

Minimal penalties for Gaussian model selection

Lucien Birgé · Pascal Massart

Received: 11 July 2004 / Revised: 24 March 2006
© Springer-Verlag 2006

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 including the classical ones. As compared to our previous paper on this topic (Birgé and Massart in *J. Eur. Math. Soc.* **3**, 203–268 (2001)), more elaborate forms of the penalties are given which are shown to be, in some sense, optimal. We indeed provide more precise upper bounds for the risk of the penalized estimators and lower bounds for the penalty terms, showing that the use of smaller penalties may lead to disastrous results. These lower bounds may also be used to design a practical strategy that allows to estimate the penalty from the data when the amount of noise is unknown. We provide an illustration of the method for the problem of estimating a piecewise constant signal in Gaussian noise when neither the number, nor the location of the change points are known.

Keywords Gaussian linear regression · Variable selection · Model selection · Mallows' C_p · Penalized least-squares

Mathematics Subject Classification (2000) Primary 62G05 · Secondary 62G07 · 62J05

L. Birgé
UMR 7599 “Probabilités et modèles aléatoires”, Laboratoire de Probabilités, boîte 188,
Université Paris VI, 4 Place Jussieu, 75252 Paris Cedex 05, France
e-mail: lb@ccr.jussieu.fr

P. Massart (✉)
UMR 8628 “Laboratoire de Mathématiques”, Bât. 425,
Université Paris Sud, Campus d’Orsay, 91405 Orsay Cedex, France
e-mail: pascal.massart@aliceadsl.fr

1 Introduction

In this paper, we pursue our study of model selection for Gaussian frameworks, as described in Birgé and Massart [11]. We recall that, given some Hilbert space \mathbf{H} with scalar product $\langle \cdot, \cdot \rangle$, a linear isonormal process indexed by a suitable linear subspace \mathbb{S} of \mathbf{H} is a centered and linear Gaussian process with covariance structure $\mathbb{E}[Z(t)Z(u)] = \langle t, u \rangle$ (we have to introduce the subspace \mathbb{S} since one cannot warrant the existence of a linear isonormal process on an arbitrary infinite dimensional Hilbert space).

We consider the statistical problem of estimating the unknown parameter $s \in \mathbf{H}$ when one observes the Gaussian Linear process Y indexed by \mathbb{S} defined by

$$Y(t) = \langle s, t \rangle + \varepsilon Z(t) \quad \text{for all } t \in \mathbb{S}, \quad (1)$$

where Z denotes a linear isonormal process and ε is a known level of noise.

1.1 Model selection from a nonasymptotic point of view

We have at hand some countable (possibly finite) collection $\{S_m, m \in \mathcal{M}\}$ of *models*, i.e. finite-dimensional linear subspaces of \mathbb{S} with respective dimensions D_m (possibly $D_m = 0$). On each model, we build the corresponding least squares estimator \hat{s}_m , i.e. the minimizer, with respect to $t \in S_m$, of the least squares criterion $\gamma(t) = \|t\|^2 - 2Y(t)$. The quality of a model S_m (or alternatively an estimator \hat{s}_m) is given by the corresponding risk

$$R_m(s) = \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] = d^2(s, S_m) + \varepsilon^2 D_m, \quad (2)$$

where $\|\cdot\|$ denotes the norm in \mathbf{H} , d the corresponding distance and \mathbb{E}_s the expectation of functions of the process $Y(\cdot)$ described by (1). An ideal model for s is one which minimizes $R_m(s)$. Unfortunately, we cannot choose this optimal model from (2) since the bias term $d^2(s, S_m)$ is unknown. Model selection consists in designing a data driven choice \hat{m} of a model $S_{\hat{m}}$ for s . It is typically impossible to choose an ideal model from the data, i.e. to design a model selection procedure $\hat{m}(Y)$ such that $\mathbb{E}_s \left[\|\hat{s}_{\hat{m}} - s\|^2 \right] = \inf_{m \in \mathcal{M}} \left\{ \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] \right\}$, but one can hope to build some \hat{m} satisfying

$$\mathbb{E}_s \left[\|\hat{s}_{\hat{m}} - s\|^2 \right] \leq C \inf_{m \in \mathcal{M}} \left\{ \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] \right\}, \quad (3)$$

with $C > 1$ independent of s . The penalization approach to model selection consists of defining

$$\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \operatorname{pen}(m) + \gamma(\hat{s}_m) \right\} = \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \operatorname{pen}(m) - \|\hat{s}_m\|^2 \right\}, \quad (4)$$

where $\operatorname{pen}(\cdot)$ denotes a suitable nonnegative function defined on \mathcal{M} .

As shown in Birgé and Massart [11], Sect. 2.1, this general Gaussian framework allows to deal with various classical model selection problems within one single framework. A typical example is the problem of variable selection in Gaussian linear regression. We observe n independent variables Y_1, \dots, Y_n with $Y_i \sim \mathcal{N}(s_i, \sigma^2)$ which can be written in vector form as

$$Y = s + \sigma \xi \quad \text{with } Y = (Y_i), \quad s = (s_i) \in \mathbb{R}^n \quad \text{and} \quad \xi \sim \mathcal{N}(0, \mathbf{Id}_n). \quad (5)$$

In this case 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

$$Y(t) = \langle Y, t \rangle_n = \langle s, t \rangle_n + \sigma \langle \xi, t \rangle_n = \langle s, t \rangle_n + \varepsilon Z(t), \quad \text{with } \varepsilon = \sigma/\sqrt{n}, \quad (6)$$

where Z is a linear isonormal process indexed by \mathbb{R}^n and $\langle \cdot, \cdot \rangle_n$ denotes the scalar product corresponding to the normalized Euclidean norm $\| \cdot \|_n$ on \mathbb{R}^n defined by $\|t\|_n^2 = n^{-1} \sum_{i=1}^n t_i^2$.

