Universités de Paris 6 & Paris 7 - CNRS (UMR 7599)

PRÉPUBLICATIONS DU LABORATOIRE

DE PROBABILITÉS & MODÈLES ALÉATOIRES

4, place Jussieu - Case 188 - 75 252 Paris cedex 05

http://www.proba.jussieu.fr

A generalized C_p criterion for Gaussian model selection

L. BIRGÉ & P. MASSART

AVRIL 2001

Prépublication n^o 647

L. Birgé : Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII, 4 place Jussieu, Case 188, F-75252 Paris Cedex 05.

P. Massart : CNRS-UMR 8628, Laboratoire de Mathématiques, Bât. 425 Université Paris Sud, Campus d'Orsay, F-91405 Orsay Cedex.

A generalized C_p criterion for Gaussian model selection

Lucien Birgé
Université Paris VI and UMR CNRS 7599

Pascal Massart
Université Paris Sud and UMR CNRS 8628

01/04/01

Abstract

This paper is mainly devoted to a precise analysis of what kind of penalties should be used in order to perform model selection via the minimization of a penalized least-squares type criterion within some general Gaussian framework. As compared to our previous paper on this topic (Birgé and Massart, 2001), more elaborate forms of the penalties are given which are shown to be, in some sense, optimal. We also provide risk bounds with explicit absolute constants and an asymptotic evaluation of the risk which generalizes the one of Shibata (1981) to our new penalties. Some applications to the estimation of change points for a signal in Gaussian noise are also developed. We finally present a practical strategy, based on sharp lower bounds for the penalty function, to design the penalty from the data when the amount of noise is unknown.

1 Introduction

1.1 Variable selection for Gaussian regression

Let us consider the following classical regression problem: we observe n independent observations $Y_1, \ldots, Y_n$ from the Gaussian linear regression set up

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^j + \sigma \xi_i \quad \text{for } 1 \leq i \leq n, \quad \text{with } \xi_1, \ldots, \xi_n \text{ i.i.d. } \mathcal{N}(0,1), \tag{1.1}$$

where X_i^j, $1 \leq i \leq n$ denote the respective values of some explanatory variable X^j. We want to estimate the mean vector $s = (s_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ of Y where $s_i = \sum_{j=1}^{p} \beta_j X_i^j$,

[0] *AMS 1991 subject classifications.* Primary 62G05; secondary 62G07, 62J05.
Key words and phrases. Gaussian linear regression, variable selection, model selection, Mallows' C_p, penalized least-squares.

assuming that σ is known. For any estimator $\hat{s}$ with values in $\mathbb{R}^n$, the normalized risk of $\hat{s}$ is given by $\mathbb{E}\left[\|s-\hat{s}\|_n^2\right]$ where $\|\cdot\|_n$ denotes the normalized Euclidean norm on $\mathbb{R}^n$ given by $\|t\|_n^2 = n^{-1}\sum_{i=1}^n t_i^2$. Typically, one uses the least squares estimator $\hat{s}_\Lambda$ which is the orthogonal projection of the vector $Y=(Y_i)_{1\le i\le n}$ onto the linear space generated by the p variables X^j, with $j\in\Lambda=\{1;2;\ldots;p\}$ (considered as vectors $X^j=(X_i^j)_{1\le i\le n}\in\mathbb{R}^n$).

Let us now consider what would happen if, instead of this classical method, we used a "wrong" or "approximate" model for Y and, although we do assume (1.1) did as if Y were actually given by

$$Y_i = \sum_{j\in m}\beta_j X_i^j + \sigma\xi_i \quad \text{for } 1\le i\le n, \tag{1.2}$$

where m denotes some subset of Λ. In this case the natural estimator becomes the least squares estimator $\hat{s}_m$ based on this new model, i.e. the orthogonal projection of Y onto the linear span S_m of the set of variables $\{X^j\}_{j\in m}$. Its risk is given by

$$\mathbb{E}\left[\|\hat{s}_m - s\|_n^2\right] = \|s_m - s\|_n^2 + \sigma^2|m|/n, \tag{1.3}$$

where s_m denotes the projection of s onto S_m. In particular, the risk of $\hat{s}_\Lambda$ is $\sigma^2 p/n$. If p is large because one has put all potentially influential variables into the model, this risk may not be negligeable, even for large values of n. On the other hand, a choice of a too parsimonious model including only a limited number $|m|$ of variables can result in a poor estimator based on a grossly wrong model if we have omitted some very influential variables resulting in a large value of $\|s_m - s\|_n$. Actually, from the point of view of minimizing the risk, the best set of explanatory variables $\{X^j\}_{j\in m}$ is the one which minimizes (1.3) and it is not necessarily the whole initial set (think of the case where some of the β_js are close to zero). Finding an optimal, or close to optimal, set m amounts, roughly speaking, to select, among a possibly large number of explanatory variables, a smaller number of them containing all influential variables.

When introducing $\hat{s}_m$ we did as if the model corresponding to (1.2) were correct, i.e. if s did belong to the linear space S_m , which may or may not be true. Therefore, such a problem of variable selection can be interpreted as a problem of "model selection": we want to choose a good model, i.e. one leading to a least squares estimator with a close to minimal risk value, among all models of the form S_m. In view of (1.3), solving this problem would be possible if we knew s hence its projection onto the linear spaces S_m, which is obviously not the case. We are therefore led to the problem of choosing m from what is available, namely the observations Y_i.

1.2 Gaussian linear processes

Another interesting problem of the same type is the following. One observes a signal in Gaussian noise at some times $x_1 = 0 < x_2 < \ldots < x_n < 1$. The signal is supposed to be constant for a while and then jumps to another value, but neither the places, nor the number of jumps are known. One wants, nevertheless, to estimate this signal. This results in the following fixed design regression set up:

$$Y_i = \sum_{j=1}^p \beta_j \mathbb{1}_{I_j}(x_i) + \sigma\xi_i \quad \text{for } 1\le i\le n, \quad \text{with } \xi_1,\ldots,\xi_n \text{ i.i.d. } \mathcal{N}(0,1), \tag{1.4}$$

where $\{I_j\}_{1\leq j\leq p}$ denotes some partition of $[0,1]$ into successive intervals. Once again, our aim is to estimate the vector $(s_i)_{1\leq i\leq n}\in\mathbb{R}^n$ where $s_i=s(x_i)$ and the function s is given by $s(x)=\sum_{j=1}^p \beta_j \mathbb{1}_{I_j}(x)$. If we define the risk of an estimator $\hat{s}$ by $\mathbb{E}\left[\|s-\hat{s}\|_n^2\right]$ with $\|t\|_n^2=n^{-1}\sum_{i=1}^n t^2(x_i)$, the problem is equivalent to the preceding one, given by (1.1) if we set $X_i^j=\mathbb{1}_{I_j}(x_i)$. In this case a model corresponds to a partition $\mathcal{I}=\{I_j\}$ and can be identified to the corresponding linear space $S_{\mathcal{I}}$ generated by the vectors $(\mathbb{1}_{I_j}(x_i))_{1\leq i\leq n}$, for $I_j\in\mathcal{I}$. The difference with the preceding example is that choosing a partition does not mean selecting a subset from a set of variables, but one has, nevertheless, to solve the same problem: choose a good model, i.e. a good partition $\mathcal{I}$ leading to a least squares estimator $\hat{s}_{\mathcal{I}}$ with a risk close to minimal, from the observations only.

Actually, both problems can be put into the framework of Gaussian Linear processes, as defined in Birgé and Massart (2001, Section 2). Let us briefly recall what we mean by that. Our regression set up (1.1) can be written in the form

$$Y=s+\sigma\xi \quad \text{with } Y,s,\xi\in\mathbb{R}^n, \quad \xi\sim\mathcal{N}(0,\boldsymbol{Id}_n) \quad \text{and} \quad s_i=\sum_{j=1}^p \beta_j X_i^j. \tag{1.5}$$

It follows that Y can be identified by duality with a linear operator on the Hilbert space $\mathbb{R}^n$, or equivalently to the Gaussian Linear process $Y(\cdot)$ indexed by $\mathbb{R}^n$ and defined by

$$Y(t)=\langle Y,t\rangle_n=\langle s,t\rangle_n+\sigma\langle\xi,t\rangle_n=\langle s,t\rangle_n+\varepsilon Z(t), \quad \text{with } \varepsilon=\sigma/\sqrt{n}, \tag{1.6}$$

where $\langle\cdot,\cdot\rangle_n$ denotes the scalar product corresponding to the norm $\|\cdot\|_n$ and Z is a centered and linear Gaussian process indexed by $\mathbb{R}^n$ with covariance structure given by $\mathbb{E}[Z(t)Z(u)]=\langle t,u\rangle_n$.

More generally, given some Hilbert space $\boldsymbol{H}$ with scalar product $\langle\cdot,\cdot\rangle$ together with some suitable linear subspace $\mathbb{S}\subset\boldsymbol{H}$, a Gaussian Linear process Y indexed by $\mathbb{S}$ is defined by

$$Y(t)=\langle s,t\rangle+\varepsilon Z(t) \quad \text{for all } t\in\mathbb{S}, \tag{1.7}$$

where Z denotes a linear isonormal process indexed by $\mathbb{S}$, i.e. a centered and linear Gaussian process with covariance structure $\mathbb{E}[Z(t)Z(u)]=\langle t,u\rangle$. The fact that we have to introduce the subspace $\mathbb{S}$ is due to the fact that one cannot warrant the existence of a linear isonormal process on an arbitrary infinite dimensional Hilbert space. It follows from the considerations developed in Section 2.1 of Birgé and Massart (2001) that this framework is not only a good representation of the classical Gaussian regression set up as we have already seen but also of the white noise framework. Indeed, given the stochastic differential equation on $\mathcal{X}=[0,1]$,

$$dX=s(x)dx+\varepsilon dW \quad \text{with } X(0)=0, \tag{1.8}$$

where W denotes a Brownian motion originating from zero and $s\in\mathbb{L}_2([0,1])$, one can define Y and Z satisfying (1.7) by

$$Y(t)=\int_0^1 t(x)dX(x) \qquad \text{and} \qquad Z(t)=\int_0^1 t(x)dW(x),$$

provided that $\mathbb{S}$ is a suitable linear subspace of $\mathbb{L}_2([0,1])$.

1.3 Projection estimators and model selection

Since the problems of estimating $s = \sum_{j=1}^{p} \beta_j X^j$ within the statistical framework given by (1.1), or $s = \sum_{j=1}^{p} \beta_j \mathbb{1}_{I_j}$ from (1.4), or the function s in (1.8), when the values of σ and ε are known, can all be reduced to the one of estimating s from (1.7) when ε is known, let us now concentrate on the latter problem. We recall that the risk of an estimator $\hat{s} = \hat{s}(Y)$ is given by $\mathbb{E}\left[\|\hat{s} - s\|^2\right]$, where $\|\cdot\|$ denotes the norm in $\boldsymbol{H}$. It is important here to notice that the restriction of this process Y to some finite dimensional linear space S of dimension D can always be written as

$$Y(t) = \langle s, t\rangle + \varepsilon\langle \xi, t\rangle \quad \text{for all } t \in S, \quad \text{with } \xi \sim \mathcal{N}(0, \boldsymbol{Id_D}).$$

Indeed, given some orthonormal basis $\varphi_1, \ldots, \varphi_D$ of S, ξ can be written as the D-dimensional Gaussian vector with independent coordinates $\xi_j = Z(\varphi_j)$. In this case, the knowledge of the restriction to S of the process Y is equivalent to the knowledge of the Gaussian vector $Y_S = s_S + \varepsilon\xi \sim \mathcal{N}(s_S, \varepsilon \boldsymbol{Id_D})$, where s_S denotes the projection of s onto S. If only Y_S is available, the best we can do is to estimate s_S, which is the mean of a Gaussian vector. Therefore a natural estimator $\hat{s}_S$ is the maximum likelihood estimator which is also the least squares estimator, i.e. the minimizer, with respect to $t \in S$ of $\|Y_S - t\|^2$. Equivalently, $\hat{s}_S$ can be defined as the minimizer, with respect to $t \in S$ of $\|t\|^2 - 2Y(t)$.

¿From now on, we shall call any finite-dimensional subspace of $\mathbb{S}$, like S, a *model* and $\hat{s}_S$ the *projection estimator* of s with respect to the model S. It is well-known that the risk of $\hat{s}_S$ is given by

$$\mathbb{E}\left[\|\hat{s}_S - s\|^2\right] = \|s_S - s\|^2 + \varepsilon^2 D,$$

which is the sum of an approximation error (bias) due to the replacement of s by s_S and an estimation error (variance term) proportional to the dimension of the model we use, which is the number of parameters to be estimated. In particular, if we know in advance that s belongs to some given linear space $\bar{S}$, with dimension $\bar{D}$, we get $\mathbb{E}\left[\|\hat{s}_{\bar{S}} - s\|^2\right] = \varepsilon^2 \bar{D}$, but this does not at all mean that $\hat{s}_{\bar{S}}$ is a good estimator, as shown by the following example. Assume, for instance, that s can be written as $\sum_{j=1}^{\bar{D}} \theta_j \varphi_j$, where $\varphi_1, \ldots, \varphi_{\bar{D}}$ is an orthonormal basis of $\bar{S}$ and $m = \{j \,|\, |\theta_j| \geq \varepsilon\}$ has a cardinality D' smaller than $\bar{D}$. Introducing the approximate model S', which is the linear span of $\{\varphi_j \,|\, j \in m\}$, we get

$$\mathbb{E}\left[\|\hat{s}_{S'} - s\|^2\right] = \sum_{j \notin m} \theta_j^2 + \varepsilon^2 D' < \varepsilon^2 D.$$

It may even happen that the risk of $\hat{s}_{S'}$ is much smaller than the risk of $\hat{s}_{\bar{S}}$. Since, in many situations, we do not even know any finite dimensional space containing s, the problem is even more delicate.

¿From a more concrete point of view, when we want to estimate s from (1.7), we have at hand, or we introduce, a suitable family of models $\{S_m, m \in \mathcal{M}\}$ and the corresponding family of projection estimators $\{\hat{s}_m, m \in \mathcal{M}\}$, with respective quadratic risks

$$R_m = \mathbb{E}\left[\|\hat{s}_m - s\|^2\right] = \|s_m - s\|^2 + \varepsilon^2 D_m, \tag{1.9}$$

where D_m denotes the dimension of S_m and s_m the orthogonal projection of s onto S_m. Since the computation of the estimators $\hat{s}_m$ only involves the random variables $Y(t)$ with $t \in \cup_{m\in\mathcal{M}} S_m$, one can always assume that $\mathbb{S}$ is the linear span of $\cup_{m\in\mathcal{M}} S_m$. We do not assume here that s belongs to any of the models, but even when this is the case, the preceding example of $\bar{S}$ and S' shows that the use of approximate models is perfectly justified.

Selecting a model leading to a minimal risk amounts to minimize R_m, which is practically impossible since, by (1.9), it depends on the unknown s. Nevertheless, one would like to build an estimator $\tilde{s}$, such that

$$\mathbb{E}\left[\|\tilde{s}-s\|^2\right] \leq C \inf_{m\in\mathcal{M}} \left\{\mathbb{E}\left[\|\hat{s}_m - s\|^2\right]\right\}, \tag{1.10}$$

at least for a large class of functions s. Moreover, it would be desirable that C be as close to one as possible. Note, as explained in Birgé and Massart (2001, Section 2.3.3), that it is generally impossible to get it whatever s.

