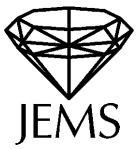
J. Eur. Math. Soc. 3, 203–268 (2001)

Digital Object Identifier (DOI) 10.1007/s100970100031



Lucien Birgé · Pascal Massart

Gaussian model selection

Received February 1, 1999 / final version received January 10, 2001
Published online April 3, 2001 –

Abstract. Our purpose in this paper is to provide a general approach to model selection via penalization for Gaussian regression and to develop our point of view about this subject. The advantage and importance of model selection come from the fact that it provides a suitable approach to many different types of problems, starting from model selection per se (among a family of parametric models, which one is more suitable for the data at hand), which includes for instance variable selection in regression models, to nonparametric estimation, for which it provides a very powerful tool that allows adaptation under quite general circumstances. Our approach to model selection also provides a natural connection between the parametric and nonparametric points of view and copes naturally with the fact that a model is not necessarily true. The method is based on the penalization of a least squares criterion which can be viewed as a generalization of Mallows' C_p. A large part of our efforts will be put on choosing properly the list of models and the penalty function for various estimation problems like classical variable selection or adaptive estimation for various types of l_p-bodies.

1. Introducing model selection from a nonasymptotic point of view

Choosing a proper parameter set is a difficult task in many estimation problems. A large one systematically leads to a large risk while a small one may result in the same consequence, due to unduly large bias. Both excessively complicated or oversimplified models should be avoided. The dilemna of the choice, between many possible models, of one which is adequate for the situation at hand, depending on both the unknown complexity of the true parameter to be estimated and the known amount of noise or number of observations, is often a nightmare for the statistician. The purpose of this paper is to provide a general methodology, namely *model selection via penalization*, for solving such problems within a unified Gaussian framework which covers many classical situations involving Gaussian variables.

L. Birgé: UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris VI, 4 Place Jussieu, 75252 Paris Cedex 05, France,
e-mail: LB@CCR.JUSSIEU.FR

P. Massart: UMR 8628 "Laboratoire de Mathématiques", Bât. 425, Université Paris Sud, Campus d'Orsay, 91405 Orsay Cedex, France,
e-mail: PASCAL.MASSART@MATH.U–PSUD.FR

Mathematics Subject Classification (1991): 62G07, 62C20, 41A46

Our approach to model selection via penalization has been inspired by the pioneering paper of Barron and Cover (1991) and first introduced in the context of density estimation in Birgé and Massart (1997). It was then developed at length by Barron et al. (1999) for various estimation problems concerning independent data but at the price of a lot of technicalities. Focusing on the simplest situation of Gaussian settings allows to describe the main specificities of our method with less technical efforts, to better emphasize the ideas underlying our approach and to get much more precise results with shorter proofs. Generalizations have been developed for general regression (possibly non-Gaussian) settings by Baraud (1997 and 2000) and Baraud et al. (1997 and 1999) and for exponential models in density estimation by Castellan (1999). A penalization method based on model complexity and which is close to ours can be found in Yang (1999).

1.1. A few classical Gaussian statistical frameworks

Let us begin our illustration of the difficulties the statistician is faced to, when choosing a proper model for an estimation problem, by a brief review of some popular Gaussian settings.

1.1.1. Gaussian linear regression

Gaussian linear regression is a statistical framework in which we observe a Gaussian vector $\boldsymbol{Y} \in \mathbb{R}^n$ with coordinates Y_i satisfying

$$Y_i = \sum_{\lambda=1}^{N} \beta_\lambda X_i^\lambda + \sigma\xi_i \quad \text{for } 1 \le i \le n, \tag{1.1}$$

where the random variables ξ_i are i.i.d. standard normal while the X_i^λs are deterministic observable quantities and the β_λs, $1 \le \lambda \le N$, unknown real parameters. We moreover assume here that σ is known. This corresponds to a situation where one observes some real variables (here "variable" is taken in its physical sense, not the probabilistic one) $X^1, \dots, X^N$ and Y at n different times or under n different circumstances. This results in n groups of values of those variables $(X_i^1, \dots, X_i^N, Y_i)$ for $1 \le i \le n$, each group corresponding to a time of observation or a particular experiment. We denote by $\boldsymbol{Y} = (Y_i)_{1\le i\le n}$ and $\boldsymbol{X}^1, \dots, \boldsymbol{X}^N$ the corresponding vectors. In this setting the main assumption is that the variable of interest Y is a linear (but otherwise unknown) function of the *explanatory* variables $X^1, \dots, X^N$ plus some random perturbation. We want to estimate the parameters β_λ or equivalently the mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}]$ of the Gaussian vector $\boldsymbol{Y}$, assuming that it belongs to the N-dimensional linear subspace of $\mathbb{R}^n$ generated by $\boldsymbol{X}^1, \dots, \boldsymbol{X}^N$.

As a particular case, corresponding to $N = 1$ and $X_i^1 = 1$ for $1 \le i \le n$, there is only one unknown parameter $\beta_1 = \theta$ and we observe n i.i.d. random variables with distribution $\mathcal{N}(\theta, \sigma^2)$. It is well-known, then, that the best we can do (from the minimax or some Bayesian point of view) is to estimate θ by the maximum likelihood estimator $\hat{\theta} = n^{-1}\sum_{i=1}^n Y_i$ and that the resulting *quadratic risk* for estimating θ is $\mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = n^{-1}\sigma^2$. Rewriting the risk in terms of the parameter

$\boldsymbol{\mu}$ and its estimator $\hat{\boldsymbol{\mu}}$ (with $\hat{\mu}_i = \hat{\theta}$ for all i), we get

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_n^2\right] = n^{-1}\sigma^2, \tag{1.2}$$

where $\|\cdot\|_n$ denotes the normalized Euclidean norm in $\mathbb{R}^n$, i.e. $\|\boldsymbol{\mu}\|_n = n^{-1}\sum_{i=1}^n \mu_i^2$, which we introduce here instead of the usual one for the sake of coherence: with this norm, $\hat{\theta}$ and $\hat{\boldsymbol{\mu}}$ have the same risk and, if all coordinates μ_i of $\boldsymbol{\mu}$ are bounded independently of i, then $\|\boldsymbol{\mu}\|_n$ also remains bounded independently of n.

1.1.2. Fixed design Gaussian regression

Analogous to the previous setting, but with a different flavour, is the *fixed design Gaussian regression*. Let s be some bounded function on $[0, 1]$ and $0 \le x_1 \le x_2 \le \cdots \le x_n \le 1$ a sequence of observation points in $[0, 1]$ (the *design*). In this setting, we observe the finite sequence

$$Y_i = s(x_i) + \sigma\,\xi_i, \quad 1 \le i \le n, \tag{1.3}$$

where the random variables ξ_i, $1 \le i \le n$ are also i.i.d. standard normal. We want to estimate the function s by $\hat{s}$, based on the set of observations $\{Y_i\}_{1\le i\le n}$ with risk $\mathbb{E}\left[\|\hat{s} - s\|_n^2\right]$ where $\|\cdot\|_n$ again denotes the normalized Euclidean norm, i.e. $\|t\|_n = n^{-1}\sum_{i=1}^n t^2(x_i)$. This normalization allows an easy comparison with the functional $\mathbb{L}_2$-norm since $\|t\|_n$ is close to the norm of t in $\mathbb{L}_2([0, 1], dx)$ provided that n is large, the design is regular and t is smooth. If we assume that s belongs to some N-dimensional linear space of functions with basis $(\varphi_1, \ldots, \varphi_N)$, we are back to the Gaussian linear regression setting (1.1) with $X_i^\lambda = \varphi_\lambda(x_i)$.

1.1.3. The white noise framework

The natural generalization of the fixed design regression, when the unknown function s is observed in continuous time, rather that at discrete points x_i, is the so-called *white noise framework*, which is supposed to give a probabilistic model for a deterministic signal observed with additional Gaussian noise. Its use has been initiated by Ibragimov and Has'minskii in the late seventies and it has then been popularized and extensively studied during the last 20 years by the "Russian school" as a toy model for many more complicated frameworks. One observes a path of the process

$$Y(z) = \int_0^z s(x)\,dx + \varepsilon W(z), \quad 0 \le z \le 1, \tag{1.4}$$

where W is a standard Brownian motion originating from 0 and s is an unknown function in $\mathbb{L}_2([0, 1], dx)$. Equivalently, Y can be viewed as the solution, originating from zero, of the stochastic differential equation $dY(z) = s(z)\,dz + \varepsilon\,dW(z)$. We look for estimators $\hat{s}$ of s (i.e. functions of Y and the known parameters like ε) belonging to $\mathbb{L}_2([0, 1], dx)$ with quadratic risk given by $\mathbb{E}\left[\|\hat{s} - s\|^2\right]$ where $\|\cdot\|$ denotes the norm in $\mathbb{L}_2([0, 1], dx)$.

1.1.4. The Gaussian sequence framework

The white noise framework is not connected to any specific orthonormal basis of $\mathbb{L}_2([0,1],dx)$ but, once we have choosen such a basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ with $\Lambda = \mathbb{N}^\star = \mathbb{N}\setminus\{0\}$, it can be transformed into the so-called *Gaussian sequence framework*. This means that we observe the filtered version of Y through the basis, i.e. the sequence of stochastic integrals $Y_\lambda = \int_0^1 \varphi_\lambda(z)\,dY(z)$, for $\lambda\in\Lambda$, where Y is the process defined by (1.4). This is a Gaussian sequence of the form

$$Y_\lambda = \beta_\lambda + \varepsilon\,\xi_\lambda, \quad \lambda\in\Lambda, \quad (\beta_\lambda)_{\lambda\in\Lambda}\in \boldsymbol{l}_2(\Lambda). \tag{1.5}$$

Here $\beta_\lambda = \langle s,\varphi_\lambda\rangle$ and the random variables ξ_λ are i.i.d. standard normal. One can identify s with the sequence $(\beta_\lambda)_{\lambda\in\Lambda}$ and estimate it by some $\hat{s}\in\boldsymbol{l}_2(\Lambda)$ with a risk $\mathbb{E}\left[\|\hat{s}-s\|^2\right]$. The norm $\|\cdot\|$ is now the norm in $\boldsymbol{l}_2(\Lambda)$ and the problem is equivalent to the previous one. Since $\mathbb{E}[Y_\lambda]=\beta_\lambda$ the problem of estimating s within the framework described by (1.5) can also be considered as an infinite-dimensional extension of the Gaussian linear regression where we want to estimate the mean $\boldsymbol{\mu}$ of $\boldsymbol{Y}\in\mathbb{R}^N$.

The interest and importance of the Gaussian sequence framework are due to the fact that, for proper choices of the basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$, smoothness properties of the parameter $s\in\mathbb{L}_2([0,1],dx)$ in (1.4) can be translated into geometric properties of $s\in\boldsymbol{l}_2(\Lambda)$ in (1.5) via the identification $s=(\beta_\lambda)_{\lambda\in\Lambda}$. Let us illustrate this fact by the following classical example. For α some positive integer and $R>0$, the Sobolev class $W^\alpha(R)$ on the torus $\mathbb{R}/\mathbb{Z}$ is defined as the set of functions s on $[0,1]$ which are the restriction to $[0,1]$ of periodic functions on the line with period 1 satisfying $\|s^{(\alpha)}\|\le R$. Given the trigonometric basis $\varphi_1=1$ and, for $j\ge 1$, $\varphi_{2j}(z)=\sqrt{2}\cos(2\pi jz)$ and $\varphi_{2j+1}(z)=\sqrt{2}\sin(2\pi jz)$, it follows from Plancherel's formula that $s=\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda$ belongs to $W^\alpha(R)$ if and only if $\sum_{j=1}^\infty(2\pi j)^{2\alpha}\left[\beta_{2j}^2+\beta_{2j+1}^2\right]\le R^2$ or equivalently if the sequence $(\beta_\lambda)_{\lambda\in\Lambda}$ belongs to the ellipsoid

$$E(\alpha,R)=\left\{(\beta_\lambda)_{\lambda\in\Lambda}\;\middle|\;\sum_{\lambda\in\Lambda}\left(\frac{\beta_\lambda}{a_\lambda}\right)^2\le 1\right\}, \tag{1.6}$$

with

$$a_1=+\infty \qquad \text{and} \qquad a_{2j}=a_{2j+1}=R(2\pi j)^{-\alpha} \quad \text{for } j\ge 1.$$

This means that, via the identification between $s\in\mathbb{L}_2([0,1],dx)$ and its coordinates vector $(\langle s,\varphi_\lambda\rangle)_{\lambda\in\Lambda}\in\boldsymbol{l}_2(\Lambda)$, one can view a Sobolev ball as a geometric object which is an infinite dimensional ellipsoid in $\boldsymbol{l}_2(\Lambda)$.

1.2. Model selection: motivations and purposes

1.2.1. Variable selection in regression

Let us go back to the framework of Sect. 1.1.1 where we want to estimate $\boldsymbol{\mu}=\mathbb{E}[\boldsymbol{Y}]$ with some estimator $\hat{\boldsymbol{\mu}}(\boldsymbol{Y})$ and loss function $\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|_n^2$. In this case, the most

classical estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is the maximum likelihood estimator, which is also the least squares estimator, i.e. the orthogonal projection of $\boldsymbol{Y}$ onto the linear space generated by the vectors $\{\boldsymbol{X}^\lambda\}_{1\leq\lambda\leq N}$, and its risk is given by $\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_n^2\right] = N\sigma^2/n$. This is satisfactory when N is small but certainly not if N is large and only a small number of the N explanatory variables are really influential. Let us for instance assume that, in (1.1), $\beta_\lambda = 0$ for $\lambda \notin m$ where m is a subset of $\{1;\ldots;N\}$. Then, using the orthogonal projection $\hat{\boldsymbol{\mu}}_m$ of $\boldsymbol{Y}$ onto the linear span S_m of $\{\boldsymbol{X}^\lambda\}_{\lambda\in m}$ as an estimator of $\boldsymbol{\mu}$, instead of $\hat{\boldsymbol{\mu}}$, leads to the substantially reduced risk $|m|\sigma^2/n$ if $|m|/N$ is small. This actually remains true if the assumption that $\beta_\lambda = 0$ for $\lambda \notin m$ only holds approximately. Indeed, for any subset m of $\{1;\ldots;N\}$,

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_m-\boldsymbol{\mu}\|_n^2\right] = |m|\sigma^2/n + \|\boldsymbol{\mu}_m-\boldsymbol{\mu}\|_n^2, \tag{1.7}$$

where $\boldsymbol{\mu}_m$ denotes the orthogonal projection of $\boldsymbol{\mu}$ onto S_m, and the risk of $\hat{\boldsymbol{\mu}}_m$ may be substantially smaller than $N\sigma^2/n$ provided that $|m|/N$ is small and $\|\boldsymbol{\mu}_m-\boldsymbol{\mu}\|_n^2$ not large as compared to σ^2/n. In any case, it would be advisable to choose a value of m which minimizes, at least approximately, the right-hand side of (1.7).

Unfortunately, this approach is definitely irrealistic from a practical point of view. Even if one suspects that only a small proportion of the N explanatory variables are really influential and that a good choice of m would lead to an improved risk, such a set is typically unknown. It follows from (1.7) that this set should be small in order to keep the so-called *variance* term $|m|\sigma^2/n$ small. But if it is too small, there is a serious chance that we omit some influential variables which would result in a possibly large *bias* term $\|\boldsymbol{\mu}_m-\boldsymbol{\mu}\|_n^2$. At the opposite, including in the model all the explanatory variables X^λ that we believe may have some influence on Y, i.e. the whole set of variables $X^1,\ldots,X^N$, makes the bias vanish, but at the price of a much larger variance term when N is not small. We are then faced to the problem of finding a suitable set m of influential variables which is neither too small nor too large. This is a typical problem of *variable selection* which can be formalized in the following way: given some family $\mathcal{M}$ of subsets of $\{1;\ldots;N\}$, how can we choose an element $m\in\mathcal{M}$ which is as good as possible, i.e. minimizes, at least approximately, the risk (1.7), although $\boldsymbol{\mu}_m$ is not known to the statistician. It may look strange, at this stage of our reflexion, to choose some $\mathcal{M}$ which is not the family of all subsets of $\{1;\ldots;N\}$, but a possibly smaller one. There are at least two good reasons for that. One is connected to some situations where the variables are naturally ordered and it is therefore meaningful to retain the first D variables for some $D\leq N$. In this case $\mathcal{M}$ is the family of all sets of the form $\{1;\ldots;D\}$ for $0\leq D\leq N$ ($D=0$ corresponding to the empty set). This is the problem of *ordered variable selection*, as opposed to *complete variable selection* when $\mathcal{M}$ contains all subsets of $\{1;\ldots;N\}$. A further reason for distinguishing between the two problems is the fact that it is much more difficult to guess an optimal value of m in the case of complete variable selection that in the case of ordered variable selection because $\mathcal{M}$ is then much larger.

1.2.2. Curve estimation and adaptation

The same problem of identifying a reasonably small number of "significant" parameters also occurs in *curve estimation*, i.e. the estimation of an unknown function like s in (1.4) or (1.5) (via the identification $s = (\beta_\lambda)_{\lambda\in\Lambda}$). Going back to the estimation of the function s from the Gaussian sequence (1.5) under the assumption that it belongs to the ellipsoid $E(\alpha, R)$ defined by (1.6), we recall the classical solution to this problem: fix some positive integer D and estimate β_λ by Y_λ for $1 \le \lambda \le D$ and by 0 for $\lambda > D$. The resulting estimator $\hat{s}_D = \sum_{\lambda=1}^{D} Y_\lambda e_\lambda$, where $(e_\lambda)_{\lambda\in\Lambda}$ denotes the canonical basis in $\boldsymbol{l}_2(\Lambda)$, is called the projection estimator on the D-dimensional linear space S_D generated by $\{e_1; \ldots ; e_D\}$. It follows from the monotonicity of the sequence $(a_\lambda)_{\lambda\in\Lambda}$ that $\sum_{\lambda>D} \beta_\lambda^2 \le a_{D+1}^2$, which implies that the risk of $\hat{s}_D$ can be bounded by

$$\mathbb{E}\left[\|s - \hat{s}_D\|^2\right] \le D\varepsilon^2 + a_{D+1}^2. \tag{1.8}$$

If we choose $D = D_{\text{opt}}$ as a function of α, R and ε in order to minimize the right-hand side of (1.8), which means setting D_{opt} approximately equal to $(R/\varepsilon)^{2/(2\alpha+1)}$, we get a risk bound of the form $C(R\varepsilon^{2\alpha})^{2/(2\alpha+1)}$. Of course, this is only an upper bound for the risk of $\hat{s}_D$ but it cannot be substantially improved, at least uniformly over $E(\alpha, R)$. To see this, let us recall that a classical and popular way of comparing estimators is to compare their maximal risk with respect to some given subset $\mathcal{T}$ of the space of parameters. From this point of view, an estimator is "good" if its maximal risk over $\mathcal{T}$ is close to the *minimax risk*.

Definition 1. *Given a random quantity Y depending on some known parameter ε and some unknown parameter s in some Hilbert space $\mathbb{H}$ with norm $\|\cdot\|$, the* minimax quadratic risk $R_M(\mathcal{T}, \varepsilon)$ *over some subset $\mathcal{T}$ of $\mathbb{H}$, is given by*

$$R_M(\mathcal{T}, \varepsilon) = \inf_{\hat{s}} \sup_{s\in\mathcal{T}} \mathbb{E}_s\left[\|\hat{s} - s\|^2\right], \tag{1.9}$$

where the infimum is taken over all possible estimators $\hat{s}$, i.e. measurable functions of Y with values in $\mathbb{H}$, which possibly also depend on $\mathcal{T}$ and ε, and $\mathbb{E}_s$ denotes the expectation of functions of Y when s obtains.

We shall see in Sect. 6.2 below that $R_M(E(\alpha, R), \varepsilon)$ is of the order of $(R\varepsilon^{2\alpha})^{2/(2\alpha+1)}$, which implies that $\hat{s}_D$ is a good estimator of s provided that D has been correctly chosen (equal to D_{opt}).

Once again, the previous approach (choosing $D = D_{\text{opt}}$) is not practically feasible since s being unknown, α and R and therefore D_{opt} are unknown too. If $\alpha = \alpha_0$ and we assume, for simplicity, that $R = 1$ is known, we get a risk bound $C\varepsilon^{4\alpha_0/(2\alpha_0+1)}$ if $D = D_{\text{opt}}$. Since α_0 is unknown, we have to guess it in some way. If our guess α is smaller than α_0 we shall choose a too large value of D resulting in a larger risk of the form $C\varepsilon^{4\alpha/(2\alpha+1)}$. At the opposite, choosing a too small value of D which is far from minimizing the right-hand side of (1.8) may lead to a risk which is much larger than the expected $C\varepsilon^{4\alpha_0/(2\alpha_0+1)}$. Let us observe that the problem we are faced with, namely, the choice of D or equivalently of a set of basis

vectors $\{e_1; \ldots ; e_D\}$ can be viewed as a problem of ordered variable selection, as defined in the previous section, but with an infinite number of variables since D is now unbounded.

The previous example is actually typical of a broad class of estimation problems where the unknown s to be estimated can be a density, a regression function, a spectral density, the intensity of a point process, the drift of a diffusion, the support of a multivariate density, the hazard rate in survival analysis, etc . . . . All these problems are curve estimation problems but, as opposed to the estimation of a distribution function from i.i.d. observations, they are ill-posed in the sense that if one does not put some restrictions on s, for instance that s belongs to some given Sobolev ball, one cannot find an estimator which is "uniformly good" for all s simultaneously. Up to the seventies, the typical approach for building uniformly good estimators was to assume that s did belong to some known set of functions $\mathcal{S}$. But then, the prior knowledge of $\mathcal{S}$ influences both the construction of the estimators and their performances.

Adaptive estimation tends to solve this dilemna by providing procedures which have good performances, more precisely, that have a risk which is of the order of the minimax risk, on some previleged family of subsets of a large parameter set. In our previous example the large parameter set was $\boldsymbol{l}_2(\Lambda)$ and the previleged subsets were the ellipsoids. Adaptive procedures include adaptive spline smoothing (see Wahba, 1990 for a review), unbiased cross-validation as proposed by Rudemo (1982) and further developed by many authors, soft and hard thresholding methods as initiated by Efroimovich and Pinsker (1984) and further developed by Donoho and Johnstone in a series of papers starting with Donoho and Johnstone (1994), (see references in Donoho and Johnstone, 1998) and by Kerkyacharian and Picard (see, for instance, Kerkyacharian and Picard, 2000 and the references therein), Lepskii's method starting with Lepskii (1990, 1991) and adaptive local polynomial fit as initiated by Katkovnik (1979) (see a detailed account and recent developments in Nemirovski, 2000). For a recent survey of various approaches to adaptation, we refer to Barron et al. (1999, Sect. 5).

1.2.3. The purpose of model selection

As illustrated by the two previous examples, a major problem in estimation is connected with the choice of a suitable set of "significant" parameters to be estimated. In the regression case, one should select some subset $\{\boldsymbol{X}^\lambda\}_{\lambda\in m}$ of the explanatory variables; for the Gaussian sequence problem we just considered, one should select a value of D and only estimate the D parameters $\beta_1, \ldots , \beta_D$. In both cases, this amounts to pretend that the unknown parameter ($\boldsymbol{\mu}$ or s) belongs to some *model* (S_m or S_D) and estimate it as if this were actually true, although we know this is not necessarily the case. In this approach, a model should therefore always be viewed as an approximate model.

In both cases, we have at hand a family of models (the linear spaces S_m or S_D) and the risk (or a risk bound) corresponding to a given model appears – see (1.7) and (1.8) – to be the sum of two components: a variance component which is proportional to the number of parameters that we have put in the model, i.e. the

dimension of the model and a bias term, which is an approximation term, resulting from the use of an approximate model and which corresponds to the square of the distance from the true parameter to the model. An optimal model is one which optimally balances the sum of these components and therefore minimizes the risk (1.7) or the risk bound (1.8). Unfortunately, such an optimal model is not available to the statistician since it depends (through the approximation term) on the unknown parameter. We shall therefore look for a genuine statistical procedure $\hat{m}(\boldsymbol{Y})$ or $\hat{D}((Y_i)_{i\geq 1})$ to select a model from the data only in such a way that the risk of the estimator corresponding to the selected model is close to the optimal risk, i.e. the minimum value of the risk among all possible models.

Model selection actually procedes in two steps: first choose some family of models S_m with $m \in \mathcal{M}$ together with estimators $\hat{s}_m$ with values in S_m. In general, $\hat{s}_m$ derives from a classical estimation procedure, here the maximum likelihood, under the assumption that the model S_m is true ($s \in S_m$). Then use the data to select a value $\hat{m}$ of m and take $\hat{s}_{\hat{m}}$ as the final estimator. A "good" *model selection procedure* is one for which the risk of the resulting estimator is as close as possible to the minimal risk of the estimators $\hat{s}_m$, $m \in \mathcal{M}$.

1.3. The nonasymptotic point of view: a link between parametric and nonparametric problems

A prototype for a *parametric* problem is the estimation of $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}]$ from (1.1) when N is small, for instance under the assumption that $\mu_i = \theta$ for all i (Problem 1), while a prototype for a *nonparametric* problem is the estimation of s from (1.4) under the assumption that it belongs to some functional class, like a Sobolev ball. Equivalently, one can estimate s in the Gaussian sequence framework (1.5) assuming that it belongs to the ellipsoid given by (1.6) (Problem 2). There are a few good reasons to distinguish between parametric and nonparametric problems: typically parametric applies to situations involving a fixed finite number of unknown real parameters (this is the case of our first example with one single parameter θ) while nonparametric refers to estimation of infinite dimensional quantities, like a function from a Sobolev ball. Another difference is connected with convergence rates of the risk from an asymptotic point of view (when n^{-1} or ε go to zero). In order to make a fair comparison between Problems 1 and 2, let us observe that Problem 1 with a risk given by $\mathbb{E}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_n^2\right]$ is equivalent to Problem 1', namely the estimation of s from (1.5) with $\varepsilon = \sigma/\sqrt{n}$, $s_i = \theta/\sqrt{n}$ for $1 \leq i \leq n$ and $s_i = 0$ otherwise, θ being unknown. The respective risks for Problems 1' and 2 are then ε^2 and $C(R\varepsilon^{2\alpha})^{2/(2\alpha+1)}$ (this last value is in fact an upper bound for the risk, but we have already mentioned that, up to the constant C, it cannot be improved uniformly over the ellipsoid). The rate ε^2 (or n^{-1}) is the typical rate of convergence of the risk for parametric problems while the rate, for nonparametric problems, is slower, depending on the "size" of the set of parameters, which, for the ellipsoids $E(\alpha, 1)$ is controlled by α.

This asymptotic approach, letting ε go to 0 in (1.4) or n go to infinity in (1.1), which explains for this distinction between parametric and nonparametric problems, is a quite popular one for curve estimation or model selection. Unfortunately,

it can be terribly misleading, even with a fairly small value of ε (for a related discussion, see Le Cam and Yang 1990, pp. 99–100). Our point of view in this paper is quite different. We want to work with ε or n^{-1} as they are and not let them go to zero. This does not mean that we do not consider in priority small values of ε or large values of n, but only that we want to measure precisely the effect of the different quantities involved in the problem, in view of their size as compared to ε or n. For instance, in Problem 2, it is important to quantify the influence of R whatever the relative sizes of R and ε since both of them are important to describe the difficulty of estimating s. Omitting the effect of R while letting ε go to zero and saying that the rate of convergence of the risk to zero is of order $\varepsilon^{(4\alpha)/(2\alpha+1)}$ can be somewhat misleading.

If we compare Problems 1' and 2 from this nonasymptotic point of view and focus on the estimation procedures we used in both situations, the difference becomes much less obvious. To estimate s in Problem 1' we project the sequence $(Y_i)_{1\le i\le n}$ onto the one-dimensional linear space S spanned by $\sum_{\lambda=1}^{n} e_\lambda$ and to estimate s in Problem 2, we do the same with the D-dimensional linear space spanned by $\{e_1; \ldots ; e_D\}$ with $D \asymp (R/\varepsilon)^{2/(2\alpha+1)}$. In the parametric case, we use, for our estimation procedure, a model S with a fixed dimension (independent of n) which contains the true parameter while the model S_D depends on ε in the nonparametric situation and is not supposed to contain s. If we consider both problems from a nonasymptotic point of view, n and ε are fixed and the difference vanishes. Indeed we treat Problem 2 as a D-dimensional parametric problem although s is infinite-dimensional. The analogy is even more visible if one introduces, for Problem 1', a second estimation procedure which consists in projecting the data onto the 0-dimensional space $S_0 = \{0\}$, with risk θ^2. The second solution is certainly better when $n\theta^2 < \sigma^2$. This shows that, in this case, one should rather use a nonparametric approach and an approximate model (here S_0) although we are faced with a truely parametric problem. Estimating θ from one observation $Y \sim \mathcal{N}(\theta, 1)$ if we suspect that θ is small is not, in a sense, more a parametric problem than estimating s in a Sobolev ball.