In order to estimate the unknown parameter s , one typically considers a set of potential variables $\{X^\lambda, \lambda \in \Lambda_N\}$, $\Lambda_N = \{1, 2, \dots, N\}$, with $X^\lambda \in \mathbb{R}^n$ and N possibly large. To each subset m of Λ_N corresponds a linear regression model

$$Y_i = \sum_{\lambda \in m} \beta_\lambda X_i^\lambda + \sigma \xi_i \quad \text{for } 1 \leq i \leq n, \quad \text{with } \xi_1, \dots, \xi_n \text{ i.i.d. } \mathcal{N}(0, 1). \quad (7)$$

Building a good model (7) amounts to select some influential variables from the set $\{X^\lambda, \lambda \in \Lambda_N\}$. To be precise, we would like to select a subset m of Λ_N which minimizes (at least approximately) the risk $\mathbb{E}[\|\hat{s}_m - s\|_n^2]$ where \hat{s}_m denotes the least squares estimator corresponding to the stochastic model (7). Using the identification given by (6), assuming that σ is known and denoting by S_m the $|m|$ -dimensional linear space generated by the vectors $X^\lambda, \lambda \in m$, this variable selection problem can be viewed as a model selection problem among the collection $\{S_m, m \subset \Lambda_N\}$, as previously defined.

1.2 Some historical remarks about model selection

Model selection via penalization is an old idea. It amounts to choosing \hat{m} as a minimizer of a penalized criterion of the form $\gamma(\hat{s}_m) + \text{pen}(m)$, where the penalty $\text{pen}(m)$ is usually proportional to the dimension D_m of S_m . In our Gaussian framework, this penalized least squares criterion corresponds to penalized maximum log-likelihood, a criterion which has been used for decades, not only for Gaussian frameworks.

The first examples we know about of such criteria are due to Mallows [29] and Akaike [[2] for FPE, [3] and [4] for AIC]. Mallows' C_p , which, according to Daniel and Wood [13] dates back to the early sixties, was designed to solve

a problem of variable selection in a linear regression problem when the variance of the errors is known (or independently estimated). Translated into the framework given by (1), Mallows' C_p corresponds to setting $\text{pen}(m) = 2\varepsilon^2 D_m$. Akaike's AIC, which has a more general scope, consists in maximizing the maximal log-likelihood on a model S_m minus the number of parameters of the model. Both criteria are based on the idea of unbiased estimation of the 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 [37] to be *asymptotically efficient*, i.e. satisfies asymptotically when $\varepsilon \rightarrow 0$,

$$\mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] \sim \inf_{m \in \mathcal{M}} \left\{ \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] \right\}, \quad (8)$$

at the price of assuming that the true s does not belong to any model in the list. Such a result has also been proved, under various assumptions, by Li [27], Polyak and Tsybakov [33] or Kneip [24]. Coming back to our Gaussian framework, this point of view amounts to define an optimal model S_{m_0} as indexed by a minimizer m_0 of the quadratic risk $\mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right]$ with respect to m . As a consequence, an optimal model does not necessarily contain s as illustrated by the example of Sect. 4.5.

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 with minimal dimension that we want to find: this is the *consistency* point of view. 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, [5] or equivalently Schwarz, [35]) and Hannan and Quinn [20]. For a recent analysis of such criteria, see Guyon and Yao [19].

The distinction between these points of view 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 [31] 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 [22], which definitely improves on AIC for small sample sizes.

In this paper, we focus on a nonasymptotic point of view. 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 . It is indeed of common practical use to introduce more explanatory variables in a regression problem when one has more observations [which corresponds to a smaller value of ε for the associated Gaussian linear process as shown by (6)] 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. The nonasymptotic approach which is based on risk evaluations allows to prove efficiency results but also leads to minimax bounds relative to each model. Such minimax results are actually not compatible with the consistency viewpoint. Indeed, it is shown in a recent paper by Yang [40] that one cannot simultaneously achieve both aims: estimates which tend to find the true model (assuming there is one) do not achieve the optimal minimax risk.

1.3 Choosing proper penalties

The main issue of this paper is what choices of penalty functions are suitable for our purposes, namely getting bounds of the form (3) with an adequate value of C . In Birgé and Massart [11], we proposed penalties of the form

$$\text{pen}(m) = K\varepsilon^2 D_m \left(1 + \sqrt{2L_m}\right)^2 \quad \text{for all } m \in \mathcal{M} \quad \text{and some } K > 1, \quad (9)$$

where the L_m are nonnegative *weights* indexed by \mathcal{M} which satisfy the condition

$$\Sigma = \sum_{\{m \in \mathcal{M} \mid D_m > 0\}} \exp[-L_m D_m] < +\infty. \quad (10)$$

This condition, which resembles Kraft’s inequality in information theory, appeared in various works devoted to the analysis of the performance of penalized criteria, in particular connected to minimum description length or minimum complexity principles. Some milestone references are Rissanen [34] and Barron and Cover [8]. It can also be interpreted in a Bayesian way as putting a prior finite measure on the list of models.

In a typical model selection problem, there is only a finite number of models of a given dimension D . Therefore, denoting by $|A|$ the cardinality of a set A and setting

$$\mathcal{M}_D = \{m \in \mathcal{M} \mid D_m = D\} \quad \text{and} \quad H(D) = D^{-1} \log |\mathcal{M}_D|, \quad (11)$$

we see that one can fix $L_m = H(D_m) + \delta$, with $\delta > 0$ arbitrarily small. Such a choice leads to the lower bound $\varepsilon^2 D_m \left(1 + \sqrt{2H(D_m)}\right)^2$ for the penalty given by (9) (corresponding to the limiting cases $K = 1$ and $\delta = 0$). Our aim is to sharpen the form of the penalty in order to define a “minimal” value of the penalty (in a suitable sense) as well as an “optimal” one and to see how they are related. This leads us to introduce a slightly different form of minimal penalty, namely

$$\text{pen}_{\min}(m) = \varepsilon^2 D_m A(D_m) \quad \text{with } A(D) = 1 + 2\sqrt{H(D)} + 2H(D). \quad (12)$$

One major concern of this paper is to show that, for various types of behaviors of $H(D)$ when D goes to infinity, (12) is actually a lower bound for an

effective penalty in the sense that smaller penalties may lead to inconsistent estimation procedures. In practice this means that penalties below this critical value lead to procedures that systematically choose models of much too large dimensions. Moreover, using penalties of the form $\text{pen}(m) = K\varepsilon^2 D_m A(D_m)$ with $K > 1$, leads to improved risk bounds for $\mathbb{E}_s [\|\hat{s}_{\hat{m}} - s\|^2]$ as compared to Birgé and Massart [11]. In view of these new upper bounds, it turns out that $K = 2$ is always a reasonable choice (non-asymptotically) and sometimes an optimal one, asymptotically. Combining these negative and positive results about the penalty allows us to propose a practical procedure for estimating ε and therefore choosing a good penalty when ε is unknown.