1.4 Penalized projection estimators

In view of the definition of $\hat{s}_m$ as the minimizer, with respect to $t \in S_m$ of $\gamma(t) = \|t\|^2 - 2Y(t)$, one could think of choosing some model m by minimizing $\gamma(\hat{s}_m)$ with respect to $m \in \mathcal{M}$, but this is obviously a bad idea if one aims at minimizing the risk, since if $S_m \subset S_{m'}$ (and $S_m \neq S_{m'}$), then $\gamma(\hat{s}_{m'}) > \gamma(\hat{s}_m)$ a.s. This implies that, if we consider an increasing sequence of models $S_{m_1} \subset \ldots \subset S_{m_p}$, the criterion will systematically choose the larger model which may be of very large dimension, and this is definitely not satisfactory in view of (1.9). One simple way out of this is to compensate the phenomenon by adding to γ a penalty depending on the model we use and which is, roughly speaking, increasing with the model dimension. This leads to choosing $\tilde{s}$ as the minimizer for all m and $t \in S_m$ of the penalized criterion $\gamma(t) + \text{pen}(m)$. Alternatively,

$$\tilde{s} = \hat{s}_{\hat{m}} \quad \text{with } \hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \{\gamma(\hat{s}_m) + \text{pen}(m)\}. \tag{1.11}$$

The method is not new at all and penalized maximum likelihood (which, in our presentation, means to choose $\gamma(t) = -$log-likelihood of t) has been used for decades. The first examples we know about of such criteria are due to Mallows (1973) and Akaike (1969 for FPE, 1973 and 1974 for AIC). Mallows' C_p, which, according to Daniel and Wood (1971) dates back to the early sixties, was designed to solve our initial problem of variable selection: estimating s in (1.5) when the variance of the errors is known (or independently estimated), while Akaike's AIC, which has a more general scope would be used when this variance is unknown. Translated into the framework given by (1.7), Mallows' C_p corresponds to setting $\text{pen}(m) = 2\varepsilon^2 D_m$ in (1.11) with the same choice of γ. Akaike's AIC criterion corresponds to $\gamma(\hat{s}_m)$ being minus the maximum likelihood on the model S_m and, in our case of a known ε, is identical to Mallows' C_p. Both criteria are based on some unbiased estimation of the quadratic risk and aim at choosing a model which minimizes this risk: this is the *efficiency* point of view. Mallows' C_p has been proved by Shibata (1981) to be

asymptotically (when ε goes to zero) efficient at the price of assuming that the true s does not belong to any model in the list.

Another point of view about model selection consists in assuming the existence of a true model of minimal size and to aim at finding it. In our framework, this means that s belongs to some model S_m in the family with minimal dimension and one wants to find it. The following criteria have been designed to find it with probability tending to one when ε goes to zero (and the list of models remains fixed): BIC (Akaike, 1978 or equivalently Schwarz 1978)) and Hannan-Quinn (1979): this is the *consistency* point of view. For a recent analysis of such criteria, see Guyon and Yao (1999).

The distinction between these points of views and the related criteria (with many more explanations and historical references) has been discussed very carefully and nicely in the first chapter of McQuarrie and Tsai (1998) to which we refer the interested reader since a more detailed discussion of the various criteria would only be a weak copy of theirs. In any case, although both points of view have their advantages, they suffer from the same drawback, which is their definitely asymptotic nature. One attempt to solve this problem has been the introduction of a modified version of AIC, namely AICc, by Hurvich and Tsai (1989), which definitely improves on AIC for small sample sizes.

In this paper, we focus on the first point of view: efficiency, but from a nonasymptotic perspective. A first reason for such a choice is that we neither want to assume that the true s does belong to one of the models (which is required for the consistency approach), nor exclude this case as requested for the asymptotic efficiency of Mallows' C_p and related criteria. Another reason is that we want to allow the list of models to depend on ε, since it is of common practical use to introduce more explanatory variables when one has more observations while one would choose parsimonious models, which are likely to be only approximately true, when one has at hand a limited number of data. In any case, the number and choice of the models depends heavily on the number of observations. A major consequence of our approach is the emergence of richer penalty structures than those involved in the classical criteria mentioned above, which are directly connected with the complexity of the family of models at hand. In this sense, the results which are presented below provide a link between those classical criteria and the general methodology of minimum description length (Rissanen, 1978) or minimum complexity (Barron and Cover, 1991) for discrete models.

In the next section we derive sufficient conditions on the penalty functions leading to nonasymptotic risk bounds for penalized estimators. When the number of models is not too large, we show that these bounds imply that (1.10) holds and allow to recover the asymptotic efficiency of Mallows' C_p. We then apply those results to the detection of change points on the mean of a Gaussian signal in Section 3. Section 4 is devoted to negative results showing that the restriction that we imposed on the penalties in our main theorem in Section 2 are actually sharp. Section 5 presents some heuristics for a data-driven choice of the penalty. The remainder of the paper is devoted to the proofs.

2 How should one choose a proper penalty function?

2.1 Introducing general penalties

In the framework of Gaussian Linear processes that we consider here, a first attempt to define general penalties, in view of designing estimates $\tilde{s}$ that achieve (1.10) in a nonasymptotic context and for arbitrary families of models has been given in Birgé and Massart (2001). There, we introduced suitable penalties which, according to the "richness" of the family of models at hand, look like either Mallows'C_p or BIC, or some mixtures of them, or have even more complicated structures. Let us now summarize the corresponding results.

Here and in the sequel, we shall stick to the following framework: we observe the process $Y(t)$ given by (1.7) where Z is a linear isonormal process on $\mathbb{S}$ and s an unknown function in $\boldsymbol{H}$ to be estimated. We consider a countable (possibly finite) collection $\{S_m, m \in \mathcal{M}\}$ of finite dimensional linear subspaces of $\mathbb{S}$ and denote by D_m the dimension of S_m, by s_m the orthogonal projection of s onto S_m and by $\hat{s}_m$ the projection estimator of s on S_m, which is the the minimizer, with respect to $t \in S_m$, of $\gamma(t) = \|t\|^2 - 2Y(t)$.

Typically, the family $\{S_m, m \in \mathcal{M}\}$ is given and one chooses $\mathbb{S}$ to be the linear span of $\cup_{m\in\mathcal{M}} S_m$. In such a case, there exists a version of Z which is linear on $\mathbb{S}$ (Birgé and Massart, 2001, Section 2.3.1). We do not assume that the correspondance $m \mapsto S_m$ is one-to-one since this may be more convenient in some cases as explained in Birgé and Massart (2001, Section 3.1) and allow 0-dimensional models ($S_m = \{0\}$).

To each model S_m with positive dimension, we associate some nonnegative *weight* L_m and assume that the family of weights satisfies the condition

$$\Sigma = \sum_{\{m\in\mathcal{M} \,|\, D_m>0\}} \exp[-D_m L_m] < +\infty, \tag{2.1}$$

which is always possible since $\mathcal{M}$ is countable. Finally given some nonnegative penalty function pen$(\cdot)$ defined on $\mathcal{M}$, we consider the penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ defined by (1.11). Since $\gamma(\hat{s}_m) = -\|\hat{s}_m\|^2$, $\hat{m}$ can alternatively by defined as

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \text{pen}(m) - \|\hat{s}_m\|^2 \right\}. \tag{2.2}$$

Suitable definitions of the penalty function imply that $\hat{m}$ is well-defined and unique almost surely as shown by the following result from Birgé and Massart (2001).

Theorem 1 *Given a family of weights $\{L_m\}_{m\in\mathcal{M}}$ satisfying (2.1) and a penalty function* pen$(\cdot)$ *such that*

$$\text{pen}(m) \geq K\varepsilon^2 D_m \left(1 + \sqrt{2L_m}\right)^2 \quad \textit{for all } m \in \mathcal{M} \textit{ and some } K > 1, \tag{2.3}$$

the penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ defined by (1.11) almost surely exists and is unique. Moreover it satisfies

$$\mathbb{E}\left[\|\tilde{s} - s\|^2\right] \leq C_1(K) \left[\inf_{m\in\mathcal{M}} \left\{ d^2(s, S_m) + \text{pen}(m) \right\}\right] + C_2(K)\varepsilon^2\Sigma, \tag{2.4}$$

where $d(s, S_m)$ denotes the distance from s to the space S_m and C_1, C_2 depend on K only.

As a consequence, if one can find a bounded family of weights ($\sup_m L_m = L < +\infty$) satisfying (2.1) (which is possible when the number of spaces S_m having the same dimension $D_m = D$ is not too large) and if equality holds in (2.9), one derives from (1.9) and (2.4) that

$$\mathbb{E}\left[\|\tilde{s} - s\|^2\right] \leq C_3(K, L)\left[\inf_{m \in \mathcal{M}}\left\{\mathbb{E}\left[\|\hat{s}_m - s\|^2\right]\right\} + \varepsilon^2\Sigma\right], \tag{2.5}$$

This means that, if no estimator in the family is close to perfect for estimating s, i.e. if

$$\inf_{m \in \mathcal{M}} \mathbb{E}\left[\|\hat{s}_m - s\|^2\right] \geq \varepsilon^2,$$

the penalized estimator $\tilde{s}$ satisfies (1.10) with $C = C_3(1 + \Sigma)$ and therefore behaves as well as the best projection estimator in the family, up to the constant C. Such a result immediately leads to various questions and in particular "how should one choose the penalty in order to minimize C?"; "is it possible to get C close to one?", and so on ...

The purpose of the present paper is twofold. First, to give (partial) answers to these questions via an improved version of Theorem 1 which will prove to be quite useful for a practical implementation of penalization methods (to be developed in subsequent papers). In particular, we shall derive a form of penalty which is slightly different from (2.3) and which will be shown to be close to optimal. Our second aim is to provide various lower bounds for the penalty term, in particular for the problem of variable selection in Gaussian regression. Such lower bounds will, in particular, allow us to explain when Mallows' C_p does or does not work and what alternative should be chosen when it does not work. Moreover, although those lower bounds arguments may look quite abstract at first sight, they do have an important practical consequence. From a practical point of view, ε is typically unknown and has to be somehow estimated. The lower bounds arguments allow us to set up a practical method for estimating the penalty function to be used when ε is unknown, as explained in Section 5.

Since the many possible applications and consequences (in particular to adaptation), of results similar to Theorem 1 have been developed at length in Birgé and Massart (2001), we shall not come back to them here and content ourselves to deal with the change point problem for fixed design regression given by (1.4).

2.2 New penalties and the corresponding risk bounds

Keeping in mind the set up and results of the previous section, we see that the performance of penalized projection estimators, with a penalty function given by (2.3), is clearly connected to the choice of the weights L_m. In view of (2.4), they should be as small as possible but nevertheless satisfy (2.1) with a reasonably small constant Σ, of the order of one (say). When going to practical examples (many of them can be found in Birgé and Massart, 2001), one typically encounters three different situations:

1. for each $D \geq 1$ the number of indices m such that $D_m = D$ is not large (bounded by a polynomial function of D, say) and one can therefore choose L_m as a small constant or even a function of D_m which goes to zero when D_m goes to infinity;

2. the number of indices m such that $D_m = D$ is moderate, leading to a choice $L_m = L$ for some constant L of moderate size;

3. the number of indices m such that $D_m = D$ is much larger, typically of order $\left(\begin{array}{c} N \\ D \end{array} \right)$ where N is a large parameter and one has to choose L_m of order $\log(N/D)$.

Such situations lead to very different types of results. As we shall see below, the most favourable case is the first one, since then, one can prove a result of the form (2.5) with C close to one asymptotically (when ε goes to zero). On the other hand, as is already visible from Theorem 1, in the third case, there is typically no hope, in general, to get (2.5) with a C smaller than $\log N$, the second case being an intermediate situation. In order to cover all standard situations, we shall give two different results. The first one gives an all purposes nonasymptotic bound while the second is purely asymptotic and specific to case 1.

Theorem 2 *Given the family of models $\{S_m\}_{m\in\mathcal{M}}$, let us consider a family of nonnegative weights $\{L_m\}_{m\in\mathcal{M}}$ satisfying (2.1), two numbers, $\theta \in (0,1)$ and $K > 2-\theta$, let us set*

$$Q_m = \varepsilon^2 D_m \left(K + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1} L_m \right) \quad \textit{for all } m \in \mathcal{M} \tag{2.6}$$

and assume that there exists a finite (possibly empty) subset $\bar{\mathcal{M}}$ of $\mathcal{M}$ such that the penalty function pen *satisfies*

$$\mathrm{pen}(m) \geq Q_m, \quad \textit{for } m \in \mathcal{M} \setminus \bar{\mathcal{M}}. \tag{2.7}$$

Then the corresponding penalized projection estimator $\tilde{s}$ defined by (1.11) exists a.s. and satisfies

$$\begin{aligned} (1-\theta)\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \quad \leq \quad & \inf_{m\in\mathcal{M}} \left\{ d^2(s,S_m) + \mathrm{pen}(m) - \varepsilon^2 D_m \right\} + \sup_{m\in\bar{\mathcal{M}}} \{Q_m - \mathrm{pen}(m)\} \\ & + \varepsilon^2 \Sigma \left[(2-\theta)^2 (K+\theta-2)^{-1} + 2\theta^{-1} \right], \end{aligned} \tag{2.8}$$

where $d(s,S_m)$ denotes the distance from s to the space S_m. If, in particular,

$$\mathrm{pen}(m) = \varepsilon^2 D_m \left(2 + 3\sqrt{L_m} + 4L_m \right) \quad \textit{whatever } m \in \mathcal{M}, \tag{2.9}$$

then

$$\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq 2 \inf_{m\in\mathcal{M}} \left\{ d^2(s,S_m) + \varepsilon^2 D_m \left[1 + 3\sqrt{L_m} + 4L_m \right] \right\} + 17\varepsilon^2\Sigma. \tag{2.10}$$

Remark: It may look strange, at first sight, to introduce a nonempty set $\bar{\mathcal{M}}$ on which (2.7) may be violated rather than assuming that it holds for all m. There are two reasons for that. The first is to show that if (2.7) does not hold for some values of m corresponding to spaces S_m of low dimension, then the consequence on the risk may be quite limited. The second reason is that it will also be necessary to introduce

it, in Section 4.1 below, in order to prove lower bounds on the penalty function and show that (2.7) has to hold, at least when D_m is large, for the penalized estimator to behave nicely.

If we assume that the upper bound (2.8) on the risk of $\tilde{s}$ is sharp (up to multiplicative constants), we see that there is no hope to get (1.10) if the L_ms are unbounded or with a small value of C when they are large. On the other hand, if one can choose a family of weights such that L_m is small when D_m becomes large, then the next proposition, which is in the spirit of Shibata (1981), Li (1987), Polyak and Tsybakov (1990) or Kneip (1994), is more suitable.

Proposition 1 *Assume that it is possible to find a family of weights $(L_m)_{m\geq 0}$ such that*

$$\sum_{m\in\mathcal{M}} \exp[-\eta L_m D_m] < +\infty \quad \textit{for all } \eta > 0 \tag{2.11}$$

and the fonction $L(D) = \sup\{L_m, m \in \mathcal{M} \,|\, D_m = D\}$ is finite and satisfies

$$\lim_{D\to+\infty} L(D) = 0. \tag{2.12}$$

Assume, moreover, that $d(s, S_m) > 0$ for all $m \in \mathcal{M}$ and that the penalty satisfies for some $a, b \geq 0$,

$$2\varepsilon^2 D_m \leq \mathrm{pen}(m) \leq \varepsilon^2 D_m \left(2 + a\sqrt{L_m} + bL_m\right) \quad \textit{for all } m \in \mathcal{M}. \tag{2.13}$$

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

$$\lim_{\varepsilon\to 0} \frac{\mathbb{E}\left[\|s - \tilde{s}\|^2\right]}{\mathbb{E}\left[\inf_{m\in\mathcal{M}} \|s - \hat{s}_m\|^2\right]} = 1.$$

The proofs of Theorem 2 and Proposition 1 being quite technical will be defered to Section 6.2.

3 Application: change points in a Gaussian signal

3.1 Two change points problems

As was previously mentioned, many applications of Theorem 1 to variable selection for the classical linear regression framework (1.1) and to the construction of minimax and adaptive estimators over various function spaces have been given in Birgé and Massart (2001, Sections 5 and 6). Although Theorem 2 and Proposition 1 improve on Theorem 1, the corresponding modifications of the theoretical treatment of variable selection are straightforward and need not be considered here. As to the practical treatment, the main problem is to define suitable numerical values for the penalty function in both cases of ordered and complete variable selection. Although the theoretical results given in this paper provide a few hints and a starting point in this direction, a much more precise analysis is needed, based on heavy simulation studies, in order to define a practically good penalty function, which means one leading to an

almost minimal value of the risk. Such studies, in the case of variable selection, will be developed in subsequent papers.

Here we shall concentrate on the following problem: we observe a signal in Gaussian white noise and the signal is constant for a while and then changes to another level. Neither the level, nor the location and number of the change points are known. In brief, the signal has the form

$$s = \sum_{j=1}^{p} \beta_j \mathbb{1}_{I_j} \quad \text{with } I_j = [a_{j-1}, a_j) \quad \text{and} \quad a_0 = 0 < a_1 < \ldots < a_p = 1 \tag{3.1}$$

and we observe on $[0,1]$ the process X given by (1.8) or equivalently, as explained in the introduction, the linear Gaussian process $Y(t) = \langle s, t\rangle + \varepsilon Z(t)$ for $t \in \mathbb{S} \subset \mathbb{L}_2([0,1])$.

