta_{p,c} \subseteq \mathscr{S}_\gamma$ *where* γ *is given by* $\gamma_D = \bigl(\sum_{\lambda > D} c_\lambda^{(1/2-1/p)^{-1}}\bigr)^{1/2-1/p}$.

COMMENTS. Instead of $\hat{\theta}^{(2)}$ we could as well consider the adaptive threshold estimator $\tilde{\theta}^{(2)}$ defined in the following way. For any $(N, D) \in (\mathbb{N}^*)^2$, we set $x_{N,D} = 3D(1 + \log(N))$, and we define

$$n\widetilde{w}(N, D) = 2D + 2\sqrt{2Dx_{N,D}} + 2x_{N,D}.$$

Let

$$\tilde{\theta}^{(2)} = \sup_{N \in \mathbb{N}^*} \sup_{A \subset \{1,2,\ldots,N\}} \biggl[\sum_{\lambda \in A} Y_\lambda^2 - \widetilde{w}(N, |A|)\biggr].$$

Since $\widetilde{w}(N, |A|)$ is proportional to the cardinality of A, $\tilde{\theta}^{(2)}$ turns out to be some adaptive threshold estimator. Namely,

$$\tilde{\theta}^{(2)} = \sup_{N \in \mathbb{N}^*} \Biggl(\sum_{\lambda=1}^{N} \Bigl[Y_\lambda^2 - \frac{2}{n}\Bigl(1 + \sqrt{6(1 + \log(N)) + 3(1 + \log(N))}\Bigr)\Bigr] \times \mathbb{1}_{Y_\lambda^2 > 2/n\left(1+\sqrt{6(1+\log(N))}+3(1+\log(N))\right)} \Biggr).$$

If we replace $\hat{\theta}^{(2)}$ by $\tilde{\theta}^{(2)}$ in the definition of $\hat{\theta}$ given in Theorem 3, then the properties of $\hat{\theta}$ are not as good as the properties of the estimator defined in Theorem 3; more precisely, the term $\log(N/D)$ appearing in the control of the quadratic risk has to be replaced by $\log(N)$. However an advantage of the adaptive threshold estimator as compared with $\hat{\theta}^{(2)}$ could be its more explicit expression.

We shall now give a corollary of Theorem 3 when c_λ is a power of λ. Let us therefore introduce, for any $p > 0$, $\alpha' > 0$ and $R > 0$, the l_p-body

$$\mathscr{E}_{p,\alpha'}(R) = \Bigl\{\beta \in l_p(\mathbb{N}^*),\ \sum_{\lambda \in \mathbb{N}^*} \lambda^{p\alpha'} |\beta_\lambda|^p \le R^p\Bigr\}. \tag{3.13}$$

The following corollary gives uniform risk bounds for the estimator $\hat{\theta}$ of θ defined in Theorem 3 over the sets $\mathscr{E}_{p,\alpha'}(R)$ which are included in $l_2(\mathbb{N}^*)$; namely, this is the case if $\alpha' > 0$ and $\alpha = \alpha' - 1/2 + 1/p > 0$.

COROLLARY 2. *One observes* $(Y_\lambda)_{\lambda \in \mathbb{N}^*}$ *given by the Gaussian sequence model* (2.3). *We set* $\beta = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$, $\theta = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda^2$, *and* $L(\beta) = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda \varepsilon_\lambda$. *Let* $\hat{\theta}^{(1)}$ *and* $\hat{\theta}^{(2)}$ *be defined by* (3.8) *and* (3.11), *respectively*, *and let*

$$\hat{\theta} = \hat{\theta}^{(1)} \vee \hat{\theta}^{(2)}.$$

Let $p > 0$, $\alpha' > 0$ *and* $R > 0$. *We define* $\alpha = \alpha' - 1/2 + 1/p$ *and assume that* $\alpha > 0$. *If* $nR^2 \geq 1$, *one has*:

(i) *If* $p < 2$,

$$\sup_{\beta \in \mathscr{E}_{p,\alpha'}(R)} \mathbb{E}_\beta \left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right)^2 \right] \leq C(p,\alpha) \inf \left\{ R^{4/(1+4\alpha')} \left(\frac{\log(1+nR^2)}{n^2} \right)^{4\alpha'/(1+4\alpha')}, R^{4/(1+2\alpha)} \left(\frac{\log(1+nR^2)}{n} \right)^{4\alpha/(1+2\alpha)} \right\}, \tag{3.14}$$

where $C(p,\alpha)$ *is a constant depending only on* p *and* α.

(ii) *Moreover*,

$$\sup_{\beta \in \mathscr{B}_{a,2,\infty}(R)} \mathbb{E}_\beta \left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right)^2 \right] \leq C(\alpha) R^{4/(1+4\alpha)} \left(\frac{\log(1+nR^2)}{n^2} \right)^{4\alpha/(1+4\alpha)} \Bigg), \tag{3.15}$$

where $C(\alpha)$ *is a constant depending only* α.

COMMENTS. (i) One can derive from (3.15) the same bounds as in Corollary 1.

(ii) If $1 < p < 2$, we obtain unusual rates of convergence, and we do not know whether these rates are optimal or not.

(iii) Let us now discuss the efficiency of $\hat{\theta}$. Comparing the right-hand side of (3.14) with $1/n$, one derives that if $4/3 \leq p \leq 2$, $\hat{\theta}$ is an efficient estimator of θ as soon as $\alpha' > 1/4$, while if $p \leq 4/3$, $\hat{\theta}$ is efficient whenever $\alpha' > 1 - 1/p$. In particular, for $p \leq 1$, $\hat{\theta}$ is always efficient.

(iv) One can also derive from (3.14) an upper bound for the uniform quadratic risk of $\hat{\theta}$. It sufficies to notice that $|\beta|^2 \leq R^2$ whenever $\beta \in \mathscr{E}_{p,\alpha'}(R)$ with $p < 2$. This leads via (3.14) to an upper bound for the quadratic risk which, up to some constant depending on p and α, is equal to

$$\inf \left\{ R^{4/(1+4\alpha')} \left(\frac{\log(1+nR^2)}{n^2} \right)^{4\alpha'/(1+4\alpha')}, R^{4/(1+2\alpha)} \left(\frac{\log(1+nR^2)}{n} \right)^{4\alpha/(1+2\alpha)} \right\} + \frac{R^2}{n}.$$

It is interesting to consider some situations where Theorem 3 applies while Corollary 2 does not. This will be the case for l_p-bodies $\Theta_{p,c}$ for which $p < 2$ and $(c_\lambda)_{\lambda \in \mathbb{N}^*}$ converge very slowly towards 0; for example, if we look at the

case where $c_\lambda = R(\log(\lambda))^{-\eta}$ for some $\eta > 0$, we obtain

$$\begin{aligned}\sup_{\beta\in\Theta_{p,c}} \mathbb{E}_\beta\left[\left(\hat\theta - \theta - \tfrac{2L(\beta)}{\sqrt{n}}\right)^2\right] \\ \le C(R,p,\eta)\inf\Big\{(\log(1+n))^{-4}, \\ n^{(2-p)((1/\eta)(1/p-1/2)-1)}\log^2(1+n)\Big\}.\end{aligned} \tag{3.16}$$

This bound shows that the rate of convergence of the estimator $\hat\theta^{(1)}$ is always logarithmic; namely, it is equal to $(\log(1+n))^{-2\eta}$, while we obtain a rate which is a negative power of n for the estimator $\hat\theta^{(2)}$, and hence for $\hat\theta$, as soon as $\eta > 1/p - 1/2$.