In the next section we introduce the new penalties and the corresponding risk bounds. Then Sect. 3 will be devoted to negative results: we show that (12) is actually a lower bound for effective penalties in typical situations and we also study the consequence of choosing too large penalties. We then deal with the important practical issue of designing data driven penalties when ε is unknown in Sect. 4 and apply the previous results to multiple change points detection in Gaussian noise in Sect. 5. The remainder of the paper is devoted to the proofs.

2 New penalties and the corresponding risk bounds

We recall that our observation is a Gaussian linear process $Y(t)$ given by (1) where Z is a linear isonormal process on the subset \mathbb{S} of the Hilbert space \mathbf{H} , with norm $\|\cdot\|$ and corresponding distance d , and s is an unknown function in \mathbf{H} to be estimated. We have at disposal a countable (possibly finite) collection $\{S_m, m \in \mathcal{M}\}$ of finite dimensional models with respective dimensions $D_m \geq 0$ and we may assume, without loss of generality, that \mathbb{S} is the linear span of $\cup_{m \in \mathcal{M}} S_m$. In this context, we first give a general non-asymptotic risk bound based on our new penalty structure.

Theorem 1 *Given the collection of models $\{S_m\}_{m \in \mathcal{M}}$, let us consider a family of nonnegative weights $\{L_m\}_{m \in \mathcal{M}}$ satisfying (10), two numbers, $\theta \in (0, 1)$ and $\kappa > 2 - \theta$ and let us assume that there exists a finite (possibly empty) subset $\overline{\mathcal{M}}$ of \mathcal{M} such that the penalty function pen satisfies*

$$\text{pen}(m) \geq Q_m \quad \text{for } m \in \mathcal{M} \setminus \overline{\mathcal{M}}, \quad (13)$$

with

$$Q_m = \varepsilon^2 D_m \left(\kappa + 2(2 - \theta)\sqrt{L_m} + 2\theta^{-1}L_m \right) \quad \text{for all } m \in \mathcal{M}. \quad (14)$$

Then the corresponding penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ with \hat{m} given by (4) exists a.s. and satisfies

$$(1 - \theta) \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \text{pen}(m) - \varepsilon^2 D_m \right\} + \sup_{m \in \overline{\mathcal{M}}} \{Q_m - \text{pen}(m)\} + \varepsilon^2 \Sigma \left[(2 - \theta)^2 (\kappa + \theta - 2)^{-1} + 2\theta^{-1} \right], \tag{15}$$

where $d(s, S_m) = \inf_{t \in S_m} \|t - s\|$ denotes the distance from s to the space S_m . If, in particular, we fix

$$\kappa = 2 \quad \text{and} \quad \text{pen}(m) = Q_m \quad \text{whatever } m \in \mathcal{M}, \tag{16}$$

then

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

The proof of Theorem 1 being rather technical, it will be deferred to Sect. 5.2.

Remarks i) A typical penalty choice would be (16). Allowing that $\overline{\mathcal{M}} \neq \emptyset$ will be useful in Sect. 3 for comparing upper and lower bounds for the penalty.

ii) As a consequence of (17), if one can find a bounded family of weights ($\sup_m L_m = L < +\infty$) satisfying (10) (which is possible when the number of spaces S_m having the same dimension $D_m = D$ is not too large) and if the penalty is given by (16) with $\theta = 1/2$, say, one derives from (2) that

$$\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] \leq 2 \left(1 + 3\sqrt{L} + 4L \right) \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] + 17\varepsilon^2 \Sigma. \tag{18}$$

This means that, if no estimator in the family is close to perfect for estimating s , i.e. if $\inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] \geq \varepsilon^2$, the penalized estimator \tilde{s} satisfies (3) with $C = 2 \left(1 + 3\sqrt{L} + 4L \right) + 17\Sigma$ and therefore behaves as well as the best least squares estimator in the family, up to the constant C .

iii) In Birgé and Massart [11], Theorem 2, we proved the following risk bound for the penalized least squares estimator \tilde{s} based on penalty (9):

$$\mathbb{E}_s \left[\|\tilde{s} - s\|^2 \right] \leq \frac{4K(K + 1)^2}{(K - 1)^3} \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + K\varepsilon^2 D_m \left(1 + \sqrt{2L_m} \right)^2 \right\} + \frac{4K(K + 1)^3}{(K - 1)^3} \varepsilon^2 \Sigma. \tag{19}$$

It is easy to see that, whatever $K > 1$, this upper bound is larger than

$$4 \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 D_m \left[1 + 2\sqrt{2L_m} + 2L_m \right] \right\} + 65\varepsilon^2 \Sigma$$

and therefore always worse than (17) with $\theta = 1/2$.

With our new risk bound, it is possible to define situations for which the penalized estimator is asymptotically efficient. Looking at (17), we see that we can hope to achieve this by letting θ tend to zero with ε and the L_m as well. We consider here a family of model selection problems with a varying noise level $\varepsilon \in (0; \varepsilon_0]$, a given collection of models independent of ε and we let ε go to zero. The assumptions involving limits when D goes to infinity should be considered as automatically satisfied if the dimensions of the models are bounded (i.e. $\sup_{m \in \mathcal{M}} D_m < +\infty$).