¿From a practical point of view, one often doesn't observe X in continuous time, but rather in discrete time, at times $x_1, \ldots, x_n$, which means that our set of observations is the set $Y_1, \ldots, Y_n$ as defined by (1.4). In this case, we consider the Hilbert space $\boldsymbol{H}$ of functions on the set $\mathcal{X} = \{x_1, \ldots, x_n\}$ with scalar product $\langle u, v\rangle = n^{-1}\sum_{i=1}^{n} u(x_i)v(x_i)$, which is an n-dimensional linear space. Clearly, any element $t \in \boldsymbol{H}$ can be identified to the vector with coordinates $t(x_i)$ in $\mathbb{R}^n$. To put it in the Gaussian Linear process form, starting from $Y_i = s(x_i) + \sigma\xi_i$, it suffices to set, for any $t \in \boldsymbol{H}$, $Y(t) = \langle \boldsymbol{Y}, t\rangle$ where $\boldsymbol{Y}$ denotes the vector with coordinates Y_i. Then $Y(t) = \langle s, t\rangle + \sigma\langle t, \boldsymbol{\xi}\rangle$ where $\boldsymbol{\xi}$, with independent coordinates ξ_i, is a standard Gaussian vector in $\mathbb{R}^n$. Obviously, $Z(t) = \sqrt{n}\langle t, \boldsymbol{\xi}\rangle = n^{-1/2}\sum_{i=1}^{n} t(x_i)\xi_i$ is a linear isonormal process and $Y(t)$ can be written in the form (1.7) with $\varepsilon = \sigma/\sqrt{n}$. Without loss of generality, one can assume that the observation times x_i are given in increasing order with $x_1 = 0$ and $x_n < 1$. Setting $x_{n+1} = 1$, one can again represent the signal s in the form (3.1) with the a_j's coinciding with some of the x_i's.

The problem of detecting the change points in a piecewise constant signal has already been considered by Yao (1988) and more recently by Lavielle and Moulines (2000) but their point of view was quite different since it was asymptotic and they assumed a fixed number of change points. Their purpose was then to detect and estimate consistently all those change points while our aim is to estimate the function s with a small quadratic risk for a given value of ε or n. This is a situation where it might be better to ignore some of the change points corresponding to small jumps of s.

3.2 Estimation in continuous time

Let us define the subset $\mathcal{J}$ of $\mathbb{N}^2$ by

$$\mathcal{J} = \{(1,1)\} \bigcup \{(N, D),\ N \geq 2,\ 2 \leq D \leq N\}$$

and consider, for each $(N, D) \in \mathcal{J}$, the set $\mathcal{M}_{N,D}$ of all subsets m of cardinality $D-1$ of $\{1; \ldots; N-1\}$, with $m = \emptyset$ when $N = D = 1$. We set $i_0 = 0$, $i_D = N$ and given $m = \{i_1 < i_2 < \ldots < i_{D-1}\} \in \mathcal{M}_{N,D}$, we define the D-dimensional linear space S_m

by

$$S_m = \left\{ \sum_{j=1}^{D} \beta_j \mathbb{1}_{I_j} \;\middle|\; \boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)^t \in \mathbb{R}^D \right\} \quad \text{with } I_j = [i_{j-1}/N, i_j/N).$$

In particular, $S_\emptyset$ is the one-dimensional linear space generated by $\mathbb{1}_{[0,1)}$. We finally set $\mathcal{M} = \cup_{(N,D)\in\mathcal{J}} \mathcal{M}_{N,D}$ and $L_m = 1 + \log(N/D) + (3\log N)/D$ when $m \in \mathcal{M}_{N,D}$. Since the cardinality of $\mathcal{M}_{N,D}$ is given by

$$|\mathcal{M}_{N,D}| = \left(\begin{array}{c} D-1 \\ N-1 \end{array} \right) \leq \left(\begin{array}{c} D \\ N \end{array} \right) \leq \left(\frac{eN}{D} \right)^D,$$

it follows that

$$\begin{aligned} \sum_{m\in\mathcal{M}} \exp(-L_m D_m) &\leq e^{-1} + \sum_{N\geq 2} \sum_{D=2}^{N} \left(\frac{eN}{D} \right)^D \exp[-D - D\log(N/D) - 3\log N] \\ &= e^{-1} + \sum_{N\geq 2} \frac{1}{N^2} = \frac{\pi^2}{6} + e^{-1} - 1. \end{aligned}$$

An immediate application of Theorem 2 shows that, if $\tilde{s}$ denotes the penalized projection estimator with penalty function given by (2.9), then

$$\mathbb{E}\left[\|s - \tilde{s}\|^2 \right] \leq C \inf_{(N,D)\in\mathcal{J}} \left\{ \left(\inf_{m\in\mathcal{M}_{N,D}} d^2(s, S_m) \right) + \varepsilon^2 D \left[1 + \log \frac{N}{D} \right] \right\}.$$

In particular, a signal of the form $a\varepsilon^2 \mathbb{1}_{[(j-1)/N; j/N)}$ for some positive integer $j \leq N$, with a large value of a will be estimated with a risk smaller than $3C\varepsilon^2(1 + \log N)$, while a signal of the form $\sum_{j=1}^{N} \beta_j \mathbb{1}_{[(j-1)/N; j/N)}$ will be estimated with a risk smaller than $C\varepsilon^2 N$.

3.3 Estimation in discrete time

In this case, given an ordered subset $m = \{i_1 < i_2 < \ldots < i_{D-1}\}$ of $\{2; \ldots; n\}$ (with $m = \emptyset$ if $D = 1$) and setting $i_0 = 1, i_D = n + 1$, we consider the associated D-dimensional linear subspace of $\boldsymbol{H}$ defined by

$$S_m = \left\{ \sum_{j=1}^{D} \beta_j \mathbb{1}_{I_j} \;\middle|\; \boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)^t \in \mathbb{R}^D \right\} \quad \text{with } I_j = [x_{i_{j-1}}, x_{i_j}).$$

Defining by $\mathcal{M}$ the set of all possible distinct subsets m when D varies from 1 to n, we use the family $\{S_m\}_{m\in\mathcal{M}}$ to define a penalized projection estimator of s. It follows that the number of models S_m such that $D_m = D$ is $\left(\begin{array}{c} D-1 \\ n-1 \end{array} \right)$. Computations quite similar to those of the previous section together with an application of Theorem 2 show that a penalized projection estimator $\tilde{s}$ associated to the models S_m, the weights

$L_m = L + \log(n/D_m)$ with $L > 1$, for $m \in \mathcal{M}, m \neq \emptyset$ and $L_\emptyset = L$ and penalty function given by (2.9), satisfies a risk bound of the form

$$\begin{aligned}&\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \\ &\quad \leq \quad C\left[\inf_{m \in \mathcal{M}\setminus\emptyset}\left\{d^2(s, S_m) + \varepsilon^2 D_m\left[1 + \log\frac{n}{D_m}\right]\right\} \bigwedge \left(d^2(s, S_\emptyset) + \varepsilon^2\right)\right]. \quad (3.2)\end{aligned}$$

The presence of the $\log(n/D_m)$ factor in the risk when $D_m \geq 2$ is indeed necessary, from the minimax point of view. This could be proved by the same arguments we used for Proposition 2 of Birgé and Massart (1998) or Theorem 5 of Birgé and Massart (2001).

¿From the minimax point of view, Bound (3.2) is unimprovable without additional information on s, apart from the value of the constant C, but it is actually possible to improve on it from a different point of view. Let us, for instance, imagine that the sequence $(x_i)_{1 \leq i \leq n+1}$ is equispaced, i.e. $x_i = (i-1)/n$, and one suspects that the change points in the representation of s are predictable since their succession follows a simple rule. One can, for instance, imagine that we observe at the discrete times $x_i, 1 \leq i \leq n$ a function $s(x)$ which jumps at times separated by some interval b/n with $1 \leq b$. Since there is no reason why the observation period $[0, (n-1)/n]$ be connected with the jumps, the discretized change points are of the form $x_{i_j} = (i_j - 1)/n$ with

$$i_j = 1 + \lceil a + (j-1)b \rceil, \quad 0 < a \leq b \quad \text{and} \quad \lceil x \rceil = \inf\{k \in \mathbb{N} \,|\, n \geq x\}. \qquad (3.3)$$

Then the set $\{x_{i_j}\}_{1 \leq j \leq D-1}$ of observable change points is entirely described by the set of integers $m = \{i_j \,|\, 1 \leq j \leq D-1\}$ with $i_1 < i_2 < \ldots < i_{D-1}$ where $x_{i_{D-1}}$ denotes the last observable change-point in $[0, (n-1)/n]$, which means that D is defined by

$$D = \inf\{j \in \mathbb{N} \,|\, a + (j-1)b > n - 1\}.$$

Then $1 \leq D \leq n$, $D = 1$ means that $m = \emptyset$ and the definition of D implies that

$$(D-1)a \leq n - 1 \qquad \text{and for } D \geq 3, \quad \frac{n-1-a}{D-2} \geq b > \frac{n-1-a}{D-1}.$$

It follows from these inequalities that, given $D \geq 3$, the number of possible choices for i_1 is bounded by $\lceil (n-1)/(D-1) \rceil \leq 2n/D$ and the number of possibilities for $\lceil b \rceil$ as well. Let $\mathcal{M}'$ be the set of those m's corresponding to these special sequences $(i_j)_{1 \leq j \leq D-1}$ given by (3.3). Since $i_j - i_{j-1}$ is either $\lceil b \rceil$ or $\lceil b \rceil - 1$, it follows that, for $D \geq 3$,

$$|\{m \in \mathcal{M}' \,|\, D_m = |m| + 1 = D\}| \leq (2n/D)^2 2^{D-2} = 2^D (n/D)^2$$

and this bound remains true when $D \leq 2$. This means that, for $m \in \mathcal{M}'$, the value of L_m can be reduced to $(2\log n)/D_m + \log 2$. The resulting risk bound (derived from Theorem 2) substancially improves on (3.2) when s belongs to some S_m such that m is generated by (3.3) and D_m is larger than $\log n$.

4 Some potential difficulties connected with bad penalty choices

It follows from Theorem 2 that a proper choice of the penalty should be of the form

$$\text{pen}(m) = \varepsilon^2 D_m \left(K + a\sqrt{L_m} + bL_m \right) \quad \text{with } K > 1, \quad a > 2 \quad \text{and} \quad b > 2,$$

and the computations of Section 6.2 show that the limiting condition

$$\text{pen}(m) > \varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m \right) \tag{4.1}$$

is required for our proof of Theorem 2 to work, which, of course, does not mean that a smaller choice of the penalty should necessarily lead to a bad estimator. Similarily, the choice of a large value of K leads to larger upper bounds for the risk, but this does not mean that the risk itself is necessarily larger. In order to choose the penalty in a satisfactory manner, it is therefore desirable to know whether the restrictions which come out from our proofs, namely that (4.1) holds and K is not too large, are indeed necessary or not. The following sections will be devoted to show that those restrictions are actually perfectly justified, in the sense that, if those conditions are violated, the penalized estimator can behave quite poorly for some values of the unknown parameter s. Although the forthcoming results are not sufficient to decide precisely what form of penalty is the most adequate in a specific situation and may therefore look rather abstract at first sight, they are indeed very useful for the practical implementation of penalized estimators since they do provide the necessary guidelines for the preliminary choices of the numerical parameters involved in the penalty and the setting of the extensive simulations that are required to optimize those parameters in a given situation.

4.1 Lower bounds for the penalty term

4.1.1 Position of the problem

Our aim here in this section, is to show that the lower bound (4.1) on the penalty term is, in some sense, necessary. It is actually not obvious to give a precise formal meaning to this claim since this lower bound depends on the weights L_m which are not uniquely defined. If, for instance, (2.1) holds with $L_m = L$ for all $m \in \mathcal{M}$ whatever $L > 0$, and we choose $L_m = 5$ for all m, it is clear, from Theorem 2, that a penalty violating (4.1) for all m, such as $\text{pen}(m) = 2\varepsilon^2 D_m$, still leads to a good penalized estimator. This emphasizes the fact that the problem of showing the necessity of (4.1) is ill-posed without further restrictions on the values of the weights L_m.

In order to overcome this difficulty, we shall first restrict our attention to some particular, although quite common, situation, where the number of models such that $D_m = D$ is finite for each integer D. The L_ms are of course allowed to be very different from one m to another, but since they are chosen by the statistician, a typical choice, in this case, is $L_m = L(D_m)$ for some positive function L. Many illustrations of this fact have been given in Birgé and Massart (2001). Setting

$$H(D) = D^{-1} \log |\{m \in \mathcal{M} \,|\, D_m = D\}|, \tag{4.2}$$

(2.1) requires that

$$\sum_{D\geq 1} \exp[-D[L(D) - H(D)]] < +\infty.$$

Choosing $L(D) = H(D) + \delta$ with $0 < \delta \leq 1/2$, $\theta = 1 - \delta$ and $K = 1 + 2\delta$, we get $\Sigma \leq \left(e^{\delta} - 1\right)^{-1}$ and

$$\begin{aligned} Q_m &= \varepsilon^2 D_m \left[1 + 2\delta + 2(1+\delta)\sqrt{H(D_m) + \delta} + 2(1-\delta)^{-1}[H(D_m) + \delta]\right] \\ &\leq \left(1 + 8\sqrt{\delta}\right) \varepsilon^2 D_m A(D_m), \end{aligned}$$

where

$$A(D) = 1 + 2\sqrt{H(D)} + 2H(D). \tag{4.3}$$

This implies, by Theorem 2, that any penalty of the form

$$\text{pen}(m) = (1+\eta)\varepsilon^2 D_m A(D_m), \quad \eta > 0 \quad \text{for all } m \in \mathcal{M} \text{ such that } D_m > \bar{D}, \tag{4.4}$$

satisfies (2.7) provided that δ is small enough and results in a risk bound of the form

$$\begin{aligned} &\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \\ &\quad \leq C(\eta) \left(\inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 [D_m A(D_m) + 1] \right\} + \varepsilon^2 \sup_{1 \leq D \leq \bar{D}} \{DA(D)\} \right). \end{aligned} \tag{4.5}$$

Our purpose in the next three sections will be to prove that if (4.4) is violated, i.e.

$$\text{pen}(m) \leq (1-\eta)\varepsilon^2 D_m A(D_m) \tag{4.6}$$

for some $\eta > 0$ and D_m sufficiently large, then the risk $\mathbb{E}\left[\|s - \tilde{s}\|^2\right]$ can be arbitrarily large, even if $s = 0$ or the estimator $\tilde{s}$ may even be undefined. The reason for focusing on large values of D_m only is that (4.6) is compatible with (4.4) provided that $D_m \leq \bar{D}$ and that the term $\sup_{1 \leq D \leq \bar{D}} \{DA(D)\}$ can be considered as an additional constant if $\bar{D}$ is not large. It is only by letting $\bar{D}$ go to infinity that we can make the bound (4.5) blow up.

The behaviour of $A(D)$ when D is large actually depends on the size of $H(D)$. If $H(D)$ is small, $A(D)$ is close to one; if $H(D)$ is large, then $A(D)$ is equivalent to $2H(D)$ while, for moderate values of $H(D)$ none of the three terms defining $A(D)$ can be ignored. This will lead us to distinguish between those three cases to prove the bad behaviour of some penalized estimators when (4.6) holds for some m for which D_m is large enough. We shall be able to exhibit this phenomenon whatever the family of models whenever $H(D)$ is small for large D and in the case of complete variable selection to illustrate the case where $H(D)$ can be arbitrarily large. The intermediate situation of moderate values for $H(D)$ is more delicate and requires a special and somehow less natural construction of a family of models but with the advantage that it shows that the structure of $A(D)$ is the right one to define a lower bound for the penalty. It is worth mentioning that each of those three cases corresponds to various interesting examples. The case of $H(D)$ tending to zero when D goes to infinity corresponds to ordered variable selection for which there is only one model per

dimension; the opposite case when $H(D)$ can be arbitrarily large arises in complete variable selection with a large number of variables, while the intermediate case occurs for the parsimonious variable selection strategy connected with adaptive estimation in Besov balls described in Birgé and Massart (2001, Section 6.4) or with the histogram selection pruning procedure associated with CART (Gey and Nédélec, 2001).

4.1.2 A small number of models

In this section, and in the following one as well, we restrict ourselves to a quite common situation: we are given an orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda_N}$ such that $|\Lambda_N| = N$ and $\{\Lambda_m\}_{m\in\mathcal{M}}$ is some family of subsets of Λ_N which includes the largest possible one Λ_N ($N \in \mathcal{M}$). Then we define S_m as the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ which gives $D_m = |\Lambda_m|$ and, in particular, $D_N = N$.