The same point of view obviously applies to the variable selection problem of Sect. 1.2.1. If, for instance, $n = 100$, there is a major difference between the case $N = 2$, which can be considered as a parametric problem with two parameters, and the case $N = 80$, which should actually be considered as a nonparametric one. More generally, one can view the situation as parametric if (1.7) is minimum when $m = \{1; \ldots ; N\}$ (all variables are really influential) and nonparametric otherwise (some variables can be omitted without damage). This is in accordance with the practical point of view tending to put more explanatory variables in (1.1) when one has more observations.

These examples show that there is indeed no difference, from a nonasymptotic point of view, between a parametric problem with a "large" (with a proper definition of this term) number of parameters and a nonparametric problem. The difficulty of estimation (the size of the risk) is not connected with the parametric nature of the problem but rather with the ratio between the number of observations and the number of "significant" parameters. Model selection, which introduces many finite dimensional models, true or not, simultaneously, treats both paramet-

ric and nonparametric problems in the same way without any distinction. It uses finite-dimensional (parametric) models to estimate infinite-dimensional objects and introduces approximate models to estimate finite-dimensional parameters, as in the nonparametric case. From this point of view, the assumption that there exists a "true" model (one containing s) becomes useless since a model is only some (possibly good but also possibly poor) approximation of the reality.

To illustrate this discussion and give a flavour of the types of results we obtain, let us conclude by providing a simplified version, restricted to the Gaussian sequence framework, of our main result to be stated in Sect. 3.2.

Theorem 1. *Given a sequence of variables $(Y_\lambda)_{\lambda\in\Lambda}$ (with $\Lambda = \mathbb{N}^\star$) satisfying (1.5) for an unknown value of the parameter $s = (\beta_\lambda)_{\lambda\in\Lambda} \in \boldsymbol{l}_2(\Lambda)$, a countable family $\mathcal{M}$ of nonvoid finite subsets m of $\mathbb{N}^\star$, a number $K > 1$ and a family of nonnegative numbers L_m for $m \in \mathcal{M}$ satisfying the condition $\sum_{m\in\mathcal{M}} \exp(-|m|L_m) = \Sigma < +\infty$, we define the function* pen *on $\mathcal{M}$, the model selector $\hat{m}$ and the estimator $\tilde{s} = (\tilde{\beta}_\lambda)_{\lambda\in\Lambda}$ by*

$$\mathrm{pen}(m) = K\varepsilon^2|m|\left(1+\sqrt{2L_m}\right)^2, \qquad \hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}}\left[-\sum_{i\in m} Y_i^2 + \mathrm{pen}(m)\right]$$

and $\tilde{\beta}_\lambda = Y_\lambda \mathbb{1}_{\lambda\in\hat{m}}$ for $\lambda \in \Lambda$. Then the quadratic risk of $\tilde{s}$ is bounded by

$$\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \le C(K)\left[\inf_{m\in\mathcal{M}}\left\{\sum_{\lambda\notin m}\beta_\lambda^2 + \mathrm{pen}(m)\right\} + \varepsilon^2\Sigma\right],$$

where $C(K)$ denotes some positive function of K.

In order to understand the meaning of this result, one should keep in mind the fact that the risk of the maximum likelihood estimator $\hat{s}_m$, with coordinates $Y_\lambda \mathbb{1}_{\lambda\in m}$, corresponding to the assumption that $s \in S_m = \{t = (\theta_\lambda)_{\lambda\in\Lambda} \mid \theta_\lambda = 0 \text{ for } \lambda \notin m\}$ is given by $\sum_{\lambda\notin m}\beta_\lambda^2 + |m|\varepsilon^2$. It follows that, if $L = \sup_m L_m < +\infty$, then

$$\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \le C(K,L,\Sigma)\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|s-\hat{s}_m\|^2\right].$$

Our theorem immediately applies to the family of models S_D, $D \in \mathbb{N}^\star$ defined in Sect. 1.2.2 with $L_D = 1$ for all D, which leads to $\Sigma = e/(e-1)$. It then easily follows from (1.8) and (1.6) that, if $K = 2$, the resulting estimator $\tilde{s}$ satisfies

$$\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \le C\left[\left(R\varepsilon^{2\alpha}\right)^{2/(2\alpha+1)} + \varepsilon^2\right],$$

for some constant C independent of α and R. This corresponds, up to some constant, to the minimax risk over the ellipsoid given by (1.6) provided that R is not too small (more precisely that $R \ge \varepsilon$) and $\tilde{s}$ is therefore adaptive over all Sobolev balls of radius $R \ge \varepsilon$.

1.4. Some historical remarks

A prototype for criteria used in model selection is Mallows' C_p, as described in Daniel and Wood (1971). A similar approach, based on penalized maximum likelihood estimation, which is due to Akaike (1973 and 1974) applies to much more general situations, each model S_m being possibly nonlinear, although defined by a finite number D_m of parameters. In the Gaussian regression framework, when σ is known and the models are linear, Akaike's and Mallows' methods coincide. In any case both approaches are based on the minimization of a penalized criterion.

Since the publication of these seminal works, many other penalized criteria for model selection have been developed for solving various types of model selection problems. For instance, several criteria essentially based on asymptotic or heuristic considerations are used in pratice to select influential variables. Besides Akaike's AIC one should in particular mention its improvement AICc by Hurvich and Tsai (1989) or Schwarz's BIC (Schwarz, 1978) among others. Many such criteria can be found in the book by McQuarrie and Tsai (1998).

In the existing literature, one can distinguish between two very different points of view, although both based on asymptotic considerations. A first one assumes a finite number of parametric models, one of which being "true" in the sense that it does contain the true unknown s. In this case, the number of observations goes to infinity while the list of models remains fixed and one looks for criteria which allow, asymptotically, to identify the true model. One then adds to the remaining sum of squares, penalties of the form $K(n)D_m$, where D_m denotes the dimension of model S_m, and one tries to recover asymptotically the "true" model. See for instance Schwarz (1978) and Nishii (1984). An opposite philosophy is to use model selection criteria to estimate s belonging to some infinite-dimensional space, which means handling a nonparametric problem. A typical result in this direction is due to Shibata (1981) in the context of (1.5). He shows, under appropriate assumptions on the list of models and s, that the quadratic risk of the estimator selected according to Mallows' C_p criterion is asymptotically equivalent to the risk of the estimator corresponding to the best model. This striking result can be obtained at the price of the following restrictions: the largest dimension of the models should tend to infinity slower than ε^{-1}, s should not belong to any model and the list of models should not be too large (the number of models of a given dimension D can be a polynomial but not an exponential function of D). Further results in this direction can be found in Li (1987) and Kneip (1994). See also Polyak and Tsybakov (1990).

1.5. A brief overview of this paper

For the sake of simplicity and to avoid redundancy, our first aim will be to define (in Sect. 2) a unified Gaussian framework to deal with all the examples we have considered in Sect. 1.1 simultaneously. Then, in Sect. 3, we shall develop, within this framework, our approach to model selection and set up a general method (via penalization) for choosing one model within a (potentially large) family of such. The properties of the resulting estimators are given below in Theorem 2, which is

a generalized version of Theorem 1. Obviously, the performances of $\tilde{s}$ depend on the *strategy*, i.e. the family of *models* $\{S_m\}_{m\in\mathcal{M}}$ and the associated *weights* L_m we have chosen. Different strategies should be selected for different needs or to take into account various types of a priori information on the unknown parameter s. Since each strategy has its advantages and disadvantages, it would be desirable to be able to mix them in order to retain the best of each one. A general way towards this aim will then be presented in Sect. 4.

The remainder of the paper will be devoted to applications of our main result (Theorem 2) including ordered and complete variable selection with applications to threshold estimators and adaption for sets of functions of increasing complexity such as Sobolev and Besov balls, among others. We shall follow here the approach of Donoho and Johnstone (1994b, 1996 and 1998) which amounts, via the choice of a convenient basis, to work within the Gaussian sequence framework, replacing families of functions by suitable geometric objects in the space $\boldsymbol{l}_2(\mathbb{N}^\star)$. We have already seen that Sobolev balls could be interpreted as ellipsoids. In the same way, Besov balls can be turned to some special cases of $\boldsymbol{l}_p$-bodies, which will lead us to study various adaptive strategies for different types of $\boldsymbol{l}_p$-bodies. Our results will, in particular, complement those of Donoho and Johnstone (1994b) and provide fully adaptive estimators for (almost) all Besov balls simultaneously. We shall conclude with some remarks on the choice of the constant K involved in the penalty, showing that $K < 1$ may lead, in some cases, to definitely poor results.

2. Model selection for Gaussian models

2.1. A generic Gaussian framework

We now want to provide a unified treatment for various problems connected with Gaussian measures and, in particular, for all the frameworks we have considered in Sect. 1.1.

Let us first consider the linear regression. Setting $\langle t, u\rangle = n^{-1}\sum_{i=1}^n t_i u_i$ for $t, u \in \mathbb{R}^n$ and $s = \sum_{\lambda=1}^N \beta_\lambda \boldsymbol{X}^\lambda$ in (1.1), we get for any $t \in \mathbb{H} = \mathbb{R}^n$,

$$\langle t, \boldsymbol{Y}\rangle = \langle s, t\rangle + \varepsilon Z(t) \quad \text{with } Z(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i t_i \quad \text{and} \quad \varepsilon = \frac{\sigma}{\sqrt{n}}.$$

Since the ξ_i's are i.i.d. $\mathcal{N}(0, 1)$, Z is a centered and linear Gaussian process indexed by $\mathbb{H}$ such that $\mathrm{Cov}(Z(t), Z(u)) = \langle t, u\rangle$.

Similarly, the discrete-time process $(Y_i)_{1\le i\le n}$ defined by (1.3) can be associated to a linear operator on the Hilbert space $\mathbb{H}$ of functions t on $\{x_1; x_2; \ldots; x_n\}$ with the scalar product $\langle t, u\rangle = n^{-1}\sum_{i=1}^n t(x_i)u(x_i)$ by the formula

$$t \mapsto \frac{1}{n}\sum_{i=1}^n Y_i t(x_i) = \langle s, t\rangle + \varepsilon Z'(t) \quad \text{with } Z'(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i t(x_i) \quad \text{and} \quad \varepsilon = \frac{\sigma}{\sqrt{n}}.$$

Since the ξ_i's are also i.i.d. $\mathcal{N}(0, 1)$, the process Z' is again linear with the same distribution as Z.

Analogously, for each $t \in \boldsymbol{l}_2(\Lambda)$ with a finite number of nonzero coordinates, one can write

$$\sum_{\lambda\in\Lambda} Y_\lambda t_\lambda = \langle s, t\rangle + \varepsilon Z''(t) \quad \text{with } Z''(t) = \sum_{\lambda\in\Lambda} \xi_\lambda t_\lambda.$$

Here $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\boldsymbol{l}_2(\Lambda)$ and Z'' is again a centered linear Gaussian process with the same covariance structure as Z since the last sum is actually a finite one. The only difference is that we have restricted Z'' to the linear subspace $\mathbb{S}$ of $\boldsymbol{l}_2(\Lambda)$ of those elements that have only a finite number of nonzero coordinates in order that the series which defines it converge in $\mathbb{R}$. Here ε^2 plays the role of σ^2/n in the previous examples.

Let us finally consider the case of (1.4). An alternative, but equivalent formulation is in the form of the stochastic differential equation

$$dY = s(x)\,dx + \varepsilon\,dW \quad \text{with } Y(0) = 0. \tag{2.1}$$

If the Brownian motion W is defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, one can deduce from the process Y given by (2.1) a linear operator from the Hilbert space $\mathbb{H} = \mathbb{L}_2([0,1], dx)$ to $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ given by

$$t \mapsto \int_0^1 t(z)\,dY(z) = \langle s, t\rangle + \varepsilon Z'''(t) \quad \text{with } Z'''(t) = \int_0^1 t(z)\,dW(z),$$

where $\langle\cdot,\cdot\rangle$ now denotes the scalar product in $\mathbb{H}$. The process $Z'''(t)$ is again a centered Gaussian process (in the $\mathbb{L}_2$ sense) indexed by $\mathbb{H}$ and with covariance structure given by the scalar product on $\mathbb{H}$. The functional $t \mapsto Z'''(t)$ is linear as an operator from $\mathbb{H}$ to $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ but, as in the previous example, we shall restrict Z''' to some linear subspace $\mathbb{S}$ of $\mathbb{H}$ in order to get a version such that $t \mapsto Z'''(t)(\omega)$ is linear on $\mathbb{S}$ for almost every $\omega \in \Omega$. Fortunately, we do not need that the process Z''' be defined on the whole Hilbert space $\mathbb{H}$ and, as we shall see below, restricting Z''' to some subspace $\mathbb{S}$ of $\mathbb{H}$ will be enough for our purposes.

This suggests to introduce the following infinite dimensional extension of a standard Gaussian vector in a Euclidean space adapted from Dudley (1967).

Definition 2. *Given a linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$ with its scalar product $\langle\cdot,\cdot\rangle$, a* linear isonormal process *Z indexed by $\mathbb{S}$ is an almost surely linear centered Gaussian process with covariance structure* $\operatorname{Cov}(Z(t), Z(u)) = \langle t, u\rangle$. *The a.s. linearity means that one can find a subset Ω' of Ω such that $\mathbb{P}(\Omega') = 1$ and $\alpha Z(t)(\omega) + \beta Z(u)(\omega) = Z(\alpha t + \beta u)(\omega)$ whatever $\omega \in \Omega'$, $\alpha, \beta \in \mathbb{R}$ and $t, u \in \mathbb{S}$.*

Proposition 1. *Given some orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ in $\mathbb{H}$, there exists a linear isonormal process indexed by the linear space $\mathbb{S} = \{t = \sum_{\lambda\in\Lambda} \theta_\lambda\varphi_\lambda \mid |\{\lambda \mid \theta_\lambda \neq 0\}| < +\infty\}$.*

Proof. Choose a set $(\xi_\lambda)_{\lambda\in\Lambda}$ of i.i.d. standard normal random variables and define Z on $\mathbb{S}$ by $Z(t) = \sum_{\lambda\in\Lambda} \theta_\lambda\xi_\lambda$, which is a finite sum. □

One can then define an infinite dimensional analogue of a Gaussian vector with covariance matrix proportional to the identity. From a statistical point of view, it provides a natural framework for generalizing to an infinite dimensional setting the problem of estimating the mean of a Gaussian vector.

Definition 3. *Given a linear subspace* $\mathbb{S}$ *of some Hilbert space* $\mathbb{H}$ *we call* Gaussian linear process *on* $\mathbb{S}$ *with mean* $s \in \mathbb{H}$ *and variance* ε^2 *any process* Y *indexed by* $\mathbb{S}$ *of the form*

$$Y(t) = \langle s, t \rangle + \varepsilon Z(t) \quad \textit{for all } t \in \mathbb{S}, \tag{2.2}$$

where Z *denotes a linear isonormal process indexed by* $\mathbb{S}$.

It follows from the preceding considerations that observation of the sets of variables $\{Y_i\}_{1 \le i \le n}$ corresponding to the statistical frameworks (1.1) and (1.3) is equivalent to observing a Gaussian linear process. This still holds true for (1.5) and also for (2.1) (see below) provided that $\mathbb{S}$ has been suitably chosen.

2.2. Back to the Gaussian sequence framework

We have seen in the previous section that the Gaussian sequence (1.5) can be turned to a Gaussian linear process, but the reverse operation will also prove extremely useful in the sequel, in particular for adaptive curve estimation. Consider the Gaussian linear process $Y(t) = \langle s, t \rangle + \varepsilon Z(t)$ where s belongs to some infinite dimensional separable Hilbert space $\mathbb{H}$ with orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \Lambda}$, $\Lambda = \mathbb{N}^\star$ and t belongs to the subspace $\mathbb{S}$ of Proposition 1, one can always write by linearity

$$Y(t) = \sum_{\lambda \in \Lambda} \beta_\lambda \theta_\lambda + \varepsilon \sum_{\lambda \in \Lambda} \theta_\lambda Z(\varphi_\lambda) \quad \text{if } s = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda.$$

This means that, once the basis $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ has been fixed, Y can be viewed, identifying s with $(\beta_\lambda)_{\lambda \in \Lambda}$ and t with $(\theta_\lambda)_{\lambda \in \Lambda}$, as a process indexed by the subspace $\mathbb{S} = \{t = (\theta_\lambda)_{\lambda \in \Lambda} \mid |\{\lambda \mid \theta_\lambda \neq 0\}| < +\infty\}$ of $\mathbb{H} = \boldsymbol{l}_2(\Lambda)$. Moreover, since the basis $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ is orthonormal, Y can be written as

$$Y(t) = \sum_{\lambda \in \Lambda} \theta_\lambda \hat{\beta}_\lambda \quad \text{with } \hat{\beta}_\lambda = \beta_\lambda + \varepsilon \xi_\lambda,$$

for some sequence $(\xi_\lambda)_{\lambda \in \Lambda}$ of i.i.d. standard normal random variables. Obviously, the process Y is entirely determined by the sequence $(\hat{\beta}_\lambda)_{\lambda \in \Lambda}$ which is a Gaussian sequence as defined by (1.5). We shall therefore call *Gaussian sequence framework associated with the basis* $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ the random sequence $(\beta_\lambda + \varepsilon \xi_\lambda)_{\lambda \in \Lambda}$ with unknown parameter $s = (\beta_\lambda)_{\lambda \in \Lambda}$. The advantage of the Gaussian linear process over the Gaussian sequences comes from the fact that it is independent of the choice of a particular basis. To a single Gaussian linear process, one can associate many Gaussian sequences and this will prove useful for curve estimation.

2.3. Linear models, projection estimators and oracles

Our purpose, given a Gaussian linear process Y on $\mathbb{S}$ with unknown mean $s \in \mathbb{H}$ and variance ε^2, is to recover s from a realization of the process Y which means building an estimator $\tilde{s}$ which is a function of Y and ε as a substitute for s. The quality of this reconstruction will be measured, in the sequel, in terms of the quantity $\mathbb{E}_s[\ell(\|\tilde{s} - s\|)]$ where $\mathbb{E}_s$ denotes the expectation of functions of Y when Y is defined by (2.2) and ℓ is a nondecreasing function on $\mathbb{R}^+$ with $\ell(0) = 0$, our reference being the classical quadratic risk which corresponds to $\ell(x) = x^2$.

2.3.1. Linear models

In order to design our estimator, we shall introduce a countable family of models $\{S_m\}_{m\in\mathcal{M}}$. By *model*, we mean hereafter a finite dimensional linear subspace of $\mathbb{H}$, possibly of dimension 0, i.e. reduced to the set $\{0\}$. We assume that all our models are subsets of $\mathbb{S}$, which is not a restriction since, if $\mathbb{S}$ is spanned by the union of the S_m's, there always exists a linear isonormal process $Z(t)$ indexed by $\mathbb{S}$. Indeed, given a linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$ which is the linear span of a countable family $\{S_m\}_{m\in\mathcal{M}}$ of finite dimensional linear subspaces of $\mathbb{H}$, one can always build, using an orthonormalization procedure, an orthonormal system $\{\varphi_\lambda\}_{\lambda\in\mathbb{N}^\star}$ such that any element $t \in \mathbb{S}$ can be written as a finite combination $t = \sum_{\lambda\in\Lambda_t} \theta_\lambda\varphi_\lambda$ with $|\Lambda_t| < +\infty$. It then suffices to apply Proposition 1.

Given some model S_m with dimension D_m, an estimator $\tilde{s}_m$ with values in S_m is any measurable application $\tilde{s}_m(Y)$ with values in S_m and its quadratic risk at s is given by

$$\mathbb{E}_s\left[\|\tilde{s}_m - s\|^2\right] = \|s_m - s\|^2 + \mathbb{E}_s\left[\|\tilde{s}_m - s_m\|^2\right], \tag{2.3}$$

where s_m denotes the projection of s onto S_m. Let now see what can be expected from such an estimator from the minimax point of view, since it is well known that one cannot base an optimality criterion on a pointwise risk comparison. Since $\|s_m - s\|$ is deterministic, an optimization of the risk amounts to an optimization of $\mathbb{E}_s\left[\|\tilde{s}_m - s_m\|^2\right]$ among estimators with values in S_m. Because of this last restriction, we have to modify slightly the definition of the minimax risk. We define the minimax risk for estimating the projection s_m of s onto S_m under the restriction that $s \in S_m + t$ where $t \in S_m^\perp$ as $\inf_{\tilde{s}_m} \sup_{s\in S_m+t} \mathbb{E}_s\left[\|\tilde{s}_m - s_m\|^2\right]$ where $\tilde{s}_m$ is restricted to take its values in S_m. A trivial modification of the classical proof of the fact that $\boldsymbol{X}$ is minimax for estimating $\boldsymbol{\mu} \in \mathbb{R}^{D_m}$ when $\boldsymbol{X} = \boldsymbol{\mu} + \varepsilon\boldsymbol{\xi}$, $\boldsymbol{\xi}$ is a D_m-dimensional standard Gaussian vector and $\varepsilon > 0$ is known, based on the fact that the restrictions of Y to S_m and $S_m^\perp$ are independent, shows that

$$\inf_{\tilde{s}_m} \sup_{s\in S_m+t} \mathbb{E}_s\left[\|\tilde{s}_m - s_m\|^2\right] \geq \varepsilon^2 D_m.$$

Therefore by (2.3),

$$\inf_{\tilde{s}_m} \sup_{s\in S_m+t} \mathbb{E}_s\left[\|\tilde{s}_m - s\|^2\right] \geq \|t\|^2 + \varepsilon^2 D_m. \tag{2.4}$$

2.3.2. Projection estimators

One can actually design a simple estimator which is minimax on the model S_m, i.e. achieves the bound (2.4) for all ts simultaneously.

Definition 4. *Let Y be a Gaussian linear process indexed by a linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$ with unknown mean $s \in \mathbb{H}$ and known variance ε^2. Let S be a finite dimensional linear subspace of $\mathbb{S}$ and let us set $\gamma(t) = \|t\|^2 - 2Y(t)$. One defines the* projection estimator *on S to be the minimizer of $\gamma(t)$ with respect to $t \in S$.*

Given a model S_m with dimension D_m, the projection estimator $\hat{s}_m$ on the model S_m is actually unique and can be computed as follows. Choose some orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ of S_m and set $t = \sum_{\lambda\in\Lambda_m} \theta_\lambda\varphi_\lambda$. By linearity, minimizing $\gamma(t)$ amounts to minimize $\sum_{\lambda\in\Lambda_m}\left[\theta_\lambda^2 - 2\theta_\lambda Y(\varphi_\lambda)\right]$ which clearly results in

$$\hat{s}_m = \sum_{\lambda\in\Lambda_m} \hat{\beta}_\lambda\varphi_\lambda \quad \text{with } \hat{\beta}_\lambda = Y(\varphi_\lambda) = \langle s, \varphi_\lambda\rangle + \varepsilon Z(\varphi_\lambda). \tag{2.5}$$

Moreover, since $s_m = \sum_{\lambda\in\Lambda_m}\langle s, \varphi_\lambda\rangle\varphi_\lambda$,

$$\hat{s}_m = s_m + \varepsilon \sum_{\lambda\in\Lambda_m} Z(\varphi_\lambda)\varphi_\lambda \quad \text{and} \quad \|\hat{s}_m - s_m\|^2 = \varepsilon^2 \sum_{\lambda\in\Lambda_m} Z(\varphi_\lambda)^2. \tag{2.6}$$

Since the $Z(\varphi_\lambda)$s are i.i.d. standard normal and $|\Lambda_m| = D_m$ we derive from (2.3) that

$$\mathbb{E}_s\left[\|\hat{s}_m - s\|^2\right] = \|s_m - s\|^2 + \varepsilon^2 D_m. \tag{2.7}$$

This shows simultaneously that (2.4) is an equality and that the projection estimator $\hat{s}_m$ is optimal from the minimax point of view.

Remark. Even if $\hat{s}_m$ is minimax, at least as an estimator of s_m, when $D_m \geq 3$ one can actually design, using Stein's method (see Stein, 1956), estimators $\bar{s}_m$ which improve on $\hat{s}_m$ in the sense that they satisfy $\mathbb{E}_s\left[\|\bar{s}_m - s\|^2\right] < \|s_m - s\|^2 + \varepsilon^2 D_m$ whatever s, although such an improvement, because of (2.4), cannot hold uniformly with respect to s. Moreover those estimators are more complicated than $\hat{s}_m$. We shall, from now on, restrict ourselves to projection estimators both because of the simplicity of their representation by (2.5) which allows an easy computation, and because of their definition as minimizers over the models of the criterion $\gamma(t) = \|t\|^2 - 2Y(t)$. This second property is indeed the milestone of our construction.

2.3.3. Ideal model selection and oracles

Let us now consider a family of models $\mathcal{F} = \{S_m\}_{m\in\mathcal{M}}$ and the corresponding family of projection estimators $\{\hat{s}_m\}_{m\in\mathcal{M}}$. Since the quadratic risk $\mathbb{E}_s\left[\|\hat{s}_m - s\|^2\right] = \|s_m - s\|^2 + \varepsilon^2 D_m$ of the estimator $\hat{s}_m$ is optimal from the minimax point of view, it can be viewed as the benchmark for the risk of an estimator with values in the

D_m-dimensional linear space S_m. We can then conclude that, from this point of view, an ideal model $S_{m(s)}$ should satisfy

$$\|s_{m(s)} - s\|^2 + \varepsilon^2 D_{m(s)} = \inf_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \varepsilon^2 D_m \right\}. \tag{2.8}$$

Such a procedure $m(s)$ which depends on the unknown parameter to be estimated cannot of course be used as a *statistical model selection procedure*. This is why, following Donoho and Johnstone (1994), we call such an ideal procedure an *oracle* and measure the statistical quality of the family $\mathcal{F}$ at s in terms of the following index:

Definition 5. *Given a family of linear models $\mathcal{F} = \{S_m\}_{m \in \mathcal{M}}$ in some Hilbert space $\mathbb{H}$ with respective dimensions D_m, a function $s \in \mathbb{H}$ and a positive number ε, we define the* oracle accuracy *of the family $\mathcal{F}$ at s as*

$$a_O(s, \mathcal{F}, \varepsilon) = \inf_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \varepsilon^2 D_m \right\},$$

where s_m is the orthogonal projection of s onto S_m.

An ideal estimator $\hat{s}$, i.e. an estimator which, for each s, is as good as the best projection estimator in the set $\{\hat{s}_m\}_{m \in \mathcal{M}}$, would be one satisfying $\mathbb{E}_s \left[\|\hat{s} - s\|^2 \right] = a_O(s, \mathcal{F}, \varepsilon)$ for all $s \in \mathbb{H}$. Unfortunately, just as oracles typically do not exist as genuine statistical procedures, ideal estimators do not exist either and we shall content ourselves to try to design almost ideal estimators, i.e. genuine statistical procedures $\tilde{s} = \hat{s}_{\hat{m}}$ based on a model selection procedure $\hat{m}(Y)$ with values in $\mathcal{M}$ which approximately mimics an oracle in the sense that one can find a constant C such that $\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] \leq C a_O(s, \mathcal{F}, \varepsilon)$ for all $\varepsilon > 0$ and $s \in \mathbb{H}$. Even this aim is too ambitious in general. There is in particular a situation for which one can easily see that it is hopeless to get such a risk bound, namely when $\{0\} \in \mathcal{F}$ and $s = 0$. Then $a_O(\mathcal{F}, 0, \varepsilon) = 0$ and such a risk bound would imply that $\tilde{s} = 0$, $\mathbb{P}_0$ a.s., $\mathbb{P}_s$ denoting the distribution of Y when (2.2) holds. Since all measures $\mathbb{P}_s$ are mutually absolutely continuous, then $\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] = \|s\|^2$ for all s. A more sophisticated argument could show that the same holds true if one excludes 0 but let $a_O(s, \mathcal{F}, \varepsilon)$ be arbitrarily small. Therefore we shall have to content ourselves with a weaker inequality, namely

$$\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] \leq C \left[a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2 \right] \quad \text{whatever } \varepsilon > 0 \quad \text{and} \quad s \in \mathbb{H}. \tag{2.9}$$

The additional term ε^2 allows s and therefore $a_O(s, \mathcal{F}, \varepsilon)$, to be arbitrarily close to zero without causing the troubles mentioned above. Let us finally notice that if $\|s\| \geq \delta\varepsilon$ for some $\delta > 0$, then $a_O(s, \mathcal{F}, \varepsilon) \geq \varepsilon^2 \left(1 \wedge \delta^2\right)$ and therefore $a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2$ is comparable to $a_O(s, \mathcal{F}, \varepsilon)$ via the inequality $a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2 \leq \left[(1 \wedge \delta^2)^{-1} + 1 \right] a_O(s, \mathcal{F}, \varepsilon)$.