Corollary 1 *Let $\{S_m, m \in \mathcal{M}\}$ be a given collection of linear subspaces of some Hilbert space \mathbf{H} with respective finite dimensions D_m , $\mathcal{M}_D = \{m \in \mathcal{M} \mid D_m = D\}$ and $\mathcal{D} = \{D \geq 0 \mid \mathcal{M}_D \neq \emptyset\}$. We assume that $|\mathcal{M}_D|$ is finite for every $D \in \mathcal{D}$, that $\log |\mathcal{M}_D| = o(D)$ when D goes to infinity in \mathcal{D} , that the true parameter s does not belong to any model (i.e. $d(s, S_m) > 0$ for all $m \in \mathcal{M}$) and that the penalty function is given, for each noise level $\varepsilon \in (0; \varepsilon_0]$, by $\text{pen}(m) = F(\varepsilon, D_m)$ with*

$$\lim_{D \rightarrow +\infty} \sup_{0 < \varepsilon \leq \varepsilon_0} \left| \left[\varepsilon^2 D \right]^{-1} F(\varepsilon, D) - 2 \right| = 0 \tag{20}$$

and

$$\lim_{\varepsilon \rightarrow 0} F(\varepsilon, D) = 0 \quad \text{for all } D \in \mathcal{D}. \tag{21}$$

Then the penalized least squares estimator \tilde{s} satisfies

$$\overline{\lim}_{\varepsilon \rightarrow 0} \frac{\mathbb{E}_s [\|s - \tilde{s}\|^2]}{\inf_{m \in \mathcal{M}} \mathbb{E}_s [\|s - \hat{s}_m\|^2]} \leq 1, \tag{22}$$

where \hat{s}_m denotes the least squares estimator corresponding to the model S_m .

Proof Let us fix $\theta \in (0, 1)$ and set $\kappa = 2 - (\theta/2)$, $L(D) = H(D) + D^{-1/2}$ for $D \geq 1$ and H given by (11) and $L_m = L(D_m)$ with $L(0) = 1$. Here, and throughout the proof, the values of D are restricted to the set \mathcal{D} . The assumption on $\log |\mathcal{M}_D|$ implies that $L(D)$ tends to 0 when D goes to infinity. Therefore if

$$Q(\varepsilon, D) = \varepsilon^2 D \left[\kappa + 2(2 - \theta)\sqrt{L(D)} + 2\theta^{-1}L(D) \right],$$

then $[\varepsilon^2 D]^{-1} Q(\varepsilon, D)$ converges to $\kappa < 2$ when D tends to infinity and (20) implies that there exists some integer D_θ , depending only on θ , such that for every $\varepsilon \in (0, \varepsilon_0]$ and every $D > D_\theta$,

$$Q(\varepsilon, D) \leq F(\varepsilon, D) \leq (2 + \theta)\varepsilon^2 D. \tag{23}$$

Setting $Q_m = Q(\varepsilon, D_m)$, we conclude that $\text{pen}(m) = F(\varepsilon, D_m) \geq Q_m$ for every $m \in \mathcal{M} \setminus \bar{\mathcal{M}}$, with $\bar{\mathcal{M}} = \{m \in \mathcal{M} \mid D_m \leq D_\theta\}$, which allows us to apply Theorem 1 and derive from (15) the risk bound

$$(1 - \theta)\mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + F(\varepsilon, D_m) - \varepsilon^2 D_m \right\} + \sup_{D \leq D_\theta} Q(\varepsilon, D) + \frac{10}{\theta} \varepsilon^2 \Sigma,$$

with

$$\Sigma = \sum_{D \in \mathcal{D}; D \geq 1} \exp[DH(D) - DL(D)] \leq \sum_{D \geq 1} \exp \left[-\sqrt{D} \right] < +\infty.$$

Setting $b(D) = \inf_{m \in \mathcal{M}_D} d(s, S_m)$, we deduce from our assumptions ($d(s, S_m) > 0$ for all m and $|\mathcal{M}_D| < +\infty$) that $b(D) > 0$ and derive from the previous risk bound that there exists a positive constant C_θ , depending only on θ , such that

$$(1 - \theta) \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq \inf_{D \in \mathcal{D}} \left\{ b^2(D) + F(\varepsilon, D) - \varepsilon^2 D \right\} + C_\theta \varepsilon^2. \tag{24}$$

Then, on the one hand, it follows from (23) that $F(\varepsilon, D) - \varepsilon^2 D \leq (1 + \theta)\varepsilon^2 D$ for $D > D_\theta$ and, on the other hand, (20) implies that $F(\varepsilon, D) \leq \theta b^2(D)$ for $D \leq D_\theta$ and $\varepsilon \leq \varepsilon_1$ for some small enough $\varepsilon_1 > 0$. It follows that, when $\varepsilon \leq \varepsilon_1$,

$$b^2(D) + F(\varepsilon, D) - \varepsilon^2 D \leq (1 + \theta) \left[b^2(D) + \varepsilon^2 D \right] \quad \text{for all } D \in \mathcal{D},$$

and by (24),

$$(1 - \theta) \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq (1 + \theta) \inf_{D \in \mathcal{D}} \left\{ b^2(D) + \varepsilon^2 D \right\} + C_\theta \varepsilon^2 = (1 + \theta) \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|s - \hat{s}_m\|^2 \right] + C_\theta \varepsilon^2. \tag{25}$$

Since $b(D) > 0$ for all D ,

$$\varepsilon^{-2} \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|s - \hat{s}_m\|^2 \right] = \inf_{m \in \mathcal{M}} \left\{ \varepsilon^{-2} d^2(s, S_m) + D_m \right\} = \inf_{D \in \mathcal{D}} \left\{ \varepsilon^{-2} b^2(D) + D \right\}$$

tends to infinity when $\varepsilon \rightarrow 0$. Therefore the right-hand side of (25) is equivalent to $(1 + \theta) \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|s - \hat{s}_m\|^2 \right]$ when $\varepsilon \rightarrow 0$ and

$$\overline{\lim}_{\varepsilon \rightarrow 0} \frac{\mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right]}{\inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|s - \hat{s}_m\|^2 \right]} \leq \frac{1 + \theta}{1 - \theta}.$$

The conclusion follows by letting θ go to 0. □

Note that the choice $F(\varepsilon, D) = 2\varepsilon^2 D$ corresponds to Mallows' C_p so recovering the asymptotic efficiency results due to Shibata [37].