We assume here that, for each $D \geq 1$, the number of elements $m \in \mathcal{M}$ such that $D_m = D$ grows at most polynomially with respect to D, or, more generally, that $H(D) \leq \bar{H}(D)$ for some function $\bar{H}(j)$ converging to zero when j goes to infinity, which implies that $\sum_{\{m\in\mathcal{M}\,|\,D_m>0\}} \exp[-D_m L] \leq \Sigma_L$ independently of N, whatever $L > 0$. It is therefore possible, at the price of a large value of Σ, to choose $L_m = L$ for all $m \in \mathcal{M}$ with L arbitrary close to zero. It follows that any penalty of the form $\mathrm{pen}(m) = (1+\eta)\varepsilon^2 D_m$ with $\eta > 0$ satisfies (2.7) with $\bar{\mathcal{M}} = \emptyset$, provided that L, $1-\theta$ and $K-1$ are small enough, depending on η, which results, by Theorem 2, in a risk bound of the form

$$\mathbb{E}\left[\|s-\tilde{s}\|^2\right] \leq C(\eta) \inf_{m\in\mathcal{M}} \left\{d^2(s,S_m) + \varepsilon^2(D_m+1)\right\},$$

where $C(\eta)$ goes to infinity with η^{-1}, but independently of N. On the other hand, if $\eta < 0$, one could get inconsistent estimation when N goes to infinity. Such a phenomenon is actually a consequence of the following proposition to be proved in Section 6.4.

Proposition 2 *Assume that there exists some positive number η such that*

$$\mathrm{pen}(N) - \mathrm{pen}(m) \leq (1-\eta)\varepsilon^2(N - D_m), \tag{4.7}$$

for any $m \in \mathcal{M}$ and that the number of elements $m \in \mathcal{M}$ such that $D_m = D$ is finite and bounded by $\exp[DH(D)]$ with $H(D) \leq \bar{H}(D)$ for some function $\bar{H}(j)$ converging to zero when j goes to infinity. Then, given $\theta, \delta \in (0,1/2)$ there exists a number N_0 depending on $\eta, \theta, \bar{H}$ and δ but neither on s nor on ε such that, for $N \geq N_0$,

$$\mathbb{P}[D_{\hat{m}} > N(1-\theta) - 1] \geq 1-\delta \quad \textit{and} \quad \mathbb{E}\left[\|s-\tilde{s}\|^2\right] \geq d^2(s,S_N) + C(\theta,\delta)\,\varepsilon^2 N,$$

where C depends only on θ and δ.

It is now easy to understand why choosing a penalty of the form $(1-\eta)\varepsilon^2 D_m$ with $\eta > 0$ leads to a bad procedure. In order to illustrate the argument, assume that we are given some orthonormal basis $\{\varphi_j\}_{j\geq 1}$ in $\boldsymbol{H}$ (the trigonometric system or a wavelet basis on $[0,1]$, for instance) and that S_m is the linear span of $\{\varphi_1,\ldots,\varphi_m\}$ for $m \in \mathbb{N}$, with $S_0 = \{0\}$. Then $D_m = m$. For $\mathcal{M}$ we have the choice among any of

the sets $\{m \leq N\}$ with $1 \leq N < \infty$. If we set $\mathrm{pen}(m) = 2\varepsilon^2 m$, for all m, it follows from Theorem 2 that, whatever s, the risk will be bounded independently of N by

$$\mathbb{E}\left[\|s - \hat{s}\|^2\right] \leq C \inf_{m \in \mathcal{M}} \left\{d^2(s, S_m) + \varepsilon^2(m+1)\right\}, \tag{4.8}$$

for a suitable constant C. In this case one would choose N to be as large as is computationally feasible (in practice, the number of models is always finite!) and get the optimal bias versus variance trade-off, apart from the constant C. The situation becomes completely different if $\mathrm{pen}(m) = (1-\eta)\varepsilon^2 m$. In this case, Proposition 2 shows that the risk becomes larger than $C' N \varepsilon^2$ for N large enough. Large values of N therefore lead to terrible results if, for instance, $s = 0$. Alternatively, if we choose a moderate value of N, in order to avoid this phenomenon there is a serious possibility that $d^2(s, S_N)$ be quite large because even the largest model is grossly wrong, resulting in an exceedingly large risk as compared to the bound given by (4.8) for a larger value of N.

4.1.3 A large number of models

We consider the same framework as in the previous section but now assume that the number of models having the same dimension D grows much faster with D. More precisely, we take for $\mathcal{M}$ the set of all subsets of Λ_N, set $\Lambda_m = m$ and we assume that $N = |\Lambda_N|$ is large. Moreover the penalty function $\mathrm{pen}(m)$ only depends on m through its cardinality $|m|$ which is the dimension D_m of S_m.

Proposition 3 *Let s be the true unknown function to estimate and set $\Lambda_1 = \{\lambda \in \Lambda_N \,|\, \langle s, \varphi_\lambda \rangle \neq 0\}$. Assume that there exist numbers δ, α, A and η with*

$$0 \leq \delta < 1, \quad 0 \leq \alpha < 1, \quad A > 0, \quad \textit{and} \quad 0 < \eta < 2(1-\alpha),$$

and some $\bar{m} \in \mathcal{M}$ with

$$|\Lambda_1| \leq \delta |\bar{m}|, \quad |\bar{m}| \leq A N^\alpha \quad \textit{and} \quad \mathrm{pen}(\bar{m}) \leq (2 - 2\alpha - \eta)(1-\delta)\varepsilon^2 |\bar{m}| \log N.$$

Then one can find two positive constants κ and N_0, depending on δ, α, A and η, such that

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq \kappa \varepsilon^2 |\bar{m}| \log N \quad \textit{for all } N \geq N_0.$$

The proof is given in Section 6.5. Let us now consider what are the consequences of this result. In the present framework, a suitable choice of weights is $L_m = \log(N/D_m) + 1 + 2(\log D_m)/D_m$ since then

$$\begin{aligned}
\sum_{\{m \in \mathcal{M} \,|\, D_m > 0\}} \exp[-L_m D_m] &= \sum_{D=1}^{N} \binom{N}{D} \frac{1}{D^2} \exp[-D \log(N/D) - D] \\
&< \sum_{D=1}^{N} \frac{1}{D^2} (eN/D)^D \exp[-D \log(N/D) - D] \\
&< \pi^2/6 - 1.
\end{aligned}$$

It then follows from Theorem 2 that, if the penalty takes the form

$$\text{pen}(m) = (1+\eta)\varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m\right) \quad \text{with } \eta > 0, \tag{4.9}$$

then

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leq C(\eta) \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 D_m [1 + \log(N/D_m)] \right\},$$

whatever s. In particular, if s satisfies the assumptions of Proposition 3 with $|\Lambda_1| \geq 3$,

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leq C(\eta)\varepsilon^2 |\Lambda_1| \log N.$$

On the other hand, by Proposition 3, under the same assumptions, if

$$\begin{aligned} \text{pen}(\bar{m}) &\leq (2 - 2\alpha - \eta)(1-\delta)\varepsilon^2 |\bar{m}| \log N \\ &= \frac{(1-\delta)(1-\alpha-\eta/2)}{1-\alpha} \varepsilon^2 |\bar{m}| [2(1-\alpha)\log N], \end{aligned} \tag{4.10}$$

then

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq \kappa \varepsilon^2 |\bar{m}| \log N,$$

when N is large enough. This implies that, for large values of N, the estimator built on some too small value of the penalty of the form $(1-\eta)\varepsilon^2 D_m \log N$ with $\eta > 0$, will have a risk which is much larger that one would get with a larger penalty, the ratio tending to infinity with N. It suffices to assume that $|\Lambda_1| = o(|\bar{m}|)$ when $N \to +\infty$ to see it. Comparing (4.9) with $m = \bar{m}$ and $|\bar{m}| \sim N^\alpha$ together with (4.10), we see that

$$\text{pen}(m) = \varepsilon^2 D_m [1 + 2\log(N/D_m)] \tag{4.11}$$

is the borderline formula for the penalty, at least when N is very large and D_m of order N^α with $0 < \alpha < 1$. Of course, such a phenomenon is definitely of an asymptotic nature. Further consequences of the choice of too small penalties in connection with threshold estimators are given in Birgé and Massart (2001, Section 6.3.4).

We are now in a position to explain to what extent classical criteria like Mallows' C_p are or are not suitable for particular situations. In order to make our discussion simple, let us focus on the problem which has motivated our study, namely the problem of variable selection connected to the Gaussian linear regression set up (1.1). Deciding which variables should enter a regression model is an important problem in Econometrics, and, in order to make our discussion precise, we should distinguish between two situations: *ordered variable selection* amounts to select only sets of variables of the form $\{X^j\}_{1 \leq j \leq k}$ with $k \leq p$, while *complete variable selection* corresponds to select any subset of the set of p variables. Although many econometric books do deal with this subject, most of them become indeed rather elusive (see for instance Chapter 2 of Amemiya, 1985) as to the choice of a suitable penalty for the second situation and some (Draper and Smith, 1981 p. 299) even suggest that one could then use Mallows' C_p (or Akaike's AIC) in this case. Even the careful study of McQuarrie and Tsai (1998) does not distinguish quite explicitely between the two situations of ordered and unordered variable selection. They do explain (p. 64) that the multiplicity of competing models of the same dimension makes a difference but do not persue their analysis further.

It follows from Proposition 3 that the use of Mallows' C_p (or more generally of underpenalized criteria) can lead to terrible results when the number of available variables is large and that a heavier penalty should be used in such a case. Even for small sample sizes and number of variables, simulation studies such as those proposed by McQuarrie and Tsai (1998, p. 62) show that stronger penalties should be prefered to C_p. This suggests that although our lower bound (4.11) for the penalty follows from asymptotic considerations, it seems to be quite relevant for practical nonasymptotic use.

4.1.4 A general lower bound

In order to deal with the intermediate case corresponding to $H(D)$, as defined by (4.2) being neither small, nor large when D is large, we have to introduce a more complicated set up. Let us consider the following situation: we have at hand a family of models $\{S_m\}_{m\in\mathcal{M}}$ such that $\mathcal{M} = \cup_{D\in\mathbb{N}}\mathcal{M}_D$. We assume that $\mathcal{M}_0$ has only one element denoted by $\emptyset$ and that $S_\emptyset = \{0\}$. For each $D \geq 1$, $\mathcal{M}_D$ is finite and nonempty with cardinality $|\mathcal{M}_D| = \kappa(D)$ and all the models S_m with $m \in \mathcal{M}_D$ are orthogonal to each other with the same dimension $D_m = D$. Moreover,

$$\exp(\alpha D) - 1 < \kappa(D) \leq \exp(\alpha D) \quad \text{for some } \alpha > 0. \tag{4.12}$$

In such a case, a suitable choice of the weights is

$$L_m = L(D_m) = \alpha + \beta D_m^{-1}\log(D_m + 1) \quad \text{with } \beta > 1, \tag{4.13}$$

which implies that $\Sigma \leq 2\sum_{n=2}^{+\infty} n^{-\beta} < +\infty$.

If $s = 0$, the ideal estimator is obviously $\hat{s}_\emptyset = 0$ since its risk is zero and it immediately follows from Theorem 2 that the risk of a suitably tuned penalized estimator will be bounded by $17\Sigma\varepsilon^2$. On the other hand, if (4.1) is violated for large values of D_m, the corresponding estimator may behave very badly in the sense that its risk may be arbitrarily large. More precisely, we will prove the following in Section 6.6.

Proposition 4 *Assume that the family of models at hand is as described just before, that L_m is given by (4.13) with $\alpha > 0$ and $\beta > 1$ and that $s = 0$. There exists some function F on $(0, +\infty)$ such that $5/6 < F(x) < 1$ for $x > 0$ and $F(x)$ converges to one when x converges either to 0 or to infinity with the following property: let λ belong to $(0, F(\alpha))$, $\bar{D}$ be some large enough integer, depending on α, β and λ and define*

$$\bar{\mathcal{M}} = \left\{ m \in \mathcal{M} \;\middle|\; D_m \geq \bar{D} \quad \text{and} \quad \mathrm{pen}(m) \leq \lambda\varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m\right) \right\}. \tag{4.14}$$

- *If $\bar{\mathcal{M}}$ is infinite, then, with a probability larger than $1/2$, $\inf_{m\in\mathcal{M}}\{\hat{\gamma}(m) + \mathrm{pen}(m)\} = -\infty$ and $\hat{m}$ is not defined.*
- *If $\bar{\mathcal{M}}$ is nonempty, finite and (2.7) holds, then*

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq C(\alpha, \lambda)\varepsilon^2 \left(\sup_{m\in\bar{\mathcal{M}}} D_m\right), \tag{4.15}$$

where C depends only on α and λ.

This proposition says that, up to some factor $F(\alpha) \in (5/6, 1)$, the lower bound (4.1) is tight. Since $F(\alpha)$ is close to one when α is either small or large, we recover, in these situations, the lower bounds $\text{pen}(m) > \varepsilon^2 D_m$ and $\text{pen}(m) > 2\varepsilon^2 D_m L(D_m)$ which derive from (4.1), but under much more restrictive assumptions than those we used in Propositions 2 and 3. For moderate values of α (around 1), there is only a slight difference between the condition (2.7) which is required in Theorem 2 and the lower bound on the penalty function given by Proposition 4. This is due to the fact that the proof of Theorem 2 relies on some large deviation inequalities based on approximations of Laplace transforms, rather than the true ones. Such approximations are justified by the fact that they lead to simple inversion formulas while the use of the true Laplace transforms would lead to untractable inversions. This is at the price of some lack of tightness in our deviation formulas which explains this loss (compare, for instance, Corollary 1 and (A.1) with $\rho = 0$ and $b = 2$).

4.2 The effect of choosing too large penalties

It follows from the preceding results that the choice $K > 1$ in (2.6) is perfectly justified. It moreover follows from Proposition 1 that $K = 2$ should often be recommended. This suggests to choose a penalty of the form (2.9) or something reasonnably close to it. In order to analyze what would be the effect of a substantially larger penalty we can use the next theorem which covers many typical examples. Its proof is given in Section 6.7.

Theorem 3 *Let us assume that the set $\mathcal{M}$ contains two specific elements 0 and 1 such that $S_0 = \{0\}$ and $D_1 = 1$ and that the weights L_m satisfy (2.1) with $\Sigma < 1$. Let $\tilde{s}$ be the penalized projection estimator corresponding to a penalty such that $\text{pen}(0) = 0$ and*

$$\text{pen}(m) \geq \varepsilon^2[(3/2)D_m + 4L_m D_m + 2A] \quad \textit{for all } m \in \mathcal{M}^\star = \mathcal{M} \setminus \{0\}. \tag{4.16}$$

Then there exists some $s \in S_1$ such that

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq A(1 - \Sigma)\varepsilon^2, \tag{4.17}$$

while

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leq \varepsilon^2 \left[2\left(1 + 3\sqrt{L_1} + 4L_1\right) + 17\Sigma\right], \tag{4.18}$$

if the penalty is given by (2.9).

Remark: The theorem assumes the existence of a model S_0 with dimension 0 in the family. It would of course be possible to prove an analogous result without it, provided that there exist some 1 and 2 dimensional models and choosing a suitable s in the two-dimensional space. The proof would be quite similar.

It immediately follows from a comparison between (4.17) and (4.18) that a value of A substantially larger that $1 \vee L_1$ would lead to a large increase of the risk for some parameters s. Two specific applications of such a result are as follows. First assume that $\mathcal{M} = \mathbb{N}$, $S_0 = \{0\}$ and for $m \geq 1$, S_m is the linear span of $\{\varphi_1, \ldots, \varphi_m\}$ where

$\{\varphi_j \,|\, j \geq 1\}$ is an orthonormal system. We can then choose $L_m = 1$ for $m \geq 1$ which implies that $\Sigma = (e-1)^{-1}$ and that $\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leq \varepsilon^2[16 + 17/(e-1)]$ if the penalty is given by (2.9), i.e. $\mathrm{pen}(m) = 9\varepsilon^2 D_m$. On the other hand, (4.17) immediately shows that penalties of the form $\mathrm{pen}(m) = C\varepsilon^2 D_m$ with a large value of C should be avoided.