Even if we exclude the preceding situation and only consider values of s such that $\|s\| \geq \varepsilon$, there are cases, depending on the choice of the class $\mathcal{F}$, for which it is impossible to obtain an inequality like (2.9), with a moderate value of C, uniformly on the set of those s such that $\|s\| \geq \varepsilon$. This is in particular true when the class

$\mathcal{F}$ is "large" in some suitable sense. In particular, complete variable selection (as defined in Sect. 1.2.1) among a set of N variables leads to an unimprovable value of C of order $\log N$ as will be shown in Sect. 5.2.

2.4. Mallows' heuristics and penalized projection estimators

Mallows, in a conference dating back to 1964, according to Daniel and Wood (1971, p. 86), proposed a method for solving the model selection problem, now refered to as Mallows' C_p (see Mallows, 1973). The heuristics underlying his method are as follows. An ideal model selection procedure minimizes over $\mathcal{M}$ the quantity $\|s_m - s\|^2 + \varepsilon^2 D_m$, or equivalently

$$\|s_m - s\|^2 + \varepsilon^2 D_m - \|s\|^2 = -\|s_m\|^2 + \varepsilon^2 D_m. \tag{2.10}$$

Since, by (2.5), $\hat{s}_m$ can be written as $s_m + \varepsilon W_m$ where W_m is a standard D_m-dimensional Gaussian vector $\mathbb{E}\left[\|\hat{s}_m\|^2\right] = \|s_m\|^2 + \varepsilon^2 D_m$ and therefore $\|\hat{s}_m\|^2 - \varepsilon^2 D_m$ is an unbiased estimator of $\|s_m\|^2$. Replacing in (2.10) $\|s_m\|^2$ by this estimator leads to Mallows' C_p criterion which amounts to minimize $-\|\hat{s}_m\|^2 + 2\varepsilon^2 D_m$ over $\mathcal{M}$. Mallows actually gave no proof of the properties of his method and one had to wait until Shibata (1981) to get a proof that such a method works, at least from an asymptotic point of view. Unfortunately, and we shall prove this precisely in Sect. 7.2 below, Mallow's C_p is only suitable for families $\{S_m\}_{m\in\mathcal{M}}$ of models which are not "too large" (in a sense that we shall make precise later). It is therefore necessary, in order to get model selection methods which are valid for arbitrary countable families of models, to consider more general criteria of the form $-\|\hat{s}_m\|^2 + \text{pen}(m)$ to be minimized with respect to $m \in \mathcal{M}$, "pen" denoting a nonnegative penalty function defined on $\mathcal{M}$. The remainder of this paper will be devoted to the evaluation of the performances of those estimators which minimize such criteria and to the discussion of those choices of the penalty function that lead to sensible and sometimes optimal results of a form similar to (2.9).

3. The performances of penalized projection estimators

3.1. The precise framework

We want to introduce and study some model selection based estimation procedures for the unknown mean of the linear Gaussian process Y given by Definition 3.

Definition 6. *Given a finite or countable family $\{S_m\}_{m\in\mathcal{M}}$ of finite dimensional linear subspaces of $\mathbb{S}$, the corresponding family of projection estimators $\hat{s}_m$ built from the same realization of the process Y according to Definition 4 and a nonnegative function* pen *defined on $\mathcal{M}$, a* penalized projection estimator *(associated to this family of models and this penalty function) is defined by $\tilde{s} = \hat{s}_{\hat{m}}$, where $\hat{m}$ is any minimizer with respect to $m \in \mathcal{M}$ (if it exists) of the penalized criterion*

$$\text{crit}(m) = -\|\hat{s}_m\|^2 + \text{pen}(m). \tag{3.1}$$

Remarks.

- We do not assume that all the models are different, i.e. that the mapping $m \mapsto S_m$ is one to one. It might look strange to allow such a redundancy in the list of models. Indeed, if m and m' are such that $S_m = S_{m'}$ with $\text{pen}(m) \neq \text{pen}(m')$, the definition of $\hat{m}$ implies that the model with the smaller penalty will always be prefered and therefore that the one with the larger penalty could be removed without affecting the estimation procedure. This remains true when both penalties are equal: one can obviously remove one of the models. Nevertheless, the consideration of redundant families of models will turn to be useful from a computational point of view, since it sometimes provides an easier description of $\hat{s}_{\hat{m}}$, as we shall see in Sect. 4.1 below.
- We allow the collection to contain zero-dimensional models $S_m = \{0\}$, in which case $D_m = 0$ for the corresponding m.
- An equivalent definition of the pair $(\hat{m}, \tilde{s})$ is

$$\operatorname*{argmin}_{m,t} \left[\left(\|t\|^2 - 2Y(t) \right) - \log(\mathbb{1}_{t \in S_m}) + \text{pen}(m) \right] \quad \text{with } \log 0 = -\infty. \tag{3.2}$$

3.2. The main result

Our aim is now to prove that a proper choice of the penalty function in (3.1) leads to some upper bounds for the risk of the corresponding penalized projection estimator that we can compare to the oracle accuracy.

Theorem 2. *Let Y be a Gaussian linear process indexed by a linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$ with unknown mean $s \in \mathbb{H}$ and known variance ε^2, $\{S_m\}_{m \in \mathcal{M}}$ be a finite or countable family of finite dimensional linear subspaces of $\mathbb{S}$ with respective dimensions D_m and $\{L_m\}_{m \in \mathcal{M}}$ be a family of* weights*, i.e. nonnegative real numbers, satisfying the condition*

$$\Sigma = \sum_{m \in \mathcal{M}^\star} \exp[-D_m L_m] < +\infty \quad \textit{with } \mathcal{M}^\star = \{m \in \mathcal{M} \mid D_m > 0\}. \tag{3.3}$$

Let us then choose a penalty function $\text{pen}(\cdot)$ *on $\mathcal{M}$ such that*

$$\text{pen}(m) \geq K\varepsilon^2 D_m \left(1 + \sqrt{2L_m}\right)^2 \quad \textit{for all } m \in \mathcal{M} \textit{ and some } K > 1. \tag{3.4}$$

The corresponding penalized projection estimator $\tilde{s}$ given by Definition 6 almost surely exists and is unique. Moreover it satisfies

$$\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] \leq \frac{4K(K+1)^2}{(K-1)^3} \left[\inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \text{pen}(m) \right\} + (K+1)\, \varepsilon^2 \Sigma \right], \tag{3.5}$$

where $d(s, S_m)$ denotes the distance of s to the space S_m. More generally, if ℓ is a nondecreasing function defined on $\mathbb{R}^+$ such that $\ell(0) = 0$ and $\ell(x + y) \leq A[\ell(x) + \ell(y)]$ for all $x, y \geq 0$ and some positive constant A, then

$$\mathbb{E}_s\left[\ell\left(\|\tilde{s} - s\|^2\right)\right] \leq C_1(A, K) \inf_{m \in \mathcal{M}} \left\{\ell\left(d^2(s, S_m)\right) + \ell(\text{pen}(m))\right\} + C_2(A, K)\, \Sigma\ell\left(\varepsilon^2\right), \quad (3.6)$$

for some suitable functions C_1 and C_2 of A and K, independent of s, ε and Σ.

Proof. Recalling from Definition 4 that $\gamma(t) = \|t\|^2 - 2[\langle s, t\rangle + \varepsilon Z(t)]$ and assuming that for any $m' \in \mathcal{M}$ we have chosen some orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$ of $S_{m'}$, we derive from (2.5) and the linearity of Z that

$$\gamma(\hat{s}_{m'}) = \sum_{\lambda \in \Lambda_{m'}} \left[\hat{\beta}_\lambda^2 - 2\hat{\beta}_\lambda[\langle s, \varphi_\lambda\rangle + \varepsilon Z(\varphi_\lambda)]\right] = -\sum_{\lambda \in \Lambda_{m'}} \hat{\beta}_\lambda^2 = -\|\hat{s}_{m'}\|^2, \quad (3.7)$$

whatever $m' \in \mathcal{M}$. We now fix some $m \in \mathcal{M}$ and define $\mathcal{M}' = \{m' \in \mathcal{M} \mid \text{crit}(m') \leq \text{crit}(m)\}$. If m' belongs to $\mathcal{M}'$, we derive from (3.7) that $\gamma(\hat{s}_{m'}) + \text{pen}(m') \leq \gamma(\hat{s}_m) + \text{pen}(m)$, which yields by Definition 4, $\gamma(\hat{s}_{m'}) + \text{pen}(m') \leq \gamma(s_m) + \text{pen}(m)$ where s_m is the projection of s onto S_m. Since for all $t \in \mathbb{S}$, $\gamma(t) + \|s\|^2 = \|s - t\|^2 - 2\varepsilon Z(t)$, we derive that, whatever $m' \in \mathcal{M}'$,

$$\|s - \hat{s}_{m'}\|^2 \leq \|s - s_m\|^2 + 2\varepsilon[Z(\hat{s}_{m'}) - Z(s_m)] - \text{pen}(m') + \text{pen}(m). \quad (3.8)$$

In order to control $Z(t) - Z(s_m)$ uniformly for $t \in S_{m'}$ and $m' \in \mathcal{M}'$, we use a classical inequality due to Cirel'son, Ibragimov and Sudakov (1976) (see Ledoux, 1996, for the specific form we use below as well as many related deviation inequalities). Since the variance of the Gaussian process $t \mapsto \|t - s_m\|^{-1}[Z(t) - Z(s_m)]$ is identically equal to one, it follows from this inequality that, whatever $\lambda_{m'} > 0$,

$$\mathbb{P}\left[\sup_{t \in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|} \geq \mathbb{E}\left[\sup_{t \in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|}\right] + \lambda_{m'}\right] \leq \exp\left[-\frac{\lambda_{m'}^2}{2}\right]. \quad (3.9)$$

Introducing the D-dimensional linear space $S = S_m + S_{m'}$ and some orthonormal basis $\psi_1, \ldots, \psi_D$ of S, we derive from the Cauchy-Schwarz Inequality and the linearity of Z on $S \subset \mathbb{S}$ that

$$\sup_{t \in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|} \leq \sup_{u \in S} \frac{Z(u)}{\|u\|} = \sup_{\alpha \in \mathbb{R}^D} \frac{\sum_{j=1}^D \alpha_j Z(\psi_j)}{\left(\sum_{j=1}^D \alpha_j^2\right)^{1/2}} = \left[\sum_{j=1}^D Z^2(\psi_j)\right]^{1/2}.$$

Setting $\lambda_{m'}^2 = 2(L_{m'} D_{m'} + \xi)$ with $\xi > 0$ in (3.9) and observing that $D \leq D_m + D_{m'}$, we get

$$\mathbb{P}\left[\sup_{t \in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|} \geq \sqrt{D_m + D_{m'}} + \lambda_{m'}\right] \leq \exp(-L_{m'} D_{m'} - \xi).$$

Summing all those inequalities with respect to $m' \in \mathcal{M}^\star$ and using (3.3) we derive that except on a set Ω_ξ of probability bounded by $\Sigma \exp(-\xi)$,

$$\begin{aligned} Z(t) - Z(s_m) &\le \|t - s_m\| \left[[D_m + D_{m'}]^{1/2} + [2(L_{m'} D_{m'} + \xi)]^{1/2} \right] \\ &\le \|t - s_m\| \left[\sqrt{D_{m'}} \left(1 + \sqrt{2L_{m'}} \right) + \sqrt{D_m} + \sqrt{2\xi} \right], \end{aligned}$$

uniformly with respect to $t \in \cup_{m' \in \mathcal{M}^\star} S_{m'} = \cup_{m' \in \mathcal{M}} S_{m'}$. Let us now fix some $\eta \in (0,1)$. Using repeatedly the fact that $2ab \le a^2 c + b^2 c^{-1}$ for any $c > 0$ and the definition of $\mathrm{pen}(\cdot)$, we derive, since $\|t - s_m\| \le \|t - s\| + \|s - s_m\|$, that, except on Ω_ξ,

$$\begin{aligned} &2\varepsilon[Z(t) - Z(s_m)] \\ &\quad \le 2\varepsilon \|t - s_m\| \left[\sqrt{D_{m'}} \left(1 + \sqrt{2L_{m'}} \right) + \sqrt{D_m} + \sqrt{2\xi} \right] \\ &\quad \le \frac{\varepsilon^2}{1-\eta} \left[(1+\eta) D_{m'} \left(1 + \sqrt{2L_{m'}} \right)^2 + \left(1 + \eta^{-1} \right) \left(\sqrt{D_m} + \sqrt{2\xi} \right)^2 \right] \\ &\qquad + (1-\eta) \left[(1+\eta) \|t - s\|^2 + \left(1 + \eta^{-1} \right) \|s - s_m\|^2 \right] \\ &\quad \le \frac{1+\eta}{(1-\eta)K} \mathrm{pen}(m') + \frac{2\varepsilon^2 \left(1 + \eta^{-1} \right)}{1-\eta} (D_m + 2\xi) \\ &\qquad + (1 - \eta^2) \|t - s\|^2 + \left(\eta^{-1} - \eta \right) \|s - s_m\|^2. \end{aligned}$$

Together with (3.8), this inequality, applied with $t = \hat{s}_{m'}$ implies that, except on Ω_ξ and whatever $m' \in \mathcal{M}'$,

$$\begin{aligned} &\eta^2 \|s - \hat{s}_{m'}\|^2 + \frac{K - 1 - \eta(1+K)}{(1-\eta)K} \mathrm{pen}(m') \\ &\qquad \le \left(1 + \eta^{-1} - \eta \right) \|s - s_m\|^2 + \frac{2\varepsilon^2 \left(1 + \eta^{-1} \right)}{1-\eta} (D_m + 2\xi) + \mathrm{pen}(m). \quad (3.10) \end{aligned}$$

Choosing η small enough to get $K > 1 + \eta(1+K)$, we derive that, on the set Ω_ξ^c, $\sup_{m' \in \mathcal{M}'} \{\mathrm{pen}(m')\} < +\infty$, which means that there exists a number y such that $\mathcal{M}' \subset \mathcal{M}(y) = \{m' \in \mathcal{M} \mid \mathrm{pen}(m') \le y\}$. Now observe that if $m' \in \mathcal{M}(y)$, then $2K\varepsilon^2 L_{m'} D_{m'} \le y$ and therefore

$$\Sigma \ge \sum_{m' \in \mathcal{M}(y)} \exp[-L_{m'} D_{m'}] \ge |\mathcal{M}(y)| \exp\left[- \frac{y}{2K\varepsilon^2} \right].$$

We then conclude that $\mathcal{M}(y)$ is finite and $\mathcal{M}'$ as well, which implies that there exists a minimizer $\hat{m}$ of $\mathrm{crit}(m')$ over $\mathcal{M}'$ and therefore over $\mathcal{M}$. Let us now turn to the unicity of the penalized projection estimator. If $S_{m'} = S_{m''}$ and $\mathrm{pen}(m') = \mathrm{pen}(m'')$, then obviously $\mathrm{crit}(m') = \mathrm{crit}(m'')$ but then $\hat{s}_{m'} = \hat{s}_{m''}$ and if $\mathrm{pen}(m') \ne \mathrm{pen}(m'')$, then $\mathrm{crit}(m') \ne \mathrm{crit}(m'')$. Therefore, in order to prove unicity, it is enough to show that $\mathrm{crit}(m') \ne \mathrm{crit}(m'')$ a.s. as soon as $S_{m'} \ne S_{m''}$. This is a consequence of the fact that, in this case, $\varepsilon^{-2} \left[\|\hat{s}_{m'}\|^2 - \|\hat{s}_{m''}\|^2 \right]$ is the difference between two

independent non-central χ^2 variables, which easily follows from (2.5). Therefore, on the set Ω_ξ^c, there exists almost surely a unique penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$. Since $\mathbb{P}_s[\Omega_\xi]$ is arbitrarily small, a.s. existence and unicity follow. Now setting $\eta = (K-1)/(K+1)$, we derive from (3.10) applied to $m' = \hat{m}$ that, on the set Ω_ξ^c,

$$\left(\frac{K-1}{K+1}\right)^2 \|s-\tilde{s}\|^2 \leq \frac{K^2+4K-1}{K^2-1}\|s-s_m\|^2 + \operatorname{pen}(m) + \frac{2K(K+1)\varepsilon^2}{K-1}(D_m+2\xi),$$

and therefore, since $K\varepsilon^2 D_m \leq \operatorname{pen}(m)$,

$$\|s-\tilde{s}\|^2 \leq \frac{(K+1)^2}{(K-1)^3}\left[\frac{K^2+4K-1}{K+1}\|s-s_m\|^2 + (3K+1)\operatorname{pen}(m) + 4K(K+1)\varepsilon^2\xi\right],$$

except on the set Ω_ξ. Consequently, there exists a nonnegative random variable V with $\mathbb{P}[V > \xi] \leq \Sigma\exp(-\xi)$ for $\xi > 0$ and therefore $\mathbb{E}[V] \leq \Sigma$, such that

$$\|s-\tilde{s}\|^2 \leq \frac{4K(K+1)^2}{(K-1)^3}\left[\|s-s_m\|^2 + \operatorname{pen}(m) + (K+1)\varepsilon^2 V\right],$$

and (3.5) follows by integration since m is arbitrary. To get (3.6) we observe that

$$\begin{aligned}\ell\left(\|s-\tilde{s}\|^2\right) &\leq \ell\left(\frac{4K(K+1)^2}{(K-1)^3}\left[\|s-s_m\|^2 + \operatorname{pen}(m) + (K+1)\varepsilon^2 V\right]\right)\\ &\leq C_1(A,K)\left[\ell\left(\|s-s_m\|^2\right) + \ell(\operatorname{pen}(m))\right] + C'(A,K)\ell\left(\varepsilon^2 V\right).\end{aligned}$$

Since $\ell\left(2^j\varepsilon^2\right) \leq (2A)^j\ell\left(\varepsilon^2\right)$ for $j \geq 1$, we derive by integration that

$$\begin{aligned}\mathbb{E}\left[\ell\left(\varepsilon^2 V\right)\right] &\leq \ell\left(\varepsilon^2\right)\mathbb{P}[0 < V \leq 1] + \sum_{j\geq 1}\ell\left(2^j\varepsilon^2\right)\mathbb{P}\left[2^{j-1} < V \leq 2^j\right]\\ &\leq \Sigma\ell\left(\varepsilon^2\right)\left[1 + \sum_{j\geq 1}(2A)^j\exp\left(-2^{j-1}\right)\right],\end{aligned}$$

and (3.6) follows since m is arbitrary. □

3.3. First comments about the choice of the penalty

Let us first observe that the values of the weights for $m \notin \mathcal{M}^\star$ are irrelevant. Their introduction is actually unnecessary and has been made for notational convenience, in order to avoid to distinguish between two cases. One can, for instance, set $L_m = 0$ when $D_m = 0$.

One then notices that the upper bounds (3.5) and (3.6) in Theorem 2 suggest, together with (3.4), a choice of penalty of the form

$$\text{pen}(m) = K\varepsilon^2 D_m \left(1 + \sqrt{2L_m}\right)^2 \quad \text{for all } m \in \mathcal{M} \text{ and some } K > 1, \tag{3.11}$$

where the weights L_m satisfy (3.3) for a reasonable value of Σ, say $\Sigma \leq 1$, although the number one has no magic meaning here. In any case, from the asymptotic point of view, Σ should remain bounded when ε tends to zero. With such a choice of the penalty, one derives from (3.5) the cruder bound

$$\mathbb{E}_s\left[\|\hat{s}_{\hat{m}} - s\|^2\right] \leq \frac{4K(K+1)^3}{(K-1)^3}\left[\inf_{m\in\mathcal{M}}\left\{d^2(s, S_m) + \varepsilon^2 D_m\left(1+\sqrt{2L_m}\right)^2\right\} + \varepsilon^2\Sigma\right]. \tag{3.12}$$

This implies that a proper choice of the penalty function leads to a risk bound which only depends, up to some constant $C(K)$, on the family of models and weights $\{(S_m, L_m)\}_{m\in\mathcal{M}}$. This suggests to introduce the following

Definition 7. *Given some linear subspace $\mathbb{S}$ of some Hilbert space $\mathbb{H}$, we call* strategy *a finite or countable family $\{(S_m, L_m)\}_{m\in\mathcal{M}}$ where for all $m \in \mathcal{M}$, S_m denotes a D_m-dimensional linear subspace of $\mathbb{S}$ and L_m a nonnegative number such that $\sum_{\{m\in\mathcal{M}\,|\,D_m>0\}} \exp(-L_m D_m) = \Sigma < +\infty$. Given a strategy $\mathcal{S}$, its* accuracy index *$a_I(s, \mathcal{S}, \varepsilon)$ at point s is then defined as*

$$a_I(s, \mathcal{S}, \varepsilon) = \inf_{m\in\mathcal{M}}\left\{d^2(s, S_m) + \varepsilon^2 D_m(L_m + 1)\right\} + \Sigma\varepsilon^2. \tag{3.13}$$

One can now rewrite (3.12) in the following form: given a strategy and a penalty function satisfying (3.11), the corresponding penalized projection estimator satisfies

$$\mathbb{E}_s\left[\|\tilde{s} - s\|^2\right] \leq C_0(K)\, a_I(s, \mathcal{S}, \varepsilon), \tag{3.14}$$

for a suitable function C_0 of K.

3.3.1. How to choose K?

Given the family $\{L_m\}_{m\in\mathcal{M}}$, one can raise two natural questions concerning the choice of K:

- is the restriction $K > 1$ necessary;
- how to choose K in order to minimize the risk of $\tilde{s}$.

One can immediately see from (3.12) that our upper bound for the quadratic risk of the penalized estimator $\tilde{s}$ converges to infinity when K tends to one which suggests that the answer to the first question is actually "yes". We shall indeed show below in the context of variable selection (see Sect. 7) that 1 is a compulsary lower bound for K if we require our estimator to have good performances. This naturally leads

to advise against a choice of K smaller than one that is very likely to cause some disaster since then the model selection procedure $\hat{m}$ systematically chooses models with close to largest dimension. This phenomenon, which will be theoretically proven in Sect. 7 is quite spectacular in simulations.

The answer to the second question is much more delicate and cannot be solved by a simple optimization procedure performed on the right-hand side of (3.12) since the factor $f(K) = 4K(K+1)^3(K-1)^{-3}$ which appears there is far from being optimal: the resulting value of K would then be irrelevant. Nevertheless, the behaviour of $f(K)$ can be used to give a rough idea of how not to choose K! The fact that $f(K)$ tends to infinity when K tends to one suggests to avoid values of K close to one. Moreover, since $f(K)$ tends to infinity with K, too large values of K should be also avoided. But the two problems are clearly not symmetrical! Indeed, our risk bound increases linearly with K for large K while K smaller than one can truly make the risk blow up as shown in Sect. 7.

The search for an optimal value of K (at least asymptotically) requires a different proof which is more complicated and also specific to this particular framework of Gaussian model selection. It will therefore be postponed to a forthcoming paper, Birgé and Massart (2001). Let us just mention here that $K = 2$ should be recommended in most situations. In this case, the only difference between our penalty and Mallows' is the introduction of the weights L_m. The advantage of the proof that we have chosen here is that it can be extended, of course with additional technicalities, to other frameworks, like density estimation, since it emphasizes the link between concentration inequalities and the calibration of the penalty.

3.3.2. The role of the weights L_m

The choice of the weights L_m appears to be much more delicate than the choice of K since there is no optimal solution to this problem. Indeed, in view of minimizing the risk bound (3.5), given a penalty of the form (3.11) and a value of K, one should choose the L_ms as small as possible such that $\Sigma \leq 1$. This is obviously an ill-posed problem since its solution requires the knowledge of $d^2(s, S_m)$ for all m. The simplest way of choosing the L_ms is to take them constant, i.e. $L_m = L > 0$ for all m. Then,

$$\Sigma = \sum_{D \geq 1} |\{m \in \mathcal{M} \mid D_m = D\}| e^{-DL}.$$

Since we have imposed $\Sigma \leq 1$, such a solution is feasible provided that the number of models having a given dimension is finite and not too large, namely if $D^{-1} \log |\{m \in \mathcal{M} \mid D_m = D\}|$ is bounded. If so, one can take

$$L = \sup_{D \geq 1} D^{-1} \log |\{m \in \mathcal{M} \mid D_m = D\}| + \log 2.$$

This strategy, which treats all dimensions in the same way, can easily be refined by choosing L_m as a function of the dimension of S_m, i.e. $L_m = L(D_m)$ for a suitable function L satisfying

$$\sum_{D \geq 1} |\{m \in \mathcal{M} \mid D_m = D\}| e^{-DL(D)} \leq 1.$$

Of course, if there is an infinite number of models of some dimension D, the preceding strategy cannot work. Even if it is not so, some more sophisticated strategies may look attractive. If one suspects that the true s is close to some particular models, one is then tempted to give small weights to these models, in order to minimize the risk if our guess is true. This approach, based on some a priori information brought by the statistician, is quite analogue to the choice of a prior distribution in a Bayesian setting. This analogy will be made more explicit in the next section.

The influence of the various strategies (constant or variable weights) we just mentioned will actually be discussed in greater details when dealing with the numerous examples below. Let us just make here the following general remark: since (3.3) does not involve any model of dimension 0, the presence of such a model in the collection has actually no influence on the choice of the penalty for those ms such that $D_m > 0$ and one can always choose $\mathrm{pen}(m) = 0$ whenever $D_m = 0$. On the other hand, since the bound in (3.6) can only be improved if one enlarges the number of models without modifying Σ and the penalty function, it is always wise to include a zero-dimensional model in the collection.