3.3. *Special strategy for Besov bodies.* We want to deal with the problem of estimating $\sum_{\lambda\in\mathbb{N}^*}\beta_\lambda^2$ provided that the sequence $(\beta_\lambda)_{\lambda\in\mathbb{N}^*}$ belongs to some (unknown) Besov body $\mathcal{B}_{\alpha,p,\infty}$. We begin with the simplest case where $p = 2$, for which we have already proposed some adaptive estimators in Section 3.1 (see Corollary 1), our aim being here to show that the level thresholding estimators considered in Johnstone (1999) can be interpreted as penalized estimators.

3.3.1. *Level thresholding estimators.* For any $j \in \mathbb{N}^*$, let $\Lambda(j) = \{2^j, \dots, 2^{j+1}-1\}$. Then $\mathbb{N}^* = \sum_{j\ge 0}\Lambda(j)$. We wish to define $\mathcal{M}$ as a collection of subsets of $\mathbb{N}^*$. Let $\overline{\mathcal{J}}$ be the family of all subsets of $\{0,\dots,J\}$ where $J = [\log_2(n^2)]$. We define for any $\mathcal{J}\in\overline{\mathcal{J}}$,

$$m_{\mathcal{J}} = \sum_{j\in\mathcal{J}}\Lambda(j).$$

Finally, let

$$\mathcal{M} = \{m_{\mathcal{J}},\ \mathcal{J}\in\overline{\mathcal{J}}\}.$$

For any $j\in\mathbb{N}$, let

$$nw(j) = 2^j + 1 + 2\sqrt{(2^j+1)2C\log(2^J)} + 4C\log(2^J),$$

C being some numerical constant larger than 1. Then, we define for any $\mathcal{J}\in\overline{\mathcal{J}}$,

$$\operatorname{pen}(m_{\mathcal{J}}) = \sum_{j\in\mathcal{J}} w(j). \tag{3.17}$$

The resulting penalized estimator can be made explicit as a level thresholding estimator which is an analogue (up to numerical constants) of the estimator used in Donoho and Johnstone (1999). Indeed,

$$\begin{aligned}\sup_{\mathcal{J}\in\overline{\mathcal{J}}}\left[\sum_{\lambda\in m_{\mathcal{J}}} Y_\lambda^2 - \operatorname{pen}(m_{\mathcal{J}})\right] &= \sup_{\mathcal{J}\subset\{0,\dots,J\}}\left[\sum_{j\in\mathcal{J}}\left(\sum_{\lambda\in\Lambda(j)} Y_\lambda^2 - w(j)\right)\right] \\ &= \sum_{j=0}^{J}\left[\sum_{\lambda\in\Lambda(j)} Y_\lambda^2 - w(j)\right]\mathbb{1}_{\sum_{\lambda\in\Lambda(j)} Y_\lambda^2 \ge w(j)}.\end{aligned}$$

On the other hand, the penalty defined by (3.17) satisfies condition (2.5) by setting for every $m \in \mathcal{M}$, $x_m = 2C \log(2^J)$.

It is easy to check that the results of Corollary 1 still hold for this level thresholding estimator. Note that a similar estimator has been introduced first by Gayraud and Tribouley (1999), the difference being that in their procedure one considers

$$\sum_{\lambda \in \hat{m}} Y_\lambda^2 - \frac{D_{\hat{m}}}{n}, \tag{3.18}$$

where $\hat{m} = \arg\max_{m \in \mathcal{M}} \big(\sum_{\lambda \in m} Y_\lambda^2 - \mathrm{pen}(m) \big)$, instead of

$$\sum_{\lambda \in \hat{m}} Y_\lambda^2 - \mathrm{pen}(\hat{m})$$

as above or in Johnstone (1999). Gayraud and Tribouley's proof relies on asymptotic arguments and specifically deals with level thresholding estimators. We do not know if in the generality of Theorem 1 the estimator (3.18) would have the same properties as our estimator.

3.3.2. *The Birgé–Massart algorithm.* Our purpose in this section is to design a new strategy which takes advantage of the fact that β belongs to some unknown Besov body. As compared to the strategy of the previous section, this prior information allows using a penalized estimator involving a smaller quantity of models. This leads to an improved risk bound. Our method relies on the compression algorithm proposed by Birgé and Massart (2000a). This algorithm provides for any $J \in \mathbb{N}$ a nonlinear approximation of β that we denote by $\tilde{\beta}(J)$ such that

$$|\beta - \tilde{\beta}(J)|_2 \le C(\alpha, p) R 2^{-J\alpha} \tag{3.19}$$

provided that $\beta \in \mathcal{B}_{\alpha, p, \infty}(R)$ with $\alpha > 1/p - 1/2$. Setting $\Lambda(j) = \{2^j, \ldots, 2^{j+1} - 1\}$, Birgé and Massart's procedure consists of retaining for each level of resolution j a prescribed number $K_J(j)$ of largest coefficients (in absolute value). For $j \le J$, one takes $K_J(j) = 2^j$ that is, one keeps all the coefficients while, for $j > J$, one defines $K_J(j) = [2^J/(j - J)^3]$. Note that the number of coefficients which are kept is of order 2^J. Let us now introduce the corresponding estimation procedure for $\sum_{\lambda \in \mathbb{N}^*} \beta_\lambda^2$.

We set, for any $J \in \mathbb{N}$,

$$n w^{(1)}(J) = 2^{J+1} + 1 + 2\sqrt{2(2^{J+1} + 1)\log(2^{J+1})} + 4\log(2^{J+1}).$$

We define

$$\hat{\theta}^{(1)} = \sup_{J \in \mathbb{N}} \left[\left(\sum_{j=0}^{J} \sum_{\lambda \in \Lambda(j)} Y_\lambda^2 \right) - w^{(1)}(J) \right]. \tag{3.20}$$

We denote by $\widehat{\Lambda}_J(j)$ a subset of $\Lambda(j)$ which contains the $K_J(j)$ indices corresponding to the largest values among the set $\{|Y_\lambda|,\ \lambda \in \Lambda(j)\}$. Let

$\Delta_J = \sum_{j=0}^{+\infty} K_J(j)$. We set

$$nw^{(2)}(J) = 10.5(\Delta_J + 1)$$

and we define

$$\hat{\theta}^{(2)} = \sup_{J \in \mathbb{N}} \left[\left(\sum_{j=0}^{+\infty} \sum_{\lambda \in \widehat{\Lambda}_J(j)} Y_\lambda^2 \right) - w^{(2)}(J) \right]. \tag{3.21}$$