Another interesting illustration is connected to the variable selection problem in linear Gaussian regression, i.e. the problem we considered in Section 1.1. Let us choose some orthogonal system $\{\varphi_1, \ldots, \varphi_N\}$ in $\boldsymbol{H}$ and let $\mathcal{M}$ be the set of all subsets of $\{1, \ldots, N\}$. For $m \in \mathcal{M}$, S_m is the linear span of $\{\varphi_j \,|\, j \in m\}$ with $S_\emptyset = \{0\}$. If $L_m = 2 + \log(N/D_m)$ for $m \neq \emptyset$, it follows from Birgé and Massart (2001, Section 5.1.2) that $\Sigma \leq (e-1)^{-1}$ and the assumptions of Theorem 3 are satisfied. If we set

$$\mathrm{pen}(m) = 5\varepsilon^2 D_m[3 + \log(N/D_m)]$$

and $s = \lambda\varphi_j$ for some j, we derive from Theorem 2 with $\theta = 1/2$ and $K = 2$ that

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] < 10\varepsilon^2[4 + \log N].$$

On the other hand, if $\mathrm{pen}(m) = C\varepsilon^2 D_m[3 + \log(N/D_m)]$ with $C > 4$, then, for $m \in \mathcal{M}^\star$,

$$\begin{aligned} \varepsilon^{-2}\,\mathrm{pen}(m) - (3/2)D_m - 4L_m D_m &= D_m[3C - 9.5 + (C-4)\log(N/D_m)] \\ &\geq (C-4)D_m[3 + \log(N/D_m)] \\ &\geq (C-4)[3 + \log N]. \end{aligned}$$

We may therefore choose $2A = (C-4)[3 + \log N]$ in (4.16) and conclude from (4.17) that $\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq (C-4)[3 + \log N]\varepsilon^2/5$ for some s of the required form. Once again, this shows that large values of C should be avoided.

Unfortunately, although the preceding results give some hints concerning "good" choices of the penalty function, they are definitely not precise enough to allow us to provide an "optimal" choice of the penalty, even in the simplest situations that we just considered. A practically efficient choice of the penalty can only be based on heavy simulations, especially for those m's that correspond to spaces S_m of small dimension.

5 Practical implementation: introducing estimated penalties

Up to now and in the companion paper, Birgé and Massart (2001), we have essentially considered the theoretical approach to penalization in regression since we always assumed that the noise level ε was known and used it freely to build our penalties. Of course, for a practical implementation of the method, we have to estimate it somehow, since in practice, it is typically unknown. We propose here a method, based on a mixture of theoretical and heuristic ideas that has been implemented and tested on various data sets and proved to be fully operative.

Rather than estimating ε, we shall try to estimate the penalty itself, or calibrate it in a close to optimal way, using the data at hand. We assume here that the family of

models $\{S_m\}_{m\in\mathcal{M}}$ contains some models of large dimension, which is not a practical restriction, since one can always add some artificial models of high dimension to those of interest. Then we choose a suitable family of weights, which means that the value of Σ, as defined by (2.1) is not large and that the L_m's have been more or less optimized i.e. are not larger than necessary. In this situation, according to Theorem 2, the risk of the penalized estimator remains under control, provided that the penalty function satisfies (2.6) with a constant K larger than one and not too close to it. Except for very special situations, which indeed correspond to a bad choice of the family of models, if K is not close to one, the infimum in (2.8) will be obtained for some value m with D_m substantially smaller than the dimension of the largest models, which implies that $D_{\hat{m}}$ will also share this property.

On the other hand, it follows from the various results developed in Section 4.1 that, if K is smaller than one, $D_{\hat{m}}$ tends to be close to the dimension of the largest models. Such a phenomenon, which is strikingly visible in practice, suggests the following method to calibrate the penalty. First, from a simulation study, in which ε is known, find a suitable form for the penalty function, i.e. a suitable function F such that a penalty given by $\mathrm{pen}(m) = \varepsilon^2 D_m F(L_m)$ leads to a good value of the risk. Then define $\mathrm{pen}_\lambda(m) = \lambda D_m F(L_m)$ and consider the corresponding model choices $\hat{m}_\lambda$ where $\hat{m}_\lambda$ denotes the minimizer with respect to $m \in \mathcal{M}$ of $\mathrm{pen}_\lambda(m) - \|\hat{s}_m\|^2$. Compute the corresponding values of $D_{\hat{m}_\lambda}$ for slowly increasing values of λ starting from $\lambda = 0$. One then typically observes that for small values of λ, those values stays very large and they suddenly jump to a much smaller value when λ reaches some threshold $\hat{\lambda}$. It follows from the considerations of the preceding section that $\mathrm{pen}_{\hat{\lambda}}$ should be close to the lower bound (4.1) on the penalty. In order to get a good penalty function, it generally suffices to choose $\mathrm{pen}(m) = \kappa\hat{\lambda} D_m F(L_m)$, where κ is a constant close to 2. In practice, one should choose, for each family of models, the value of κ from simulated data.

6 Proofs

6.1 Proving the existence of $\tilde{s}$

We recall that our observation is the process $Y(t)$ given by (1.7) where Z is a linear isonormal process on $\mathbb{S}$ and s an unknown function in $\boldsymbol{H}$. To each $m \in \mathcal{M}$, we associate some orthonormal basis $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ of S_m with $|\Lambda_m| = D_m$. Then the restriction to S_m of the process Z can be written by linearity as

$$Z(t) = \sum_{\lambda\in\Lambda_m} \langle t, \varphi_\lambda\rangle Z(\varphi_\lambda) = \langle t, \zeta_m\rangle \quad \text{with } \zeta_m = \sum_{\lambda\in\Lambda_m} Z(\varphi_\lambda)\varphi_\lambda \in S_m,$$

from which it follows that

$$\zeta_m \sim \mathcal{N}(0, \boldsymbol{I}_m) \qquad \text{and} \qquad V_m = \|\zeta_m\|^2 \sim \chi^2(D_m), \tag{6.1}$$

where $\mathcal{N}(0, \boldsymbol{I}_m)$ denotes the D_m-dimensional standard Gaussian distribution and $\chi^2(D_m)$ the chi-square distribution with D_m degrees of freedom. Recalling that s_m denotes the orthogonal projection of s onto S_m, we derive that the projection

estimator $\hat{s}_m$ on S_m is the minimizer, with respect to $t \in S_m$ of

$$\begin{aligned} \gamma(t) &= \|t\|^2 - 2Y(t) = \|s-t\|^2 - \|s\|^2 - 2\varepsilon Z(t) \\ &= \|s - s_m\|^2 + \|t - s_m\|^2 - \|s\|^2 - 2\varepsilon\langle t, \zeta_m\rangle. \end{aligned} \tag{6.2}$$

Therefore $\hat{s}_m$ is the minimizer with respect to $t \in S_m$ of $\|t - s_m\|^2 - 2\varepsilon\langle t - s_m, \zeta_m\rangle$, which leads to

$$\hat{s}_m = s_m + \varepsilon\zeta_m = s_m + \varepsilon \sum_{\lambda \in \Lambda_m} Z(\varphi_\lambda)\varphi_\lambda,$$

hence

$$\gamma(\hat{s}_m) = \|s - s_m\|^2 - \|s\|^2 - \varepsilon^2 V_m - 2\varepsilon Z(s_m) \tag{6.3}$$

and

$$\|\hat{s}_m - s\|^2 = \|s - s_m\|^2 + \varepsilon^2 V_m. \tag{6.4}$$

Since

$$2\varepsilon|Z(s_m)| = 2|\langle s_m, \varepsilon\zeta_m\rangle| \le \eta^{-1}\|s_m\|^2 + \eta\varepsilon^2 V_m \quad \text{whatever } \eta > 0,$$

it follows from (6.3) that

$$\gamma(\hat{s}_m) \ge -\|s\|^2 - \eta^{-1}\|s_m\|^2 - \varepsilon^2(1+\eta)V_m \ge -\left(1 + \eta^{-1}\right)\|s\|^2 - \varepsilon^2(1+\eta)V_m$$

and from Lemma 1 in the Appendix with $\rho = 0$, $b = 2$ and $x = L_m D_m + \xi$ that

$$\mathbb{P}\left[V_m \ge D_m\left(1 + 2\sqrt{L_m + \xi/D_m} + 2L_m + 2\xi/D_m\right)\right] \le \exp(-L_m D_m - \xi).$$

Under the assumption (2.1), we derive that, on some set Ω_ξ of probability larger than $1 - \Sigma e^{-\xi}$, for all m's simultaneously,

$$\gamma(\hat{s}_m) \ge -\left(1 + \eta^{-1}\right)\|s\|^2 - \varepsilon^2(1+\eta)D_m\left(1 + 2\sqrt{L_m} + 2L_m\right)[1 + \xi/(L_m D_m)].$$

Consequently, if (2.7) holds and η is small enough, depending on K and θ, one gets

$$\gamma(\hat{s}_m) + \mathrm{pen}(m) \ge -\left(1 + \eta^{-1}\right)\|s\|^2 + \eta\varepsilon^2 D_m\left(1 + 2\sqrt{L_m} + 2L_m\right),$$

for all $m \notin \bar{\mathcal{M}}$ such that $L_m D_m \ge \xi\eta^{-1}$. Since $\bar{\mathcal{M}}$ is finite, this implies that $\gamma_n(\hat{s}_m) + \mathrm{pen}(m)$ tends to infinity with $L_m D_m$. By (2.1), there is only a finite number of m's such that $L_m D_m \le n$, whatever the integer n. One therefore concludes that (2.1) and (2.7) imply that there exists a minimizer $\hat{m}$ of $\gamma(\hat{s}_m) + \mathrm{pen}(m)$ on the set Ω_ξ and therefore a.s. since ξ is arbitrary.

6.2 Proof of Theorem 2

Since $\tilde{s} = \hat{s}_{\hat{m}}$ exists a.s., it follows from the definition of $\hat{m}$ that

$$\|s\|^2 + 2\varepsilon Z(s) + \gamma(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) = \inf_{m \in \mathcal{M}}\left\{\|s\|^2 + 2\varepsilon Z(s) + \gamma(\hat{s}_m) + \mathrm{pen}(m)\right\} \tag{6.5}$$

and from (6.3) and (6.4) that

$$\|s\|^2 + \gamma(\hat{s}_{\hat{m}}) + 2\varepsilon Z(s) = \|s - \tilde{s}\|^2 - 2\varepsilon^2 V_{\hat{m}} - 2\varepsilon Z(s_{\hat{m}} - s).$$

Using (6.3) again to evaluate $\gamma(\hat{s}_m)$ we derive from (6.5) that

$$\begin{aligned}\|s-\tilde{s}\|^2 &= 2\varepsilon^2 V_{\hat{m}} + 2\varepsilon Z(s_{\hat{m}} - s) - \mathrm{pen}(\hat{m}) \\ &\quad + \inf_{m\in\mathcal{M}} \left\{ \|s-s_m\|^2 - 2\varepsilon Z(s_m - s) - \varepsilon^2 V_m + \mathrm{pen}(m) \right\}.\end{aligned}$$

Setting $d_m = \|s - s_m\|$, $U_m = d_m^{-1} Z(s_m - s)$ and noticing that $\|s-\tilde{s}\|^2 = \varepsilon^2 V_{\hat{m}} + d_{\hat{m}}^2$ by (6.4), we finally get

$$\begin{aligned}(1-\theta)\|s-\tilde{s}\|^2 &= (2-\theta)\varepsilon^2 V_{\hat{m}} - \theta d_{\hat{m}}^2 + 2\varepsilon d_{\hat{m}} U_{\hat{m}} - \mathrm{pen}(\hat{m}) \\ &\quad + \inf_{m\in\mathcal{M}} \left\{ d_m^2 - 2\varepsilon d_m U_m - \varepsilon^2 V_m + \mathrm{pen}(m) \right\},\end{aligned}$$

or equivalently,

$$\|s-\tilde{s}\|^2 = (1-\theta)^{-1} \left(\Delta_{\hat{m}} + \inf_{m\in\mathcal{M}} R_m \right), \tag{6.6}$$

where

$$\Delta_m = (2-\theta)\varepsilon^2 V_m + 2\varepsilon d_m U_m - \theta d_m^2 - \mathrm{pen}(m) \tag{6.7}$$

and

$$R_m = d_m^2 + \mathrm{pen}(m) - \varepsilon^2 V_m - 2\varepsilon d_m U_m. \tag{6.8}$$

Since $\hat{m}$ can, in principle, take any value in $\mathcal{M}$, we need, in order to control $\|s-\tilde{s}\|^2$, to control Δ_m uniformly with respect to m. To do this, we fix some positive number ξ and set $A_m = V_m + 2d_m U_m [\varepsilon(2-\theta)]^{-1}$, $x_m = L_m D_m + \xi$,

$$\Omega_{\xi,m} = \left\{ A_m < D_m + \frac{\theta d_m^2}{\varepsilon^2(2-\theta)} + 2\sqrt{D_m x_m} + \frac{2x_m}{\theta(2-\theta)} \right\} \quad \text{and} \quad \Omega_\xi = \bigcap_{m\in\mathcal{M}} \Omega_{\xi,m}.$$

Since $\langle \varphi_\lambda, s - s_m \rangle = 0$ for any $\lambda \in \Lambda_m$, ζ_m and $Z(s-s_m)$ are independent and the random variables V_m and U_m are also independent with respective distributions $\chi^2(D_m)$ and $\mathcal{N}(0,1)$. It then follows from Lemma 1 in the Appendix with $\rho = 2d_m[\varepsilon(2-\theta)]^{-1}$ and $b = 2[\theta(2-\theta)]^{-1} > 2$ that $\mathbb{P}\left[\Omega_{\xi,m}^c\right] \le \exp(-x_m)$ and therefore

$$\mathbb{P}\left[\Omega_\xi^c\right] \le \sum_{m\in\mathcal{M}} \exp(-L_m D_m - \xi) = \Sigma \exp(-\xi). \tag{6.9}$$

Using the inequalities $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $2ab \le \delta a^2 + \delta^{-1} b^2$, we derive that

$$2\sqrt{D_m(L_m D_m + \xi)} \le 2D_m \sqrt{L_m} + \alpha D_m + \alpha^{-1}\xi, \quad \text{for } \alpha > 0.$$

It therefore follows from the definition of Ω_ξ that, whatever $m \in \mathcal{M}$, on the set Ω_ξ,

$$A_m \le (1+\alpha) D_m + \frac{\theta d_m^2}{\varepsilon^2(2-\theta)} + 2D_m\sqrt{L_m} + \left(\alpha^{-1} + \frac{2}{\theta(2-\theta)} \right) \xi + \frac{2L_m D_m}{\theta(2-\theta)}.$$

If we define α by $K = (1+\alpha)(2-\theta)$, then $\alpha > 0$ since $K > 2-\theta$ and

$$(2-\theta)\varepsilon^2 A_m \le Q_m + \theta d_m^2 + \varepsilon^2 \xi \left[(2-\theta)\alpha^{-1} + 2\theta^{-1} \right].$$

It then follows from (6.7) and our definition of the penalty function that

$$\begin{aligned}\Delta_m \mathbb{1}_{\Omega_\xi} &= \left[(2-\theta)\varepsilon^2 A_m - \theta d_m^2 - \mathrm{pen}(m)\right] \mathbb{1}_{\Omega_\xi} \\ &\leq \left(Q_m - \mathrm{pen}(m) + \varepsilon^2 \xi \left[(2-\theta)\alpha^{-1} + 2\theta^{-1}\right]\right) \mathbb{1}_{\Omega_\xi}.\end{aligned}$$

Since this inequality holds whatever $m \in \mathcal{M}$ one can conclude from (2.7) that

$$\Delta_{\hat{m}} \mathbb{1}_{\Omega_\xi} \leq \left(\varepsilon^2 \xi \left[(2-\theta)\alpha^{-1} + 2\theta^{-1}\right] + \sup_{m \in \bar{\mathcal{M}}} \{Q_m - \mathrm{pen}(m)\}\right) \mathbb{1}_{\Omega_\xi} \tag{6.10}$$

and therefore, by (6.9), for all $\xi > 0$,

$$\mathbb{P}\left[\Delta_{\hat{m}} > \varepsilon^2 \left((2-\theta)\alpha^{-1} + 2\theta^{-1}\right)\xi + \sup_{m \in \bar{\mathcal{M}}} \{Q_m - \mathrm{pen}(m)\}\right] \leq \Sigma \exp(-\xi).$$

Integrating with respect to ξ, we get

$$\mathbb{E}[\Delta_{\hat{m}}] \leq \Sigma\varepsilon^2 \left[(2-\theta)\alpha^{-1} + 2\theta^{-1}\right] + \sup_{m \in \bar{\mathcal{M}}} \{Q_m - \mathrm{pen}(m)\}. \tag{6.11}$$

Since it follows from (6.8) that

$$\mathbb{E}\left[\inf_{m \in \mathcal{M}} R_m\right] \leq \inf_{m \in \mathcal{M}} \mathbb{E}[R_m] = \inf_{m \in \mathcal{M}} \left(d_m^2 + \mathrm{pen}(m) - \varepsilon^2 D_m\right),$$

we conclude from (6.6) and (6.11) that (2.8) holds.