3.4. A Bayesian interpretation of penalization

In order to discuss this point and for the sake of simplicity, let us forget the Gaussian linear processes for a while and go back to the simpler problem of estimating the mean s of a multidimensional Gaussian vector $Y \in \mathbb{R}^n$ with covariance matrix $\sigma^2 I_n$, where I_n denotes the identity matrix. We assume that we have at hand an at most countable collection $\{S_m\}_{m\in\mathcal{M}}$ of linear subspaces of $\mathbb{R}^n$ and that all those spaces are distinct. As we mentioned in Sect. 3.1, one can always remove the duplicate models without changing the value of the penalized estimator. We denote by ν the Lebesgue measure on $\mathbb{R}^n$, by ν_m the Lebesgue measure on S_m (which is the Dirac measure δ_0 when $D_m = 0$) and set $\mu = \sum_{m\in\mathcal{M}} \nu_m$. Let us then define

$$\mathcal{M}_m = \{m' \in \mathcal{M} \mid S_{m'} \not\supseteq S_m\} \qquad \text{and} \qquad S'_m = \left(\bigcup_{m'\in\mathcal{M}_m} S_{m'}\right) \bigcap S_m.$$

Since $\nu_m(S_{m'} \cap S_m) = 0$ when $m' \in \mathcal{M}_m$, then $\nu_m(S'_m) = 0$. As a consequence, a version of the density $d\nu_m/d\mu$ can be taken as $\mathbb{1}_{S_m\setminus S'_m}$. Let us now specify the prior "distribution" that we want to put on the parameter space $\cup_{m\in\mathcal{M}} S_m$. We first choose some prior θ on $\mathcal{M}$ with $\theta(\{m\}) = \theta_m$ and $\sum_{m\in\mathcal{M}} \theta_m = 1$. Then we assume that, given the value of m, s is "uniformly" distibuted on S_m, which means that it has the density 1 with respect to ν_m. This is obviously an improper prior. In other words, the prior "distribution" of s is taken as $\sum_{m\in\mathcal{M}} \theta_m \nu_m$, which has a density $\sum_{m\in\mathcal{M}} \theta_m \mathbb{1}_{S_m\setminus S'_m}$ with respect to μ. We recall that given s, the observation $Y \in \mathbb{R}^n$ is normal with mean s and covariance matrix $\sigma^2 I_n$. Therefore, the joint density of Y and s with respect to $\nu \otimes \mu$ is given by

$$\left(2\pi\sigma^2\right)^{-n/2} \sum_{m\in\mathcal{M}} \theta_m \mathbb{1}_{S_m\setminus S'_m}(s) \exp\left(-\frac{n}{2\sigma^2}\|y - s\|_n^2\right),$$

where $\|\cdot\|_n$ is the normalized Euclidean norm as in Sect. 1.2 and the posterior density of s given Y is proportional to

$$\sum_{m\in\mathcal{M}} \theta_m \mathbb{1}_{S_m\setminus S'_m}(s) \exp\left(-\frac{n}{2\sigma^2}\|s-Y\|_n^2\right).$$

Now observe that if $s \in (S_m \setminus S'_m) \cap S_{m'}$ with $m' \neq m$, then $S_{m'} \supset S_m$ and therefore $s \in S'_{m'}$. This implies that the sets $S_m \setminus S'_m$ are all disjointed. Moreover,

$$\sup_{s\in S_m\setminus S'_m}\left(-\|s-Y\|_n^2\right) = -\|\hat{s}_m - Y\|_n^2 = \|\hat{s}_m\|_n^2 - \|Y\|_n^2, \quad \text{a.s. with respect to } \nu.$$

Therefore the posterior mode is almost surely equal to $\hat{s}_{\hat{m}}$ with

$$\hat{m} = \operatorname*{argmax}_{m\in\mathcal{M}}\left\{\log(\theta_m) + n\|\hat{s}_m\|_n^2/(2\sigma^2)\right\}.$$

Equivalently, it can be written as

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}}\left[-\|\hat{s}_m\|_n^2 - 2n^{-1}\sigma^2\log(\theta_m) + C\right],$$

where C is an arbitrary constant, which means that $\hat{s}_{\hat{m}}$ is exactly the penalized projection estimator with penalty function $\operatorname{pen}(m) = -2n^{-1}\sigma^2\log(\theta_m) + C$, or equivalently

$$\theta_m = \exp\left(\frac{C - \operatorname{pen}(m)}{2\varepsilon^2}\right) \quad \text{with } \varepsilon = \frac{\sigma}{\sqrt{n}}.$$

Since θ is a probability distribution on $\mathcal{M}$,

$$C = -2\varepsilon^2\log\left(\sum_{m\in\mathcal{M}}\exp\left[-\operatorname{pen}(m)/\left(2\varepsilon^2\right)\right]\right).$$

The assumptions (3.3) and (3.4) clearly imply the convergence of the series which may very well diverge if $K < 1$ in (3.4).

The previous comparison shows that the penalized projection estimator is the mode of the posterior distribution in a Bayesian framework with an improper "uniform" prior distribution on each model and a prior probability for model S_m proportional to $\exp\left[-\operatorname{pen}(m)/\left(2\varepsilon^2\right)\right]$. The choice of the weights therefore amounts to the choice of a prior distribution on the family of models. We shall not go further in this direction and investigate the properties of the posterior distribution.

4. How to select good strategies?

4.1. Mixing several strategies

The choice of a strategy heavily depends on the type of problem we consider or the type of result we are looking for. For instance, as we shall see below, one should use different strategies for solving the problems of ordered and complete variable

selection, as defined in Sect. 1.2.1. Going back to our initial example, developed in Sects. 1.1.4 and 1.2.2, of a function s belonging to some unknown Sobolev ball $W^\alpha(R)$, we have seen in Sect. 1.5 that a good strategy to estimate s is based on the family of models S_D, $D \in \mathbb{N}$ where S_D is the linear span of the D first elements of the trigonometric basis $\{\varphi_i\}_{i\geq 1}$ defined in Sect. 1.1.4 ($S_0 = \{0\}$), with weights $L_D = 1$ for all $D \geq 1$. The resulting estimator is minimax, up to constants, over all Sobolev balls of radius $R \geq \varepsilon$. Unfortunately, such a strategy is good if s belongs to some Sobolev ball, but it may be definitely inadequate when s belongs to some particular Besov ball. In this case, one should use quite different strategies, for instance a thresholding method (which, as we shall see, is a specific strategy for complete variable selection) in connection with a wavelet basis, rather than the trigonometric one.

These examples are illustrations of a general recipe for designing simple strategies in view of solving the most elementary problems of adaptation: choose some orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ and a countable family $\mathcal{M}$ of finite subsets m of Λ, then define S_m to be the linear span of $\{\varphi_\lambda\}_{\lambda\in m}$ and find a family of weights L_m satisfying (3.3). Once again, the choice of a proper value of m can be viewed as a problem of variable selection from an infinite set of variables which are the coordinates vectors in the Gaussian sequence framework associated with the basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$. Obviously, the choice of a basis influences the approximation properties of the induced families of models. For instance the Haar basis is not suitable for approximating functions s which are "too smooth" (such that $\int_0^1 [s''(x)]^2\,dx$ is not large, say). If we have at hand a collection of bases, the choice of a "best" basis given s, ε and a strategy for estimating within each of the bases corresponds to the minimal value of the accuracy index at s and this "best basis" typically depends on the unknown s. Therefore one would like to be able to use all bases simultaneously rather than choosing one in advance. This is, in particular, a reason for prefering the Gaussian linear process approach to the Gaussian sequence framework.

The problem of the basis choice has been first considered and solved by Donoho and Johnstone (1994c) for selecting among the different threshold estimators built on the various bases. The following theorem provides a generic data driven way of mixing several strategies in order to retain the "best one".

Theorem 3. *Let $\mathcal{J}$ be a finite or countable set and μ a probability distribution on $\mathcal{J}$. For each $j \in \mathcal{J}$ we are given a collection $\{S_m\}_{m\in\mathcal{M}_j}$ of finite dimensional linear models with respective dimensions D_m and a collection of weights $\{L_{m,j}\}_{m\in\mathcal{M}_j}$ and we assume that the distribution μ satisfies*

$$\sum_{j\in\mathcal{J}} \mu(\{j\}) \left(\sum_{\{m\in\mathcal{M}_j \mid D_m>0\}} \exp[-D_m L_{m,j}] \right) = \Sigma < +\infty.$$

Let us consider for each $j \in \mathcal{J}$ a penalty function $\mathrm{pen}_j(\cdot)$ on $\mathcal{M}_j$ such that

$$\mathrm{pen}_j(m) \geq K\varepsilon^2 D_m \left(1 + \sqrt{2L_{m,j}}\right)^2 \quad \text{with } K > 1,$$

and the corresponding penalized projection estimator $\tilde{s}_j = \hat{s}_{\hat{m}_j}$ *where* $\hat{m}_j$ *minimizes the penalized criterion* $-\|\hat{s}_m\|^2 + \mathrm{pen}_j(m)$ *over* $\mathcal{M}_j$. *Let* $\hat{\jmath}$ *be a minimizer with respect to* $j \in \mathcal{J}$ *of*

$$-\|\tilde{s}_j\|^2 + \mathrm{pen}_j(\hat{m}_j) + \frac{2xK}{1-x}\varepsilon^2 l_j \quad \text{with } K^{-1} < x < 1 \quad \text{and} \quad l_j = -\log[\mu(\{j\})].$$

The resulting estimator $\tilde{s} = \tilde{s}_{\hat{\jmath}}$ *then satisfies*

$$\mathbb{E}_s\left[\|\tilde{s} - s\|^2\right] \le C(x, K)\left[\inf_{j\in\mathcal{J}}\left\{R_j + \frac{2xK}{1-x}\varepsilon^2 l_j\right\} + (xK+1)\,\varepsilon^2\Sigma\right],$$

with

$$C(x, K) = \frac{4xK(xK+1)^2}{(xK-1)^3} \quad \text{and} \quad R_j = \inf_{m\in\mathcal{M}_j}\left\{d^2(s, S_m) + \mathrm{pen}_j(m)\right\}.$$

Proof. Let $\mathcal{M} = \bigoplus_{j\in\mathcal{J}} \mathcal{M}_j \times \{j\}$ and set for all $(m, j) \in \mathcal{M}$ such that $D_m > 0$, $L'_{(m,j)} = L_{m,j} + D_m^{-1} l_j$. Then

$$\sum_{\{(m,j)\in\mathcal{M}\,|\,D_m>0\}} \exp[-D_m L'_{(m,j)}] = \Sigma.$$

Let $\mathrm{pen}((m, j)) = \mathrm{pen}_j(m) + [(2xK)/(1-x)]\varepsilon^2 l_j$, for all $(m, j) \in \mathcal{M}$. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we derive that

$$\left(\sqrt{D_m} + \sqrt{2L_{m,j}D_m + 2l_j}\right)^2 \le D_m\left(1 + \sqrt{2L_{m,j}}\right)^2 + 2l_j + 2\sqrt{2l_j D_m},$$

which implies since $2\sqrt{2l_j D_m} \le 2l_j x/(1-x) + D_m(1-x)/x$ that

$$\left(\sqrt{D_m} + \sqrt{2L_{m,j}D_m + 2l_j}\right)^2 \le x^{-1} D_m\left(1 + \sqrt{2L_{m,j}}\right)^2 + 2l_j/(1-x).$$

It then follows that

$$\begin{aligned}\mathrm{pen}((m, j)) &\ge xK\varepsilon^2\left[x^{-1}D_m\left(1 + \sqrt{2L_{m,j}}\right)^2 + 2l_j/(1-x)\right]\\ &\ge xK\varepsilon^2\left(\sqrt{D_m} + \sqrt{2L_{m,j}D_m + 2l_j}\right)^2,\end{aligned}$$

and therefore

$$\mathrm{pen}((m, j)) \ge xK\varepsilon^2 D_m\left(1 + \sqrt{2L'_{(m,j)}}\right)^2. \tag{4.1}$$

We can now apply Theorem 2 to the strategy defined for all $(m, j) \in \mathcal{M}$ by the model S_m and the penalty $\mathrm{pen}((m, j))$. By definition, the resulting estimator is clearly $\tilde{s}$ and the risk bound follows from (3.5) with K replaced by $xK > 1$ because of (4.1). □

Remarks.

- The definition of $\mathcal{M}$ that we used in the proof of the theorem may lead to situations where the same model S_m appears several times with possibly different weights. This is why we emphasized, in the presentation of the framework preceding Theorem 2, the fact that such a redundancy was allowed.
- Note that the choice of a suitable value of x leads to the same difficulties as the choice of K and one should avoid to take xK close to 1 (see Sect. 3.3.1).

The preceding theorem gives indeed a solution to the problems we considered before. If one wants to mix a moderate number of strategies one can build a "superstrategy" as indicated in the theorem, with μ the uniform distribution on $\mathcal{J}$, and the price to pay in the risk is an extra term of order $\varepsilon^2 \log(|\mathcal{J}|)$. In this case, the choice of $\hat{\jmath}$ is particularily simple since it should merely satisfy $\|\tilde{s}_{\hat{\jmath}}\|^2 - \mathrm{pen}_{\hat{\jmath}}(\hat{m}_{\hat{\jmath}}) = \sup_{j\in\mathcal{J}} \left\{\|\tilde{s}_j\|^2 - \mathrm{pen}_j(\hat{m}_j)\right\}$. If $\mathcal{J}$ is too large, one should take a different "prior" than the uniform on the set of available strategies. One should put larger values of $\mu(\{j\})$ for the strategies corresponding to values of s we believe are more likely and smaller values for the other strategies. As for the choice of the weights L_m (see Sect. 3.4), the choice of μ has some Bayesian flavour.

As a matter of conclusion, let us mention that the problem of mixing several estimation methods in order to get the best of each is not new. Our approach to this problem seems to be new but it is limited to a specific class of estimators, namely penalized projection estimators. More general points of view appear in Yang (2000) and Catoni (2000), based on previous ideas of Barron (1987), as well as in Nemirovski (2000, Chaps. 5 and 6).

4.2. Adaptation in the minimax sense

Now that we have at hand a powerful tool (Theorem 3) to mix strategies associated with different bases or with a single one, it remains to decide what are the good strategies within a given basis. As we already noticed in Sect. 2.3.3 a natural benchmark for measuring the performance of penalized projection estimators, is the oracle accuracy given by Definition 5. If $\mathcal{S}$ is a strategy with bounded weights, i.e. $L_m \leq L$ for all $m \in \mathcal{M}$, it follows from (3.13) that the accuracy index is comparable to the oracle accuracy via the inequality

$$a_I(s, \mathcal{S}, \varepsilon) \leq [(1 \vee \Sigma) + L]\left[a_O(s, \mathcal{S}, \varepsilon) + \varepsilon^2\right]. \tag{4.2}$$

Note here that although the oracle accuracy does not depend on the weights L_m, the notation $a_O(s, \mathcal{S}, \varepsilon)$ is perfectly meaningful. An advantage of this approach is that this comparison with the oracle makes sense for all s. On the other hand, it has at least two serious drawbacks:

- this comparison becomes meaningless when the family $\{L_m\}_{m\in\mathcal{M}}$ is unbounded or is misleading when either L or Σ is large since the right-hand side of (4.2) can then be substantially larger than the accuracy index;
- it provides no information on the comparison between $\tilde{s}$ and some arbitrary estimator, which typically does not belong to the family $\{\hat{s}_m\}_{m\in\mathcal{M}}$.

The above criticisms of "oracle inequalities" such as (4.2) suggest to consider other optimality criteria in order to judge of the quality of penalized projection estimators. As mentioned in Sect. 1.2.2, one such criterion is the minimax risk over suitable subsets $\mathcal{T}$ of $\mathbb{H}$, as defined by (1.9). From this point of view, the performance of an estimator $\hat{s}_\varepsilon$ (generally depending on ε) can then be measured by the ratio

$$\mathcal{R}(\hat{s}_\varepsilon, \mathcal{T}, \varepsilon) = \sup_{s\in\mathcal{T}} \frac{\mathbb{E}_s\left[\|\hat{s}_\varepsilon - s\|^2\right]}{R_M(\mathcal{T}, \varepsilon)},$$

and the closer this ratio to one, the better the estimator. In particular, if this ratio is bounded independently of ε, the family of estimators $\{\hat{s}_\varepsilon\}_{\varepsilon>0}$ will be called *approximately minimax* with respect to $\mathcal{T}$. Many approximately minimax estimators have been constructed for various sets $\mathcal{T}$. As for the case of Sobolev balls, they typically depend on $\mathcal{T}$ which is a serious drawback. One would like to design estimators which are approximately minimax for many $\mathcal{T}$s simultaneously, for instance all Sobolev balls $W^\alpha(R)$, with $\alpha > 0$ and $R \geq \varepsilon$. The construction of such adaptive estimators has been the concern of many statisticians (see Barron et al. 1999, Sect. 5 for a detailed discussion of adaptation with many bibliographic citations).

In order to see to what extent our method allows to build adaptive estimators in various situations, we shall consider below a number of examples and for any such example, use the same construction. Given a class of sets $\{\mathcal{T}_\theta\}_{\theta\in\Theta}$ we choose a family of models $\{S_m\}_{m\in\mathcal{M}}$ which adequately approximate those sets. This means that we choose the models in such a way that any s belonging to some $\mathcal{T}_\theta$ can be closely approximated by some model of the family. Then we choose a family of weights $\{L_m\}_{m\in\mathcal{M}}$ satisfying the condition (3.3) for some reasonably small value of Σ. Theses choices completely determine the construction of the penalized projection estimator $\tilde{s}$ (up to the choice of K which is irrelevant in term of rates since it only influences the constants). In order to analyze the performances of the resulting estimator, it is necessary to evaluate, for each $\theta \in \Theta$ and each $s \in \mathcal{T}_\theta$ the accuracy index $a_I(s, \mathcal{S}, \varepsilon)$ since

$$\mathcal{R}\big(\hat{s}_\varepsilon, \mathcal{T}_\theta, \varepsilon\big) \leq C \sup_{s\in\mathcal{T}_\theta} \frac{a_I(s, \mathcal{S}, \varepsilon)}{R_M(\mathcal{T}_\theta, \varepsilon)}.$$

In order to bound the accuracy index, we first have to compute the distances $d(s, S_m)$ for each $m \in \mathcal{M}$, which derive from Approximation Theory, then procede to the minimization with respect to $m \in \mathcal{M}$.

4.3. Choice of collections of models and Approximation Theory

In order to understand how to choose "good" families of models, let us consider some particular weighting strategy for which it is especially easy to understand the behaviour of the accuracy index. If for any integer D there is only a finite number of models of dimension D one can choose L_m as a function $L(D_m)$ of the dimension which satisfies

$$\sum_{D\geq 1} |\{m \in \mathcal{M} \mid D_m = D\}| \exp[-DL(D)] = \Sigma < +\infty. \tag{4.3}$$

This leads to

$$a_I(s, \mathcal{S}, \varepsilon) = \inf_{D \geq 0} \left\{ \varepsilon^2 [DL(D) + D + \Sigma] + \inf_{\{m \in \mathcal{M} \mid D_m = D\}} d^2(s, S_m) \right\}. \quad (4.4)$$

Taking (4.3) into account, we see that controlling the accuracy index forces us to make some compromises between the number of models of the same dimension and their approximation capabilities since a large number of models of dimension D potentially reduces the value of $\inf_{\{m \in \mathcal{M} \mid D_m = D\}} d^2(s, S_m)$ but requires a large value for $L(D)$ or Σ. Moreover we would like to control the accuracy index for as many s simultaneously as possible. It is precisely one of the main purposes of Approximation Theory to provide linear or nonlinear approximation procedures for various types of regular functions. Most of the constructive methods of approximation that we know amount to select some coefficients from infinite dimensional expansions on one single basis such as polynomials, piecewise polynomials, trigonometric polynomials, wavelets, splines, ... (and it is here that the basis choice comes in) and naturally lead to collections of finite dimensional linear models S_m for which $\inf_{\{m \in \mathcal{M} \mid D_m = D\}} d^2(s, S_m)$ can be controlled in term of the various moduli of smoothness of s. Results from Approximation Theory will therefore be at the heart of our choices of suitable families of models for curve estimation.

5. Finite-dimensional variable selection

In order to define some basic strategies that will be the milestones for further developments and to analyze in a more concrete way the role of the weights and the difference between our criterion and Mallows' C_p, let us go back to one of the problems that motivated this work: variable selection for a Gaussian linear regression.

Translating (1.1) into our framework, we start from an observation of the process $Y(t) = \langle s, t \rangle + \varepsilon Z(t)$, where t varies in the linear span $\mathbb{S}_N$ of some system $\{\varphi_1, \ldots, \varphi_N\}$ of linearly independent (but not necessarily orthonormal) vectors and s is unknown in $\mathbb{H}$. Variable selection then amounts to choose a proper value of Λ_m as a subset of $\Lambda = \{1; \ldots; N\}$ (with $N \geq 2$ in order to avoid trivialities), from the observation of Y. In short, we look for an efficient reconstruction of s, taking into account the noise level and the number N of available variables, in the form $\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda$ which means that S_m is the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$. We focus here on two typical and, in some sense, extremal situations:

- Ordered variable selection amounts to restrict to subsets Λ_m of Λ of the form $\{1; \ldots; m\}$ with $1 \leq m \leq N$. In this case, one takes $\mathcal{M} = \{1; \ldots; N\}$. Such a variable selection problem is especially meaningful for time-depending variables or variables which are algebraic or trigonometric polynomials.
- Complete variable selection considers all subsets $\Lambda_m = m$ of $\{1; \ldots; N\}$. In this case $\mathcal{M} = \mathcal{P}(\{1; \ldots; N\})$.

5.1. Oracle type inequalities

Here, we want to compare the risk bounds (3.5) to the oracle accuracy when the penalty is chosen according to (3.11). Of course, this comparison heavily depends

on the choice of the weights L_m but, in principle, on K also. Since the dependence with respect to K of bound (3.5) is very crude, we shall make no attempt to be precise in the following evaluations and shall content ourselves with comparisons of risks bounds up to multiplicative constants depending on K. Taking (3.14) into account, we shall focus on the evaluation of the accuracy index for various strategies. In the sequel we shall use various quantities depending on the parameters involved in our strategies. Such a quantity will systematically be denoted by C or more precisely $C(\cdot, \cdots, \cdot)$ to indicate that its value only depends on the parameters appearing as its arguments. Its value may change from line to line.

5.1.1. Ordered variable selection

Constant weights Let us begin with the simplest weighting strategy. Choosing $L_m = L > 0$ for $m \in \mathcal{M}$ gives (3.3) with $\Sigma < \left(e^L - 1\right)^{-1}$ and (3.13) then leads to a strategy $\mathcal{S}$ with accuracy index bounded by

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 m \, (1 + L) \right\} + \varepsilon^2 \left(e^L - 1\right)^{-1}. \tag{5.1}$$

Since $a_O(s, \mathcal{F}, \varepsilon) \geq \varepsilon^2$, one gets $a_I(s, \mathcal{S}, \varepsilon) \leq C(L)\, a_O(s, \mathcal{F}, \varepsilon)$. It should also be noted that all penalties of the form $K' m \varepsilon^2$ with $K' > 1$, which correspond to suitable choices of the pair (K, L), are allowed, including that of Mallows, namely $K' = 2$.

Variable weights In order to improve on (5.1), we can introduce weights which depend on the dimension. Let us, for instance, set $L_m = \theta^2 m^{-1/2}$ for some $\theta > 0$. Then (3.3) holds with $\Sigma < \Sigma_\theta = \sum_{D=1}^{+\infty} \exp\left(- \theta^2 \sqrt{D}\right)$ which leads to a strategy $\mathcal{S}$ with accuracy index bounded by

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 m \left(1 + \theta^2 m^{-1/2}\right) \right\} + \Sigma_\theta \varepsilon^2. \tag{5.2}$$

Straightforward computations show that, if $L = \log\left[1 + \exp\left(\theta^2\right)/2\right]$ and $\theta^2 \geq 3$, then $\Sigma_\theta < \left(e^L - 1\right)^{-1}$ which allows an easy comparison with (5.1): if m_0 denotes a minimizer of $d^2(s, S_m) + \varepsilon^2 m(1 + L)$ over $\mathcal{M}$, bound (5.2) is better than bound (5.1) whenever $L > \theta^2 m_0^{-1/2}$. It should also be noted that, when $K = 2$, our penalty can be viewed as a corrected version of Mallows' C_p, which is equivalent to it when m goes to infinity.

In both situations (constant or variable weights), we have proved that an oracle inequality of the form (2.9) holds for suitable choices of the penalty. This shows that those penalized projection estimators are optimal in the sense that, up to constants, they are minimax over any space S_m with $m \in \mathcal{M}$. Nevertheless, the preceding study tends to indicate that suitably chosen variable weights should be prefered to constant weights.

5.1.2. Complete variable selection and thresholding

A quite different situation occurs when one allows $m = \Lambda_m$ to be any subset of $\{1; \ldots; N\}$, S_m being the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ with $D_m = |m|$. When the

system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ is orthonormal, we shall denote by $\hat{\beta}_\lambda$ the estimated coefficient $Y(\varphi_\lambda)$. Then by (2.5) $\hat{s}_m = \sum_{\lambda\in m} \hat{\beta}_\lambda \varphi_\lambda$ for all $m \in \mathcal{M}$.

Constant weights If we choose constant weights, namely $L_m = L$ for all $m \in \mathcal{M}$, we get

$$\Sigma = \sum_{D=1}^{N} \binom{N}{D} e^{-LD} = \left(1 + e^{-L}\right)^N - 1.$$

In order that Σ remains bounded independently of N one has to take $L/\log N \geq 1 + o(1)$ when $N \to +\infty$. This also means that the use of Mallows' C_p criterion, which requires that L be smaller than $3/2 - \sqrt{2}$ leads to $\Sigma > 1.9^N - 1$ and the resulting risk bound is therefore irrelevant when N is large. At this point, one can suspect that Mallows' criterion might not be suitable for complete variable selection with a large number of variables and should rather be replaced by a penalty of the form $\text{pen}(m) = K|m|\varepsilon^2\left(1 + \sqrt{2L}\right)^2$ with $L = \log N$ which warrants that $\Sigma < (1 + \log N) \wedge (e - 1)$ and that accuracy index of the corresponding strategy is bounded by

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m\in\mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2|m|[1 + \log N] \right\} + \Sigma\varepsilon^2 \tag{5.3}$$

$$\leq [1 + \log N] \left[a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2 \right]. \tag{5.4}$$

We then miss a factor $1 + \log N$ with respect to the oracle accuracy.

Let us now turn to computational issues when the system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ is orthonormal and first show that

$$a_O(s, \mathcal{F}, \varepsilon) = \sum_{\lambda\in\Lambda} \left(\beta_\lambda^2 \wedge \varepsilon^2 \right). \tag{5.5}$$

Indeed,

$$a_O(s, \mathcal{F}, \varepsilon) = \inf_{m\in\mathcal{M}} \left\{ \sum_{\lambda\notin m} \beta_\lambda^2 + \varepsilon^2|m| \right\} = \|s\|^2 + \inf_{m\in\mathcal{M}} \sum_{\lambda\in m} \left(-\beta_\lambda^2 + \varepsilon^2 \right). \tag{5.6}$$

The infimum is reached by $m^\star = \left\{ \lambda \;\middle|\; \beta_\lambda^2 > \varepsilon^2 \right\}$. Plugging this value of $m^\star$ in (5.6) gives (5.5) and (5.4) therefore becomes

$$a_I(s, \mathcal{S}, \varepsilon) \leq [1 + \log N] \left[\sum_{\lambda\in\Lambda} \left(\beta_\lambda^2 \wedge \varepsilon^2 \right) + \varepsilon^2 \right]. \tag{5.7}$$

As to the penalized projection estimator, while its computation apparently requires an optimization over a family of models of cardinality 2^N which does not seem numerically feasible when N is large, it turns out, when the system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ is orthonormal, to be easily computed. Indeed, since the penalty is proportional to the dimension of the model, i.e. $\text{pen}(m) = T^2|m|$ for some positive number $T = \sqrt{K}\left(1 + \sqrt{2\log N}\right)\varepsilon$, the minimization with respect to $m \in \mathcal{M}$ of the

penalized criterion $-\|\hat{s}_m\|^2 + \text{pen}(m)$ is similar to the one involved in the computation of the oracle accuracy. It amounts, according to (2.5) to the minimization of $\sum_{\lambda\in\Lambda_m}\left[-\hat{\beta}_\lambda^2 + T^2\right]$ and $\hat{m}$ is therefore given by $\hat{m} = \{\lambda \in \Lambda \,|\, |\hat{\beta}_\lambda| > T\}$. This results in a *threshold estimator* of the form

$$\tilde{s}_T = \hat{s}_{\hat{m}} = \sum_{\lambda\in\Lambda} \hat{\beta}_\lambda \mathbb{1}_{\{|\hat{\beta}_\lambda|>T\}}\varphi_\lambda, \tag{5.8}$$

which means that one only keeps in the expansion of $\tilde{s}$ the coefficients $\hat{\beta}_\lambda$ which have a large enough absolute value, larger than the threshold T. These estimators have been investigated in great details by Donoho and Johnstone (1994a), at least from an asymptotic point of view (when N tends to infinity). They have shown (Theorem 4, p. 439) that the choice $T = T_N = (2\log N)^{1/2}\varepsilon$ leads to

$$\mathbb{E}_s\left[\|s - \tilde{s}_{T_N}\|^2\right] \le \kappa_N\left[\sum_{\lambda\in\Lambda}\left(\beta_\lambda^2 \wedge \varepsilon^2\right) + \varepsilon^2\right] \quad \text{when } s = \sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda, \tag{5.9}$$

where the ratio $\kappa_N/(2\log N)$ converges to 1 as N goes to infinity. Apart from the multiplicative factor $C_0(K)[1+\log N]/\kappa_N$, which is bounded, the inequality resulting from the combination of (3.14) and (5.7) is the same as (5.9) .