THEOREM 4. *Assume that one observes* $(Y_\lambda)_{\lambda \in \mathbb{N}^*}$ *given by the Gaussian sequence model* (2.3) *and set* $\beta = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$, $\theta = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda^2$, *and* $L(\beta) = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda \epsilon_\lambda$.
Let $\hat{\theta}^{(1)}$ *and* $\hat{\theta}^{(2)}$ *be defined by* (3.20) *and* (3.21), *respectively. We define*

$$\hat{\theta} = \hat{\theta}^{(1)} \vee \hat{\theta}^{(2)}.$$

Let $0 < p \leq +\infty$, $\alpha > 0$, $R > 0$ *and assume that* $\alpha' = 1/2 + \alpha - 1/p > 0$. *As soon as* $nR^2 \geq 1$, *the following inequalities hold*:

(i) *If* $p < 2$,

$$\sup_{\beta \in \mathscr{B}_{\alpha, p, \infty}(R)} \mathbb{E}_\beta \left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right)^2 \right]$$
$$\leq C(p, \alpha) \inf \left\{ R^{4/(1+4\alpha')} \left(\frac{\log(1 + nR^2)}{n^2} \right)^{4\alpha'/(1+4\alpha')} ; R^{4/(1+2\alpha)} n^{-4\alpha/(1+2\alpha)} \right\},$$

where $C(p, \alpha)$ *is a constant depending on* p *and* α;
(ii) *if* $p \geq 2$, $\mathscr{B}_{\alpha, p, \infty}(R) \subseteq \mathscr{B}_{\alpha, 2, \infty}(R)$ *and*

$$\sup_{\beta \in \mathscr{B}_{\alpha, 2, \infty}(R)} \mathbb{E}_\beta \left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right)^2 \right] \leq C(\alpha) R^{4/(1+4\alpha)} \left(\frac{\log(1 + nR^2)}{n^2} \right)^{4\alpha/(1+4\alpha)},$$

where $C(\alpha)$ *is a constant depending on* α.

COMMENTS. (i) If $p = q$, the Besov body $\mathscr{B}_{\alpha, p, q}(R)$ coincides with the l_p-body $\Theta_{p, c}$ if we set $\forall\ j \in \mathbb{N}$, $\forall\ \lambda \in \Lambda(j)$, $c_\lambda = 2^{-j\alpha'}$. It is therefore interesting to compare the results of Corollary 2 and Theorem 4 in this situation. Since $\mathscr{B}_{\alpha, p, p}(R) \subset \mathscr{B}_{\alpha, p, \infty}(R)$, Theorem 4 ensures that, for $p < 2$,

$$\sup_{\beta \in \mathscr{B}_{\alpha, p, p}(R)} \mathbb{E}_\beta \left[\left(\hat{\theta} - \theta - \frac{2L(\beta)}{\sqrt{n}} \right)^2 \right]$$
$$\leq C(p, \alpha) \inf \left\{ R^{4/(1+4\alpha')} \left(\frac{\log(1 + nR^2)}{n^2} \right)^{4\alpha'/(1+4\alpha')} ; R^{4/(1+2\alpha)} n^{-4\alpha/(1+2\alpha)} \right\},$$

while in Corollary 2, the term $n^{-4\alpha/(1+2\alpha)}$ is replaced by $(n/\log(1+nR^2))^{-4\alpha/(1+2\alpha)}$. Therefore, the rate obtained in Theorem 4 is a little bit better

than the rate obtained in Corollary 2 since we save a logarithmic factor. Nevertheless, Corollary 2 is in some sense more general; for example, it allows considering situations where the sequence $(c_\lambda)_{\lambda\in\mathbb{N}^*}$ converges very slowly towards 0 as shown by (3.16).

(ii) Since there is no major difference of behavior (in terms of risk bound) between the estimator studied in Theorem 4 and the one studied in the previous section, the comments that we made about Corollary 2 are still valid here.

4. Proof of the main theorem. The key tool for proving Theorem 1 is an exponential inequality for chi-square distributions.

4.1. *An exponential inequality for chi-square distributions.* We indeed prove a slightly more general inequality than what is really necessary for the proof of Theorem 1. This generalization is painless and can prove to be helpful for other purposes.

LEMMA 1. *Let* $(Y_1,\dots,Y_D)$ *be i.i.d. Gaussian variables, with mean* 0 *and variance* 1. *Let* $a_1,\dots,a_D$ *be nonnegative. We set*

$$|a|_\infty = \sup_{i=1,\dots,D} |a_i|, \qquad |a|_2^2 = \sum_{i=1}^D a_i^2.$$

Let

$$Z = \sum_{i=1}^D a_i(Y_i^2 - 1).$$

Then, the following inequalities hold for any positive x:

$$\mathbb{P}(Z \ge 2|a|_2\sqrt{x} + 2|a|_\infty x) \le \exp(-x), \tag{4.1}$$

$$\mathbb{P}(Z \le -2|a|_2\sqrt{x}) \le \exp(-x). \tag{4.2}$$

COMMENTS. As an immediate corollary of Lemma 1, one obtains an exponential inequality for chi-square distributions. Let U be a χ^2 statistic with D degrees of freedom. For any positive x,

$$\mathbb{P}\big(U - D \ge 2\sqrt{Dx} + 2x\big) \le \exp(-x), \tag{4.3}$$

$$\mathbb{P}\big(D - U \ge 2\sqrt{Dx}\big) \le \exp(-x). \tag{4.4}$$

PROOF OF LEMMA 1. Let Y a random variable with $\mathcal{N}(0,1)$ distribution. Let ψ denote the logarithm of the Laplace transform of $Y^2 - 1$,

$$\psi(u) = \log\Big[\mathbb{E}\Big[\exp(u(Y^2-1))\Big]\Big] = -u - \tfrac{1}{2}\log(1-2u).$$

Then, for $0 < u < 1/2$,

$$\psi(u) \le \frac{u^2}{(1-2u)}.$$

Indeed,

$$\psi(u) = 2u^2 \sum_{k\geq 0} \frac{(2u)^k}{k+2} \quad \text{and} \quad \frac{u^2}{1-2u} = u^2 \sum_{k\geq 0} (2u)^k.$$

Therefore,

$$\log(\mathbb{E}[e^{uZ}]) = \sum_{i=1}^{D} \log\left(\mathbb{E}\left[\exp(a_i u(Y_i^2 - 1))\right]\right) \leq \sum_{i=1}^{D} \frac{a_i^2 u^2}{1-2a_i u}$$
$$\leq \frac{|a|_2^2 u^2}{1-2|a|_\infty u}.$$

We now refer to Birgé and Massart (1998), where it is proved that if

$$\log(\mathbb{E}[e^{uZ}]) \leq \frac{vu^2}{2(1-cu)},$$

then, for any positive x,

$$\mathbb{P}\left(Z \geq cx + \sqrt{2vx}\right) \leq e^{-x}.$$

Therefore (4.1) holds.

In order to prove (4.2), we just notice that for $-1/2 < u < 0$, $\psi(u) \leq u^2$. This concludes the proof of Lemma 1. □

We are now in position to prove Theorem 1.