It is actually possible to prove a better result than (22), with $\mathbb{E}_s [\inf_{m \in \mathcal{M}} \|s - \hat{s}_m\|^2]$ in the denominator, at the price of a substantially more complicated proof. We refer the interested reader to Birgé and Massart [12], Proposition 1.

3 Some potential difficulties connected with bad penalty choices

It follows from Theorem 1 that a proper choice of the penalty should be of the form

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

and the limiting condition

$$\text{pen}(m) > \varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m\right) \quad (26)$$

is required for our proof of Theorem 1 to work, which, of course, does not mean that a smaller choice of the penalty should necessarily lead to a bad estimator. Similarly, the choice of a large value of κ in (14) leads to larger upper bounds for the risk in (15), but this does not mean that the risk itself is necessarily larger. It is therefore desirable to know whether these restrictions which come out from our proofs are indeed necessary or not. This section is devoted to showing that, for some families of models, these restrictions are actually perfectly justified in the sense that, if these conditions are violated, the penalized estimator can behave quite poorly for some values of the unknown parameter s .

3.1 Lower bounds for the penalty term

3.1.1 Position of the problem

It is actually not obvious to give a precise formal meaning to what is a lower bound on the penalty. If, for instance, (10) 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 1, that a penalty violating (26) 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 (26) is ill-posed without further restrictions on the values of the weights L_m .

In order to overcome this difficulty, we shall 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_m 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 [11]. With $H(D)$ given by (11), we see that (10) 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 $\kappa = 1 + 2\delta$, leads to $\Sigma \leq (e^\delta - 1)^{-1}$ and

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

with $A(D)$ given by (12). This implies, by Theorem 1, that any penalty of the form

$$\text{pen}(m) = (1 + \eta)\varepsilon^2 D_m A(D_m), \quad \eta > 0 \quad \text{for all } m \in \mathcal{M} \text{ such that } D_m > \bar{D}, \tag{27}$$

satisfies (13) provided that δ is small enough and results in a risk bound of the form

$$\begin{aligned} &\mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \\ &\leq C(\eta) \left(\inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2 [D_m A(D_m) + 1] \right\} + \varepsilon^2 \sup_{1 \leq D \leq \bar{D}} \{DA(D)\} \right). \end{aligned} \tag{28}$$

Our purpose in the next three sections will be to prove that if (27) is violated, i.e.

$$\text{pen}(m) \leq (1 - \eta)\varepsilon^2 D_m A(D_m) \tag{29}$$

for some $\eta > 0$ and D_m sufficiently large, then the risk $\mathbb{E}_s [\|s - \tilde{s}\|^2]$ 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 (29) is compatible with (27) 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 (28) blow up.

The behaviour of $A(D)$ when D is large actually depends on the size of $H(D)$. When going to practical examples (many of them can be found in Birgé and Massart [11]), one typically encounters three different situations:

1. For each $D \geq 1$ the number $e^{DH(D)}$ of those indices m such that $D_m = D$ is not large (bounded by a polynomial function of D , say), which means that $H(D)$ goes to zero when D goes to infinity;

2. The number of indices m such that $D_m = D$ is much larger, typically of order $\binom{N}{D}$ where N is a large parameter and then $H(D)$ is of order $\log(N/D)$;
3. The number of indices m such that $D_m = D$ is moderate, which means that $H(D)$ remains bounded away from zero and infinity when D goes to infinity.

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 (29) holds for some m for which D_m is large enough.

3.1.2 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 (i.e. $N \in \mathcal{M}$). 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 which, in particular includes the case of nested models. More generally, we consider the case when $H(D) \leq \bar{H}(D)$ for some function $\bar{H}(j)$ converging to zero when j goes to infinity, which implies that $\sum_{\{m \in \mathcal{M} \mid D_m > 0\}} \exp[-LD_m] \leq \Sigma_L$ 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 $\text{pen}(m) = (1 + \eta)\varepsilon^2 D_m$ with $\eta > 0$ satisfies (13) with $\bar{\mathcal{M}} = \emptyset$, provided that $L, 1 - \theta$ and $\kappa - 1$ are small enough, depending on η , which results, by Theorem 1, in a risk bound of the form

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

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 Sect. 5.3.

Proposition 1 *Assume that there exists some positive number η such that*

$$\text{pen}(N) - \text{pen}(m) \leq (1 - \eta)\varepsilon^2(N - D_m), \tag{30}$$

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 η, θ, \bar{H} and δ but neither on s nor on ε such that, for $N \geq N_0$,

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

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 \mathbf{H} (the trigonometric system or a wavelet basis on $[0, 1]$, for instance) and that S_m is the linear span of $\{\varphi_1, \dots, \varphi_m\}$ 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 $\text{pen}(m) = 2\varepsilon^2 m$, for all m , it follows from Theorem 1 that, whatever s , the risk will be bounded independently of N by

$$\mathbb{E}_s [\|s - \hat{s}\|^2] \leq C \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \varepsilon^2(m + 1) \right\}, \tag{31}$$

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 we set $\text{pen}(m) = (1 - \eta)\varepsilon^2 m$. In this case, Proposition 1 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 (31) for a larger value of N .

3.1.3 A large number of models: variable selection

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. This actually corresponds to the situation of variable selection in Gaussian linear regression described in Sect. 1.3. We also assume that the penalty function $\text{pen}(m)$ only depends on m through its cardinality $|m|$ which is the dimension D_m of S_m .