Note that we were unable to prove an oracle inequality of the form (2.9) with a universal constant C. It is actually impossible to get such an inequality due to the following result of Donoho and Johnstone (1994a, Theorem 3) which, according to (5.5), can be written as

$$\liminf_{N\to+\infty}\frac{1}{\log N}\left[\inf_{\hat{s}}\sup_{s\in\mathbb{S}_N}\frac{\mathbb{E}_s\left[\|s-\hat{s}\|^2\right]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2}\right] \ge 2,$$

$\hat{s}$ denoting an arbitrary estimator. This shows that complete variable selection is definitely more difficult than ordered variable selection. A further consequence of this lower bound is that (5.9) is asymptotically optimal in this minimax sense, i.e.

$$\limsup_{N\to+\infty}\frac{1}{\log N}\left[\sup_{s\in\mathbb{S}_N}\frac{\mathbb{E}_s\left[\|s-\tilde{s}_{T_N}\|^2\right]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2}\right] = 2. \tag{5.10}$$

It should be noticed that bound (5.7) may be too pessimistic. Indeed, using (5.3) one gets instead

$$a_I(s,\mathcal{S},\varepsilon) \le \sum_{\lambda\in\Lambda}\left(\beta_\lambda^2 \wedge \varepsilon^2[1+\log N]\right) + (e-1)\varepsilon^2, \tag{5.11}$$

which can be substantially better than (5.7) as can be seen when $s = \varepsilon\sum_{\lambda\in m}\varphi_\lambda$ where $m \in \mathcal{M}\setminus\emptyset$. As a conclusion, we see that it is more clever to use bounds (3.5) or (3.13) as they stand rather than trying to put them in the form (2.9), which may be misleading when C is large.

Variable weights As in the case of ordered variable selection, it is possible to improve on (5.3) by simply introducing weights which depend on the dimension, i.e. $L_m = L(|m|)$. This leads to

$$\Sigma = \sum_{D=1}^{N} \binom{N}{D} \exp[-DL(D)] \leq \sum_{D=1}^{N} \left(\frac{eN}{D}\right)^{D} \exp[-DL(D)]$$
$$\leq \sum_{D=1}^{N} \exp\left[-D\left[L(D) - 1 - \log\left(\frac{N}{D}\right)\right]\right].$$

Hence the choice $L(D) = 1+\theta+\log(N/D)$ with $\theta > 0$ leads to $\Sigma \leq \sum_{D=1}^{\infty} e^{-D\theta} = \left[e^{\theta} - 1\right]^{-1}$. Choosing $\theta = \log 2$ for the sake of simplicity we derive the following bound for the accuracy index:

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 |m| \left[1 + \log(2N/|m|)\right] \right\} + \varepsilon^2, \tag{5.12}$$

which is better than (5.3) when $|m| \neq 1$ and only slightly worse (from a factor 1.7) when $|m| = 1$. This implies that the accuracy index of our new strategy satisfies analogues of (5.11) and (5.7) and therefore the corresponding estimator $\tilde{s}$ satisfies an analogue of (5.10), namely

$$\limsup_{N \to +\infty} \frac{1}{\log N} \left[\sup_{s \in \mathbb{S}_N} \frac{\mathbb{E}_s\left[\|s - \tilde{s}\|^2\right]}{a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2} \right] \leq C(K). \tag{5.13}$$

On the other hand, the variable weights penalized projection estimator is also rather easy to compute when the system $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ is orthonormal. Indeed

$$\inf_{m \in \mathcal{M}} \left\{ -\sum_{\lambda \in m} \hat{\beta}_\lambda^2 + K\varepsilon^2 |m| \left(1 + \sqrt{2L(|m|)}\right)^2 \right\}$$
$$= \inf_{0 \leq D \leq N} \left\{ - \sup_{\{m \,|\, |m| = D\}} \sum_{\lambda \in m} \hat{\beta}_\lambda^2 + K\varepsilon^2 |D| \left(1 + \sqrt{2L(|D|)}\right)^2 \right\}$$
$$= \inf_{0 \leq D \leq N} \left\{ -\sum_{j=1}^{D} \hat{\beta}_{\tau(j)}^2 + K\varepsilon^2 |D| \left(1 + \sqrt{2L(|D|)}\right)^2 \right\} \tag{5.14}$$

where $\hat{\beta}_{\tau(1)}^2 \geq \ldots \geq \hat{\beta}_{\tau(N)}^2$ are the squared estimated coefficients of s in decreasing order. This suggests to introduce the following notations:

Definition 8. *Given a set of real numbers $\{b_i\}_{i \in I}$ indexed by some finite set I with cardinality N, one denotes by $\{b_{(j)(I)}\}_{1 \leq j \leq N}$ the same set of numbers in decreasing order of their absolute values which means that $|b_{(1)(I)}| \geq |b_{(2)(I)}| \geq \ldots \geq |b_{(N)(I)}|$. For $1 \leq D \leq N$, $\overline{I}[D]$ is then the subset of I of those indices corresponding to the elements $\{b_{(j)(I)}\}_{1 \leq j \leq D}$, i.e.*

$$|\overline{I}[D]| = D \quad \text{and} \quad |b_i| \geq |b_j| \quad \text{for all } i, j \text{ with } i \in \overline{I}[D],\ j \notin \overline{I}[D]. \tag{5.15}$$

Formally, the definition of $\overline{I}[D]$ depends primarily on the numbers b_i although they do not appear in the notation, for the sake of simplicity. This will not cause any ambiguity in the sequel since we shall always consider quantities of the form $\sum_{i\notin\overline{I}[D]} b_i^2$ (say) and the indices then systematically apply to the elements which were used to define the corresponding sets. Going back to (5.14) we see that minimizing crit(m) amounts to select a value $\hat{D}$ of D which minimizes

$$-\sum_{j=1}^{D} \hat{\beta}^2_{(j)(\Lambda)} + K\varepsilon^2|D|\left(1+\sqrt{2L(|D|)}\right)^2.$$

This finally leads to $\hat{m} = \overline{\Lambda}[\hat{D}]$.

5.2. *Minimax and adaptive properties of the variable weights strategies*

From the global minimax point of view introduced by Donoho and Johnstone, the threshold estimator which is a penalized projection estimator with constants weights and the penalized projection estimator with variable weights have similar performances (apart from the asymptotic constants) – compare (5.10) and (5.13) –. However this point of view may be somewhat misleading since inequalities like (5.9) or (5.11) can be substantially improved for some values of s when replacing the threshold estimator which corresponds to the constant weights strategy by the penalized projection estimator $\tilde{s}$ which corresponds to the variable weights strategy. In order to understand what type of improvement is possible, let us now consider a less global minimax point of view. For this purpose we introduce the spaces $\mathbb{S}_D$ of those functions s that have at most $D \geq 1$ nonzero coefficients, namely $\mathbb{S}_D = \cup_{\{m\in\mathcal{M}\,|\,|m|=D\}} S_m$. Comparing the performances of estimators with respect to $\mathbb{S}_D$ appears to be rather natural in the context of complete variable selection. In order to get precise results, we assume all along this section the system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ to be orthonormal.

Defining $\tilde{s}$ as the variable weights penalized projection estimator described in the previous section, one derives from (5.12) and (3.14) that

$$\sup_{s\in\mathbb{S}_D} \mathbb{E}_s\left[\|\tilde{s}-s\|^2\right] \leq C_0(K)\varepsilon^2[D + D\log(2N/D) + 1]. \tag{5.16}$$

It is interesting to notice that such a bound is not achievable with constant weights. To see this, we recall that the corresponding penalized projection estimator is a threshold estimator $\tilde{s}_T$ given by (5.8). When $T = \sqrt{K}\left(1+\sqrt{2\log N}\right)\varepsilon$, (5.11) combined with (3.14) leads to a risk bound which is weaker than (5.16) since

$$\sup_{s\in\mathbb{S}_D} \sum_{\lambda\in\Lambda}\left(\beta_\lambda^2 \wedge \varepsilon^2[1+\log N]\right) = D\varepsilon^2[1+\log N].$$

One can prove a more concrete result in the form of a lower bound for the risk of any threshold estimator $\tilde{s}_T$ (i.e. any penalized projection estimator with constant weights). Very similar computations appear in the Appendix of Donoho and Johnstone (1994a) and the lectures of Johnstone (1998), but since they aim at getting upper bounds, they are not stated in a form which is suitable to our needs.

Proposition 2. *Let $T > 0$, $\tilde{s}_T$ be the threshold estimator defined by (5.8), m an arbitrary subset of $\{1; \dots ; N\}$ and $\delta_\lambda = \pm 1$ for $\lambda \in m$. If $s = T \sum_{\lambda \in m} \delta_\lambda \varphi_\lambda$, then*

$$\mathbb{E}_s\left[\|\tilde{s}_T - s\|^2\right] \geq \frac{|m|}{2}\left(T^2 + \varepsilon^2\right) + \frac{N - |m|}{2}\varepsilon^2 \left(\frac{T\sqrt{2}}{\varepsilon\sqrt{\pi}} \vee 1\right) \exp\left(-\frac{T^2}{2\varepsilon^2}\right). \tag{5.17}$$

Proof. We can assume without loss of generality, that $s = T \sum_{\lambda \in m} \varphi_\lambda$. Then

$$\begin{aligned}
\mathbb{E}_s\left[\|\tilde{s}_T - s\|^2\right] &= \sum_{\lambda \in m} \mathbb{E}_s\left[\left(T - \hat{\beta}_\lambda \mathbb{1}_{\{|\hat{\beta}_\lambda| > T\}}\right)^2\right] + \sum_{\lambda \notin m} \mathbb{E}_s\left[\hat{\beta}_\lambda^2 \mathbb{1}_{\{|\hat{\beta}_\lambda| > T\}}\right] \\
&= |m| \mathbb{E}\left[\left(T - (T + \varepsilon\xi)\mathbb{1}_{\{|(T+\varepsilon\xi)| > T\}}\right)^2\right] \\
&\quad + (N - |m|)\varepsilon^2 \mathbb{E}\left[\xi^2 \mathbb{1}_{\{|\varepsilon\xi| > T\}}\right].
\end{aligned}$$

In order to conclude, it suffices to observe that

$$\left(T - (T + \varepsilon\xi)\mathbb{1}_{\{|(T+\varepsilon\xi)| > T\}}\right)^2 \geq T^2 \mathbb{1}_{\{\xi \leq 0\}} + \varepsilon^2 \xi^2 \mathbb{1}_{\{\xi > 0\}},$$

and apply the next elementary lemma. □

Lemma 1. *If ξ is standard normal and $t \geq 0$, then*

$$\mathbb{E}\left[\xi^2 \mathbb{1}_{\{\xi > t\}}\right] \geq \left(\frac{t}{\sqrt{2\pi}} \vee \frac{1}{2}\right) \exp\left(-\frac{t^2}{2}\right).$$

It follows from Proposition 2 with $m = \emptyset$ that whenever $T \leq \varepsilon\sqrt{2c \log N}$ with $c < 1$, the risk at zero of the threshold estimator is at least $\left(\varepsilon^2/2\right) N^{1-c}$ which is much larger than the expected $\varepsilon^2 \log N$. This suggests to focus on threshold estimators with a large enough level T of thresholding, say $T \geq \varepsilon\sqrt{\log N}$, in which case it follows from (5.17) that

$$\sup_{s \in \mathbb{S}_D} \mathbb{E}_s\left[\|\tilde{s}_T - s\|^2\right] \geq \varepsilon^2 (D/2)(1 + \log N). \tag{5.18}$$

In particular, when $D = N$, the threshold estimator looses a $\log N$ factor from the risk of $\tilde{s}$ as shown by (5.16). More generally, when $T \geq \varepsilon\sqrt{\log N}$ and $D \geq 1$,

$$\frac{\sup_{s \in \mathbb{S}_D} \mathbb{E}_s\left[\|\tilde{s}_T - s\|^2\right]}{\sup_{s \in \mathbb{S}_D} \mathbb{E}_s\left[\|\tilde{s} - s\|^2\right]} \geq C(K) \frac{1 + \log N}{1 + \log(N/D)}. \tag{5.19}$$

This last inequality suggests that variable weights should be prefered to constant weights but one can even prove a more convincing result demonstrating the superiority of the variables weights strategy, namely the fact that it is adaptive over the family of all spaces $\mathbb{S}_D$. The following theorem shows that it is minimax, up to constants, over all the spaces $\mathbb{S}_D$ with $D \geq 1$ simultaneously and that (5.16) cannot be improved.

Theorem 4. *There exist two positive universal constants κ and κ' such that the minimax risk $R_M(\mathbb{S}_D, \varepsilon)$ over $\mathbb{S}_D$ – as defined by (1.9) – satisfies*

$$\kappa\varepsilon^2 D[1 + \log(N/D)] \leq R_M(\mathbb{S}_D, \varepsilon) \leq \kappa'\varepsilon^2 D[1 + \log(N/D)], \tag{5.20}$$

for all $\varepsilon > 0$, $N \geq 1$ and $1 \leq D \leq N$.

Since the upper bound derives from (5.16), we only have to prove the lower bound result, which is an immediate consequence of the following theorem (the proof of which is given in Sect. 8.2) with $b^2 = 1 + \log(N/D)$.

Theorem 5. *Let $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ with $\Lambda = \{1, 2, \ldots, N\}$ be an orthonormal system in $\mathbb{H}$ and for $1 \leq D \leq N$, $\mathcal{M}_D$ be the collection of all subset of cardinality D of Λ. We denote by $\mathcal{B}(N, D, b)$ with $b > 0$ the subset of $\mathbb{H}$ containing all the points s of the form $s = \sum_{\lambda\in m} \beta_\lambda\varphi_\lambda$ where m is any element of $\mathcal{M}_D$ and $|\beta_\lambda| \leq b\varepsilon$ for all $\lambda \in m$. Then the minimax risk over $\mathcal{B}(N, D, b)$ satisfies*

$$R_M(\mathcal{B}(N, D, b), \varepsilon) \geq \frac{\varepsilon^2 D}{216}\left[\left(18b^2\right) \wedge 5\log\left(\frac{N}{D} \vee 650\right)\right].$$

6. Infinite-dimensional variable selection

6.1. From function spaces to sequence spaces

It is now part of the statistical folklore that, for a suitable choice of an orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ ($\Lambda = \mathbb{N}^\star$) of some Hilbert space $\mathbb{H}$ of functions, properties of the elements of $\mathbb{H}$ can be translated into properties of their coefficients in the space $\boldsymbol{l}_2(\Lambda)$. One should look at Meyer (1990) for the basic ideas and Donoho and Johnstone (1998, Sect. 2) for a review. Many classical functional classes in some $\mathbb{L}_2$ space $\mathbb{H}$ can therefore be turned to specific geometric objects in $\boldsymbol{l}_2(\Lambda)$ via the natural isometry between $\mathbb{H}$ and $\boldsymbol{l}_2(\Lambda)$ given by $s \leftrightarrow (\beta_\lambda)_{\lambda\in\Lambda}$ if $s = \sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda$. In particular Sobolev balls could be interpreted in terms of ellipsoids (with respect to the trigonometric basis) and balls in Besov spaces can be turned to special types of $\boldsymbol{l}_p$-bodies when expanded on suitable wavelet bases (see Sect. 8.1 below). As shown in Sect. 2.2, once we have chosen a suitable basis, a Gaussian linear process can be turned to an associated Gaussian sequence of the form

$$\hat{\beta}_\lambda = \beta_\lambda + \varepsilon\xi_\lambda, \quad \lambda \in \Lambda, \tag{6.1}$$

for some sequence of i.i.d. standard normal variables ξ_λ. We shall therefore concentrate here on the search for good strategies for estimating $s = (\beta_\lambda)_{\lambda\in\Lambda} \in \boldsymbol{l}_2(\Lambda)$ from the sequence $(\hat{\beta}_\lambda)_{\lambda\in\Lambda}$ under the assumption that it belongs to various types of $\boldsymbol{l}_p$-bodies.

The study of minimax and adaptive estimation in the Gaussian sequence framework has been mainly developed by Pinsker (1980) and Efroimovich and Pinsker (1984) for ellipsoids and by Donoho and Johnstone (1994a, b, 1995, 1996 and 1998) for $\boldsymbol{l}_p$-bodies. Let us now recall the corresponding definitions.

Definition 9. *Let p be some positive real number and $a = (a_\lambda)_{\lambda\in\Lambda}$ ($\Lambda = \mathbb{N}^\star$) be a nonincreasing sequence of numbers in $[0, +\infty]$, converging to 0 when $\lambda \to +\infty$ and such that*

$$\sum_{\lambda\in\Lambda} a_\lambda^{2p/(p-2)} < +\infty \quad \text{if } p > 2. \tag{6.2}$$

The $\boldsymbol{l}_p$-body $\mathcal{E}(p, a)$ is the subset of $\mathbb{R}^\Lambda$ given by

$$\mathcal{E}(p, a) = \left\{ s = (\beta_\lambda)_{\lambda\in\Lambda} \;\middle|\; \sum_{\lambda\in\Lambda} \left|\frac{\beta_\lambda}{a_\lambda}\right|^p \le 1 \right\}, \tag{6.3}$$

with the convention that $0/0 = 0$ and $x/(+\infty) = 0$ whatever $x \in \mathbb{R}$. An $\boldsymbol{l}_2$-body is called an ellipsoid.

It is important here to notice that the ordering, induced by Λ, that we have chosen on $\{\varphi_\lambda\}_{\lambda\in\Lambda}$, plays an important role since $\boldsymbol{l}_p$-bodies are not invariant under permutations of Λ.

If follows from classical inequalities between the norms in $\boldsymbol{l}_2(\Lambda)$ and $\boldsymbol{l}_p(\Lambda)$ that $\mathcal{E}(p, a) \subset \boldsymbol{l}_2(\Lambda)$ when $p \le 2$. If $p > 2$, (6.2) warrants that $\mathcal{E}(p, a) \subset \boldsymbol{l}_2(\Lambda)$. More precisely, it follows from Hölder's Inequality that if $s \in \mathcal{E}(p, a)$,

$$\begin{aligned}\sum_{\lambda>N} \beta_\lambda^2 = \sum_{\lambda>N} \left(\frac{\beta_\lambda^2}{a_\lambda^2}\right) a_\lambda^2 \;&\le\; \left(\sum_{\lambda>N} \left|\frac{\beta_\lambda}{a_\lambda}\right|^p\right)^{2/p} \left(\sum_{\lambda>N} a_\lambda^{2p/(p-2)}\right)^{1-2/p} \\ &\le \left(\sum_{\lambda>N} a_\lambda^{2p/(p-2)}\right)^{1-2/p}.\end{aligned} \tag{6.4}$$

The results developed in the next sections essentially parallel and complement those obtained by Donoho and Johnstone in a series of papers devoted to asymptotic evaluation of the minimax risk for various $\boldsymbol{l}_p$-bodies and adaptation (see Donoho and Johnstone 1994a, b, 1995, 1996 and 1998). Their approach, based on thresholding methods, is essentially asymptotic while ours is not. The asymptotic viewpoint allows them to get precise asymptotic values while we have to content ourselves with rougher evaluations, up to more or less precise multiplicative constants. As a counterpart, we are able to deal with more general situations that their assumptions exclude (see the case of $\boldsymbol{l}_p$-balls below or the case $\alpha = 1/p - 1/2$ for Besov bodies). Moreover, we shall see in Sect. 6.3.4 below that the search for exact asymptotic minimaxity may lead to serious drawbacks.

The methods we use are also different. While hard thresholding, which is at the heart of their results, is a particular case of penalization, penalization includes many other strategies which will allow us to derive adaptation results over much larger classes. As far as we know, such results could not be derived using standard thresholding methods. In any case, it will immediately be clear to any reader who is familiar with the works of Donoho and Johnstone that our point of view has been strongly influenced by theirs.

In our treatment of the Gaussian sequence model associated with the basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$, we shall stick to the following notations: the family $\{\Lambda_m\}_{m\in\mathcal{M}}$ is a countable family of finite subsets of $\Lambda = \mathbb{N}^\star$ and for each $m \in \mathcal{M}$, S_m is the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$; if $\Lambda_m = \emptyset$, then $S_m = \{0\}$. Selecting a value of m amounts to select a set Λ_m or equivalently some finite subset of the coordinates. Our purpose will be to define proper collections $\{\Lambda_m\}_{m\in\mathcal{M}}$ of subsets of Λ together with weights L_m satisfying (3.3) for which the accuracy index $a_I(s, \mathcal{S}, \varepsilon)$ defined by (3.13) can be bounded when s belongs to some typical $\boldsymbol{l}_p$-bodies. Such computations, which require the evaluation of $d^2(s, S_m)$ involve the approximation properties of the models S_m in the collection.

6.2. *Adaptation with respect to $\boldsymbol{l}_p$-bodies for $p \geq 2$*

We first introduce a strategy which is suitable when s belongs to some unknown ellipsoid (case $p = 2$) or more generally some unknown $\boldsymbol{l}_p$-body with $p \geq 2$. This strategy is given by $\mathcal{M} = \mathbb{N}$, $\Lambda_0 = \emptyset$, $\Lambda_m = \{1, 2, \ldots, m\}$ for $m > 0$ and $(L_m)_{m\geq 1}$ is any bounded nonnegative sequence satisfying (3.3) (such as $L_m = L > 0$ for all m or $L_m = \theta^2 m^{-1/2}$). If $L = \sup_{m\geq 1}\{L_m\}$, one immediately derives that, whatever $s \in \boldsymbol{l}_2(\Lambda)$,

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m\in\mathbb{N}} \left\{ \sum_{\lambda>m} \beta_\lambda^2 + \varepsilon^2 m(1 + L) \right\} + \varepsilon^2\Sigma.$$

Since $\sum_{\lambda>m} \beta_\lambda^2$ converges to zero when m goes to infinity, it follows that $a_I(s, \mathcal{S}, \varepsilon)$ goes to zero with ε and our strategy leads to consistent estimators for all $s \in \boldsymbol{l}_2(\Lambda)$.

Let us now assume that s belongs to some $\boldsymbol{l}_p$-body $\mathcal{E}(p, a) \subset \boldsymbol{l}_2(\Lambda)$ such that (6.2) holds if $p > 2$. If we define

$$a'_\lambda = a_\lambda \quad \text{if } p = 2 \quad \text{and} \quad a'_\lambda = \left(\sum_{j\geq\lambda} a_j^{2p/(p-2)} \right)^{1/2-1/p} \geq a_\lambda \quad \text{if } p > 2, \tag{6.5}$$

we deduce from the monotonicity of the sequence a when $p = 2$ and from (6.4) otherwise that $d^2(s, S_m) \leq a'^2_{m+1}$. Consequently we get for all $s \in \mathcal{E}(p, a)$, $p \geq 2$

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m\in\mathbb{N}} \left\{ a'^2_{m+1} + \varepsilon^2 m(1 + L) \right\} + \varepsilon^2\Sigma \quad \text{with } L = \sup_{m\geq 1}\{L_m\}. \tag{6.6}$$

The following proposition, the proof of which is essentially based on the results by Donoho et al. (1990), implies that, given $0 < \eta \leq 1$, our penalized projection estimator $\tilde{s}$ is, up to a constant depending only on η, K, L and Σ, simultaneously minimax among all possible $\boldsymbol{l}_p$-bodies $\mathcal{E}(p, a)$ which satisfy $p \geq 2$ and $a_1 \geq \eta\varepsilon$ (and therefore $a'^{-1}_1\varepsilon \leq \eta^{-1}$).

Proposition 3. *Let $R_M(\mathcal{E}(p, a), \varepsilon)$ be the minimax risk over the $\boldsymbol{l}_p$-body $\mathcal{E}(p, a)$ with $p \geq 2$ and $a_1 > 0$, $\mathcal{S}$ be the above strategy and* pen(m) *be given by (3.11).*

Then the resulting penalized projection estimator $\tilde{s}$ satisfies

$$\sup_{s\in\mathcal{E}(p,a)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \le C'(K)\left[(1+L)\vee\Sigma\right]\left[1\vee\left(a_1'^{-1}\varepsilon\right)\right]^2 R_M(\mathcal{E}(p,a),\varepsilon)$$

for some positive constant $C'(K)$.

Proof. On the one hand, it follows from (3.14) and (6.6) that

$$\sup_{s\in\mathcal{E}(p,a)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \le C(K)\left[(1+L)\vee\Sigma\right]\inf_{m\in\mathbb{N}}\left\{a_{m+1}'^2+\varepsilon^2(m+1)\right\}. \quad (6.7)$$

On the other hand, since $p \ge 2$, $\mathcal{E}(p,a)$ is orthosymmetric, compact, convex and quadratically convex, according to the terminology of Donoho et al. (1990). Now let A be an arbitrary finite subset of Λ and $\hat{s}_A$ the corresponding projection estimator $\hat{s}_A = \sum_{\lambda\in A}\hat{\beta}_\lambda\varphi_\lambda$. The maximal risk of such an estimator then derives from (2.7):

$$\sup_{s\in\mathcal{E}(p,a)} \mathbb{E}_s\left[\|s-\hat{s}_A\|^2\right] = |A|\varepsilon^2 + \sup_{s\in\mathcal{E}(p,a)}\sum_{\lambda\in\Lambda\setminus A}\beta_\lambda^2.$$

It follows from Donoho et al. (1990, Corollary p. 1428) that the minimax risk over $\mathcal{E}(p,a)$ satisfies

$$\begin{aligned} 4.44R_M(\mathcal{E}(p,a),\varepsilon) &\ge \inf_{A\subset\Lambda}\sup_{s\in\mathcal{E}(p,a)} \mathbb{E}_s\left[\|s-\hat{s}_A\|^2\right] \\ &= \inf_{A\subset\Lambda}\left\{|A|\varepsilon^2+\sup_{s\in\mathcal{E}(p,a)}\sum_{\lambda\in\Lambda\setminus A}\beta_\lambda^2\right\}. \end{aligned}$$

If $|A|=m$, it follows from the monotonicity of the sequence (a_λ) that

$$\sup_{s\in\mathcal{E}(p,a)}\sum_{\lambda\in\Lambda\setminus A}\beta_\lambda^2 \ge \sup_{s\in\mathcal{E}(p,a)}\sum_{\lambda>m}\beta_\lambda^2 = a_{m+1}'^2,$$

since the bound (6.4) is sharp in the sense that there exists an $s\in\mathcal{E}(p,a)$ for which the equality holds. We can therefore derive that

$$4.44R_M(\mathcal{E}(p,a),\varepsilon) \ge \inf_{m\in\mathbb{N}}\left\{m\varepsilon^2+a_{m+1}'^2\right\}. \quad (6.8)$$

The conclusion follows from a comparison between (6.7) and (6.8). □

6.3. Estimation in arbitrary $\mathbf{l}_p$-bodies

6.3.1. Presentation of a new strategy and the corresponding estimator

We now only want to assume that $s\in\mathcal{E}(p,a)$ for some unknown values of the sequence $a=(a_\lambda)_{\lambda\in\Lambda}$ and the positive parameter p.