4.2. *Proof of Theorem* 1. The main issue is to prove inequality (2.8). Let $V_m = \hat{\theta}_m - \theta - 2L(s)/\sqrt{n}$. By definition of $\hat{\theta}$,

$$\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}} = \sup_{m\in\mathscr{M}} V_m.$$

Moreover, since

$$\left|\sup_{m\in\mathscr{M}} V_m\right| \leq \left[\sup_{m\in\mathscr{M}} (V_m)_+\right] \vee \left[\inf_{m\in\mathscr{M}} (V_m)_-\right],$$

the following inequality holds:

$$\mathbb{E}_s\left[\left|\sup_{m\in\mathscr{M}} V_m\right|^r\right] \leq \sum_{m\in\mathscr{M}} \mathbb{E}_s\left[(V_m)_+^r\right] + \inf_{m\in\mathscr{M}} \mathbb{E}_s\left[(V_m)_-^r\right]. \tag{4.5}$$

We turn now to the control of $\mathbb{E}_s[(V_m)_+^r]$ for all $m \in \mathscr{M}^*$. Let us consider an orthonormal basis of S_m denoted by $(\phi_\lambda, \lambda \in \Lambda_m)$ where the cardinality of Λ_m equals D_m. Let, for $\lambda \in \Lambda_m$, $\beta_\lambda = \langle s, \phi_\lambda\rangle$. We recall that

$$s_m = \sum_{\lambda\in\Lambda_m} \beta_\lambda \phi_\lambda, \qquad \hat{s}_m = \sum_{\lambda\in\Lambda_m} Y(\phi_\lambda)\phi_\lambda.$$

By orthogonality, we obtain

$$V_m = \|\hat{s}_m - s_m\|^2 - \mathrm{pen}(m) + 2\langle s_m, \hat{s}_m - s_m\rangle - \frac{2}{\sqrt{n}}L(s) - \|s - s_m\|^2$$
$$= \frac{1}{n}\sum_{\lambda\in\Lambda_m} L^2(\phi_\lambda) - \mathrm{pen}(m) + \frac{2}{\sqrt{n}}L(s_m - s) - \|s - s_m\|^2.$$

Using the inequality $2ab \le a^2 + b^2$ leads to

$$V_m \le \frac{1}{n}\sum_{\lambda\in\Lambda_m} L^2(\phi_\lambda) - \mathrm{pen}(m) + \frac{L^2(s - s_m)}{n\|s - s_m\|^2}.$$

The variable $Z_m = \sum_{\lambda\in\Lambda_m} L^2(\phi_\lambda)$ is a χ^2 statistic with D_m degrees of freedom. Let

$$W_m = \frac{L(s - s_m)}{\|s - s_m\|}.$$

W_m has a $\mathcal{N}(0, 1)$ distribution and is independent of $(L(\phi_\lambda), \lambda \in \Lambda_m)$; hence, if we set $U_m = Z_m + W_m^2$, then U_m is a χ^2 statistic with $D_m + 1$ degrees of freedom and $V_m \le U_m/n - \mathrm{pen}(m)$. Therefore, it remains to control the deviations of U_m. This can be performed via inequality (4.3) taking into account condition (2.5) which in turn yields

$$\mathbb{P}(nV_m \ge h(\xi)) \le e^{-x_m}e^{-\xi},$$

where $h(\xi) = 2\sqrt{(D_m + 1)\xi} + 2\xi$. Using the identity

$$\mathbb{E}_s\left[(V_m)_+^r\right] = \frac{r}{n^r}\int_0^\infty t^{r-1}\mathbb{P}(nV_m \ge t)\,dt,$$

and the elementary inequality

$$h^{-1}(t) \ge \frac{t^2}{4((D_m + 1) + t)} \ge \frac{t^2}{8(D_m + 1)} \wedge \frac{t}{8},$$

we obtain

$$\int_0^{+\infty} t^{r-1}\mathbb{P}(nV_m \ge t)\,dt \le \exp(-x_m)\Bigg[(D_m + 1)^{r/2}\int_0^{+\infty} y^{r-1}\exp(-y^2/8)\,dy + \int_0^{+\infty} y^{r-1}\exp(-y/8)\,dy\Bigg].$$

Hence,

$$\mathbb{E}_s\left[(V_m)_+^r\right] \le \frac{C(r)}{n^r}e^{-x_m}D_m^{r/2}.$$

For $m = 0$, we similarly get $V_m \le W_0^2/n$ where W_0 is a standard Gaussian variable. Therefore, if we define Γ_r by $\Gamma_r = (1/\sqrt{2\pi})\int_0^\infty x^r\exp(-x^2/2)\,dx$, then

$$\mathbb{E}_s\left[(V_0)_+^r\right] \le \frac{\Gamma_{2r}}{n^r},$$

which, by (4.5), concludes the proof of (2.8). Let us now prove (2.9). We recall that for all $m \in \mathcal{M}$,

$$-V_m = -\frac{Z_m}{n} + \text{pen}(m) + \|s - s_m\|^2 + \frac{2}{\sqrt{n}}(L(s - s_m)).$$

Using the convexity, or the subadditivity of the function $x \to x^r$, whether $r \geq 1$ or $r < 1$, we obtain

$$(V_m)_-^r \leq 4^{(r-1)_+}\left[\frac{(D_m - Z_m)_+^r}{n^r} + \left(\text{pen}(m) - \frac{D_m}{n}\right)^r\right] + 4^{(r-1)_+}\left[\|s - s_m\|^{2r} + \left(\frac{2}{\sqrt{n}}\right)^r \mathbb{E}_s\left[(L(s - s_m))_+^r\right]\right].$$

Using (4.4), we get

$$\mathbb{E}_s\left[(D_m - Z_m)_+^r\right] \leq \sqrt{2\pi} 2^r D_m^{r/2} r\Gamma_{r-1}.$$

Moreover,

$$\mathbb{E}_s\left[(L(s - s_m))_+^r\right] = \|s - s_m\|^r \Gamma_r.$$

Using again the inequality $2ab \leq a^2 + b^2$,

$$\left(\frac{2}{\sqrt{n}}\right)^r \mathbb{E}_s\left[(L(s - s_m))_+^r\right] \leq 2^{r-1}\Gamma_r\left[n^{-r} + \|s - s_m\|^{2r}\right]$$

and (2.9) follows. □

5. Proofs of the results about the Gaussian sequence model. In order to prove Theorems 2, 3 and 4, we shall apply Theorem 1. We now introduce some notations that will be used throughout these proofs. We consider the Hilbert space $\mathbb{H} = l_2(\mathbb{N}^*)$ with its canonical basis $(\phi_\lambda,\ \lambda \in \mathbb{N}^*)$ and recall that when one observes $(Y_\lambda)_{\lambda \in \mathbb{N}^*}$, as defined by (2.3), one can define a Gaussian linear process $Y(\cdot)$ with mean $s = \beta = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$ and variance $1/n$ by setting $Y(t) = \sum_{\lambda \in \mathbb{N}^*} t_\lambda Y_\lambda$. When applying Theorem 1, we shall consider collections of models $(S_m)_{m \in \mathcal{M}}$ where S_m is defined as the linear span of $(\phi_\lambda,\ \lambda \in \Lambda_m)$ for some subset Λ_m of $\mathbb{N}^*$ and therefore has dimension $D_m = |\Lambda_m|$. The precise description of the collection of sets $(\Lambda_m)_{m \in \mathcal{M}}$ will depend on the theorem to be proved. It should be noticed that for every $m \in \mathcal{M}$, the orthogonal projection of s over S_m and accordingly, the projection estimator of s over S_m, can be described by their expansions on the basis $(\phi_\lambda,\ \lambda \in \mathbb{N}^*)$ more preceisely,

$$s_m = \sum_{\lambda \in \Lambda_m} \beta_\lambda \phi_\lambda, \qquad \hat{s}_m = \sum_{\lambda \in \Lambda_m} Y_\lambda \phi_\lambda.$$

In the sequel, we shall denote by C some constants whose values may vary from one line to another; we shall always mention the dependency of these constants with respect to the parameters involved in the problem, that is, $C(\alpha)$ stands for a constant depending only on α.