Proposition 2 *Let s be the true unknown function to estimate and set $\Lambda_1 = \{\lambda \in \Lambda_N \mid \langle s, \varphi_\lambda \rangle \neq 0\}$. Assume that there exist numbers δ, α, A and η with*

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

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

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

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

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

The proof is given in Sect. 5.4. 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

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

It then follows from Theorem 1 that, if the penalty takes the form

$$\text{pen}(m) = (1 + \eta) \varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m \right) \quad \text{with } \eta > 0, \tag{32}$$

then

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

whatever s . In particular, if s satisfies the assumptions of Proposition 2 with $|\Lambda_1| \geq 3$,

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

On the other hand, by Proposition 2, under the same assumptions, if

$$\begin{aligned} \text{pen}(\bar{m}) &\leq (2 - 2\alpha - \eta)(1 - \delta) \varepsilon^2 |\bar{m}| \log N \\ &= \frac{(1 - \delta)(1 - \alpha - \eta/2)}{1 - \alpha} \varepsilon^2 |\bar{m}| [2(1 - \alpha) \log N], \end{aligned} \tag{33}$$

then

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

when N is large enough. This implies that, for large values of N , the estimator associated to some too small value of the penalty of the form $(1 - \eta')2(1 - \alpha) \varepsilon^2 D_m \log N$ with $\eta' > 0$, will have a risk which is much larger than 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 $N \rightarrow +\infty$ to see it. Comparing (32) with $m = \bar{m}$ and $|\bar{m}| \sim N^\alpha$ together with (33), we see that

$$\text{pen}(m) = \varepsilon^2 D_m [1 + 2 \log(N/D_m)] \quad (34)$$

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 [11], Sect. 6.3.4. Penalties of the same order of magnitude as the one given in (34) have been introduced recently by several authors among which George and Foster [17], Barron, Birgé and Massart [7], Birgé and Massart [11], Efron, Hastie, Johnstone and Tibshirani [15] — see the discussion by Loubes and Massart [28] following the paper—and Abramovich, Benjamini, Donoho and Johnstone [1]. We refer to Sect. 1.9 of the latter paper for a detailed discussion on this matter.

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 of variable selection in the Gaussian linear regression set up (7) of Sect. 1.3. Deciding which variables pertaining to a given set $\{X^\lambda, \lambda \in \Lambda_N\}$, with $\Lambda_N = \{1, 2, \dots, N\}$, 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 $\{X^\lambda\}_{1 \leq \lambda \leq k}$ with $k \leq N$, while *complete variable selection* corresponds to select any subset of Λ_N . Although many Econometrics books do deal with this subject, most of them become indeed rather elusive (see for instance Chapter 2 of Amemiya, [6]) as to the choice of a suitable penalty for the second situation and some ([14] 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 [31] does not distinguish quite explicitly 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 pursue their analysis further.

It follows from Proposition 2 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 ([31], p. 62) show that stronger penalties should be preferred to C_p . This suggests that although our lower bound (34) for the penalty follows from asymptotic considerations, it seems to be quite relevant for practical nonasymptotic use.

3.1.4 The intermediate case

In order to deal with the intermediate case corresponding to $H(D)$ being neither small nor large when D is large, we have to introduce a more complicated

set up: 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 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,

$$\exp(\alpha D) - 1 < |\mathcal{M}_D| \leq \exp(\alpha D) \quad \text{for some } \alpha > 0. \tag{35}$$

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

$$L_m = L(D_m) \quad \text{with} \quad L(D) = \alpha + \beta D^{-1} \log(D + 1) \quad \text{for some } \beta > 1, \tag{36}$$

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 1 that the risk of a suitably tuned penalized estimator will be bounded by $17\Sigma\varepsilon^2$. On the other hand, if (26) 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 shall prove the following in Sect. 5.5.

Proposition 3 *Assume that the family of models at hand is as described just before, that L_m is given by (36) and that $s = 0$. Let $0 < \lambda < F(\alpha)$ where F denotes some specific function defined on $(0, +\infty)$ which satisfies $5/6 < F(x) < 1$ for $x > 0$ and $F(x)$ converges to one when x converges either to 0 or to infinity. Let \bar{D} be some large enough integer depending on α, β and λ and $\bar{\mathcal{M}} \subset \{m \in \mathcal{M} \mid D_m \geq \bar{D}\}$. Assume that*

$$\text{pen}(m) \leq \lambda\varepsilon^2 D_m \left(1 + 2\sqrt{L_m} + 2L_m\right) \quad \text{for all } m \in \bar{\mathcal{M}}. \tag{37}$$

- If $\bar{\mathcal{M}}$ is infinite, then, with a probability larger than 1/2, $\inf_{m \in \mathcal{M}} \{\hat{\gamma}(m) + \text{pen}(m)\} = -\infty$ and \hat{m} is not defined.
- If $\bar{\mathcal{M}}$ is nonempty, finite and the function pen satisfies the assumptions of Theorem 1, i.e. $\text{pen}(m) \geq Q_m$ for $m \in \mathcal{M} \setminus \bar{\mathcal{M}}$ with Q_m given by (14), then there exists a constant $C > 0$, depending only on α and λ , such that

$$\mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \geq C\varepsilon^2 \left(\sup_{m \in \bar{\mathcal{M}}} D_m \right). \tag{38}$$

This proposition says that (up to some factor) the lower bound (26) is tight if the only assumption we have on the family of models is that $|\mathcal{M}_D|$ behaves as an exponential function of D . This holds for some examples of interest like the parsimonious variable selection strategy connected with adaptive estimation in Besov balls described in Birgé and Massart [11], Sect. 6.4, or with the histogram selection pruning procedure associated with CART (see Gey and Nédélec [18]). Unfortunately we were unable to prove an analogue of Proposition 3 for these

motivating examples. Instead, we have considered a somehow artificial family of models which has the virtue to provide a proveable counter-example.

The comparison between the lower bound (26) and (37) shows that (26) is tight up to a factor $F(\alpha) \in (5/6, 1)$. The suboptimal factor $F(\alpha)$ (instead of one) is due to the fact that the proof of Theorem 1 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 2 and (74) with $\rho = 0$ and $b = 2$).