The strategy We choose for $\mathcal{M}$ the collection of all finite subsets m of Λ and set $\Lambda_m = m$ and $N_m = \sup m$; then, if $m \neq \emptyset$, $1 \leq D_m = |m| \leq N_m$. Finally, in order to define the weights, fix some $\theta > 0$ and set for all $m \neq \emptyset$, $L_m = L(D_m, N_m)$ with

$$L(D, N) = \log\left(\frac{N}{D}\right) + (1+\theta)\left(1 + \frac{\log N}{D}\right).$$

Let us now check that (3.3) is satisfied with Σ bounded by some Σ_θ depending only on θ. We first observe that $\mathcal{M} \setminus \emptyset$ is the disjoint union of all the sets $\mathcal{M}(D, N)$, $1 \leq D \leq N$, where

$$\mathcal{M}(D, N) = \{m \in \mathcal{M} \mid D_m = D \text{ and } N_m = N\}, \tag{6.9}$$

and that

$$|\mathcal{M}(D, N)| = \binom{N-1}{D-1} \leq \binom{N}{D} \leq \left(\frac{eN}{D}\right)^D,$$

from which we derive that

$$\begin{aligned}
\Sigma &\leq \sum_{N\geq 1} \sum_{D=1}^{N} |\mathcal{M}(D, N)| \exp[-D\log(N/D) - (1+\theta)(D + \log N)] \\
&\leq \sum_{N\geq 1} \sum_{D\geq 1} \exp[-\theta D] N^{-\theta-1} \quad \leq \quad \frac{e^{-\theta}}{1-e^{-\theta}} \int_{1/2}^{+\infty} x^{-\theta-1}\, dx \\
&\leq \frac{e^{-\theta}}{1-e^{-\theta}} \frac{2^\theta}{\theta} \quad = \quad \Sigma_\theta.
\end{aligned} \tag{6.10}$$

Computation of the estimator If one chooses $\mathrm{pen}(m)$ as in (3.11), it is a function of D_m and N_m and can therefore be written as $\mathrm{pen}'(D_m, N_m)$. In order to compute the penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ one has to find the minimizer $\hat{m}$ of

$$\mathrm{crit}(m) = -\|\hat{s}_m\|^2 + \mathrm{pen}(m) = -\sum_{\lambda\in m} \hat{\beta}_\lambda^2 + \mathrm{pen}'(D_m, N_m).$$

Given N and D, the minimization of $\mathrm{crit}(m)$ over the set $\mathcal{M}(D, N)$ amounts to the maximization of $\sum_{\lambda\in m} \hat{\beta}_\lambda^2$ over this set. Since by definition all such m's contain N and $D-1$ elements of the set $\{1, 2, \ldots, N-1\}$, it follows that the minimizer $m(D, N)$ of $\mathrm{crit}(m)$ over $\mathcal{M}(D, N)$ is the set containing N and the indices of the $D-1$ largest elements $\hat{\beta}_\lambda^2$ for $1 \leq \lambda \leq N-1$ or more formally, using Definition 8,

$$\inf_{m\in\mathcal{M}(D,N)} \mathrm{crit}(m) = \mathrm{crit}(m(D, N)) = -\sum_{\lambda\in m(D,N)} \hat{\beta}_\lambda^2 + \mathrm{pen}'(D, N),$$

with $m(D, N) = \{N\} \cup \overline{\{1, 2, \ldots, N-1\}}[D-1]$. The computation of $\hat{m}$ then results from an optimization with respect to N and D. In order to perform this optimization, let us observe that if $J = \max \overline{\{1, 2, \ldots, N\}}[D] < N$, then

$\sum_{\lambda\in m(D,N)} \hat{\beta}_\lambda^2 \leq \sum_{\lambda\in m(D,J)} \hat{\beta}_\lambda^2$. On the other hand, it follows from the definition of $L(D,\cdot)$ that $L(D,J) < L(D,N)$ and therefore $\mathrm{crit}(m(D,N)) > \mathrm{crit}(m(D,J))$. This implies that, given D, the optimization with respect to N should be restricted to those N's such that $\max \overline{\{1,2,\ldots,N\}}[D] = N$. It can easily be deduced from an iterative computation of the sets $\{\hat{\beta}_\lambda^2\}_{\lambda\in\overline{\{1,2,\ldots,N\}}[D]}$ starting with $N = D$. It then remains to optimize our criterion with respect to D.

6.3.2. Bounding the accuracy index

In this section we want to prove various upper bounds for the accuracy index which will prove useful in the sequel. As usual, $C(\theta)$ will denote some constant depending only on θ, but which may vary from line to line. First applying (3.13) with $m = \emptyset$ and therefore $D_m = 0$, we get

$$a_I(s,\mathcal{S},\varepsilon) \leq C(\theta)\left(\|s\|^2 + \varepsilon^2\right), \tag{6.11}$$

which is a useful bound when $\varepsilon^{-1}\|s\|$ is not large.

In order to deal with the other cases, we observe that $\log N/D < \log(N/D) + 0.37$ for any pair of positive integers $D \leq N$, which implies that

$$L(D,N) \leq (2+\theta)\log\left(\frac{N}{D}\right) + 1.37(1+\theta). \tag{6.12}$$

If we restrict to those ms such that $N_m = D_m$, then $L_m \leq 1.37(1+\theta)$. Moreover,

$$\sum_{\lambda>N} \beta_\lambda^2 \leq \left(\sum_{\lambda>N} |\beta_\lambda|^p\right)^{2/p} \leq a_{N+1}^2 \quad \text{for } 0 < p \leq 2, \tag{6.13}$$

which leads to the following analogue of (6.6),

$$a_I(s,\mathcal{S},\varepsilon) \leq C(\theta) \inf_{N\geq 1}\left\{a_{N+1}^2 + \varepsilon^2 N\right\}. \tag{6.14}$$

By the arguments used in the preceding section, this bound remains valid when $p > 2$ provided that a_{N+1} is replaced by a'_{N+1} as defined by (6.5). This means that an analogue of Proposition 3 still holds for the new estimator $\tilde{s}$, namely

$$\sup_{s\in\mathcal{E}(p,a)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(K,\theta)\left[1 \vee \left(a_1'^{-1}\varepsilon\right)\right]^2 R_M(\mathcal{E}(p,a),\varepsilon) \quad \text{for } p \geq 2,$$

and $\tilde{s}$ is minimax, up to some constant $C(K,\theta,\eta)$, over all $\boldsymbol{l}_p$-bodies such that $p \geq 2$ and $a_1 \geq \eta\varepsilon > 0$.

Let us now turn to a more general bound which will allow us to deal with $\boldsymbol{l}_p$-bodies for $p < 2$. This bound is based on the following:

Lemma 2. *Given N nonnegative numbers $\{b_i\}_{i\in I}$ such that $\sum_{i\in I} b_i^p \le R^p$ with $0 < p \le 2$, an integer Q satisfying $0 \le Q \le N-1$ and the set $\overline{I}[Q]$ given by Definition 8, one gets, recalling that the numbers $b_{(1)(I)} \ge \ldots \ge b_{(N)(I)}$ represent a permutation of the set $\{b_i\}_{i\in I}$,*

$$\sum_{i\notin \overline{I}[Q]} b_i^2 = \sum_{j=Q+1}^{N} b_{(j)(I)}^2 \le R^2 (Q+1)^{1-2/p}.$$

Proof. The result being clearly true when $Q = 0$, we can assume that $Q \ge 1$. Let $b = b_{(Q+1)(I)}$. Then $b \le b_{(j)(I)}$ whatever $j \le Q$ and therefore $(1+Q)b^p \le R^p$. We then conclude from

$$\sum_{j=Q+1}^{N} b_{(j)(I)}^2 \le b^{2-p} \sum_{j=Q+1}^{N} b_{(j)(I)}^p \le \left(\frac{R^p}{1+Q}\right)^{2/p-1} R^p. \qquad \square$$

We can now derive the following upper bound for the accuracy index:

Proposition 4. *Let s belong to some $\mathbf{l}_p$-body $\mathcal{E}(p,a)$ with $p < 2$ and $\mathcal{S}$ be the strategy defined in Sect. 6.3.1. Then*

$$a_I(s,\mathcal{S},\varepsilon) \le C(\theta) \inf_{\{(D,N)\,|\,1\le D\le N\}} \left\{ a_{N+1}^2 + a_D^2 D^{1-2/p} + \varepsilon^2 D \left[\log\left(\frac{N}{D}\right) + 1 \right] \right\}. \tag{6.15}$$

Proof. Setting $\bar{\mathcal{M}}(J,M) = \cup_{J\le N\le M} \mathcal{M}(J,N)$ where $\mathcal{M}(J,N)$ is defined by (6.9), we derive from (6.12), (6.10) and (3.13) that

$$\begin{aligned} &a_I(s,\mathcal{S},\varepsilon) \\ &\quad \le \inf_{\{(J,N)\,|\,1\le J\le N\}} \left\{ \left(\inf_{m\in\mathcal{M}(J,N)} d^2(s,S_m) \right) + \varepsilon^2 J[L(J,N)+1] \right\} + \Sigma_\theta \varepsilon^2 \\ &\quad \le C(\theta) \inf_{\{(J,M)\,|\,1\le J<M\}} \left\{ \left(\inf_{m\in\bar{\mathcal{M}}(J,M)} d^2(s,S_m) \right) + \varepsilon^2 J \left[\log\left(\frac{M}{J}\right) + 1 \right] \right\}. \end{aligned}$$

Let us fix some pair (J,M) and some $m \in \bar{\mathcal{M}}(J,M)$. It follows from (6.13) that

$$d^2(s,S_m) = \sum_{\lambda>M} \beta_\lambda^2 + \sum_{\substack{1\le\lambda\le M \\ \lambda\notin m}} \beta_\lambda^2 \le a_{M+1}^2 + \sum_{\substack{1\le\lambda\le M \\ \lambda\notin m}} \beta_\lambda^2,$$

and therefore, setting $I_M = \{1,2,\ldots,M\}$,

$$a_I(s,\mathcal{S},\varepsilon) \le C(\theta) \inf_{J\ge 1} \inf_{M>J} F(J,M), \tag{6.16}$$

with

$$\begin{aligned} F(J,M) &= \inf_{m\in\bar{\mathcal{M}}(J,M)} d^2(s,S_m) + \varepsilon^2 J \left[\log\left(\frac{M}{J}\right) + 1 \right] \\ &\le a_{M+1}^2 + \sum_{\lambda=J+1}^{M} \beta_{(\lambda)(I_M)}^2 + \varepsilon^2 J \left[\log\left(\frac{M}{J}\right) + 1 \right]. \end{aligned}$$

Let us now observe that if $1 \le D \le J+1 \le M$,

$$\sum_{\lambda=D}^{M} \beta_{(\lambda)(I_M)}^p \le \sum_{j=D}^{M} |\beta_j|^p \le a_D^p.$$

It then follows from Lemma 2 with $N = M - D + 1$, $R = a_D$ and $Q = J - D + 1$ that

$$\sum_{\lambda=J+1}^{M} \beta_{(\lambda)(I_M)}^2 \le a_D^2 (J - D + 2)^{1-2/p}.$$

Let us now define

$$\lceil x \rceil = \inf\{n \in \mathbb{N} \mid n \ge x\}, \tag{6.17}$$

and fix $D = \lceil (J+1)/2 \rceil$ and $N = \lceil (M-1)/2 \rceil$. Then $J - D + 2 \ge D$ and $J/2 < D \le J$, which implies that

$$F(M, J) \le a_{N+1}^2 + a_D^2 D^{1-2/p} + 2\varepsilon^2 D \left[\log\left(\frac{2N+1}{D} \right) + 1 \right].$$

Finally, since $N \ge D$ implies that $M > J$ and $\log(2N+1)+1 \le (1+\log 3)(\log N + 1)$, (6.15) follows from (6.16). □

6.3.3. Adaptation over l_p-balls

Following the terminology of Donoho and Johnstone (1994b) we define the l_p-ball $\mathcal{L}(p, N, R) \subset l_2(\Lambda)$ of dimension N and radius R, with $N \in \mathbb{N}^\star$, $R > 0$ and $0 < p \le 2$ as

$$\mathcal{L}(p, N, R) = \left\{ s = (\beta_\lambda)_{\lambda \in \Lambda} \;\middle|\; \sum_{\lambda=1}^{N} |\beta_\lambda|^p \le R^p \quad \text{and} \quad \beta_\lambda = 0 \quad \text{for } \lambda > N \right\}. \tag{6.18}$$

The performances of our strategy when s belongs to some unknown l_p-ball $\mathcal{L}(p, N, R)$ of dimension $N \ge 2$ (in order to avoid trivialities) are described by the following proposition.

Proposition 5. *Let s belong to some l_p-ball $\mathcal{L}(p, N, R)$ with $N \ge 2$ and $0 < p < 2$. Let $\rho_p > 1.76$ be the unique solution of the equation $\rho_p \log \rho_p = 2/p$. The accuracy index of our strategy can then be bounded by*

$$a_I(s, \mathcal{S}, \varepsilon) \le C(\theta) R^p \varepsilon^{2-p} \left[1 + \log\left(\frac{N\varepsilon^p}{R^p} \right) \right]^{1-p/2}, \tag{6.19}$$

when

$$\sqrt{\log N} \le R/\varepsilon \le \rho_p^{-1/2} N^{1/p}; \tag{6.20}$$

by

$$a_I(s, \mathcal{S}, \varepsilon) \le C(\theta)\left(R^2 + \varepsilon^2\right) \quad \textit{when } R < \varepsilon\sqrt{\log N}; \tag{6.21}$$

and by

$$a_I(s, \mathcal{S}, \varepsilon) \le C(\theta)\,\varepsilon^2 N \quad \textit{when } R > \varepsilon\rho_p^{-1/2}N^{1/p}. \tag{6.22}$$

Moreover the minimax risk over $\mathcal{L}(p, N, R)$ is bounded from below by

$$R_M(\mathcal{L}(p, N, R), \varepsilon) \ge \kappa_1 R^p \varepsilon^{2-p}\left[1 + \log\left(\frac{N\varepsilon^p}{R^p}\right)\right]^{1-p/2} \quad \textit{when (6.20) holds};$$

$$R_M(\mathcal{L}(p, N, R), \varepsilon) \ge \kappa_1 R^2 \quad \textit{when } R < \varepsilon\sqrt{\log N};$$

and

$$R_M(\mathcal{L}(p, N, R), \varepsilon) \ge \kappa_1 \rho_p^{-1} N\varepsilon^2 \quad \textit{when } R > \varepsilon\rho_p^{-1/2}N^{1/p},$$

where κ_1 denotes some universal constant.

Proof. Since the $\boldsymbol{l}_p$-ball $\mathcal{L}(p, N, R)$ is a particular case of an $\boldsymbol{l}_p$-body $\mathcal{E}(p, a)$ with $a_\lambda = R$ for $1 \le \lambda \le N$ and $a_\lambda = 0$ for $\lambda > N$, it follows from (6.15) that

$$a_I(s, \mathcal{S}, \varepsilon) \le C(\theta) \inf_{1\le D\le N}\left\{R^2 D^{1-2/p} + \varepsilon^2 D[\log(N/D) + 1]\right\}. \tag{6.23}$$

In order to minimize the right-hand side of (6.23), one should choose some D which approximately equates the two terms $R^2D^{1-2/p}$ and $\varepsilon^2 D[\log(N/D) + 1]$. This leads to the choice

$$D = \left\lceil \left(\frac{R}{\varepsilon}\right)^p \left[\log\left(\frac{N\varepsilon^p}{R^p}\right)\right]^{-p/2} \right\rceil \quad \text{with } \lceil x\rceil \text{ given by (6.17)}, \tag{6.24}$$

provided that D satisfies to $1 \le D \le N$. Since (6.24) defines a nondecreasing function of R/ε, this condition is satisfied when (6.20) holds and we then derive (6.19) from (6.23). Otherwise (6.21) follows from (6.11) since $p < 2$ and (6.22) from (6.14).

The proof of the lower bounds for the minimax risk is based on Theorem 5. If we choose $b = R\varepsilon^{-1}D^{-1/p}$, the set $\mathcal{B}(N, D, b)$ defined in this theorem is contained in $\mathcal{L}(p, N, R)$ and since D is arbitrary between 1 and N, we derive from Theorem 5 that

$$R_M(\mathcal{L}(p, N, R), \varepsilon) \ge \kappa \sup_{1\le D\le N}\left[\left(R^2D^{1-2/p}\right) \wedge \left(D\varepsilon^2[1 + \log(N/D)]\right)\right].$$

Since the minimum is obtained, as before, by approximately equating $R^2D^{1-2/p}$ and $\varepsilon^2 D[\log(N/D) + 1]$, the same computations lead to the lower bounds results. □

6.3.4. A paradox about sharp asymptotic minimaxity

It is interesting to compare these results with the evaluations given by Theorem 3 and Corollary 4 of Donoho and Johnstone (1994b). They only consider the case $R = 1$ but a proper rescaling of the observations easily reduces the general situation to this one. Since they keep the radius of the balls fixed, their asymptotics let ε go to zero but it is equivalent to keep ε fixed and let R go to infinity. From this alternative point of view their result can be restated as follows.

Theorem 6 (Donoho and Johnstone). *Let $\mathcal{L}(p, N, R_N)$, $N \geq 1$ be a sequence of $\boldsymbol{l}_p$-balls with $0 < p < 2$ such that, when N goes to infinity,*

$$R_N \longrightarrow \infty; \qquad NR_N^{-p} \longrightarrow \infty \qquad \text{and} \qquad R_N^{-2} \log\left(NR_N^{-p}\right) \longrightarrow 0. \quad (6.25)$$

Then

$$R_M(\mathcal{L}(p, N, R_N), \varepsilon) = R_N^p \left[2\varepsilon^2 \log\left(N\varepsilon^p / R_N^p\right)\right]^{1-p/2} [1 + o(1)] \qquad (6.26)$$

and

$$\sup_{s \in \mathcal{L}(p,N,R_N)} \mathbb{E}_s \left[\|s - \tilde{s}_{T_N}\|^2\right] = R_M(\mathcal{L}(p, N, R_N), \varepsilon)[1 + o(1)],$$

where $\tilde{s}_{T_N}$ denotes the threshold estimator defined by (5.8) with threshold

$$T_N = \varepsilon \left[2 \log\left(N\varepsilon^p / R_N^p\right) + \alpha \log\left[2 \log\left(N\varepsilon^p / R_N^p\right)\right]\right]^{1/2}, \quad \alpha > p - 1. \quad (6.27)$$

This is an extremely precise result on the one hand and also a truly asymptotic one on the other hand in the sense that it definitely rules out the situations described by (6.21) and (6.22) and is even more restrictive than (6.20). Under the assumptions of the theorem, (6.19) holds and produces, together with (3.14), a nonasymptotic analogue of (6.26), namely, fixing K and θ, the bound

$$R_M(\mathcal{L}(p, N, R_N), \varepsilon) \leq K' R_N^p \varepsilon^{2-p} \left[1 + \log\left(N\varepsilon^p / R_N^p\right)\right]^{1-p/2} \quad \text{for all } N \geq 1.$$

We actually get from Proposition 5 a complete nonasymptotic counterpart to the results of Donoho and Johnstone in the form of the following

Corollary 1. *The estimator $\tilde{s}$ derived from the strategy described in Sect. 6.3.1, with a penalty given by (3.11) satisfies*

$$\sup_{s \in \mathcal{L}(p,N,R)} \mathbb{E}_s \left[\|\tilde{s} - s\|^2\right] \leq C(K, \theta) \left[1 \vee \frac{\varepsilon}{R} \vee \rho_p \mathbb{1}_{(0,R)}\left(\rho_p^{-1/2} \varepsilon N^{1/p}\right)\right] R_M(\mathcal{L}(p, N, R), \varepsilon).$$

Since the bracketted factor, which controls the ratio between the risk of our estimator and the minimax risk remains bounded unless R is too close to zero which implies that $\varepsilon^{-2}R_M$ is small, as we already noticed, or when R is large and p close to zero since $\rho_p \to +\infty$ when $p \to 0$, it follows that $\tilde{s}$ is uniformly minimax for almost all $\boldsymbol{l}_p$-balls. One should also notice that the construction of $\tilde{s}_{T_N}$ requires the knowledge of R_N and p, which is not the case for $\tilde{s}$.

There is another rather surprising phenomenon that occurs concerning the sharp asymptotically minimax threshold estimators $\tilde{s}_{T_N}$ of Theorem 6. On the one hand, it immediately follows that, if $s_N \in \mathcal{L}(p, N, R_N)$ and s_N has no more that D_N nonzero coordinates, the risk of $\tilde{s}$ is, by (3.13) and (6.12) bounded by $C(K, \theta)\varepsilon^2 D_N[1 + \log(N/D_N)]$, independently of R_N. On the other hand, an application of Proposition 2 with T_N given by (6.27) shows that the risk of $\tilde{s}_{T_N}$ at s_N is bounded from below by

$$\mathbb{E}_{s_N}\left[\|\tilde{s}_{T_N} - s_N\|^2\right] \geq C'(1 - D_N/N)R_N^p \varepsilon^{2-p}\left[\log\left(N\varepsilon^p/R_N^p\right)\right]^{(1-\alpha)/2}.$$

Choosing $R_N^p = N^\delta$ for some $\delta \in (0, 1)$, which is compatible with (6.25), and $D_N/N \leq c < 1$, we derive from (6.26) that

$$R_M(\mathcal{L}(p, N, R_N), \varepsilon) = N^\delta \varepsilon^{2-p}[2(1-\delta)\log N]^{1-p/2}[1 + o(1)],$$

while

$$\mathbb{E}_{s_N}\left[\|\tilde{s}_{T_N} - s_N\|^2\right] \geq C'' N^\delta \varepsilon^{2-p}(\log N)^{(1-\alpha)/2}[1 + o(1)].$$

Therefore, apart from some power of $\log N$ which is arbitrary small when α is close enough to $p - 1$, the risk at s_N is equal to the minimax risk while $\tilde{s}$ has a substantially better performance at s_N when $N^{-\delta}D_N$ is small.

It is indeed not necessary to build a sophisticated estimator like $\tilde{s}$ to get such an improvement over $\tilde{s}_{T_N}$. A simple change of the level of thresholding would do. Set, for instance, $U_N = \sqrt{K}\left(1 + \sqrt{2\log N}\right)\varepsilon$ with $K > 1$. If $R_N^p = N^\delta$, an application of (3.13) together with Lemma 2 shows that

$$\sup_{s \in \mathcal{L}(p,N,R_N)} \mathbb{E}_s\left[\|\tilde{s}_{U_N} - s\|^2\right] \leq C(K)N^\delta \varepsilon^{2-p}(\log N)^{1-p/2}$$

and

$$\mathbb{E}_{s_N}\left[\|\tilde{s}_{U_N} - s_N\|^2\right] \leq C'(K)D_N\varepsilon^2 \log N.$$

This means that $\tilde{s}_{U_N}$ is again minimax, up to constants, but substantially improves over $\tilde{s}_{T_N}$ when D_N is not too large. This result can be viewed as a serious advertisement against the use of the minimax point of view without extreme caution. It also shows that the search for sharp asymptotically minimax estimators may be a bad idea, leading to otherwise vastly suboptimal performances. In the particular situation at hand, the search for sharp asymptotic minimaxity forces to choose a level of thresholding which is clearly too small in many other respects. A larger one would preserve the asymptotic minimaxity, only loosing the sharp asymptotic constant, and improving the estimator otherwise.

6.3.5. The case of extended Besov bodies

In the case of general $\boldsymbol{l}_p$-bodies, we cannot, unfortunately, handle the minimization of the right-hand side of (6.15) as we did for (6.6) since it involves a_D and a_{N+1} simultaneously. We now need to be able to compare $a_D^2 D^{1-2/p}$ with a_{N+1} which requires a rather precise knowledge about the rate of decrease of a_λ as a function of λ. This is why we shall restrict ourselves to some particular $\boldsymbol{l}_p$-bodies.

Definition 10. *Given parameters M', α, p, r and R with $M' \in \mathbb{N}^\star, 0 < p < 2$, $\alpha \geq 1/p - 1/2$, $R > 0, r \in \mathbb{R}$ and $r > 0$ when $\alpha = 1/p - 1/2$, we define the* extended Besov body $\mathcal{B}(M', \alpha, p, r, R)$ *as the $\boldsymbol{l}_p$-body $\mathcal{E}(p, a)$ with coefficients*

$$a_{M'+k} = \begin{cases} Rk^{-(\alpha+1/2-1/p)}[b + \log k]^{-r} & \text{for } k \geq 1, \\ +\infty & \text{for } 1 - M' \leq k \leq 0, \end{cases} \tag{6.28}$$

and

$$b = \frac{-r}{\alpha + 1/2 - 1/p} \vee 1.$$

The restrictions on r and the definition of b are made in order to ensure that the sequence $(a_\lambda)_{\lambda \in \Lambda}$ be nonincreasing. We exclude the case $p = 2$ since then a Besov body is merely an ellipsoid and this case has already been considered. When $r = 0$ we shall speak of *classical Besov bodies* (compare with Donoho and Johnstone, 1998), since they are the geometric objects which correspond to balls in Besov spaces when $(\varphi_\lambda)_{\lambda \in \Lambda}$ is a suitable wavelet basis (see Sect. 8.2 for details). Extended Besov bodies are natural extensions which allow to handle more general objects without additional efforts.

The classical case: $\alpha > 1/p - 1/2$ As far as we know, statistical estimation in Besov bodies, up to now, has been limited to the case $\alpha > 1/p - 1/2$ which is the easiest one. In order to avoid to deal with exceptional cases, which would make the conclusions unnecessarily long and complicated, we shall assume that the ratio R/ε is not too small. We can then prove:

Proposition 6. *Let $\mathcal{S}$ be the strategy defined in Sect. 6.3.1 and assume that $0 < p < 2$, $\alpha + 1/2 - 1/p > 0$, and that R/ε is large enough, namely that*

$$R/\varepsilon \geq e \qquad \text{and} \qquad \Delta = (R/\varepsilon)^{\frac{2}{2\alpha+1}} [\log(R/\varepsilon)]^{\frac{-(2r+1)}{2\alpha+1}} \geq 2M'. \tag{6.29}$$

Then,

$$\sup_{s \in \mathcal{B}(M', \alpha, p, r, R)} a_I(s, \mathcal{S}, \varepsilon) \leq CR^2 (\varepsilon/R)^{\frac{4\alpha}{2\alpha+1}} [\log(R/\varepsilon)]^{\frac{2(\alpha-r)}{2\alpha+1}}, \tag{6.30}$$

and

$$R_M(\mathcal{B}(M', \alpha, p, r, R), \varepsilon) \geq C' R^2 (\varepsilon/R)^{\frac{4\alpha}{2\alpha+1}} [\log(R/\varepsilon)]^{\frac{-2r}{2\alpha+1}}, \tag{6.31}$$

where the constants C, C' depend on M', α, r, p (and C also depends on θ).

Proof. From (6.15) and (6.28) we derive, restricting ourselves to $N \geq D \geq 2M'$, the following bound for the accuracy index

$$a_I(s, \mathcal{S}, \varepsilon) \leq C \inf_{\{(D,N)\,|\,2M' \leq D \leq N\}} \Big\{ N^{-2(\alpha+1/2-1/p)} R^2 [b + \log N]^{-2r} + D^{-2\alpha} R^2 [b + \log D]^{-2r} + \varepsilon^2 D[\log(N/D) + 1] \Big\}.$$

We then choose $D = \lceil \Delta \rceil$ and $N = \left\lceil \Delta^{\frac{\alpha}{\alpha+1/2-1/p}} \right\rceil$. Then $N \geq D \geq 2M' \geq 2$ from which we derive (6.30). On the other hand, $\mathcal{B}(M', \alpha, p, r, R)$ contains the $\boldsymbol{l}_p$-body with coefficients satisfying $a_\lambda = R' = RD^{-(\alpha+1/2-1/p)}[b + \log D]^{-r}$ for $M' + 1 \leq \lambda \leq M' + D$ with $D \geq 1$ and $a_\lambda = 0$ otherwise, which can be identified to some $\boldsymbol{l}_p$-ball $\mathcal{L}(p, D, R')$. Let us now set

$$A = R/\varepsilon \qquad \text{and} \qquad D^{\alpha+1/2} = cA(\log A)^{-r} \quad \text{with } c \geq 1,$$

where c denotes a suitably chosen constant. Then $D \geq \Delta \geq 2$ and it follows from (6.29) that $(r + 1/2)\log(\log A) < \log A$, hence

$$\log c + (1 + 2r)^{-1} \log A \leq (\alpha + 1/2) \log D \leq \log c + (1 + |r|) \log A.$$

This implies that, for $c \geq c_0(\alpha, r, p)$,

$$\frac{R'}{\varepsilon} = c^{-1} \left(\frac{\log A}{b + \log D} \right)^r D^{1/p} \leq \rho_p^{-1/2} D^{1/p}.$$

Let us choose c to be the smallest value such that D is an integer and $c \geq c_0$. Then,

$$c_1(\alpha, r, p)\varepsilon \leq R' \leq \varepsilon \rho_p^{-1/2} D^{1/p}.$$

It then follows from Proposition 5 that

$$\begin{aligned} R_M(\mathcal{B}(M', \alpha, p, r, R)) &\geq R_M(\mathcal{L}(p, D, R')) \\ &\geq \kappa_1 R'^2 \left[(\varepsilon/R')^{2-p} \wedge 1 \right] \geq c_2(\alpha, r, p) R'^p \varepsilon^{2-p} \\ &= c_2 R^2 (\varepsilon/R)^{2-p} D^{-p(\alpha+1/2-1/p)} [b + \log D]^{-pr}, \end{aligned}$$

which gives (6.31) from our lower bounds on D and R/ε. □

We can conclude that the upper bound (6.30) matches the lower bound (6.31), up to a power of $\log(R/\varepsilon)$. As we shall see below, the lower bound is actually sharp and a refined strategy, especially designed for estimation in Besov bodies, can improve the upper bound. For classical Besov bodies, $r = 0$ and the minimiax risk is known to be of the order of $R^2(\varepsilon/R)^{(4\alpha)/(2\alpha+1)}$ as indicated by our lower bound (see Donoho and Johnstone, 1998).