6.3 Proof of Proposition 1

Proposition 1 is actually the consequence of a more general result which is as follows.

Theorem 4 *Under the settings of Theorem 2 with Q_m given by (2.6), assume that the weights L_m are bounded by some finite number L and that one can find nonnegative numbers $K' \geq K, a'$ and b' such that*

$$Q_m \leq \mathrm{pen}(m) \leq \varepsilon^2 D_m \left(K' + a'\sqrt{L_m} + b' L_m\right), \quad \text{for all } m \in \mathcal{M}. \tag{6.12}$$

i) If $L < 1/4$ and $\lambda, \beta > 0$ satisfy

$$(2+\lambda)(1-\beta) - K' - \sqrt{L}(4 + 2\lambda + a') - L\left(2\lambda^{-1} + b'\right) = 0, \tag{6.13}$$

then

$$\begin{aligned}(1-\theta)\,\mathbb{E}\left[\|s - \tilde{s}\|^2\right] &\leq (1+\lambda)\,\mathbb{E}\left[\inf_{m \in \mathcal{M}} \|s - \hat{s}_m\|^2\right] \\ &\quad + \Sigma\varepsilon^2 \left[\frac{(2-\theta)^2}{K + \theta - 2} + \frac{2}{\theta} + \frac{2+\lambda}{\beta} + \frac{2}{\lambda}\right].\end{aligned} \tag{6.14}$$

ii) If the family of weights satisfies (2.12), $K' = 2$ and $d(s, S_m) > 0$ for all $m \in \mathcal{M}$, then

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}\left[\|s - \tilde{s}\|^2\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}} \|s - \hat{s}_m\|^2\right]} \leq (1-\theta)^{-1}. \tag{6.15}$$

Remark: It is easily seen that the condition $L < 1/4$ is necessary and sufficient for the existence of pairs λ, β satisfying (6.13).

Proof: To prove these results, we shall derive a sharper bound for $\mathbb{E}\left[\inf_{m\in\mathcal{M}} R_m\right]$. To do this, we select some $\lambda > 0$ and define $\rho' = 2d_m[\varepsilon(2+\lambda)]^{-1}$, $A'_m = V_m + \rho' U_m$,

$$\Omega'_{\xi,m} = \left\{ A'_m > D_m - 2\sqrt{(D_m + \rho'^2/2)\,(L_m D_m + \xi)} \right\} \quad \text{and} \quad \Omega'_\xi = \bigcap_{m\in\mathcal{M}} \Omega'_{\xi,m}. \tag{6.16}$$

It follows from (A.2) below that $\mathbb{P}\left[\Omega'^c_{\xi,m}\right] \leq \exp(-L_m D_m - \xi)$ and therefore that

$$\mathbb{P}\left[\Omega'^c_\xi\right] \leq \sum_{m\in\mathcal{M}} \exp(-L_m D_m - \xi) = \Sigma \exp(-\xi). \tag{6.17}$$

Since for $\beta > 0$,

$$\begin{aligned} 2\sqrt{(D_m + \rho'^2/2)\,(L_m D_m + \xi)} &\leq 2\sqrt{(\rho'^2/2)\,(L_m D_m + \xi)} + 2D_m\sqrt{L_m} + 2\sqrt{D_m \xi} \\ &\leq \frac{\rho'^2\lambda(2+\lambda)}{4} + \frac{2(L_m D_m + \xi)}{\lambda(2+\lambda)} \\ &\quad + 2D_m\sqrt{L_m} + \beta D_m + \beta^{-1}\xi, \end{aligned}$$

we derive that, on the set Ω'_ξ and for all $m \in \mathcal{M}$,

$$A'_m \geq (1-\beta)D_m - \frac{\lambda d_m^2}{\varepsilon^2(2+\lambda)} - 2D_m\sqrt{L_m} - \left(\frac{1}{\beta} + \frac{2}{\lambda(2+\lambda)}\right)\xi - \frac{2L_m D_m}{\lambda(2+\lambda)}.$$

It then follows from (6.12) that

$$\begin{aligned} (2+\lambda)\varepsilon^2 V_m + 2\varepsilon d_m U_m &= (2+\lambda)\varepsilon^2 A'_m \\ &\geq \varepsilon^2 D_m\left[(2+\lambda)(1-\beta) - 2(2+\lambda)\sqrt{L_m} - 2\lambda^{-1}L_m\right] \\ &\quad - \lambda d_m^2 - \left(\frac{2+\lambda}{\beta} + \frac{2}{\lambda}\right)\varepsilon^2\xi \\ &\geq \mathrm{pen}(m) - \lambda d_m^2 - \left(\frac{2+\lambda}{\beta} + \frac{2}{\lambda}\right)\varepsilon^2\xi - \varepsilon^2 D_m G_m, \end{aligned}$$

with

$$G_m = K' + [2(2+\lambda) + a']\sqrt{L_m} + \left(2\lambda^{-1} + b'\right)L_m - (2+\lambda)(1-\beta).$$

Together with (6.8) and (6.4), the last inequality implies that

$$R_m \leq (1+\lambda)\left(\|s - \hat{s}_m\|^2\right) + \varepsilon^2 D_m G_m + \left(\frac{2+\lambda}{\beta} + \frac{2}{\lambda}\right)\varepsilon^2\xi.$$

Since this holds on the set Ω'_ξ for all $m \in \mathcal{M}$, we deduce from (6.17) that,

$$\mathbb{P}\left[\inf_{m\in\mathcal{M}} R_m > \inf_{m\in\mathcal{M}}\left\{(1+\lambda)\|s-\hat{s}_m\|^2 + \varepsilon^2 D_m G_m\right\} + \varepsilon^2\left((2+\lambda)\beta^{-1} + 2\lambda^{-1}\right)\xi\right] \leq \Sigma\exp(-\xi).$$

Integrating with respect to ξ, we conclude that

$$\begin{aligned}\mathbb{E}\left[\inf_{m\in\mathcal{M}} R_m\right] &\leq \mathbb{E}\left[\inf_{m\in\mathcal{M}}\left\{(1+\lambda)\|s-\hat{s}_m\|^2+\varepsilon^2 D_m G_m\right\}\right] \\ &\quad + \Sigma\varepsilon^2\left((2+\lambda)\beta^{-1}+2\lambda^{-1}\right). \end{aligned} \tag{6.18}$$

If λ and β are chosen in order to satisfy (6.13), then $G_m \leq 0$ for all $m \in \mathcal{M}$ and (6.14) follows from (6.6), (6.11) and the fact that $\bar{\mathcal{M}}$ is empty.

Let us now prove (6.15). Obviously, setting $\mathcal{M}_1 = \{m \in \mathcal{M} \,|\, G_m > 0\}$ and $\mathcal{M}_2 = \mathcal{M} \setminus \mathcal{M}_1$, we get

$$\begin{aligned}&\inf_{m\in\mathcal{M}}\left\{(1+\lambda)\|s-\hat{s}_m\|^2+\varepsilon^2 D_m G_m\right\} \\ &\qquad \leq \inf_{m\in\mathcal{M}_1}\left\{(1+\lambda)\|s-\hat{s}_m\|^2+\varepsilon^2 D_m G_m\right\} \bigwedge (1+\lambda)\inf_{m\in\mathcal{M}_2}\|s-\hat{s}_m\|^2 \\ &\qquad \leq \left[(1+\lambda)\inf_{m\in\mathcal{M}_1}\|s-\hat{s}_m\|^2+\varepsilon^2\sup_{m\in\mathcal{M}_1} D_m G_m\right] \\ &\qquad\quad \bigwedge (1+\lambda)\inf_{m\in\mathcal{M}_2}\|s-\hat{s}_m\|^2. \end{aligned} \tag{6.19}$$

If $K' = 2$, $\lambda < 1$ and $\beta = \lambda/3$, then $G_m \leq 0$ provided that L_m is small enough, which, by assumption, is true as soon as $D_m \geq D$, where D depends on λ, a' and b'. In view of (2.1) and the boundedness of the weights L_m, this implies that the set $\mathcal{M}_1$ is finite. Then $\sup_{m\in\mathcal{M}_1} D_m G_m < +\infty$ while, by our assumption on s, $\inf_{m\in\mathcal{M}_1} d_m > 0$. It then follows from (6.4) that, for ε small enough,

$$\lambda \inf_{m\in\mathcal{M}_1}\|s-\hat{s}_m\|^2 \geq \lambda \inf_{m\in\mathcal{M}_1} d_m^2 \geq \varepsilon^2 \sup_{m\in\mathcal{M}_1} D_m G_m$$

and finally, by (6.19),

$$\inf_{m\in\mathcal{M}}\left\{(1+\lambda)\|s-\hat{s}_m\|^2+\varepsilon^2 D_m G_m\right\} \leq (1+2\lambda)\inf_{m\in\mathcal{M}}\|s-\hat{s}_m\|^2.$$

Putting this inequality together with (6.18), (6.6) and (6.11), we conclude that, for ε small enough, depending on λ,

$$\begin{aligned}(1-\theta)\,\mathbb{E}\left[\|s-\tilde{s}\|^2\right] &\leq (1+2\lambda)\,\mathbb{E}\left[\inf_{m\in\mathcal{M}}\|s-\hat{s}_m\|^2\right] \\ &\quad + \Sigma\varepsilon^2\left[\frac{(2-\theta)^2}{K+\theta-2}+\frac{2}{\theta}+\frac{3(2+\lambda)}{\lambda}+\frac{2}{\lambda}\right].\end{aligned}$$

Letting ε, then λ go to zero gives (6.15). □

Let us now turn to the proof of Proposition 1. Since, by definition, $\|s-\tilde{s}\| \geq \inf_{m\in\mathcal{M}}\|s-\hat{s}_m\|$, it suffices to show that (6.15) holds for θ arbitrarily close to zero and therefore to check that the assumptions required to apply Theorem 4 are satisfied whatever $\theta > 0$. Let us set

$$Q_m = \varepsilon^2 D_m\left(2-\theta/2+2(2-\theta)\sqrt{L'_m}+2\theta^{-1}L'_m\right) \quad \text{with } L'_m = \eta L_m.$$

Since $\sup_{m\in\mathcal{M}} L_m < +\infty$, $\mathrm{pen}(m) \geq Q_m$ for all $m \in \mathcal{M}$ by (2.13) if $\eta > 0$ is small enough. Moreover, the weights L'_m satisfy (2.1) by assumption. Using the upper bound in (2.13), we also see that (6.12) holds with $K' = 2$, $a' = \eta^{-1/2}a$, $b' = \eta^{-1}b$ and L_m replaced by L'_m. Since the new weights L'_m also satisfy (2.12), (6.15) holds whatever $\theta > 0$.

6.4 Proof of Proposition 2

Let m be given in $\mathcal{M}$. It follows from (4.7) that

$$\begin{aligned}\Delta(m,N) &= \|\hat{s}_N\|^2 - \|\hat{s}_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(N)\\ &\geq \|\hat{s}_N - \hat{s}_m\|^2 - \varepsilon^2(1-\eta)(N - D_m),\end{aligned}$$

with

$$\hat{s}_N - \hat{s}_m = s_N - s_m + \varepsilon(\zeta_N - \zeta_m),$$

where $\zeta_N - \zeta_m$ is a standard normal vector with dimension $N - D_m$. This implies that $U = \|\varepsilon^{-1}(\hat{s}_N - \hat{s}_m)\|^2$ has the distribution of a non-central chi-square with $N - D_m$ degrees of freedom and noncentrality parameter $\mu = \varepsilon^{-1}\|s_N - s_m\|$. Then

$$\Delta(m,N) \geq \varepsilon^2\left[U - (1-\eta)E_m\right] \quad \text{with } E_m = N - D_m,$$

and by (A.2) (with $\rho = 0$ and $D = E_m$) and the fact that U is stochastically larger than a chi-square variable with E_m degrees of freedom,

$$\mathbb{P}\left[U \leq E_m - 2\sqrt{xE_m}\right] \leq e^{-x} \quad \text{for } x > 0.$$

Setting $x = \eta^2 E_m/4$, we conclude that $\Delta(m,N) > 0$ with probability at least $1 - \exp\left[-\eta^2 E_m/4\right]$. Defining the integer D by $N(1-\theta) - 1 < D \leq N(1-\theta)$, we get

$$\begin{aligned}\mathbb{P}\left[\inf_{m\in\mathcal{M}_n \,|\, D_m \leq D} \Delta(m,N) \leq 0\right] &\leq \sum_{j=0}^{D} \exp\left[jH(j) - \frac{\eta^2}{4}(N-j)\right]\\ &\leq \exp\left[-\frac{\theta\eta^2 N}{4}\right]\sum_{j=0}^{D}\exp[jH(j)].\end{aligned}$$

By assumption, there exists some integer k depending on $\bar{H}, \theta$ and η such that $H(j) \leq \bar{H}(j) \leq \eta^2\theta/[8(1-\theta)]$ as soon as $j \geq k$. Assuming that $D \geq k$, we then derive that

$$\sum_{j=0}^{D}\exp[jH(j)] \leq \sum_{j=0}^{k-1}\exp[j\bar{H}(j)] + \sum_{j=k}^{D}\exp\left[\frac{\theta\eta^2 j}{8(1-\theta)}\right] \leq C_1 + C_2\exp\left[\frac{\theta\eta^2 N}{8}\right],$$

with constants C_1 and C_2 depending only on $\bar{H}, \theta$ and η. Therefore for N large enough (depending on $\bar{H}, \theta, \eta$ and δ), $\Delta(m,N) > 0$ for all m such that $D_m \leq D$ with probability at least $1 - \delta$. In view of the definition of Δ, we conclude that $\mathbb{P}[D_{\hat{m}} > D] \geq 1 - \delta$.