3.2 The effect of choosing too large penalties

It follows from the preceding results that the choice $\kappa > 1$ in (14) is perfectly justified. Moreover the risk bounds in Theorem 1 suggest to choose penalties of the form $\text{pen}(m) = Q_m$ with a moderate value of κ like in (16). In order to analyze what would be the effect of choosing a substantially larger penalty we can use the next theorem which covers many typical examples. Its proof is given in Sect. 5.6.

Theorem 2 *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 (10) with $\Sigma < 1$. Let \tilde{s} be the penalized least squares estimator corresponding to a penalty such that $\text{pen}(0) = 0$ and*

$$\text{pen}(m) \geq \varepsilon^2[(3/2)D_m + 4L_mD_m + 2A] \quad \text{for all } m \in \mathcal{M}^* = \mathcal{M} \setminus \{0\}. \quad (39)$$

Then

$$\sup_{s \in S_1} \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \geq A(1 - \Sigma)\varepsilon^2, \quad (40)$$

while, if the penalty is given by (16),

$$\sup_{s \in S_1} \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq \varepsilon^2 \left[2 \left(1 + 3\sqrt{L_1} + 4L_1 \right) + 17\Sigma \right]. \quad (41)$$

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 this assumption, 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 (40) and (41) that a value of A substantially larger than $1 \vee L_1$ 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 $\{\varphi_1, \dots, \varphi_m\}$ 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 $\sup_{s \in \mathcal{S}_1} \mathbb{E}_s [\|s - \tilde{s}\|^2] \leq \varepsilon^2 [16 + 17/(e - 1)]$ if the penalty is given by (16), i.e. $\text{pen}(m) = 9\varepsilon^2 D_m$. On the other hand, (40) immediately shows that penalties of the form $\text{pen}(m) = C\varepsilon^2 D_m$ with a too large value of C should be avoided.

Another interesting illustration is connected to the variable selection problem of Sect. 1.3 and 3.1.3 where S_m is the linear span of $\{\varphi_j | j \in m\}$ for m some arbitrary nonempty subset of $\{1, \dots, N\}$ and $\{\varphi_1, \dots, \varphi_N\}$ some orthogonal system in \mathbf{H} . By convention $S_\emptyset = \{0\}$. If $L_m = 2 + \log(N/D_m)$ for $m \neq \emptyset$, then $\Sigma \leq (e - 1)^{-1} < 1$ (see Birgé and Massart [11], Sect. 5.1.2) and the assumptions of Theorem 2 are satisfied. Let $s = \lambda\varphi_j$ for some j . If we choose $\text{pen}(m) = 5\varepsilon^2 D_m [3 + \log(N/D_m)]$, we derive from Theorem 1 with $\theta = 1/2$ and $\kappa = 2$ that

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

On the other hand, if we set $\text{pen}(m) = C\varepsilon^2 D_m [3 + \log(N/D_m)]$ with $C > 4$, then, for $m \neq \emptyset$,

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

which corresponds to $2A = (C - 4)[3 + \log N]$ in (39). We conclude from (40) that $\mathbb{E}_s [\|s - \tilde{s}\|^2] \geq (C - 4)[3 + \log N]\varepsilon^2/5$ for some s of the required form. Once again, this shows that too large values of C should be avoided.

4 Introducing estimated penalties

Up to now we have 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. One could try to estimate it independently from the model selection procedure and plug the resulting estimator in the penalty term. This will work in simple situations like selection of some variables among a set of cardinality substantially smaller than the number of observations. In more complicated situations, such an estimator may be difficult to find or grossly overestimate the true level of noise ε because of a high bias (think of the case of more variables than observations). We propose here a method to solve this problem which is based on a mixture of theoretical and heuristic ideas: rather than estimating ε , we shall actually try to estimate the penalty itself, or calibrate it using

the data at hand. This clearly results in a data-based choice of the penalty. The introduction of data-driven penalties has been considered in several papers based on different types of arguments and motivations. A Bayesian point of view appears in George and Foster [17] using hierarchical Bayesian models introduced by Mitchell and Beauchamp [32]. Their prior measures depend on some hyperparameters which are estimated by the empirical Bayes principle, which results in a random penalty. Another approach is developed in Shen and Ye [36] who consider penalties of the form $\lambda \varepsilon^2 D$ and estimate λ by the minimization of some estimated risk. In their paper, Efron, Hastie, Johnstone and Tibshirani [15] introduce the algorithm of least angle regression and propose some C_p -type model selection criterion, which can be interpreted, as shown by Loubes and Massart [28], as a randomly penalized least squares criterion. In all these papers, the resulting penalty takes the form $\hat{C}(\varepsilon, D_m)$ and since the level of noise is unknown, it has to be estimated separately by a plug-in method. On the contrary, our method consists in calibrating the penalty directly without estimating ε .

4.1 Minimal penalties

We consider a family of models $\{S_m\}_{m \in \mathcal{M}}$ which 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. We assume that the number of models of a given dimension D is finite and restrict ourselves to penalties which depend on the dimension only. The results of Sect. 3.1 justify the introduction of the *minimal* penalty given by (12), minimal meaning here that if we choose $\text{pen}(m) = K \text{pen}_{\min}(m)$ with $K < 1$, then $D_{\hat{m}}$ tends to be close to the dimension of the largest models. This phenomenon, which emerges from the theoretical results of Sect. 3.1, is also strikingly visible on simulated data.