The borderline case: $\alpha = 1/p - 1/2$ The case $\alpha = 1/p - 1/2$ and $r > 0$ which, to our knowledge, has never been previously considered in statistics, is more delicate to handle because N cannot be taken as a power of R/ε any more but should be much larger. Still assuming that R/ε is large enough, we procede to the minimization of the right-hand side of (6.15) which amounts to minimize, since then $b = 1$,

$$[1 + \log N]^{-2r} + D^{-2\alpha}[1 + \log D]^{-2r} + (\varepsilon/R)^2 D[\log(N/D) + 1].$$

It follows from monotonicity arguments that one should approximately equate those three terms, which leads to $\log N \asymp \left(D(\varepsilon/R)^2\right)^{-1/(1+2r)}$ and finally to

$$D = \left\lceil (R/\varepsilon)^{\frac{2r}{\alpha(1+2r)+r}} [\log(R/\varepsilon)]^{\frac{-r(1+2r)}{\alpha(1+2r)+r}} \right\rceil;$$
$$\log N = \left\lceil (R/\varepsilon)^{\frac{2\alpha}{\alpha(1+2r)+r}} [\log(R/\varepsilon)]^{\frac{r}{\alpha(1+2r)+r}} \right\rceil.$$

We conclude that, if R/ε is large enough to ensure that $N \geq D \geq 2M'$,

$$\sup_{s \in \mathcal{B}(M',\alpha,p,r,R)} a_I(s, \mathcal{S}, \varepsilon) \leq CR^2(\varepsilon/R)^{\frac{4r\alpha}{\alpha(1+2r)+r}} [\log(R/\varepsilon)]^{\frac{-2r^2}{\alpha(1+2r)+r}}. \quad (6.32)$$

One can immediately notice that this rate is not at all the analogue of (6.30) when $\alpha = 1/p - 1/2$, the rate (neglecting the logarithmic terms) becoming $4r\alpha/[\alpha(1+2r) + r]$ instead of $4\alpha/[2\alpha + 1]$, which is worse. This is a situation where we are not able to get the corresponding lower bounds and therefore we have no idea about the optimality of (6.32).

6.4. A special strategy for extended Besov bodies

6.4.1. The strategy and the estimator

Let us recall from the previous section that we have at hand a strategy for model selection in the Gaussian sequence model which is, up to constants, minimax over $\boldsymbol{l}_p$-bodies for $p \geq 2$ and all $\boldsymbol{l}_p$-balls whatever p, but fails to be minimax for classical Besov bodies since its risk contains some extra $\log(R/\varepsilon)$ factors. We want here to design a new strategy, especially directed towards estimation in extended Besov bodies, which will be minimax for all extended Besov bodies with coefficients given by (6.28) when $\alpha > 1/p - 1/2$.

The strategy The construction of the models is based on a decomposition of $\Lambda = \mathbb{N}^\star$ into a partition $\Lambda = \cup_{j \geq -1} \Lambda(j)$ with $\Lambda(-1) = \{1, \ldots, M'\}$, $\mu_0 = 1$ and

$$\Lambda(j) = \{M' + \mu_j, \ldots, M' + \mu_{j+1} - 1\} \text{ with } 2^j \leq \mu_{j+1} - \mu_j \leq M2^j \text{ for } j \geq 0. \quad (6.33)$$

Typically, the basis $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ is a wavelet basis and such a partition is induced by the structure of this basis as will be recalled in Sect. 8.1, but this is definitely not necessary and decompositions based on piecewise polynomials could be considered as well (see for instance Birgé and Massart, 2000). We also have to choose a real parameter $\theta > 2$ (the choice $\theta = 3$ being quite reasonable) and set for $J, k \in \mathbb{N}$,

$$K(J,k) = \left\lfloor 2^{-k}(k+1)^{-\theta}|\Lambda(J+k)| \right\rfloor \quad \text{with } \lfloor x \rfloor = \sup\{j \in \mathbb{N} \mid j \le x\}.$$

It follows that

$$\left\lfloor M2^J(k+1)^{-\theta} \right\rfloor \ge K(J,k) > 2^J(k+1)^{-\theta} - 1, \tag{6.34}$$

which in particular implies that $K(J,k) = 0$ for k large enough (depending on J). Let us now set for $J \in \mathbb{N}$

$$\mathcal{M}_J = \left\{ m \subset \Lambda \;\middle|\; m = \left[\bigcup_{-1\le j\le J-1} \Lambda(j) \right] \bigcup \left[\bigcup_{k\ge 0} \Lambda'(J+k) \right] \right\},$$

with

$$\Lambda'(J+k) \subset \Lambda(J+k) \qquad \text{and} \qquad |\Lambda'(J+k)| = K(J,k).$$

Clearly, each $m \in \mathcal{M}_J$ is finite with cardinality $D_m = M(J)$ satisfying

$$M(J) = M' + \sum_{j=0}^{J-1} |\Lambda(j)| + \sum_{k\ge 0} K(J,k)$$

and therefore by (6.34)

$$M' + 2^J \le M(J) \le M' + M2^J \left[1 + \sum_{n\ge 1} n^{-\theta} \right]. \tag{6.35}$$

It also follows from Proposition 4 of Birgé and Massart (2000) that

$$|\mathcal{M}_J| \le \exp\left[c_\theta M 2^J \right], \tag{6.36}$$

with some constant c_θ depending only on θ. Let us now set $\mathcal{M} = \cup_{J\ge 0}\mathcal{M}_J$ and $L_m = c_\theta M + L$ with $L > 0$ for all m. Then by (6.35) and (6.36)

$$\sum_{m\in\mathcal{M}} e^{-L_m D_m} \le \sum_{J\ge 0} |\mathcal{M}_J| \exp\left[-c_\theta M 2^J - L2^J \right] \le \sum_{J\ge 0} \exp\left[-L2^J \right] = \Sigma_L,$$

and it follows that (3.3) is satisfied with $\Sigma \le \Sigma_L$.

The construction of the estimator Since $\tilde{s} = \hat{s}_{\hat{m}}$ one has to compute the minimizer $\hat{m}$ of $\mathrm{crit}(m) = \mathrm{pen}(m) - \sum_{\lambda \in m} \hat{\beta}_\lambda^2$. First observe that m always includes $\Lambda(-1)$. Therefore the $\hat{\beta}_\lambda$s with $\lambda \in \Lambda(-1)$ can be omitted in the sum. Second, the penalty function, as defined by (3.11), only depends on J when $m \in \mathcal{M}_J$ since L_m is constant and $D_m = M(J)$. Setting $\mathrm{pen}(m) = \mathrm{pen}'(J)$ when $m \in \mathcal{M}_J$, we see that $\hat{m}$ is the minimizer with respect to J of $\mathrm{pen}'(J) - \sum_{\lambda \in \hat{m}_J} \hat{\beta}_\lambda^2$ where $\hat{m}_J \in \mathcal{M}_J$ maximizes $\sum_{k \geq 0} \sum_{\lambda \in \Lambda'(J+k)} \hat{\beta}_\lambda^2$ with respect to $m \in \mathcal{M}_J$. Since the cardinality $K(J,k)$ of $\Lambda'(J+k)$ only depends of J and k, one should choose for the $\Lambda'(J+k)$ corresponding to $\hat{m}_J$ the subset of $\Lambda(J+k)$ of those $K(J,k)$ indices corresponding to the $K(J,k)$ largest values of $\hat{\beta}_\lambda^2$ for $\lambda \in \Lambda(J+k)$. In practice of course, the number of coefficients $\hat{\beta}_\lambda$ at hand, and therefore the maximal value of J is bounded. A practical implementation of this estimator is therefore feasible and has actually been completed by Misiti, Misiti, Oppenheim and Poggi (1996).

6.4.2. The performances of the estimator

We are now in a position to prove the following

Proposition 7. *Let the sequence $(a_\lambda)_{\lambda \in \Lambda}$ be given by (6.28) and let $s = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda$ be an element of $\mathbb{H}$ which satisfies either*

$$\sup_{j \geq 0} \left\{ \sum_{\lambda \in \Lambda(j)} \left| \frac{\beta_\lambda}{a_\lambda} \right|^p \right\} \leq 1 \quad \text{if } \alpha > 1/p - 1/2, \tag{6.37}$$

or

$$\sum_{j \geq 0} \sum_{\lambda \in \Lambda(j)} \left| \frac{\beta_\lambda}{a_\lambda} \right|^p \leq 1 \quad \text{if } \alpha = 1/p - 1/2. \tag{6.38}$$

Let $\mathcal{S}$ be the strategy defined in Sect. 6.4.1. Then, assuming that $R/\varepsilon \geq \delta > 1$ and that either (6.38) with $r > \theta\alpha$ or (6.37) holds, we get

$$a_I(s, \mathcal{S}, \varepsilon) \leq C R^2 (\varepsilon/R)^{\frac{4\alpha}{2\alpha+1}} [\log(R/\varepsilon)]^{\frac{-2r}{2\alpha+1}}. \tag{6.39}$$

If (6.38) holds with $0 < r \leq \theta\alpha$, we get

$$a_I(s, \mathcal{S}, \varepsilon) \leq C R^2 (\varepsilon/R)^{\frac{4r}{2r+\theta}}. \tag{6.40}$$

In both cases, the constant C depends on θ, δ and the parameters α, p, r, M and M'.

Proof. For each $j \geq 0$ we set $B_j = \left[\sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^p \right]^{1/p}$ and, following Definition 8, denote the coefficients $|\beta_\lambda|$ in decreasing order, for $\lambda \in \Lambda(J+k)$, $k \geq 0$ by $|\beta_{(j)(\Lambda(J+k))}|$. The arguments we just used to define $\hat{m}_J$ immediately show that

$$\inf_{m \in \mathcal{M}_J} d^2(s, S_m) = \sum_{k \geq 0} \sum_{j=K(J,k)+1}^{\mu_{J+k+1} - \mu_{J+k}} \beta^2_{(j)(\Lambda(J+k))},$$

and it follows from Lemma 2 and (6.34) that

$$\sum_{j=K(k,J)+1}^{\mu_{J+k+1}-\mu_{J+k}} \beta^2_{(j)(\Lambda(J+k))} \le B^2_{J+k}[K(J,k)+1]^{1-2/p}$$
$$\le B^2_{J+k}2^{J(1-2/p)}\left[2^J \wedge (k+1)^\theta\right]^{(2/p-1)},$$

from which we get

$$\inf_{m\in\mathcal{M}_J} d^2(s,S_m) \le \sum_{k\ge 0} B^2_{J+k}2^{J(1-2/p)}\left[2^J \wedge (k+1)^\theta\right]^{(2/p-1)}. \tag{6.41}$$

We also observe, using (6.28), that

$$a_{M'+\mu_{J+k}} \le R2^{-(J+k)(\alpha+1/2-1/p)}[b+(J+k)\log 2]^{-r}, \tag{6.42}$$

since by (6.33) $\mu_{J+k} \ge 2^{J+k}$. Now, under (6.37) $B_{J+k} \le \sup_{\lambda\in\Lambda(J+k)} a_\lambda = a_{M'+\mu_{J+k}}$ and it then follows from (6.41) that

$$\inf_{m\in\mathcal{M}_J} d^2(s,S_m)$$
$$\le R^2 2^{-2J\alpha}\sum_{k\ge 0} 2^{-2k(\alpha+1/2-1/p)}[b+(J+k)\log 2]^{-2r}\left[2^J \wedge (k+1)^\theta\right]^{(2/p-1)}. \tag{6.43}$$

Now, distinguishing between the cases $r \ge 0$ and $r < 0$, we observe that

$$[b+(J+k)\log 2]^{-r} \le (J+1)^{-r}\left([b(k+1)]^{-r} \vee [\log 2]^{-r}\right),$$

which implies that the series in (6.43) converges with a sum bounded by $C(J+1)^{-2r}$ where $C = C(\alpha, p, r, \theta)$. Using (6.35) and (3.13), we can then bound the accuracy index by

$$a_I(s,\mathcal{S},\varepsilon) \le C(\alpha,p,r,\theta,M,M')\inf_{J\ge 0}\left\{2^J\varepsilon^2 + R^2(J+1)^{-2r}2^{-2J\alpha}\right\}. \tag{6.44}$$

We then set $\Delta(x) = x^{\frac{2}{2\alpha+1}}(\log x)^{\frac{-2r}{2\alpha+1}}$ and $J = \inf\left\{j \ge 0 \,\middle|\, 2^j \ge \Delta(R/\varepsilon)\right\}$. Then $2^J = \rho\Delta(R/\varepsilon)$ with $1 \le \rho < 2 \vee \left[\inf_{x\ge\delta}\Delta(x)\right]^{-1}$ from which we derive (6.39).

If (6.38) holds, then $b = 1$ and we use the fact that

$$\sum_{k\ge 0}\left[\frac{B_{J+k}}{a_{M'+\mu_{J+k}}}\right]^2 \le \sum_{k\ge 0}\left[\frac{B_{J+k}}{a_{M'+\mu_{J+k}}}\right]^p \le 1$$

to derive from (6.41) and (6.42) that

$$\inf_{m\in\mathcal{M}_J} d^2(s,S_m) \le 2^{J(1-2/p)}\sum_{k\ge 0}\left[\frac{B_{J+k}}{a_{M'+\mu_{J+k}}}\right]^2\left[2^J \wedge (k+1)^\theta\right]^{(2/p-1)} a^2_{M'+\mu_{J+k}}$$
$$\le R^2 2^{J(1-2/p)}\sup_{k\ge 0}\left\{\left[2^J \wedge (k+1)^\theta\right]^{(2/p-1)}[1+(J+k)\log 2]^{-2r}\right\}$$
$$\le CR^2 2^{-2J\alpha}\sup_{k\ge 0}\left\{\left[2^J \wedge (k+1)^\theta\right]^{2\alpha}[J+k+1]^{-2r}\right\},$$

since $r > 0$ and $\alpha = 1/p - 1/2$. We conclude by noticing that

$$\sup_{k\geq 0}\left\{\left[2^J \wedge (k+1)^\theta\right]^{2\alpha}[1+J+k]^{-2r}\right\} \leq C\begin{cases}(J+1)^{-2r} & \text{if } r > \theta\alpha;\\ 2^{2J(\alpha-r/\theta)} & \text{if } 0 < r \leq \theta\alpha.\end{cases}$$

This implies that (6.44) and therefore (6.39) remain valid if $r > \theta\alpha$ and that

$$a_I(s,\mathcal{S},\varepsilon) \leq CR^2 \inf_{J\geq 0}\left\{2^J(\varepsilon/R)^2 + 2^{-2Jr/\theta}\right\}$$

when $0 < r \leq \theta\alpha$, which leads to (6.40). □

Let us now analyze what are the consequences of this result. We first notice that the set of elements s satisfying (6.37) contains the extended Besov body $\mathcal{B}(M', \alpha, p, r, R)$ which means that (6.39) holds when $s \in \mathcal{B}(M', \alpha, p, r, R)$ with $\alpha > 1/2 - 1/p$. On the other hand, the lower bound (6.31) also holds when s satisfies either (6.37) or (6.38) and (6.29) holds. This implies the following

Corollary 2. *The strategy defined in Sect. 6.4.1 is minimax, up to constants, over all extended Besov bodies $\mathcal{B}(M', \alpha, p, r, R)$ such that either $\alpha > 1/p - 1/2$ or $\alpha = 1/p - 1/2$ and $r > \theta\alpha$, provided that R/ε is large enough for (6.29) to hold.*

Let us conclude this section by two remarks. First, when $\alpha > 1/p - 1/2$, it is not more difficult to estimate on the larger set of those s which satisfy (6.37) than on the $\boldsymbol{l}_p$-body $\mathcal{B}(M', \alpha, p, r, R)$. We shall see in Sect. 8.1 below that this larger set corresponds, in term of function spaces, to some Besov balls of the form $\left\{t \mid |t|_{B^\alpha_\infty(L_p)} \leq R\right\}$. This means that our method is indeed adaptive over all the balls of this type, provided that $\alpha > 1/p - 1/2$ and R/ε is not too small. On the other hand, as opposed to the strategy developed in Sect. 6.3.1, the present one is not suitable for estimation over $\boldsymbol{l}_p$-balls. This is indeed not a problem since Theorem 3 allows us to mix both strategies and get a new one which will be simultaneously adaptive for $\boldsymbol{l}_p$-balls and Besov bodies.

7. Lower bounds for the penalty term

Our aim in this section is to show that a choice of $K < 1$ in (3.11) may lead to penalized projection estimators which behave in a quite unsatisfactory way. This means that the restriction $K > 1$ in Theorem 2 is, in some sense, necessary. It actually follows from the forthcoming results that the condition $K > 1$ in Theorem 2 is sharp and that a choice of K smaller than one should be avoided. The study of the limiting case $K = 1$ is definitely more involved and beyond the scope of this paper.

7.1. A small number of models

We first assume that, for each D, the number of elements $m \in \mathcal{M}$ such that $D_m = D$ grows at most subexponentially with respect to D. In such a case, (3.3) holds with

$L_m = L$ for all $L > 0$ and one can apply Theorem 2 with a penalty of the form $\mathrm{pen}(m) = K\varepsilon^2\big(1+\sqrt{2L}\big)^2 D_m$, where $K-1$ and L are positive but arbitrarily close to 0. This means that, whatever $K' > 1$, the penalty $\mathrm{pen}(m) = K'\varepsilon^2 D_m$ is allowed. Alternatively, the following result shows that if the penalty function satisfies $\mathrm{pen}(\bar m) = K'\varepsilon^2 D_{\bar m}$ with $K' < 1$, even for *one* single model $S_{\bar m}$, provided that the dimension of this model is large enough (depending on K'), the resulting procedure behaves quite poorly if $s = 0$.

Proposition 8. *Consider some collection of models $\{S_m\}_{m\in\mathcal{M}}$ such that*

$$\sum_{m\in\mathcal{M}} e^{-xD_m} < \infty, \quad \text{for any } x > 0. \tag{7.1}$$

For any pair of real numbers $K, \delta \in (0,1)$, there exists some integer $\bar N$, depending only on K and δ, with the following property. If $s = 0$ and

$$\mathrm{pen}(\bar m) \le K\varepsilon^2 D_{\bar m} \quad \text{for some } \bar m \in \mathcal{M} \text{ with } D_{\bar m} \ge \bar N, \tag{7.2}$$

then, whatever the value of the penalty $\mathrm{pen}(m)$ *for $m \neq \bar m$, either*

i) $\inf_{m\in\mathcal{M}} \mathrm{crit}(m) = -\infty$ *and $\hat m$ is not defined;*

or

ii) $\hat m = \mathrm{argmin}_{m\in\mathcal{M}} \mathrm{crit}(m)$ *is well-defined but then*

$$\mathbb{P}_0\left[D_{\hat m} \ge \frac{(1-K)}{2} D_{\bar m}\right] \ge 1-\delta \ \text{ and } \ \mathbb{E}_0\big[\,\|\tilde s\|^2\,\big] \ge \frac{(1-\delta)(1-K)}{4} D_{\bar m}\varepsilon^2.$$

Proof. Let us define, for any $m \in \mathcal{M}$, the nonnegative random variable χ_m by $\chi_m^2 = \varepsilon^{-2}\left\|\hat s_m\right\|^2$. Then,

$$\mathrm{crit}(m) - \mathrm{crit}(\bar m) = \left\|\hat s_{\bar m}\right\|^2 - \left\|\hat s_m\right\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\bar m) \quad \text{for all } m \in \mathcal{M},$$

and therefore, by (7.2),

$$\varepsilon^{-2}[\mathrm{crit}(m) - \mathrm{crit}(\bar m)] \ge \chi_{\bar m}^2 - \chi_m^2 - KD_{\bar m}. \tag{7.3}$$

The following proof relies on an argument about the concentration of the variables χ_m^2 around their expectations. As in the proof of Theorem 2, we can use the Cirel'son-Ibragimov-Sudakov concentration inequality for χ_m. Indeed choosing some orthonormal basis $\{\varphi_\lambda, \lambda \in \Lambda_m\}$ of S_m and recalling that $s = 0$, we have $\chi_m^2 = \sum_{\lambda\in\Lambda_m} Z^2(\varphi_\lambda)$, which means that χ_m can be interpreted as the supremum of $\sum_{\lambda\in\Lambda_m} a_\lambda Z(\varphi_\lambda)$ over the set of all vectors a in $\mathbb{R}^{\Lambda_m}$ with $\sum_{\lambda\in\Lambda_m} a_\lambda^2 = 1$. Hence, the Cirel'son-Ibragimov-Sudakov inequality implies that for any positive x,

$$\mathbb{P}_0\left[\chi_m \ge \mathbb{E}_0[\chi_m] + \sqrt{2x}\right] \le e^{-x}, \tag{7.4}$$

and

$$\mathbb{P}_0\left[\chi_m \le \mathbb{E}_0[\chi_m] - \sqrt{2x}\right] \le e^{-x}. \tag{7.5}$$

A proper integration of the resulting tail inequality $\mathbb{P}_0\left[|\chi_m - \mathbb{E}_0[\chi_m]| \geq u\right] \leq 2e^{-u^2/2}$ with respect to u, leads to

$$\mathbb{E}_0\left[\chi_m^2\right] - \left(\mathbb{E}_0[\chi_m]\right)^2 = \mathbb{E}_0\left[\left(\chi_m - \mathbb{E}_0[\chi_m]\right)^2\right] \leq 2\int_0^\infty e^{-z/2}dz = 4,$$

and therefore, since $\mathbb{E}_0\left[\chi_m^2\right] = D_m$,

$$D_m - 4 \leq \left(\mathbb{E}_0[\chi_m]\right)^2 \leq D_m. \tag{7.6}$$

Let us now set

$$\eta = (1-K)/4 < 1/4; \qquad D = 2D_{\bar{m}}\eta < D_{\bar{m}}/2; \qquad L = \eta^2/12 \tag{7.7}$$

and assume that $\bar{N}$ is large enough for the following inequalities to hold:

$$e^{-LD}\sum_{m\in\mathcal{M}} e^{-LD_m} \leq \delta; \qquad LD \geq 2/3. \tag{7.8}$$

Since the last inequality implies that $D > 128$, we can introduce the event

$$\bar{\Omega} = \left[\bigcap_{D_m<D}\left\{\chi_m \leq \sqrt{D_m} + \sqrt{2L\left(D_m + D\right)}\right\}\right] \bigcap \left[\bigcap_{D_m\geq D}\left\{\chi_m \geq \sqrt{D_m - 4} - \sqrt{2L\left(D_m + D\right)}\right\}\right].$$

Combining (7.6) with either (7.4) if $D_m < D$ or (7.5) if $D_m \geq D$, we get by (7.8)

$$\mathbb{P}_0\left[\bar{\Omega}^c\right] \leq \sum_{m\in\mathcal{M}} e^{-L(D_m+D)} \leq \delta.$$

Moreover, on $\bar{\Omega}$, $\chi_m^2 \leq \left(1 + 2\sqrt{L}\right)^2 D$, for all m such that $D_m < D$ and, by (7.7) and (7.8), $\chi_{\bar{m}} \geq \sqrt{D_{\bar{m}} - 4} - \sqrt{3LD_{\bar{m}}}$ and $LD_{\bar{m}} > 4/3$. Therefore $\chi_{\bar{m}}^2 \geq D_{\bar{m}}\left(1 - 2\sqrt{3L}\right)$. Hence, on $\bar{\Omega}$, (7.3) and (7.7) yield

$$\begin{aligned}\varepsilon^{-2}[\mathrm{crit}(m) - \mathrm{crit}(\bar{m})] &\geq D_{\bar{m}}\left(1 - 2\sqrt{3L}\right) - \left(1 + 2\sqrt{L}\right)^2 D - KD_{\bar{m}} \\ &> (1-\eta)D_{\bar{m}} - 3\eta D_{\bar{m}} - (1-4\eta)D_{\bar{m}} \;=\; 0,\end{aligned}$$

for all m such that $D_m < D$. This immediately implies that, if $\hat{m}$ is well-defined, $D_{\hat{m}}$ cannot be smaller than D on $\bar{\Omega}$ and therefore,

$$\mathbb{P}_0[D_{\hat{m}} \geq D] \geq \mathbb{P}_0[\bar{\Omega}] \geq 1 - \delta. \tag{7.9}$$

Moreover, on the same set $\bar{\Omega}$, $\chi_m \geq \sqrt{D_m - 4} - \sqrt{2L\left(D_m + D\right)}$ if m is such that $D_m \geq D$. Setting $D'_m = D_{\bar{m}} \vee D_m$,we derive, since $L \leq \eta/48$ and $2\eta D'_m \geq D > 128$, that

$$\chi_m \geq \sqrt{2\eta D'_m - 4} - \sqrt{3LD'_m} > (5/4)\sqrt{\eta D'_m} - (1/4)\sqrt{\eta D'_m} = \sqrt{\eta D'_m} \geq \sqrt{\eta D_{\bar{m}}}.$$

Hence, on $\bar{\Omega}$, $D_{\hat{m}} \geq D$ and $\chi_m \geq \sqrt{\eta D_{\tilde{m}}}$ for all m such that $D_m \geq D$. Therefore $\chi_{\hat{m}} \geq \sqrt{\eta D_{\tilde{m}}}$. Finally,

$$\mathbb{E}_0\left[\|\tilde{s}\|^2\right] = \varepsilon^2 \mathbb{E}_0\left[\chi^2_{\hat{m}}\right] \geq \varepsilon^2 \eta D_{\tilde{m}} \mathbb{P}_0\left[\chi_{\hat{m}} \geq \sqrt{\eta D_{\tilde{m}}}\right] \geq \varepsilon^2 \eta D_{\tilde{m}} \mathbb{P}_0[\bar{\Omega}],$$

which, together with (7.7) and (7.9) concludes the proof. □

Remark. It may look strange to use concentration inequalities like (7.4) and (7.5) to derive an inequality like (7.6) in a situation where $\mathbb{E}_0[\chi_m]$ can be computed exactly and is known to be $\sqrt{2}\Gamma[(D_m+1)/2][\Gamma(D_m/2)]^{-1}$. Nevertheless, it is not clear at all that one can derive (7.6) from this exact value by a three lines proof as we did above. This is actually a good illustration of the power and usefulness of concentration inequalities.

In order to illustrate the meaning of this proposition, let us assume that we are given some orthonormal basis $\{\varphi_j\}_{j\geq 1}$ of $\mathbb{H}$ and that S_m is the linear span of $\varphi_1, \ldots, \varphi_m$ for $m \in \mathbb{N}$. Assume that $s = 0$ and $\text{pen}(m) = K\varepsilon^2 m$ with $K < 1$. If $\mathcal{M} = \mathbb{N}$, then Proposition 8 applies with $D_{\tilde{m}}$ arbitrarily large and letting $D_{\tilde{m}}$ go to infinity and δ to zero, we conclude that $\inf_{m\in\mathcal{M}} \text{crit}(m) = -\infty$ a.s. If we set $\mathcal{M} = \{0, 1, \ldots, N\}$, then $\hat{m}$ is defined but, setting $D_{\tilde{m}} = N$, we see that $\mathbb{E}_0\left[\|\tilde{s}\|^2\right]$ is of the order of $N\varepsilon^2$ when N is large. In order to avoid this phenomenon, we have to restrict drastically our family of models to small enough values of m, say $m \leq n$. But then functions of the form $\lambda\varphi_{n+1}$ cannot be estimated with a risk smaller than λ^2, which may be arbitrarily large. If, on the contrary, we choose $\mathcal{M} = \mathbb{N}$ and $\text{pen}(m) = Km\varepsilon^2$ with $K = 2$, for instance, as in Mallows' C_p, it follows from Theorem 2 that

$$\mathbb{E}_s\left[\|\tilde{s} - s\|^2\right] \leq C(m+1)\varepsilon^2 \quad \text{for all } s \in S_m, \quad \text{whatever } m.$$

This means that choosing a penalty of the form $\text{pen}(m) = K\varepsilon^2 m$ with $K < 1$ is definitely not advisable.