Let us now prove the second part of the proposition. We first recall from (6.4) that

$$\|s - \tilde{s}\|^2 = \varepsilon^2 V_{\hat{m}} + \|s - s_{\hat{m}}\|^2 \geq \varepsilon^2 V_{\hat{m}} + \|s - s_N\|^2 \tag{6.20}$$

and set

$$M = \sum_{\lambda \in \Lambda_N} \mathbb{1}_{[0,\tau)} \left([Z(\varphi_\lambda)]^2 \right) \quad \text{with} \quad \mathbb{P}\left[\chi^2(1) < \tau \right] = \theta/2.$$

Noticing that the variables $[Z(\varphi_\lambda)]^2$ for $\lambda \in \Lambda_N$ are i.i.d. with distribution $\chi^2(1)$, we derive that M is binomial with parameters N and $\theta/2$ and get, using a classical binomial inequality (see Hoeffding, 1963)

$$\mathbb{P}\left[M \geq N\theta \right] = \mathbb{P}\left[M - N\theta/2 \geq N\theta/2 \right] \leq \exp\left[-N\theta^2/8 \right].$$

Once again, this is bounded by δ for N large enough and therefore, except on a set of probability bounded by 2δ we get simultaneously $D_{\hat{m}} > N(1-\theta) - 1$ and $M < N\theta$, which implies that

$$V_{\hat{m}} = \sum_{\lambda \in \Lambda_{\hat{m}}} [Z(\varphi_\lambda)]^2 \geq [N(1-2\theta) - 1]\tau.$$

The conclusion follows from (6.20) since $\Phi\left(\sqrt{\tau}\right) = (\theta+2)/4$ and therefore

$$\mathbb{E}[V_{\hat{m}}] \geq (1-2\delta)[N(1-2\theta)-1] \left[\Phi^{-1}\left(\frac{\theta+2}{4} \right) \right]^2.$$

6.5 Proof of Proposition 3

Setting $\Lambda_2 = \Lambda \setminus \Lambda_1$, we recall that the variables $W_\lambda = [Y(\varphi_\lambda)]^2$ for $\lambda \in \Lambda_2$ are i.i.d. with distribution $\chi^2(1)$. We denote by $W_{(1)} < \ldots < W_{(n)}$ with $n = N - |\Lambda_1|$ the corresponding order statistics and, as usual, by $\hat{m}$ the minimizer with respect to $m \in \mathcal{M}$ of

$$\begin{aligned} \gamma(\hat{s}_m) + \mathrm{pen}(m) &= -\|\hat{s}_{m\cap\Lambda_1}\|^2 - \|\hat{s}_{m\cap\Lambda_2}\|^2 + \mathrm{pen}(m) \\ &= -\|\hat{s}_{m\cap\Lambda_1}\|^2 - \varepsilon^2 \sum_{\lambda \in m \cap \Lambda_2} W_\lambda + \mathrm{pen}(m). \end{aligned}$$

Since $\mathrm{pen}(m)$ only depends on $|m|$, we deduce that

$$\gamma(\hat{s}_{\hat{m}}) = -\|\hat{s}_{\hat{m}\cap\Lambda_1}\|^2 - \varepsilon^2 \sum_{j=1}^{k} W_{(n+1-j)} \quad \text{with } k = |\hat{m} \cap \Lambda_2| \tag{6.21}$$

and that

$$\|s - \hat{s}_{\hat{m}}\|^2 = \|s - \hat{s}_{\hat{m}\cap\Lambda_1}\|^2 + \varepsilon^2 \sum_{j=1}^{k} W_{(n+1-j)}. \tag{6.22}$$

Now, let us consider the subset m' of Λ defined by

$$m' = (\hat{m} \cap \Lambda_1) \cup \{\lambda \in \Lambda_2 \,|\, W_\lambda = W_{(n+1-j)} \quad \text{for some } j,\ 1 \leq j \leq J = |\bar{m}| - |\hat{m} \cap \Lambda_1|\}.$$

Since $|m'| = |\bar{m}|$, $\mathrm{pen}(m') = \mathrm{pen}(\bar{m})$ and

$$\begin{aligned} \gamma(\hat{s}_{\hat{m}}) &\leq \gamma(\hat{s}_{m'}) + \mathrm{pen}(m') \\ &\leq -\|\hat{s}_{\hat{m}\cap\Lambda_1}\|^2 - \varepsilon^2 \sum_{j=1}^{J} W_{(n+1-j)} + (2 - 2\alpha - \eta)(1-\delta)\varepsilon^2 |\bar{m}| \log N, \end{aligned}$$

then from (6.21)

$$\sum_{j=1}^{k} W_{(n+1-j)} \geq \sum_{j=1}^{J} W_{(n+1-j)} - (2-2\alpha-\eta)(1-\delta)|\bar{m}| \log N. \tag{6.23}$$

Since

$$n \geq N\left(1-\delta A N^{\alpha-1}\right) \quad \text{and} \quad (1-\delta)|\bar{m}| \leq J \leq |\bar{m}| \leq A N^{\alpha}, \tag{6.24}$$

we derive that n/J goes to infinity with N. It then follows from Lemma 3 with $\theta = 3$ that there exists a set Ω' with

$$\mathbb{P}[\Omega'] \geq 1 - \left[\exp\left(\frac{9}{8}\right) - 1\right]^{-1} > 1/2, \tag{6.25}$$

such that on Ω' and uniformly for $1 \leq j \leq J$,

$$W_{(n+1-j)} \geq -2\log(2j/n)[1+o(1)] \geq [2\log(n/J)][1+o(1)],$$

since $n/j \geq n/J$ goes to infinity with N. Therefore by (6.23) and (6.24), when $N \to +\infty$,

$$\begin{aligned}
\sum_{j=1}^{k} W_{(n+1-j)} &\geq 2J\left[\log N - \log J\right][1+o(1)] - (2-2\alpha-\eta)J\log N \\
&\geq \eta J \log N[1+o(1)].
\end{aligned}$$

It then follows from (6.22) and (6.25) that

$$\begin{aligned}
\mathbb{E}\left[\|s-\tilde{s}\|^2\right] &\geq \varepsilon^2 \mathbb{E}\left[\mathbb{1}_{\Omega'} \sum_{j=1}^{k} W_{(n+1-j)}\right] \geq (\eta/2) J\varepsilon^2 \log N[1+o(1)] \\
&\geq (\eta/2)(1-\delta)|\bar{m}|\varepsilon^2 \log N[1+o(1)],
\end{aligned}$$

which concludes the proof.

6.6 Proof of Proposition 4

For $D \geq 1$, the variables V_m, defined by (6.1), for $m \in \mathcal{M}_D$, are i.i.d. with a chi-square distribution with D degrees of freedom, in view of the orthogonality of the spaces S_m. Therefore, if we denote by $\chi^2(D)$ a random variable with such a distribution, for any $z > 0$,

$$\log\left(\mathbb{P}\left[\sup_{m\in\mathcal{M}_D} V_m < z\right]\right) = \kappa(D)\log\left(1-\mathbb{P}\left[\chi^2(D) \geq z\right]\right). \tag{6.26}$$

An application of (A.1) with $x = \alpha D + 2\log(D+3)$, $\rho = 0$ and $b = 2$ gives,

$$\mathbb{P}\left[\chi^2(D) \geq (1+2\alpha)D + 2D\sqrt{\alpha}\sqrt{1+\frac{2\log(D+3)}{\alpha D}} + 4\log(D+3)\right] \leq \frac{\exp(-\alpha D)}{(D+3)^2}.$$

Setting

$$G(\alpha) = 1 + 2\sqrt{\alpha} + 2\alpha, \qquad z = G(\alpha)D + 2\left(2 + \alpha^{-1/2}\right)\log(D+3)$$

and using $\sqrt{1+u} \le 1 + u/2$, we derive that

$$\mathbb{P}\left[\chi^2(D) \ge z\right] \le \frac{\exp(-\alpha D)}{(D+3)^2} \le \frac{1}{16}.$$

For $u \le 1/16$, $u^{-1}\log(1-u) \ge 16\log(1-1/16) > -1.033$. It then follows from (6.26) and (4.12), that

$$\log\left(\mathbb{P}\left[\sup_{m\in\mathcal{M}_D} V_m < z\right]\right) \ge \exp(\alpha D)\log\left(1 - \frac{\exp(-\alpha D)}{(D+3)^2}\right) \ge -\frac{1.033}{(D+3)^2}$$

and therefore,

$$\mathbb{P}\left[\sup_{m\in\mathcal{M}_D} V_m \ge z\right] \le 1 - \exp\left(-\frac{1.033}{(D+3)^2}\right) \le \frac{1.033}{(D+3)^2}.$$

Finally,

$$\mathbb{P}\left[\sup_{m\in\mathcal{M}\setminus\emptyset}\left\{V_m - G(\alpha)D_m - 2\left(2+\alpha^{-1/2}\right)\log(D_m+3)\right\} \ge 0\right]$$
$$\le \quad 1.033\sum_{D\ge 1}(D+3)^{-2} \quad < \quad 0.3. \quad (6.27)$$

We now want to prove an inequality in the opposite direction. In order to do this, we set

$$\theta(x) = 1 + x^{-1/2} - \frac{\log[G(x)]}{2x}, \qquad g(x) = \begin{cases} 5/6 & \text{if } 0 < x < 3, \\ [\theta(5x/12)]^{-1} & \text{if } x \ge 3, \end{cases}$$

$a = \alpha g(\alpha)/2$ and define $D(\alpha)$ to be the smallest integer $n \ge 3$ such that

$$\frac{\alpha n}{4} \ge \log n \ge \frac{1}{2}\log(4\pi) + \log\left(2a + \sqrt{2a}\right) + \frac{1}{n}\left[\frac{1}{6} + \frac{1}{2\left(a+\sqrt{a}\right)^2} + \frac{1}{a+\sqrt{a}}\right], \quad (6.28)$$

$$\sum_{j\ge n}\exp\left(-\sqrt{j}\left(1 - e^{-\alpha j}\right)\right) \le 0.2 \quad \text{and} \quad G(\alpha) \ge 8\left(1 + (2/3)\alpha^{-1/2}\right)\frac{\log n}{n}. \quad (6.29)$$

If $D \ge D(\alpha)$, then by (6.28) $y = g(\alpha)(\alpha D - 2\log D) \ge aD$,

$$\sqrt{D}\left(a + \sqrt{a}\right) \le \left(y/\sqrt{D}\right) + \sqrt{y} \le \sqrt{D}\left(2a + \sqrt{2a}\right)$$

and Corollary 1 below together with (6.28) imply that if $z = D + 2\sqrt{Dy} + 2y$,

$$\begin{aligned}\log\left(\mathbb{P}\left[\chi^2(D) \ge z\right]\right) \quad &\ge \quad -y\theta\left(\frac{y}{D}\right) - \frac{1}{2}\log(4\pi D) - \log\left(2a + \sqrt{2a}\right) \\ &\qquad - \frac{1}{D}\left[\frac{1}{6} + \frac{1}{2\left(a+\sqrt{a}\right)^2} + \frac{1}{a+\sqrt{a}}\right] \\ &\ge \quad -y\theta\left(\frac{y}{D}\right) - \frac{3}{2}\log D.\end{aligned}$$

It follows from Proposition 5 below that the function $x \mapsto \theta(x)$ is bounded by $6/5$ and decreasing for $x \geq 5/4$. Consequently $\theta(y/D) \leq 1/g(\alpha)$ for $\alpha < 3$ and if $\alpha \geq 3$, then $y/D \geq a > 5\alpha/12 \geq 5/4$ hence $\theta(y/D) \leq \theta(5\alpha/12) = 1/g(\alpha)$. Therefore $y\theta(y/D) \leq \alpha D - 2\log D$, $\log\left(\mathbb{P}\left[\chi^2(D) \geq z\right]\right) \geq -\alpha D + (\log D)/2$ and it follows from (6.26) and (4.12) that

$$\log\left(\mathbb{P}\left[\sup_{m\in\mathcal{M}_D} V_m < z\right]\right) \leq -\kappa(D)\mathbb{P}\left[\chi^2(D) \geq z\right] \leq -\sqrt{D}\left(1 - e^{-\alpha D}\right). \tag{6.30}$$

Since $\sqrt{1-u} > 1 - 0.6u$ for $u \leq 1/2$, we derive from (6.28) that

$$\sqrt{y} = \sqrt{\alpha g(\alpha) D}\sqrt{1 - 2\log D/(\alpha D)} > \sqrt{\alpha g(\alpha) D} - 1.2(\log D)\sqrt{g(\alpha)/(\alpha D)},$$

which implies that

$$z > DG[\alpha g(\alpha)] - \left(2.4\sqrt{g(\alpha)/\alpha} + 4g(\alpha)\right)\log D.$$

Setting $F(\alpha) = G[\alpha g(\alpha)]/G(\alpha)$, we easily derive from the properties of θ that $F(\alpha)$ converges to one when α converges to zero or to infinity and that $1 > F(\alpha) > g(\alpha) \geq 5/6$ for all $\alpha > 0$. It follows that

$$z > G(\alpha)F(\alpha)D - 4F(\alpha)\log D\left(1 + (2/3)\alpha^{-1/2}\right)$$

and we conclude from (6.30) and (6.29) that

$$\mathbb{P}\left[\sup_{\{m\in\mathcal{M}\,|\,D_m\geq D(\alpha)\}}\left\{V_m - G(\alpha)F(\alpha)D_m - 4F(\alpha)\log D_m\left(1 + (2/3)\alpha^{-1/2}\right)\right\} < 0\right]$$
$$\leq \sum_{j\geq D(\alpha)} \exp\left[-\sqrt{j}\left(1 - e^{-\alpha j}\right)\right] \leq 0.2.$$

Together with (6.27), this means that $\mathbb{P}[\Omega] \geq 1/2$, if we denote by Ω the event defined by the set of inequalities

$$V_m < G(\alpha)D_m + 2\left(2 + \alpha^{-1/2}\right)\log(D_m + 3), \quad \text{for all } m \in \mathcal{M}, \tag{6.31}$$

since $V_\emptyset = 0$ and

$$V_m \geq G(\alpha)F(\alpha)D_m - 4F(\alpha)\left(1 + (2/3)\alpha^{-1/2}\right)\log D_m \quad \text{if } D_m \geq D(\alpha). \tag{6.32}$$

Let us now analyze what happens on the event Ω, provided that $\bar{D}$ satisfies

$$\bar{D}[F(\alpha) - \lambda] \geq 4D(\alpha) \tag{6.33}$$

and

$$\bar{D}G(\alpha)[F(\alpha) - \lambda] \geq 2\log(\bar{D}+1)\left[4F(\alpha)\left(1 + (2/3)\alpha^{-1/2}\right) + \lambda\beta\left(\alpha^{-1/2} + 2\right)\right]. \tag{6.34}$$

For any $m \in \bar{\mathcal{M}}$, it follows from (6.33) that $D_m \geq \bar{D} > 4D(\alpha)$. Moreover, by (4.14) and (4.13),

$$\mathrm{pen}(m) \leq \lambda\varepsilon^2 \left[D_m G(\alpha) + \beta \left(\alpha^{-1/2} + 2 \right) \log(D_m + 1) \right]$$

and, since $s = s_m = 0$, $\gamma(\hat{s}_m) = -\varepsilon^2 V_m$ by (6.3). Therefore, by (6.32) and (6.34),

$$\begin{aligned} \gamma(\hat{s}_m) + \mathrm{pen}(m) &\leq -\varepsilon^2 D_m G(\alpha)[F(\alpha) - \lambda] \\ &\quad + \varepsilon^2 \log(D_m + 1) \left[4F(\alpha) \left(1 + (2/3)\alpha^{-1/2} \right) + \lambda\beta \left(\alpha^{-1/2} + 2 \right) \right] \\ &\leq -(\varepsilon^2/2) D_m G(\alpha)[F(\alpha) - \lambda]. \end{aligned} \tag{6.35}$$

— If $\bar{\mathcal{M}}$ is infinite, then D_m can be taken arbitrarily large and

$$\mathbb{P}\left[\inf_{m \in \mathcal{M}} \{\gamma(\hat{s}_m + \mathrm{pen}(m)\} = -\infty \right] \geq \mathbb{P}[\Omega] \geq 1/2.$$

— If $\bar{\mathcal{M}}$ is finite, Theorem 2 applies, implying that $\tilde{s}$ exists. On the other hand, if $D' = \lfloor (F(\alpha) - \lambda) D_m / 4 \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x and $m \in \bar{\mathcal{M}}$, then $D' \geq D(\alpha) \geq 3$ by (6.33) and it follows from (6.31) and (6.29) that,

$$\begin{aligned} \inf_{D \leq D'} \inf_{m \in \mathcal{M}_D} (\gamma(\hat{s}_m) + \mathrm{pen}(m)) &\geq -\varepsilon^2 \sup_{D \leq D'} \sup_{m \in \mathcal{M}_D} V_m \\ &> -\varepsilon^2 \left[G(\alpha) D' + 2 \left(2 + \alpha^{-1/2} \right) \log(D' + 3) \right] \\ &> -2\varepsilon^2 G(\alpha) D' \\ &> -\varepsilon^2 G(\alpha)[F(\alpha) - \lambda] D_m / 2. \end{aligned} \tag{6.36}$$

Comparing (6.35) with (6.36) and taking into account (6.33), one concludes, since m is arbitrary in $\bar{\mathcal{M}}$, that, on the set Ω,

$$D_{\hat{m}} > \frac{1}{4}[F(\alpha) - \lambda] \left(\sup_{m \in \bar{\mathcal{M}}} D_m \right) \geq D(\alpha).$$

Since, by (6.4), $\|s - \tilde{s}\|^2 = \varepsilon^2 V_{\hat{m}}$, it follows from (6.32) and (6.29) that

$$\varepsilon^{-2} \|s - \tilde{s}\|^2 \geq D_{\hat{m}} G(\alpha) F(\alpha) - 4F(\alpha) \left(1 + (2/3)\alpha^{-1/2} \right) \log D_{\hat{m}} \geq D_{\hat{m}} G(\alpha) F(\alpha)/2$$

and (4.15) follows since $\mathbb{P}[\Omega] \geq 1/2$.