5.1. *Proof of Theorem* 2. We set here $\mathcal{M} = \mathbb{N}^*$ and for every $m \in \mathcal{M}$, $\Lambda_m = \{1, 2, \ldots, m\}$. We consider the penalties $\text{pen}(m)$ and the weights x_m of Definition 3 and apply Theorem 1 to the corresponding penalized estimator. We notice that

$$\Sigma_r = \sum_{m \geq 1} m^{r/2} \exp(-K \log(m+1)) \leq \sum_{m \geq 1} m^{-K+r/2} < \infty$$

since $K > 1 + r/2$, and so assumption (2.7) is fulfilled.

It follows from (2.8) and (2.9) that for any $r > 0$, and for any $s \in \mathcal{S}_\gamma$,

$$\mathbb{E}_s\left[\left|\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}}\right|^r\right] \leq C(r)\left(T_n + \frac{\Sigma_r}{n^r}\right),$$

where

$$T_n = \inf_{m \in \mathbb{N}^*}\left[\|s - s_m\|^{2r} + \left(\frac{m \log(m+1)}{n^2}\right)^{r/2}\right].$$

Assuming that the sequence $s = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$ belongs to the set $\mathcal{S}_\gamma$ implies that

$$\|s - s_m\|^2 = \sum_{\lambda > m} \beta_\lambda^2 \leq \gamma_m^2.$$

This concludes the proof of Theorem 2 by possibly enlarging $C(r)$. □

5.2. *Proof of Corollary* 1. If the sequence $s = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$ belongs to the Besov body $\mathcal{B}_{\alpha, 2, \infty}(R)$, then $\forall\, m \in \mathbb{N}^*$,

$$\sum_{\lambda > m} \beta_{\lambda^2} \leq R^2 m^{-2\alpha} \frac{2^{4\alpha}}{2^{2\alpha} - 1}.$$

Hence, by Theorem 2, $\forall\, r \leq 2(K-1)$,

$$\sup_{s \in \mathcal{B}_{\alpha, 2, \infty}(R)} \mathbb{E}_s\left[\left|\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}}\right|^r\right]$$

$$\leq C(r, \alpha) \inf_{m \in \mathbb{N}^*}\left[R^{2r} m^{-2r\alpha} + \left(\frac{m \log(m+1)}{n^2}\right)^{r/2}\right].$$

We set

$$m_n = \left[\left(\frac{n^2 R^4}{\log(1 + n^2 R^4)}\right)^{1/1+4\alpha}\right].$$

Since for every positive x, $x \geq \log(1+x)$, $m_n \geq 1 \vee (1/2)(n^2 R^4 / \log(1 + n^2 R^4))^{1/1+4\alpha}$. Moreover, since $nR^2 \geq 1$, $m_n \leq 2n^2 R^4$ and

$$\log(m_n + 1) \leq \log(1 + 2n^2 R^4) \leq 2 \log(1 + n^2 R^4).$$

Therefore,

$$\sup_{s\in\mathscr{B}_{\alpha,2,\infty}(R)}\mathbb{E}_s\left[\left|\hat{\theta}-\theta-\frac{2L(s)}{\sqrt{n}}\right|^r\right]\le C(r,\alpha)\left[R^{2r/(1+4\alpha)}\left(\frac{\log(1+n^2R^4)}{n^2}\right)^{2r\alpha/(1+4\alpha)}\right]$$
$$\le C(r,\alpha)\left[R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)}\right],$$

hence, (3.4) is proved. This implies that

$$\sup_{s\in\mathscr{B}_{\alpha,2,\infty}(R)}\mathbb{E}_s\left[\left|\hat{\theta}-\theta\right|^r\right]\le C(r,\alpha)\left[R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)}+R^rn^{-r/2}\right]$$

since for any $s\in\mathscr{B}_{\alpha,2,\infty}(R)$, $\mathbb{E}_s(|L(s)/\sqrt{n}|^r)\le C(r,\alpha)R^rn^{-r/2}$.

Conditions $\alpha\le 1/4$ and $nR^2\ge 1$ imply that

$$R^rn^{-r/2}\le C(r,\alpha)R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)},$$

hence, (3.4) holds.

If $\alpha>1/4$ then

$$R^{2r/(1+4\alpha)}\left(\frac{\log(1+nR^2)}{n^2}\right)^{2r\alpha/(1+4\alpha)}\le C(r,\alpha)R^rn^{-r/2},$$

and therefore (3.5) holds.

If R and α are given with $R>0$, $\alpha>1/4$ and n goes to infinity, then $(n^2/\log(n))^{-2r\alpha/(1+4\alpha)}=o(n^{-r/2})$; hence for any $s\in\mathscr{B}_{\alpha,2,\infty}(R)$, $\mathbb{E}_s[|\sqrt{n}(\hat{\theta}-\theta)-2L(s)|^r]\to 0$ as $n\to\infty$. Since $L(s)$ has a $\mathscr{N}(0,\theta)$ distribution, this implies that

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{\mathscr{D}}\mathscr{N}(0,4\theta)\quad\text{as } n\to\infty.$$

If $r\ge 1$, by the triangle inequality,

$$n^{r/2}\mathbb{E}_s\left[|\hat{\theta}-\theta|^r\right]\to 2^r\mathbb{E}_s(|L(s)|^r)=2^r\theta^{r/2}\mathbb{E}(|\xi|^r)\quad\text{as } n\to\infty,$$

where ξ is a standard normal variable. □

5.3. *Proof of Theorem* 3. We define

$$\mathscr{M}^{(1)}=\mathbb{N}^*,$$
$$\mathscr{M}^{(2)}=\{m=(N,A_N),\ A_N\in\mathscr{P}(1,2,\ldots,N),\ N\in\mathbb{N}^*\},$$

where $\mathscr{P}(1,2,\ldots,N)$ denotes the set of all nonempty subsets of $\{1,2,\ldots,N\}$.