7.2. A large number of models

The preceding result corresponds to a situation where the number of models having the same dimension D is moderate in which case we can choose the weights L_m in Theorem 2 all equal to an arbitrary small positive constant. This means that the influence of the weights on the penalty is limited in the sense that they only play the role of a correction to the main term $K\varepsilon^2 D_m$. This remains true for the variable weights strategy, when $L_m = cD_m^{-1/2}$. The situation becomes quite different when the number of models having the same dimension D grows much faster with D. More precisely, if we turn back to the case of complete variable selection as described at the beginning of Sect. 5 and assume that $L_m = L$ for all $m \in \mathcal{M}$, then Theorem 2 applies when $L = \log N$ and $\text{pen}(m) = K\varepsilon^2|m|\left(1 + \sqrt{2\log N}\right)^2$ with $K > 1$. In such a situation, L_m tends to infinity with N and directly determines the order of magnitude of the penalty. The penalized projection estimator is then

the threshold estimator $\tilde{s}_T$ with $T = \varepsilon\sqrt{K}\left(1+\sqrt{2\log N}\right)$ as defined by (5.8). If $s = 0$, we derive from Proposition 2 that

$$\mathbb{E}_0\left[\|\tilde{s}_T\|^2\right] \geq \varepsilon^2 N \exp\left[-K\left(1/\sqrt{2}+\sqrt{\log N}\right)^2\right]$$
$$= \varepsilon^2 \exp\left[(1-K)\log N - K\left(\sqrt{2\log N}+1/2\right)\right].$$

If $K < 1$, this grows like ε^2 times a power of N when N goes to infinity, as compared to the risk bound $C(K)\varepsilon^2 \log N$ which holds when $K > 1$. Clearly the choice $K < 1$ should be avoided. In particular penalization procedures based on Mallows' C_p, which are of the form $\text{pen}(m) = 2\varepsilon^2 D_m$ are definitely not suitable for complete variable selection involving a large number of variables, although it is a rather common practice to use them in this situation, as more or less suggested for instance by Draper and Smith (1981, p. 299).

8. Appendix

8.1. From function spaces to sequence spaces, a reminder

Our purpose here is to briefly recall, following more or less Donoho and Johnstone (1998), why it is natural to search for adaptive procedures over various types of $\boldsymbol{l}_p$-bodies and particularily Besov bodies.

We recall that, given three positive numbers $p, q \in (0, +\infty]$ and $\alpha > 1/p - 1/2$ one defines the Besov semi-norm $|t|_{B_q^\alpha(L_p)}$ of any function $t \in \mathbb{L}_2([0,1])$ by

$$|t|_{B_q^\alpha(L_p)} = \begin{cases} \left(\sum_{j=0}^{\infty}\left[2^{j\alpha}\omega_r(t, 2^{-j}, [0,1])_p\right]^q\right)^{1/q} & \text{when } q < +\infty, \\ \sup_{j\geq 0} 2^{j\alpha}\omega_r(t, 2^{-j}, [0,1])_p & \text{when } q = +\infty, \end{cases} \tag{8.1}$$

where $\omega_r(t, x, [0,1])_p$ denotes the modulus of smoothness of t, as defined by DeVore and Lorentz (1993, p. 44) and $r = \lfloor\alpha\rfloor + 1$. Since $\omega_r(t, 2^{-j}, [0,1])_p \geq \omega_r(t, 2^{-j}, [0,1])_2$ when $p > 2$, then $\left\{t \mid |t|_{B_q^\alpha(L_p)} \leq R\right\} \subset \left\{t \mid |t|_{B_q^\alpha(L_2)} \leq R\right\}$ for $p \geq 2$. Keeping in mind that we are interested in adaptation and therefore comparing the risk of our estimators to the minimax risk over such Besov balls, we can restrict our study to the case $p \leq 2$. Indeed, our nonasymptotic computations can only be done up to constants and it is known that the influence of p on the minimax risk is limited to those constants. It is therefore natural to ignore the smaller balls corresponding to $p > 2$.

Modulo the choice of a convenient wavelet basis, the Besov balls $\left\{t \mid |t|_{B_q^\alpha(L_p)} \leq R\right\}$ are contained in subsets of $\boldsymbol{l}_2(\Lambda)$ that have some nice geometrical properties. Given a pair (father and mother) of compactly supported orthonormal wavelets $(\bar{\psi}, \psi)$, any $t \in \mathbb{L}_2([0,1])$ can be written on $[0,1]$ as

$$t = \sum_{k\in\Lambda(-1)} \alpha_k \bar{\psi}_k + \sum_{j=0}^{\infty}\sum_{k\in\Lambda(j)} \beta_{j,k}\psi_{j,k}, \tag{8.2}$$

with

$$|\Lambda(-1)| = M' < +\infty \quad \text{and} \quad 2^j \le |\Lambda(j)| \le M2^j \quad \text{for all } j \ge 0. \tag{8.3}$$

For a suitable choice of the wavelet basis and provided that the integer r satisfies $1 \le r \le \bar{r}$ with $\bar{r}$ depending on the basis,

$$2^{j(1/2-1/p)} \left(\sum_{k\in\Lambda(j)} |\beta_{j,k}|^p \right)^{1/p} \le C\omega_r(t, 2^{-j}, [0,1])_p, \tag{8.4}$$

for all $j \ge 0$, $p \ge 1$, with a constant $C > 0$ depending only on the basis. See Cohen, Daubechies and Vial (1993) and Theorem 2 of Donoho and Johnstone (1998). This result remains true if one replaces the wavelet basis by a piecewise polynomial basis generating dyadic piecewise polynomial expansions as shown in Birgé and Massart (2000, Sect. 4.1.1). With some suitable restrictions on ω_r, this inequality still holds for $0 < p < 1$ and C depending on p (see DeVore et al. 1993 or Birgé and Massart, 2000). In particular, if we fix $p \in [1, 2]$, $q \in (0, +\infty]$, $\alpha > 1/p - 1/2$ and $R' > 0$ and consider those ts satisfying $|t|_{B_q^\alpha(L_p)} \le R'$, one derives from (8.4) that the coefficients $\beta_{j,k}$ of t in the expansion (8.2) satisfy

$$\sum_{j=0}^{\infty} \left[(R'C)^{-1} 2^{j(\alpha+1/2-1/p)} \left(\sum_{k\in\Lambda(j)} |\beta_{j,k}|^p \right)^{1/p} \right]^q \le 1 \quad \text{when } q < +\infty, \tag{8.5}$$

$$\sup_{j\ge 0} (R'C)^{-1} 2^{j(\alpha+1/2-1/p)} \left(\sum_{k\in\Lambda(j)} |\beta_{j,k}|^p \right)^{1/p} \le 1 \quad \text{when } q = +\infty, \tag{8.6}$$

and one can show that such inequalities still hold for $p < 1$ (with C depending on p). Clearly, if (8.5) is satisfied for some q, it is also satisfied for all $q' > q$. The choice $q = +\infty$ dominates all other choices but does not allow us to deal with the limiting case $\alpha = 1/p - 1/2$ (when $p < 2$) since, with such a choice of α, (8.6) does not warrant that the coefficients $\beta_{j,k}$ belong to $\boldsymbol{l}_2(\Lambda)$. It is therefore necessary, in this case, to restrict to $q = p$. For this reason, only two values of q are of interest for us: $q = p$ and $q = +\infty$, results for other values deriving from the results concerning those two ones. For the sake of simplicity, we shall actually focus on the case $q = p$, only minor modifications being needed to extend the results, when $\alpha > 1/p - 1/2$, to the case $q = +\infty$.

If $q = p \le 2$, (8.5) becomes

$$\sum_{j=0}^{\infty} \left[2^{jp(1/2-1/p)} \left[\omega(2^{-j})\right]^{-p} \sum_{k\in\Lambda(j)} |\beta_{j,k}|^p \right] \le 1, \tag{8.7}$$

with $\omega(x) = Rx^\alpha$ and $R = R'C$. Apart from the fact that it corresponds to some smoothness of order α in the usual sense, there is no special reason to restrict to

functions ω of this particular form. If for instance,

$$\sum_{j=0}^{\infty}\left[\frac{\omega_r(t,2^{-j},[0,1])_p}{\omega(2^{-j})}\right]^p \le C^{-1}$$

for some nonnegative continuous function ω such that $x^{1/2-1/p}\omega(x)$ is bounded on $[0,1]$, it follows from (8.4) that (8.7) still holds and the set of βs satisfying (8.7) is a subset of $\boldsymbol{l}_2(\Lambda)$. If the function $x \mapsto x^{1/2-1/p}\omega(x)$ is nondecreasing and tends to zero when $x \to 0$, this set is an ellipsoid for $p = 2$. These considerations were the main motivation for our Definition 10 of extended Besov bodies, the particular case $r = 0$ of classical Besov bodies (compare with Donoho and Johnstone, 1998, Sect. 2) being suitable for analyzing sets of functions of the form $\left\{t \mid |t|_{B_p^\alpha(L_p)} \le R\right\}$. Indeed if we consider some well-chosen orthonormal wavelet basis $\{\bar{\psi}_k \mid k \in \Lambda(-1)\} \cup \left(\cup_{j\in\mathbb{N}}\{\psi_{j,k} \mid k \in \Lambda(j)\}\right)$ in $\mathbb{L}_2([0,1])$ with sets $\Lambda(j)$ satisfying (8.3), one can order it according to lexical order as $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ and this correspondence gives $k \in \Lambda(-1) \longleftrightarrow \lambda$ with $1 \le \lambda \le M'$ and

$$\text{if } j \ge 0, k \in \Lambda(j),\ (j,k) \longleftrightarrow \lambda \text{ with } M' + 2^j \le \lambda \le M' + M(2^{j+1} - 1). \quad (8.8)$$

Assuming that (8.4) holds and $\alpha+1/2-1/p > 0$, we derive that $\left\{t \mid |t|_{B_p^\alpha(L_p)} \le R\right\}$ is included in the set of t's with coefficients satisfying

$$C^{-p}\sum_{j=0}^{\infty}\left[2^{jp(\alpha+1/2-1/p)}R^{-p}\sum_{k\in\Lambda(j)}|\beta_{j,k}|^p\right]$$
$$=\sum_{j=0}^{\infty}\sum_{k\in\Lambda(j)}\left|\frac{\beta_{j,k}}{2^{-j(\alpha+1/2-1/p)}RC}\right|^p \le 1,$$

and it follows from (8.8) that $\left\{t \mid |t|_{B_p^\alpha(L_p)} \le R\right\} \subset \mathcal{B}(M', \alpha, p, R', 0)$ with $R' = RC(2M)^{\alpha+1/2-1/p}$.

8.2. *Proof of Theorem 5*

As often for minimax lower bounds, the proof relies on an application of some version of Fano's Lemma. We shall use here the following one which is proved in Birgé (2001).

Proposition 9. *Let $(\mathbb{S}, d)$ be a metric space, and $\{P_s\}_{s\in\mathbb{S}}$ a set of probability distributions indexed by $\mathbb{S}$. Let $\mathcal{C}$ be a finite subset of $\mathbb{S}$ with $|\mathcal{C}| \ge 6$, such that for all pairs $s \ne s' \in \mathcal{C}$, $d(s,s') \ge \eta > 0$. Assume, moreover, that there exists an element $s_0 \in \mathcal{C}$ such that*

$$K(P_s, P_{s_0}) = \int \log\left(\frac{dP_s}{dP_{s_0}}\right) dP_s \le H < \log(|\mathcal{C}|) \quad \textit{for all } s \in \mathcal{C}.$$

Then, for any estimator $\tilde{s}$ with values in $\mathbb{S}$ and any nondecreasing function ℓ,

$$\sup_{s\in\mathcal{C}} \mathbb{E}_s\left[\ell(d(\tilde{s},s))\right] \geq \ell\left(\frac{\eta}{2}\right)\left[1-\left(\frac{2}{3}\vee\frac{H}{\log|\mathcal{C}|}\right)\right].$$

The preceding proposition involves Kullback-Leibler information numbers between the underlying probability distributions and, in order to use it, we have to compute the mutual Kullback-Leibler information numbers between the possible distributions of the process Y.

Lemma 3. *Let P_s be the distribution of the process $t \mapsto \langle s,t\rangle + \varepsilon Z(t)$ where Z is a linear isonormal process indexed by some subspace $\mathbb{S}$ of a Hilbert space $\mathbb{H}$. Given two elements s and s' in $\mathbb{H}$, the Kullback-Leibler information number $K(P_s, P_{s'})$ satisfies*

$$K(P_s, P_{s'}) = \int \log\left(\frac{dP_s}{dP_{s'}}(y)\right) dP_s(y) = \frac{\|s-s'\|^2}{2\varepsilon^2}. \tag{8.9}$$

Sketch of proof. If E is the linear space spanned by s and s' and $u \in E$, we denote by Q_u the Gaussian distribution on E with mean u and covariance operator $\varepsilon^2 I$. Since the restrictions of Z to E and $E^\perp$ respectively are independent, the distribution, under P_s, of the likelihood ratio $dP_s/dP_{s'}$ is the same as the distribution, in E, of the likelihood ratio $dQ_s/dQ_{s'}$ under Q_s. The conclusion follows straightforwardly. □

In order to apply Fano's Lemma, we have to exhibit a suitable subset $\mathcal{C}$ of $\mathbb{S}_D$. To get the non-trivial logarithmic factor $\log(N/D)$ in the lower bound, we need to build a large enough set $\mathcal{C}$, the existence of which will derive from the following corollary of Lemma 9 of Birgé and Massart (1998).

Lemma 4. *Let N and n be two positive integers such that $N \geq 6n$. Given a finite set Λ with cardinality N and $\mathcal{M}$ the set of all subsets of cardinality $2n$ of Λ we consider the distance δ on $\mathcal{M}$ defined by*

$$\delta(m,m') = \frac{1}{2}\int |\mathbb{1}_m(\lambda) - \mathbb{1}_{m'}(\lambda)|\, d\mu(\lambda) = 2n - |m\cap m'|,$$

where μ denotes the counting measure on Λ. Then there exists a subset $\mathcal{C}$ of $\mathcal{M}$ such that $\delta(m,m') \geq n+1$ for all $m \neq m' \in \mathcal{C}$ and

$$\log(|\mathcal{C}|) > n[\log(N/n) - \log 16 + 1]. \tag{8.10}$$

Proof. It follows from Lemma 9 in Birgé and Massart (1998) with $M = N$, $C = 2n$ and $q = n$ that there exists a subset $\mathcal{C}$ of $\mathcal{M}$ such that $\delta(m,m') > n$ (and therefore $\delta(m,m') \geq n+1$ since δ is integer valued) for all $m, m' \in \mathcal{C}$ with $m \neq m'$ and

$$|\mathcal{C}| \geq \frac{N-4n}{N-3n}\,\frac{\binom{N}{2n}}{\binom{2n}{n}\binom{N-2n}{n}} = \frac{N-4n}{N-3n}\,\frac{N!(n!)^3(N-3n)!}{[(2n)!(N-2n)!]^2}.$$

A direct computation shows that (8.10) holds when $n = 1$. Let us now assume that $n \geq 2$. Recalling from Stirling's formula (Feller, 1968, p. 54) that

$j! = j^j e^{-j}\sqrt{2\pi j}\psi(j)$ with $(12j+1)^{-1} < \log[\psi(j)] < (12j)^{-1}$, which implies that ψ is decreasing, we derive that $|\mathcal{C}| \geq \sqrt{F_0}F_1F_2$ with

$$F_0 = 2\pi \frac{Nn^3(N-3n)}{(2n)^2(N-2n)^2}\left[\frac{N-4n}{N-3n}\right]^2 = \frac{\pi n}{2}\,\frac{(N/n)(N/n-4)^2}{(N/n-2)^2(N/n-3)},$$

$$F_1 = \frac{N^N n^{3n}[N-3n]^{N-3n}}{(2n)^{4n}(N-2n)^{2(N-2n)}} = 2^{-4n}\frac{N^N[N-3n]^{N-3n}}{n^n(N-2n)^{2(N-2n)}},$$

and

$$F_2 = \frac{\psi(N)\psi^3(n)\psi[N-3n]}{\psi^2(2n)\psi^2(N-2n)} > \frac{\psi(n)}{\psi(N-2n)}\frac{\psi^2(n)}{\psi^2(2n)}\frac{\psi(N-3n)}{\psi(N-2n)}.$$

We first observe that F_0 is an increasing function of N/n for $N/n \geq 4$, and therefore $F_0 \geq 1$ since $N \geq 6n$. Then it follows from the monotonicity of ψ that $F_2 > 1$. Setting $x = n/N$, we finally get

$$\log(|\mathcal{C}|) > \log F_1 = n\left[-\log 16 + \log(N/n) + x^{-1}G(x)\right],$$

where $G(x) = (1-3x)\log(1-3x) - 2(1-2x)\log(1-2x)$. From the expansion $(1-x)\log(1-x) = -x + \sum_{i\geq 2}[i(i-1)]^{-1}x^i$ we derive that $G(x) > x$ and (8.10) follows. □

Let us now prove Theorem 5, distinguishing between three cases.

Case 1. $N < 650D$
If $N = D$, $\mathcal{B}(D, D, b)$ is a D-dimensional cube with edges of length $b\varepsilon$ and it follows from Donoho et al. (1990) that

$$R_M(\mathcal{B}(D, D, b), \varepsilon) \geq \frac{4D}{5}\,\frac{b^2\varepsilon^4}{b^2\varepsilon^2+\varepsilon^2} \geq \frac{2D\varepsilon^2}{5}\big(b^2 \wedge 1\big).$$

Now observe that $\mathcal{B}(D, D, b)$ can be considered as a subset of $\mathcal{B}(N, D, b)$, hence

$$R_M(\mathcal{B}(N, D, b), \varepsilon) \geq R_M(\mathcal{B}(D, D, b), \varepsilon) \geq \frac{2D\varepsilon^2}{5}\big(b^2 \wedge 1\big). \tag{8.11}$$

Case 2. $D = 1$ and $N \geq 650$
Set $a = \big[b \wedge \sqrt{2(\log N)/3}\big]$ and, for any $\lambda \in \Lambda$, define $s_\lambda = a\varepsilon\varphi_\lambda \in \mathcal{B}(N, 1, b)$. Then if $\lambda \neq \lambda' \in \Lambda$,

$$\|s_\lambda - s_{\lambda'}\|^2 = 2a^2\varepsilon^2 \qquad \text{and} \qquad K(P_{s_\lambda}, P_{s_{\lambda'}}) = a^2$$

by (8.9). Since $N \geq 6$, we can apply Proposition 9 to the set $\{s_\lambda\}_{\lambda\in\Lambda} \subset \mathcal{B}(N, 1, b)$ and derive that, whatever the estimator $\tilde{s}$,

$$\sup_{\lambda\in\Lambda}\mathbb{E}_{s_\lambda}\left[\|\tilde{s} - s_\lambda\|^2\right] \geq \left(\frac{a^2\varepsilon^2}{2}\right)\left[1 - \left(\frac{2}{3} \vee \frac{a^2}{\log N}\right)\right] \geq \frac{a^2\varepsilon^2}{6},$$

from which we conclude that

$$R_M(\mathcal{B}(N, D, b), \varepsilon) \geq \frac{\varepsilon^2}{6}\left[b^2 \wedge \frac{2}{3}\log N\right] = \frac{D\varepsilon^2}{6}\left[b^2 \wedge \frac{2}{3}\log\left(\frac{N}{D}\right)\right]. \tag{8.12}$$

Case 3. $D \geq 2$ and $N \geq 650D$
Let n be the positive integer defined by $D \geq 2n \geq D-1$ and set

$$c = \log(N/D) - \log 8 + 1 \qquad \text{and} \qquad a = \left[b \wedge \sqrt{c/3}\right].$$

It follows from Lemma 4 that we can find a subset $\mathcal{C}$ of $\mathcal{M}_{2n}$ such that

$$\log(\mathcal{C}) > nc \quad \text{and} \quad 2n \geq \delta(m, m') \geq n+1 \quad \text{for all } m \neq m' \in \mathcal{C}. \tag{8.13}$$

For any $m \in \mathcal{C}$, define $s(m) = a\varepsilon \sum_{\lambda \in m} \varphi_\lambda \in \mathcal{B}(N, 2n, b)$. Then

$$\|s(m) - s(m')\|^2 = 2a^2\varepsilon^2\delta(m, m') \quad \text{for all } m, m' \in \mathcal{C},$$

and by (8.9) and (8.13),

$$K(P_{s(m)}, P_{s(m')}) \leq 2na^2 \quad \text{and} \quad \|s(m) - s(m')\|^2 \geq 2(n+1)a^2\varepsilon^2.$$

Applying Proposition 9 to the set $\{s(m)\}_{m\in\mathcal{C}} \subset \mathcal{B}(N, 2n, b)$ we derive that, whatever the estimator $\tilde{s}$,

$$\sup_{m\in\mathcal{C}} \mathbb{E}_{s(m)}\left[\|\tilde{s} - s(m)\|^2\right] \geq \left(\frac{(n+1)a^2\varepsilon^2}{2}\right)\left[1 - \left(\frac{2}{3} \vee \frac{2na^2}{\log|\mathcal{C}|}\right)\right] \geq \frac{(n+1)a^2\varepsilon^2}{6}.$$

Since $\log(N/D) \geq \log(650) > 6(\log 8 - 1)$, $c > (5/6)\log(N/D)$ and

$$R_M(\mathcal{B}(N, 2n, b), \varepsilon) \geq \frac{(n+1)\varepsilon^2}{6}\left[b^2 \wedge \frac{c}{3}\right] > \frac{D\varepsilon^2}{12}\left[b^2 \wedge \frac{5}{18}\log\left(\frac{N}{D}\right)\right].$$

Since $\mathcal{B}(N, 2n, b) \subset \mathcal{B}(N, D, b)$, we can conclude that $R_M(\mathcal{B}(N, 2n, b), \varepsilon) \leq R_M(\mathcal{B}(N, D, b), \varepsilon)$ and therefore

$$R_M(\mathcal{B}(N, D, b), \varepsilon) > \frac{D\varepsilon^2}{12}\left[b^2 \wedge \frac{5}{18}\log\left(\frac{N}{D}\right)\right].$$

Putting this together with (8.11) and (8.12) gives the result. □

Acknowledgements. We would like to thank an anonymous referee for his constructive suggestions.

References

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

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automatic Control **19**, 716–723

Baraud, Y. (1997). Model selection for regression on random design. Technical Report No. 97.74. Mathématiques, Université Paris-Sud, Orsay

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

Baraud, Y., Comte, F., Viennet, G. (1997). Adaptive estimation in an autoregression and geometrical β-mixing framework. Technical Report No. 97.75. Mathématiques, Université Paris-Sud, Orsay

Baraud, Y., Comte, F., Viennet, G. (1999). Model selection for (auto-)regression with dependent data. Technical Report LMENS-99-12. Ecole Normale Supérieure, Paris

Barron, A.R. (1987). Are Bayes rules consistent in information. In: Open Problems in Communication and Computation, T.M. Cover, B. Gopinath (Eds.), pp. 85–91. Berlin: Springer

Barron, A.R., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. Probab. Theory Relat. Fields **113**, 301–415

Barron, A.R., Cover, T.M. (1991). Minimum complexity density estimation. IEEE Transact. Inf. Theory **37**, 1034–1054

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

Birgé, L. (2001). A new look at an old result: Fano's Lemma. Technical Report 632. Lab. de Probabilités, Université Paris VI

Birgé, L., 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, G. Yang (Eds.), pp. 55–87. New York: Springer

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

Birgé, L., Massart, P. (2000). An adaptive compression algorithm in Besov spaces. Constructive Approximation **16**, 1–36

Birgé, L., Massart, P. (2001). A generalized C_p criterion for Gaussian model selection. Technical Report. Lab. de Probabilités, Université Paris VI

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 99.61. Mathématiques, Université Paris-Sud, Orsay

Catoni, O. (2000). Universal aggregation rules with sharp oracle inequalities. Ann. Statist. (to appear)

Cirel'son, B.S., Ibragimov, I.A., Sudakov, V.N. (1976). Norm of gaussian sample function. In: Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory, Springer Lecture Notes in Mathematics 550, pp. 20–41. Berlin: Springer

Cohen, A., Daubechies, I., Vial, P. (1993). Wavelets and fast wavelet transform on an interval. Appl. Comput. Harmon. Anal. **1**, 54–81

Daniel, C., Wood, F.S. (1971). Fitting Equations to Data. New York: Wiley

DeVore, R.A., Kyriazis, G., Leviatan, D., Tikhomirov, V.M. (1993). Wavelet compression and nonlinear n-widths. Adv. Computat. Math. **1**, 197–214

DeVore, R.A., Lorentz, G.G. (1993). Constructive Approximation. Berlin: Springer

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

Donoho, D.L., Johnstone, I.M. (1994b). Minimax risk over l_p-balls for l_q-error. Probab. Theory Relat. Fields **99**, 277–303

Donoho, D.L., Johnstone, I.M. (1994c). Ideal denoising in an orthonormal basis chosen from a library of bases. C. R. Acad. Sc. Paris Sér. I Math. **319**, 1317–1322

Donoho, D.L., Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. JASA. **90**, 1200–1224

Donoho, D.L., Johnstone, I.M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. Bernoulli **2**, 39–62

Donoho, D.L., Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. Ann. Statist. **26**, 879–921

Donoho, D.L., Liu, R.C., MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. Ann. Statist. **18**, 1416–1437

Draper, N.R., Smith, H. (1981). Applied regression analysis, second edition. New York: Wiley

Dudley, R.M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. J. Funct. Anal. **1**, 290–330
Efroimovich, S.Yu., Pinsker, M.S. (1984). Learning algorithm for nonparametric filtering. Automat. Remote Control **11**, 1434–1440, translated from Avtomatika i Telemekhanika **11**, 58–65
Hurvich, K.L., Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika **76**, 297–307
Johnstone, I.M. (1998). Function estimation and wavelets. Lectures at Ecole Normale Supérieure, Paris. Unpublished manuscript, HTTP://WWW-STAT.STANFORD.EDU/~IMJ/MONOGRAPH.PS
Katkovnik, V.Ya. (1979). Linear and nonlinear methods of nonparametric regression analysis. Automatika **5**, 35–46
Kerkyacharian, G., Picard, D. (2000). Thresholding algorithms, maxisets and well concentrated bases. Test **9**, 283–344
Kneip, A. (1994). Ordered linear smoothers. Ann. Statist. **22**, 835–866
Le Cam, L.M., Yang, G.L. (1990). Asymptotics in statistics: some basic concepts. New York: Springer
Ledoux, M. (1996). Isoperimetry and Gaussian analysis. In: Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXIV-1994 (P. Bernard ed.). Lecture Note in Mathematics 1648, pp. 165–294. Berlin: Springer
Lepskii, O.V. (1990). On a problem of adaptive estimation in Gaussian white noise. Theory Probab. Appl. **36**, 454–466
Lepskii, O.V. (1991). Asymptotically minimax adaptive estimation I: upper bounds. Optimally adaptive estimates. Theory Probab. Appl. **36**, 682–697
Li, K.C. (1987). Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set. Ann. Statist. **15**, 958–975
Mallows, C.L. (1973). Some comments on C_p. Technometrics **15**, 661–675
McQuarrie, A.D.R., Tsai, C.-L. (1998). Regression and time series model selection. Singapore: World Scientific
Meyer, Y. (1990). Ondelettes et Opérateurs I. Paris: Hermann
Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.M. (1996). Matlab Wavelet Toolbox. Natick: The Math Works Inc.
Nemirovski, A.S. (2000). Topics in non-parametric statistics. In: Lecture on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXVIII – 1998 (P. Bernard, ed.). Lecture Note in Mathematics 1738, pp. 85–297. Berlin: Springer
Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. **12**, 758–765
Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. Problems Information Transmission **16**, 120–133
Polyak, B.T., Tsybakov, A.B. (1990). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. Theory Probab. Appl. **35**, 293–306
Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. Scand. J. Statist. **9**, 65–78
Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. **6**, 461–464
Shibata, R. (1981). An optimal selection of regression variables. Biometrika **68**, 45–54
Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. of the 3rd Berkeley Symp. **1**, 197–206
Yang, Y. (1999). Model selection for nonparametric regression. Statistica Sinica **9**, 475–499
Yang, Y. (2000). Mixing strategies for density estimation. Ann. Statist. **28**, 75–87
Wahba, G. (1990). Spline models for observational data. Philadelphia: S.I.A.M.