6.7 Proof of Theorem 3

Let S_1 be any one-dimensional model in the family and s an element of S_1 such that $\|s\| = \varepsilon\sqrt{A}$. If $\hat{m} = 0$, then $\tilde{s} = 0$, hence

$$\mathbb{E}\left[\|s - \tilde{s}\|^2 \right] \geq A\varepsilon^2 \mathbb{P}[\hat{m} = 0].$$

Since $\hat{s}_0 = 0$ and $\mathrm{pen}(0) = 0$, it follows from (2.2) that $\hat{m} = 0$ if $\mathrm{pen}(m) > \|\hat{s}_m\|^2$ for all $m \neq 0$. Setting $U_m = \varepsilon^{-2} \|\hat{s}_m\|^2$, we know that U_m has the distribution of a

non-central chi-square with parameters D_m and $\|s_m\|/\varepsilon$ and by Lemma 1 of Birgé (2001), since $\|s_m\|$ is either $\varepsilon\sqrt{A}$ or 0,

$$\mathbb{P}\left[U_m \geq D_m + A + 2\sqrt{(D_m+2A)x_m} + 2x_m\right] \leq \exp(-x_m) \quad \text{for all } m \in \mathcal{M}.$$

Setting $x_m = L_m D_m$, we derive that if

$$\Omega = \left\{U_m < D_m + A + 2\sqrt{(D_m+2A)L_m D_m} + 2L_m D_m \quad \text{for all } m \in \mathcal{M}^\star\right\},$$

then

$$\mathbb{P}[\Omega] \geq 1 - \sum_{m\in\mathcal{M}^\star} \exp(-L_m D_m) = 1 - \Sigma.$$

Putting everything together we can conclude that if

$$\varepsilon^{-2}\,\mathrm{pen}(m) \geq D_m + A + 2\sqrt{(D_m+2A)L_m D_m} + 2L_m D_m \quad \text{for all } m \in \mathcal{M}^\star, \quad (6.37)$$

then

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \geq A\varepsilon^2 \mathbb{P}[\Omega] \geq A\varepsilon^2(1-\Sigma).$$

Since (6.37) is an immediate consequence of (4.16), (4.17) holds while the upper bound for the risk of $\tilde{s}$ when $\mathrm{pen}(m)$ is given by (2.9) follows from (2.10).

APPENDIX

Lemma 1 *Let V and U be independent random variables with respective distributions $\chi^2(D)$ and $\mathcal{N}(0,1)$ and ρ be some real number. Then, for any positive x, the following probability bounds hold*

$$\mathbb{P}\left[V + \rho U \geq D + \rho^2/(2b) + 2\sqrt{Dx} + bx\right] \leq \exp(-x) \quad \textit{for any } b \geq 2 \qquad (A.1)$$

and

$$\mathbb{P}\left[V + \rho U \leq D - 2\sqrt{(D+\rho^2/2)x}\right] \leq \exp(-x). \qquad (A.2)$$

Proof of Lemma 1: Let us first observe that the Laplace transform of a centered $\chi^2(1)$ variable $U^2 - 1$ satisfies

$$\log \mathbb{E}\left[e^{y(U^2-1)}\right] = -\frac{1}{2}\log(1-2y) - y \leq \frac{y^2}{1-2y} \quad \text{for } y < \frac{1}{2},$$

which implies by independence that

$$\log \mathbb{E}\left[e^{y(V-D+\rho U)}\right] \leq \frac{Dy^2}{1-2y} + \frac{y^2\rho^2}{2}, \qquad (A.3)$$

since $\mathbb{E}[tU] = \exp(t^2/2)$. If $b \geq 2$ the right-hand side of (A.3) can be bounded by $Dy^2/(1-by) + y\rho^2/(2b)$ for $0 < y < b^{-1}$ which implies that

$$\log \mathbb{E}\left[e^{y\left[V-D+\rho U-\rho^2/(2b)\right]}\right] \leq \frac{Dy^2}{1-by} \quad \text{for } 0 < y < b^{-1}.$$

Inequality (A.1) then follows from Lemma 2 below with $a^2 = D$. Its proof is part of the proof of Lemma 8 of Birgé and Massart (1998).

On the other hand, setting $a^2 = D + \rho^2/2$ and $A^2 = 4a^2x$, we get

$$\begin{aligned}
\mathbb{P}[V + \rho U \leq D - A] &= \mathbb{P}[-V - \rho U + D - A \geq 0] \\
&\leq \inf_{t \geq 0} \mathbb{E}[\exp(t(-V - \rho U + D - A))] \\
&= \inf_{y \leq 0} e^{Ay} \mathbb{E}[\exp(y(V + \rho U - D))] \\
&\leq \inf_{y \leq 0} \exp\left(Ay + a^2 y^2\right) = \exp\left(-A^2 a^{-2}/4\right),
\end{aligned}$$

and (A.2) follows. □

Lemma 2 *Let X be a random variable such that*

$$\log\left(\mathbb{E}[\exp(yX)]\right) \leq \frac{(ay)^2}{1 - by} \quad \textit{for } 0 < y < b^{-1},$$

where a and b are positive constants. Then

$$\mathbb{P}[X \geq 2a\sqrt{x} + bx] \leq \exp(-x) \quad \textit{for all } x > 0.$$

Lemma 3 *Let $W_{(1)} < \ldots < W_{(n)}$ be an ordered sample of size n from the chi-square distribution with one degree of freedom, j be a positive integer, θ a positive number such that $j(1+\theta) \leq n$ and Φ the standard normal c.d.f. Then*

$$\mathbb{P}\left[W_{(n+1-j)} \leq \left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2\right] \leq \exp\left[-\frac{j\theta^2}{2(1+\theta)}\right], \tag{A.4}$$

and consequently if $\theta \geq 2.06$

$$W_{(n+1-j)} > \left[\Phi^{-1}\left(1 - \frac{j(1+\theta)}{2n}\right)\right]^2 \quad \textit{for } 1 \leq j \leq \frac{n}{(1+\theta)}, \tag{A.5}$$

apart from a set of probability bounded by

$$\left[\exp\left(\frac{\theta^2}{2(1+\theta)}\right) - 1\right]^{-1} < 1.$$

Moreover, uniformly for $0 < y \leq x$,

$$\left[\Phi^{-1}(1-y)\right]^2 = -(2\log y)[1 + o(1)] \quad \textit{when } x \to 0.$$

Proof: Let us first observe that if $F(t)$ is the cumulative distribution function of the absolute value of a normal variable and U is uniform on $[0,1]$, then $W = \left[F^{-1}(U)\right]^2$ has the chi-square distribution with one degree of freedom. It follows that $W_{(j)}$ can be written as $\left[F^{-1}(U_{(j)})\right]^2$ where $U_{(1)} < \ldots < U_{(n)}$ is an ordered sample of size n of the uniform distribution. Now set $x = j(1+\theta)/n$. Since (A.4) clearly holds when

$x = 1$, we may assume that $x < 1$. Denoting by $\mathcal{B}(n,p)$ a binomial random variable with parameters n and p we notice that

$$\mathbb{P}[U_{(n+1-j)} \leq 1-x] = \mathbb{P}[U_{(n+1-j)} < 1-x] = \mathbb{P}[\mathcal{B}(n,x) < j] = \mathbb{P}[\mathcal{B}(n,x) < nx - j\theta]. \tag{A.6}$$

Recalling from Massart (1990, Theorem 2) that, for $0 < y \leq p$,

$$\mathbb{P}[\mathcal{B}(n,p) - np < -ny] \leq \exp\left[-\frac{ny^2}{2(p-y/3)(1-p+y/3)}\right] < \exp\left[-\frac{ny^2}{2p}\right], \tag{A.7}$$

we derive from (A.6) that

$$\mathbb{P}\left[W_{(n+1-j)} \leq [F^{-1}(1-x)]^2\right] = \mathbb{P}\left[U_{(n+1-j)} \leq 1-x\right] \leq \exp\left[-\frac{j\theta^2}{2(1+\theta)}\right] \tag{A.8}$$

and (A.4) follows since $F(t) = 2\Phi(t) - 1$. Summing the different probabilities gives (A.5). The last result follows from Feller (1968, Lemma 2 p. 175). □

Proposition 4, which is our most general result concerning lower bounds for the penalty, is based on some corollary of the following proposition which is of interest by itself since it evaluates rather precisely the probabilities of large deviations of gamma random variables from their mean. A similar result appeared as Lemma 6.1 in Johnstone (2001) and our proof follows the same lines as his. In particular, the upper bound part in the next lemma is implicit in its proof. Unfortunately, we cannot use his result since we do need a lower bound for the deviations of chi-square variables, while he only established upper bounds. Moreover, his result is only valid for $x + \sqrt{x} \leq 1/4$ which is not enough for our purpose.

Proposition 5 *Let X be a random variable with gamma distribution $\Gamma(t,1)$. If $x > 0$ then*

$$\log\left(\mathbb{P}\left[X \geq t\left(1 + 2x + 2\sqrt{x}\right)\right]\right) = -2tx\theta(x) - (1/2)\log(2\pi/\lambda) - \Phi, \tag{A.9}$$

with

$$\theta(x) = 1 + x^{-1/2} - (2x)^{-1}\log\left(1 + 2x + 2\sqrt{x}\right); \tag{A.10}$$

$$\lambda = t\left[2t\left(x + \sqrt{x}\right) + 1\right]^{-2} \qquad \text{and} \qquad 0 < \Phi < 1/(12t) + \log(1+\lambda).$$

Moreover $\theta(x)$ is decreasing for $x \geq 5/4$,

$$1 < \theta(x) < 1.196 \qquad \text{and} \qquad \lim_{x\to 0}\theta(x) = \lim_{x\to+\infty}\theta(x) = 1. \tag{A.11}$$

Remark: Bound (A.9) is only useful for $\lambda < 2\pi$. Otherwise, since $2tx\theta(x) < 1.2/(2\lambda)$, (A.9) becomes non significant since Φ is not precisely known.

The proof of this proposition is mainly based on the following elementary lemma which controls the tails of gamma integrals (see Johnstone, 2001, proof of Lemma 6.1).

Lemma 4 *The following inequality holds for all $z > t > 0$:*

$$\frac{z^{t+1}e^{-z}}{z-t} > I(z) = \int_z^{+\infty} x^t e^{-x}\,dx > \left(1 + \frac{t}{(z-t)^2}\right)^{-1}\frac{z^{t+1}e^{-z}}{z-t}.$$

Proof: One merely notices that the derivative of the function $-x^{t+1}e^{-x}/(x-t)$ is $x^t e^{-x}\left(1+t(x-t)^{-2}\right)$, which implies, for $z > t$, that

$$I(z) < \int_z^{+\infty} x^t e^{-x}\left(1+\frac{t}{(x-t)^2}\right)dx = \frac{z^{t+1}e^{-z}}{z-t} < \left(1+\frac{t}{(z-t)^2}\right)I(z). \qquad \square$$

Proof of Proposition 5: Given $u > 0$, it follows from the preceding lemma that

$$\mathbb{P}[X \geq t+u] = \frac{1}{\Gamma(t)}\int_{t+u}^{+\infty} x^{t-1}e^{-x}\,dx = \frac{(t+u)^t e^{-(t+u)}}{(u+1)\Gamma(t)}\Delta',$$

with $1 > \Delta' > \left[1+t(u+1)^{-2}\right]^{-1}$. Since by Stirling's Formula (see Whittaker and Watson, 1927 p. 258),

$$\Gamma(t) = t^{t-1/2}e^{-t}\sqrt{2\pi}\exp[\theta_t/(12t)] \quad \text{with } 0 < \theta_t < 1,$$

it follows that

$$\mathbb{P}[X \geq t+u] = \Delta\left(1+ut^{-1}\right)^t e^{-u}\sqrt{\delta/(2\pi)}, \tag{A.12}$$

with

$$\delta = t(u+1)^{-2} \qquad \text{and} \qquad \left[(1+\delta)e^{1/(12t)}\right]^{-1} < \Delta < 1.$$

Applying this result with $u = 2t\,(x+\sqrt{x})$, we derive that

$$\begin{aligned}\log\left(\mathbb{P}\left[X \geq t\left(1+2x+2\sqrt{x}\right)\right]\right) &= t\left[\log\left(1+2x+2\sqrt{x}\right) - 2\left(x+\sqrt{x}\right)\right] \\ &\quad - (1/2)\log[2\pi/\lambda] - \Phi,\end{aligned}$$

with $0 < \Phi < (12t)^{-1} + \log(1+\lambda)$, which proves (A.9). As to (A.11), it can be derived from some elementary analytical considerations and numerical computations. $\square$

Corollary 1 *Let Y be a chi-square random variable with D degrees of freedom and $y > 0$. Then*

$$\log\left(\mathbb{P}\left[Y \geq D+2\sqrt{Dy}+2y\right]\right) = -y\theta\left(\frac{y}{D}\right) - \log\left(\frac{y}{\sqrt{D}}+\sqrt{y}\right) - \frac{1}{2}\log(4\pi) - \Psi,$$

where the function θ defined by (A.10) satisfies (A.11) and

$$0 < \Psi < \frac{1}{6D} + \frac{1}{2}\left[\frac{y}{\sqrt{D}}+\sqrt{y}\right]^{-2} + \left(y+\sqrt{Dy}\right)^{-1}.$$

Proof: Since Y has a distribution $\Gamma(D/2, 1/2)$, $X = Y/2$ has a distribution $\Gamma(D/2, 1)$. Applying Proposition 5 with $t = D/2$ and $x = y/D$, we get

$$\begin{aligned}\log\left(\mathbb{P}\left[Y \geq D+2\sqrt{Dy}+2y\right]\right) &= \log\left(\mathbb{P}\left[X \geq (D/2)\left(1+2x+2\sqrt{x}\right)\right]\right) \\ &= -Dx\theta(x) - (1/2)\log(2\pi/\lambda) - \Phi \\ &= -y\theta(y/D) + (1/2)\log(2\lambda) - (1/2)\log(4\pi) - \Phi,\end{aligned}$$

with $0 < \Phi < 1/(6D) + \lambda$ and $\lambda = (D/2)\left[y + \sqrt{Dy} + 1\right]^{-2}$. Moreover

$$2\lambda = \left(\frac{1}{1 + (y + \sqrt{Dy})^{-1}}\right)^2 \left[y/\sqrt{D} + \sqrt{y}\right]^{-2} < \left[y/\sqrt{D} + \sqrt{y}\right]^{-2},$$

and therefore

$$-\left(y + \sqrt{Dy}\right)^{-1} - \log\left(y/\sqrt{D} + \sqrt{y}\right) < (1/2)\log(2\lambda) < -\log\left(y/\sqrt{D} + \sqrt{y}\right),$$

hence our result. □

References

AKAIKE, H. (1969). Statistical predictor identification. *Annals Inst. Statist. Math.* **22**, 203-217.

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

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

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals Inst. Statist. Math.* **30**, Part A, 9-14.

AMEMIYA, T (1985). *Advanced Econometrics*. Basil Blackwell, Oxford.

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

BARRON, A.R. and COVER, T.M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37**, 1034-1054.

BIRGÉ, L. (2001). An alternative point of view on Lepski's method, in *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet* (Mathisca C.M. de Gunst, Chris A.J. Klaassen, Aad W. van der Vaart, eds.), Institute of Mathematical Statistics, Lecture Notes–Monograph Series **36**, 113-133.

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

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

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. To appear in *JEMS*.

DANIEL, C. and WOOD, F.S. (1971). *Fitting Equations to Data*. John Wiley, New York.

DONOHO, D.L. and JOHNSTONE, I.M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Relat. Fields* **99**, 277-303.

DRAPER, N.R. and SMITH, H. (1981). *Applied Regression Analysis, second edition*. Wiley, New York.

FELLER, W. (1968). *An Introduction to Probability Theory and its Applications, Vol. I (Third Ed.)*. John Wiley, New York.

GUYON, X. and YAO, J.-f. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *Jour. Multivariate Analysis* **70**, 221-249.

HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of an autoregression. *J.R.S.S., B* **41**, 190-195.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J.A.S.A.* **58**, 13-30.

HURVICH, K. L. and TSAI, C.-L. (1989). Regression and time series model selection in small samples *Biometrika* **76**, 297-307.

JOHNSTONE, I. (2001). Chi-square oracle inequalities. To appear in *State of the Art in Probability and Statistics; Festschrift for Willem R. van Zwet* (Mathisca C.M. de Gunst, Chris A.J. Klaassen, Aad W. van der Vaart, eds.), Institute of Mathematical Statistics, Lecture Notes–Monograph Series.

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835-866.

LAVIELLE, M. and MOULINES, E. (2000) Least Squares estimation of an unknown number of shifts in a time series. *Jour. of Time Series Anal.* **21**, 33-59.

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. and TSAI, C.-L. (1998). *Regression and Time Series Model Selection.* World Scientific, Singapore.

POLYAK, B.T. and TSYBAKOV, A.B. (1990). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, 293-306.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465-471.

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.

WHITTAKER, E.T. and WATSON, G.N. (1927). *A Course of Modern Analysis.* Cambridge University Press, London.

YAO, Y.C. (1988). Estimating the number of change points via Schwarz criterion. *Stat. and Probab. Letters* **6**, 181-189.

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

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