Let $\mathscr{M}=\mathscr{M}^{(1)}\times\{1\}\oplus\mathscr{M}^{(2)}\times\{2\}$. Let $m\in\mathscr{M}$, if $m=m_1\times\{1\}$, we set $\Lambda_m=\Lambda_{m_1}=\{1,2,\ldots,m_1\}$, and if $m=m_2\times\{2\}$, where $m_2=(N,A_N)$, we set $\Lambda_m=\Lambda_{m_2}=A_N$. If $m=m_1\times\{1\}$, we consider the penalty $\mathrm{pen}(m)=\mathrm{pen}(m_1)$ and the weight $x_m=x_{m_1}$ of Definition 3 with $K=3$. If $m=m_2\times\{2\}$, with $m_2=(N,A_N)$, we set $x_m=x_{N,|A_N|}$ [where $x_{N,D}$ is defined by (3.9)] and

$\mathrm{pen}(m) = w(N, |A_N|)$ [where $w(N, D)$ is given by (3.10)]. It is clear that

$$\hat{\theta}^{(1)} = \sup_{m_1 \in \mathscr{M}^{(1)}} \left[\sum_{\lambda \in \Lambda_{m_1}} Y_\lambda^2 - \mathrm{pen}(m_1) \right]$$

and since when N is given, for $m_2 = (N, A_N) \in \mathscr{M}^{(2)}$, the penalty of m_2 depends on A_N only through its cardinality, one has

$$\hat{\theta}^{(2)} = \sup_{m_2 \in \mathscr{M}^{(2)}} \left[\sum_{\lambda \in \Lambda_{m_2}} Y_\lambda^2 - \mathrm{pen}(m_2) \right]$$

and therefore

$$\hat{\theta} = \hat{\theta}^{(1)} \vee \hat{\theta}^{(2)} = \sup_{m \in \mathscr{M}} \left[\sum_{\lambda \in \Lambda_m} Y_\lambda^2 - \mathrm{pen}(m) \right].$$

Hence we can apply Theorem 1 to $\hat{\theta}$. To do so we have to check assumption (2.7) with $r = 2$. We note that $\Sigma_2 = S^{(1)} + S^{(2)}$ with

$$S^{(i)} = \sum_{m \in \mathscr{M}^{(i)}} D_m e^{-x_m}.$$

We have to control the series $S^{(i)}$ for $i = 1, 2$. We first note that

$$S^{(1)} = \sum_{m \geq 2} m^{-2} \leq 1.$$

Moreover,

$$S^{(2)} \leq \sum_{N \in \mathbb{N}^*} \sum_{1 \leq D \leq N} \binom{N}{D} D \exp\left[-3D\left(1 + \log\left(\frac{N}{D}\right)\right)\right].$$

Since

$$\binom{N}{D} \leq \left(\frac{eN}{D}\right)^D,$$

we derive that

$$S^{(2)} \leq \sum_{N \in \mathbb{N}^*} \sum_{1 \leq D \leq N} D e^{-2D} \left(\frac{N}{D}\right)^{-2D}.$$

Using the fact that $De^{-2D} \leq 1$, for $D \leq N^{1/4}$ and $(N/D)^{-D} \leq 1$ for $N^{1/4} < D \leq N$, one has

$$\sum_{1 \leq D \leq N} D e^{-2D} \left(\frac{N}{D}\right)^{-2D} \leq \sum_{1 \leq D \leq N^{1/4}} \left(\frac{N}{D}\right)^{-2D} + \sum_{N^{1/4} < D \leq N} D e^{-2D},$$

which leads to

$$\sum_{1 \leq D \leq N} D e^{-2D} \left(\frac{N}{D}\right)^{-2D} \leq \frac{N^{-3/2}}{1 - N^{-3/2}} + N^2 e^{-N/2}.$$

Therefore the series $S^{(2)}$ is convergent.

It follows from Theorem 1 that for any $s \in \Theta_{p,c}$,

$$\mathbb{E}_s\left[\left(\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}}\right)^2\right] \leq C\left(T_n^{(1)} \wedge T_n^{(2)} + \frac{\Sigma_2}{n^2}\right) \tag{5.1}$$

and for $i = 1, 2$,

$$T_n^{(i)} = \inf_{m \in \mathcal{M}^{(i)}} \left\{\left(\sum_{\lambda \notin \Lambda_m} \beta_\lambda^2\right)^2 + \frac{D_m}{n^2} + \left(\mathrm{pen}(m) - \frac{D_m}{n}\right)^2\right\},$$

(5.1) ensures that $\hat{\theta}$ performs as well as $\hat{\theta}^{(1)}$ and therefore (3.12) can be derived from Theorem 2 and Comment (ii) following Theorem 2.

Let $p < 2$; in order to control $T_n^{(1)}$, we notice that $s = \beta$ belongs to the l_p-body $\Theta_{p,c}$, which means that $\sum_{\lambda > D} |\beta_\lambda|^p \leq c_D^p$. Using the subadditivity of the function $x \mapsto x^{p/2}$ for $p \leq 2$, one derives that

$$\sum_{\lambda > D} \beta_\lambda^2 \leq \left(\sum_{\lambda > D} |\beta_\lambda|^p\right)^{2/p} \leq c_D^2.$$

Hence,

$$T_n^{(1)} \leq \inf_{D \in \mathbb{N}^*}\left[c_D^4 + \left(\frac{D\log(D)}{n^2}\right)\right].$$

It remains to control $T_n^{(2)}$. For $D \in \mathbb{N}^*$ we define $\varepsilon_D = D^{-1/p} c_D$ and $G_D = \{\lambda \in \{1, 2, \ldots, N\}, |\beta_\lambda| \geq \varepsilon_D\}$. Then, since $\sum_{\lambda > D} |\beta_\lambda|^p \leq c_D^p$,

$$\begin{aligned}
\sum_{\lambda \notin G_D} \beta_\lambda^2 &\leq \sum_{\lambda \notin G_D,\, \lambda \leq D} \beta_\lambda^2 + \sum_{\lambda \notin G_D,\, D < \lambda \leq N} \beta_\lambda^2 + \sum_{\lambda > N} \beta_\lambda^2 \\
&\leq D\varepsilon_D^2 + \varepsilon_D^{2-p} \sum_{\lambda > D} |\beta_\lambda|^p + c_N^2 \\
&\leq D\varepsilon_D^2 + \varepsilon_D^{2-p} c_D^p + c_N^2 \\
&\leq 2D^{1-\frac{2}{p}} c_D^2 + c_N^2
\end{aligned}$$

by definition of ε_D. To bound the cardinality of G_D, we note that

$$c_D^p \geq \sum_{\lambda \in G_D,\, \lambda > D} |\beta_\lambda|^p \geq \varepsilon_D^p |\{G_D \cap \{\lambda, \lambda > D\}\}|,$$

which implies that $|\{G_D \cap \{\lambda, \lambda > D\}\}| \leq c_D^p \varepsilon_D^{-p} = D$. Therefore, the cardinality of G_D is bounded by 2D. This leads to

$$T_n^{(2)} \leq C \inf_{N \in \mathbb{N}^*} \inf_{1 \leq D \leq N} \left\{\left(D^{1-2/p} c_D^2\right)^2 + \left(\frac{D(1 + \log(N/D))}{n}\right)^2 + c_N^4\right\}.$$

This concludes the proof of Theorem 3. □

5.4. *Proof of Corollary* 2. We first notice that (3.12) ensures that $\hat{\theta}$ behaves as well as the estimator studied in Corollary 1; therefore (3.15) follows from (3.4). We turn now to the proof of (3.14).

Let $p \le 2$; we derive from Theorem 3 that

$$\sup_{s\in\mathscr{C}_{p,\alpha'}(R)} \mathbb{E}_s\left[\left(\hat{\theta}-\theta-\frac{2L(s)}{\sqrt{n}}\right)^2\right] \le C\inf\{v_1(n); v_2(n)\},$$

where

$$v_1(n) = \inf_{N\in\mathbb{N}^*}\left\{\inf_{D\in\mathbb{N}^*}\left[\left(D^{1-(2/p)}c_D^2\right)^2+\left(\frac{D(1+\log(N/D))}{n}\right)^2\right]+c_N^4\right\},$$

$$v_2(n) = \inf_{D\in\mathbb{N}^*}\left[c_D^4+\frac{D\log(D+1)}{n^2}\right].$$

Here, $c_\lambda = R\lambda^{-\alpha'}$. Let

$$D_1(n) = \left[\left(\frac{nR^2}{\log(1+nR^2)}\right)^{1/(1+2\alpha)}\right] \quad \text{and} \quad N(n) = \left[(nR^2)^{\alpha/(\alpha'(1+2\alpha))}\right].$$

Note that $D_1(n) \ge 1$ and that $N(n) \ge 1$. Since for $p \le 2$, $\alpha > \alpha'$,

$$\log\left(\frac{N(n)}{D_1(n)}\right) \le C(p,\alpha)\log(1+nR^2)$$

which ensures that

$$v_1(n) \le C(p,\alpha)R^{4/(1+2\alpha)}\left(\frac{\log(1+nR^2)}{n}\right)^{4\alpha/(1+2\alpha)}.$$

Moreover, we set

$$D_2(n) = \left[\left(\frac{n^2R^4}{\log(1+n^2R^4)}\right)^{1/1+4\alpha'}\right].$$

Since $D_2(n) \ge 1$, we obtain by similar computations as in the proof of Corollary 1 that

$$v_2(n) \le C(p,\alpha)R^{4/(1+4\alpha')}\left(\frac{\log(1+nR^2)}{n^2}\right)^{4\alpha'/(1+4\alpha')}.$$

This concludes the proof of Corollary 2. □

PROOF OF (3.16). When $c_\lambda = R(\log(\lambda))^{-\alpha'}$, let $N\in\mathbb{N}^*$ satisfy $N\,n^{n^{(1/\alpha')(1/p-1/2)}} \le N \le Cn^{n^{(1/\alpha')(1/p-1/2)}}$. Setting $D_1(n) = [n^{(p/2-p/2\alpha')(1/p-1/2)}]$ leads to $v_1(n) \le C(R,p,\alpha)n^{(2-p)((1/\alpha')(1/p-1/2)-1)}\log^2(1+n)$. Moreover, let $D_2(n) = [n^2/(\log(1+n))^{1+4\alpha'}]$; we get $v_2(n) \le C(R,\alpha)(\log(1+n))^{-4\alpha'}$.

5.5. *Proof of Theorem* 4. We set

$$\mathscr{M}^{(1)} = \mathbb{N}$$

and for any $J \in \mathbb{N}$,

$$\mathscr{M}_J^{(2)} = \left\{ m \subset \mathbb{N}^*, \forall\, j \geq 0, |m \cap \Lambda(j)| = K_J(j) \right\}$$

which leads to the definition of $\mathscr{M}^{(2)}$ as

$$\mathscr{M}^{(2)} = \bigcup_{J \in \mathbb{N}} \mathscr{M}_J^{(2)}.$$

Let $\mathscr{M} = \mathscr{M}^{(1)} \times \{1\} \oplus \mathscr{M}^{(2)} \times \{2\}$. Let $m \in \mathscr{M}$.

(i) If $m = J \times \{1\}$, with $J \in \mathbb{N}$, we set $\Lambda_m = \Lambda_J = \cup_{j=0}^{J} \Lambda(j)$, $x_m = x_J = 2\log(D_m)$ and $\mathrm{pen}(m) = \mathrm{pen}(J) = w^{(1)}(J)$.

(ii) If $m = m_2 \times \{2\}$ with $m_2 \in \mathscr{M}_J^{(2)}$, we set $\Lambda_m = \Lambda_{m_2} = m_2$, $x_m = x_{m_2} = 3D_m$ and $\mathrm{pen}(m) = \mathrm{pen}(m_2) = w^{(2)}(J)$.

Note that with our definitions of the penalties and the weights, it is easy to check that for any $m \in \mathscr{M}$,

$$n\mathrm{pen}(m) \geq D_m + 1 + 2\sqrt{(D_m + 1)x_m} + 2x_m,$$

which is the required assumption on the penalty function in Theorem 1. Moreover, by definition,

$$\hat{\theta}^{(1)} = \sup_{m_1 \in \mathscr{M}^{(1)}} \left(\sum_{\lambda \in m_1} Y_\lambda^2 - \mathrm{pen}(m_1) \right)$$

and since, when J is given, for $m_2 \in \mathscr{M}_J^{(2)}$ the supremum of $\sum_{\lambda \in m_2} Y_\lambda^2$ is achieved for $m_2 = \sum_{j=0}^{+\infty} \widehat{\Lambda}_J(j)$, one has

$$\begin{aligned} \hat{\theta}^{(2)} &= \sup_{J \geq 0} \sup_{m_2 \in \mathscr{M}_J^{(2)}} \left(\sum_{\lambda \in m_2} Y_\lambda^2 - w^{(2)}(J) \right) \\ &= \sup_{m_2 \in \mathscr{M}^{(2)}} \left(\sum_{\lambda \in m_2} Y_\lambda^2 - \mathrm{pen}(m_2) \right). \end{aligned}$$

Hence,

$$\hat{\theta} = \sup_{m \in \mathscr{M}} \left(\sum_{\lambda \in m} Y_\lambda^2 - \mathrm{pen}(m) \right)$$

and therefore we can apply Theorem 1 to $\hat{\theta}$ provided that assumption (2.7) is fulfilled with $r = 2$. In order to check (2.7), we notice that $\Sigma_2 = S^{(1)} + S^{(2)}$ with

$$S^{(i)} = \sum_{m \in \mathscr{M}^{(i)}} D_m e^{-x_m},$$

$$S^{(1)} = \sum_{J \geq 0} \left(2^{J+1} - 1\right)^{-1} \leq 2, \tag{5.2}$$

$$S^{(2)} = \sum_{J \geq 0} \Delta_J |\mathscr{M}_J^{(2)}| e^{-3\Delta_J},$$

where we recall that $\Delta_J = \sum_{j=0}^{\infty} K_J(j)$. Now,

$$|\mathscr{M}_J^{(2)}| = \prod_{j > J} \binom{2^j}{K_J(j)},$$

where the product is indeed finite since $\binom{2^j}{K_J(j)} = 1$ for j large enough. Using the inequality

$$\log \binom{k}{[kx]} \leq kx\left(1 + \log\left(\frac{1}{x}\right)\right),$$

which holds for any $x \in]0, 1]$ and $k \in \mathbb{N}^*$, we derive that

$$\begin{aligned}
\log\left(|\mathscr{M}_J^{(2)}|\right) &\leq \sum_{j > J} \frac{2^J}{(j - J)^3}\left[1 + \log\left(\frac{(j - J)^3}{2^{J-j}}\right)\right] \\
&\leq 2^J\left[\sum_{l \geq 1} \frac{1}{l^3} + \log(2) \sum_{l \geq 1} \frac{1}{l^2} + 3 \sum_{l \geq 1} \frac{\log(l)}{l^3}\right] \\
&\leq C_3 2^J,
\end{aligned}$$

where

$$C_3 = \sum_{l \geq 1} \frac{1}{l^3} + \log(2) \sum_{l \geq 1} \frac{1}{l^2} + 3 \sum_{l \geq 1} \frac{\log(l)}{l^3} < 3.$$

Hence,

$$|\mathscr{M}_J^{(2)}| \leq \exp(C_3 2^J). \tag{5.3}$$

Since $\Delta_J \geq 2^J$ and $x \to xe^{-3x}$ is decreasing on $[1, \infty[$, combining (5.2) and (5.3) yields

$$S^{(2)} \leq \sum_{J \geq 0} 2^J \exp(C_3 2^J) \exp(-32^J) < \infty.$$

It follows that the series Σ_2 is convergent. We get by Theorem 1 the following risk bound:

$$\mathbb{E}_s\left[\left(\hat{\theta} - \theta - \frac{2L(s)}{\sqrt{n}}\right)^2\right] \leq C\left[T_n^{(1)} \wedge T_n^{(2)} + \frac{\Sigma_2}{n^2}\right],$$

where

$$T_n^{(1)} = \inf_{J\geq 0}\left[\left(\sum_{\lambda\notin\Lambda_J}\beta_\lambda^2\right)^2 + \left(w^{(1)}(J) - \frac{|\Lambda_J|}{n}\right)^2\right],$$

$$T_n^{(2)} = \inf_{J\geq 0}\inf_{m\in\mathscr{M}_J^{(2)}}\left[\left(\sum_{\lambda\notin m}\beta_\lambda^2\right)^2 + \left(w^{(2)}(J) - \frac{\Delta_J}{n}\right)^2\right].$$

Let $\beta \in \mathscr{B}_{\alpha,\,p,\,\infty}(R)$.

1. If $p \geq 2$, then by convexity of the function $x \mapsto x^{p/2}$,

$$\sum_{\lambda\in\Lambda(j)}\beta_\lambda^2 \leq \left(|\Lambda(j)|^{p/2-1}\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^p\right)^{2/p} \leq R^2 2^{-2j\alpha}$$

   which ensures that $\mathscr{B}_{\alpha,\,p,\,\infty}(R) \subseteq \mathscr{B}_{\alpha,\,2,\,\infty}(R)$.
2. If $p < 2$, by subadditivity of the function $x \mapsto x^{p/2}$,

$$\sum_{\lambda\in\Lambda(j)}\beta_\lambda^2 \leq \left(\sum_{\lambda\in\Lambda(j)}|\beta_\lambda|^p\right)^{2/p} \leq R^2 2^{-2j\alpha'},$$

   where $\alpha' = 1/2 + \alpha - 1/p$.

This implies that for every $J \in \mathscr{M}^{(1)} = \mathbb{N}$ and any $p > 0$,

$$\sum_{\lambda\notin\Lambda_J}\beta_\lambda^2 = \sum_{j>J}\sum_{\lambda\in\Lambda(j)}\beta_\lambda^2 \leq C(\alpha')R^2 2^{-2J\alpha''},$$

where $\alpha'' = \inf(\alpha, \alpha')$ and since $|\Lambda_J| \leq 2^{J+1}$,

$$\left(w^{(1)}(J) - \frac{|\Lambda_J|}{n}\right)^2 \leq C\left(\frac{2^J(J+1)}{n^2}\right).$$

Hence

$$T_n^{(1)} \leq C\inf_{J\geq 0}\left[R^4 2^{-4J\alpha''} + \frac{2^J(J+1)}{n^2}\right].$$

Let

$$J_n^{(1)} = \left[\frac{1}{1+4\alpha''}\log_2\left(\frac{n^2R^4}{\log(1+n^2R^4)}\right)\right].$$

This leads to

$$T_n^{(1)} \leq C(p,\alpha)R^{4/(1+4\alpha'')}\left(\frac{\log(1+nR^2)}{n^2}\right)^{4\alpha''/(1+4\alpha'')}.$$

We now turn to the control of $T_n^{(2)}$ for $p < 2$.

It follows from Birgé and Massart (2000a) that for any $J \in \mathbb{N}$ there exists $m \in \mathcal{M}_J^{(2)}$ such that $\sum_{\lambda \notin m} \beta_\lambda^2 \leq C(p,\alpha) R^2 2^{-2J\alpha}$. Therefore, since $\Delta_J \leq \kappa 2^J$ where κ is an absolute constant,

$$T_n^{(2)} \leq C(p,\alpha) \inf_{J \geq 0} \left(R^4 2^{-4J\alpha} + \frac{2^{2J}}{n^2} \right).$$

Let

$$J_n^{(2)} = \left[\frac{1}{1+2\alpha} \log_2(nR^2) \right].$$

This leads to

$$T_n^{(2)} \leq C(p,\alpha) R^{4/(1+2\alpha)} n^{-4\alpha/(1+2\alpha)}.$$

This concludes the proof of Theorem 4. □

REFERENCES

BARAUD, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117** 467–493.

BARRON, A. R., BIRGÉ, L. and MASSART, P. (1999). Risk bound for model selection via penalization. *Probab. Theory Related Fields* **113** 301–415.

BICKEL, P. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.) 55–87. Springer, New York.

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

BIRGÉ, L. and MASSART, P. (2000a). An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16** 1–36.

BIRGÉ, L. and MASSART, P. (2000b). Gaussian model selection. Technical Report 2000.05, Univ. Paris Sud.

DEVORE, R. A., JAWERTH, B. and POPOV, V. (1992). Compression of wavelet decompositions. *Amer. J. Math.* **114** 737–785.

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer, New York.

DEVORE, R. A., KYRIAZIS, G. LEVIATAN, D. and TIKHOMIROV, V. M. (1993). Wavelet compression and nonlinear n-widths. *Adv. Comput. Math.* **1** 197–214.

JOHNSTONE, I. (1999). Chi-square oracle inequalities. Preprint.

DONOHO, D. and JOHNSTONE, I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

DONOHO, D. and LIU, R. (1991). Geometrizing rates of convergence II. *Ann. Statist.* **19** 633–668.

DONOHO, D. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

DUDLEY, R. M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1** 66–103.

EFROÏMOVICH, S. and LOW, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24** 1106–1125.

GAYRAUD, G. and TRIBOULEY, K. (1999). Wavelet methods to estimate an integrated quadratic functional: adaptivity and asymptotic law. *Statist. Probab. Lett.* **44** 109–122.

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681.

LEPSKII, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKII, O. V. (1992). On problems of adaptive estimation in Gaussian white noise. *Adv. Soviet Math.* **12** 87–106.

LABORATOIRE DE MATHÉMATIQUES
BAT. 425
UNIVERSITÉ PARIS SUD
F-91405 ORSAY CÉDEX
FRANCE
E-MAIL: Beatrice.Laurent@math.u-psud.fr
Pascal.Massart@math.u-psud.fr