4.2 Optimal penalties

On the other hand, if $K > 1$, the model selection procedure based on the penalty $\text{pen} = K \text{pen}_{\min}$ works in the sense that one can apply Theorem 1 which leads to the risk bound (28). This results in a value of $D_{\hat{m}}$ which is substantially smaller than the dimension of the largest models, at least when $\text{argmin}_{m \in \mathcal{M}} \mathbb{E}_s [\|s - \hat{s}_m\|^2]$ corresponds to a model of moderate dimension. Moreover, choosing $K = 2$ (or close to two) appears to be a reasonable and in some cases optimal strategy. Indeed, in the situation described by Corollary 1, where there are only few models of a given dimension, choosing $K = 2$ leads to an asymptotically optimal penalty. Another extremal case occurs when the function H is constant with a large value L . If we fix $K > 1$ and apply Theorem 1, the risk bound (15) becomes when L goes to infinity

$$\mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq \frac{K}{K-1} \left[\inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + 2KL\varepsilon^2 D_m \right\} + \varepsilon^2 \mathcal{O}_L(1) \right].$$

If we try to minimize this upper bound with respect to K , which amounts, when L goes to infinity, to minimize $K^2/(K-1)$, we find the solution $K=2$. In any case, it follows from Theorem 1 that the choice $K=2$, which means $\text{pen} = 2 \text{pen}_{\min}$, is always reasonable. Therefore, finding a reasonable penalty essentially amounts to estimate pen_{\min} from the data when ε is unknown.

4.3 Data driven penalties

We shall retain from the preceding theoretical results the following facts: there exists a minimal penalty of the form $\text{pen}_{\min}(m) = \varepsilon^2 D_m F(D_m)$ and, if we consider the continuous family of penalties $\text{pen}_\alpha(m) = \alpha D_m F(D_m)$ for $\alpha > 0$, the choice $\alpha < \varepsilon^2$ leads to an explosion of the model selection procedure, while $\alpha = 2\varepsilon^2$ provides a good (sometimes nearly optimal) penalty. Therefore, to design a good penalty, we shall proceed in two steps. First fix a basic functional form for $F(D_m)$. The theory tells us that $F(D_m) = 1 + 2\sqrt{H(D_m)} + 2H(D_m)$ is adequate but we do not pretend that it is the only reasonable choice. In particular, taking $F(D_m)$ as a constant may be a more attractive choice because of its simplicity, or one could think of optimizing the shape of F from simulation studies with a known value of ε .

The second step consists in varying the parameter α and computing the corresponding model choices \hat{m}_α where \hat{m}_α denotes the minimizer with respect to $m \in \mathcal{M}$ of $\text{pen}_\alpha(m) - \|\hat{s}_m\|^2$. Considering the values of $D_{\hat{m}_\alpha}$ for slowly increasing values of α starting from $\alpha = 0$, one typically observes that, for small values of α , those values stay very large and they suddenly jump to a much smaller value when α reaches some threshold $\hat{\alpha}$. In other words the explosion phenomenon occurs for $\alpha < \hat{\alpha}$, which suggests to retain $\hat{\alpha}$ as our estimator for ε^2 and to define our final data driven penalty as $\text{pen}(m) = 2\hat{\alpha} D_m F(D_m)$.

In this context, it is important, in order to evaluate the quality of a given procedure with a data driven penalty, to define a proper benchmark for the risk of a penalized estimator. From this point of view, $\inf_{m \in \mathcal{M}} \mathbb{E}_s [\|\hat{s}_m - s\|^2]$ is not adequate when the function H is such that $H(D)$ becomes large with D . We indeed know (see Birgé and Massart [11], Sect. 5.2) that (3) can then only hold with a large value of C that we cannot evaluate sharply. Since our penalization strategy gives the same penalty to all models of a given dimension D , the selection procedure actually selects an estimator of the form $\hat{s}_D = \operatorname{argmin}_{t \in \cup_{m \in \mathcal{M}_D} S_m} \mathcal{Y}(t)$, with \mathcal{M}_D as in (11). Hence, a natural benchmark for a selection procedure of that kind, which aims at choosing the best \hat{s}_D , is

$$\inf_{D \in \mathcal{D}} \mathbb{E}_s [\|\hat{s}_D - s\|^2], \quad \text{with } \mathcal{D} = \{D_m, m \in \mathcal{M}\}. \quad (42)$$

Now, one can hope that a properly designed penalty will lead to an estimator $\hat{s}_{\hat{D}}$ with a risk close to (42). This is helpful when one wants to calibrate penalized procedures from simulated data since the benchmark (42) can then be approximated by Monte-Carlo.

4.4 Application: change points in a Gaussian signal

In order to illustrate the results of the previous section, we shall consider the following change points problem: we observe an unknown signal s at times $x_1 < x_2 < \dots < x_n$, ($n \geq 2$) with homoscedastic Gaussian errors assuming that the signal is piecewise constant on the interval $[x_1, x_n]$, i.e. $s = \sum_{j=0}^p \beta_j \mathbb{1}_{J_j}$ where $\{J_0, \dots, J_p\}$ is a partition of $[x_1, x_n]$ into $p + 1 \geq 1$ successive intervals. Neither the levels β_j nor the location and number of the change points are known. Equivalently, we observe a Gaussian vector Y with n independent coordinates Y_1, \dots, Y_n given by

$$Y_i = \sum_{j=0}^p \beta_j \mathbb{1}_{J_j}(x_i) + \sigma \xi_i \quad \text{for } 1 \leq i \leq n, \quad \text{with } \xi_1, \dots, \xi_n \text{ i.i.d. } \mathcal{N}(0, 1). \quad (43)$$

This corresponds to the regression framework (5) with $s_i = s(x_i)$. As explained in Sect. 1.3, it can be turned to a Gaussian Linear process $Y(t) = \langle Y, t \rangle_n = n^{-1} \sum_{i=1}^n Y_i t_i$ indexed by $\mathbf{H} = \mathbb{R}^n$ with its normalized Euclidean norm given by $\|t\|_n^2 = n^{-1} \sum_{i=1}^n t_i^2$. Our aim is now to estimate the vector $(s_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ that we shall again denote by s .

The problem of detecting the change points in a piecewise constant signal has already been considered by Yao [41] and more recently by Lavielle and Moulines [25] 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 vector s with a small quadratic risk for a given value of σ and n in (43). This is a situation where it might be better to ignore some of the change points corresponding to small jumps of s .

Given a number D of change points, $0 \leq D \leq n - 1$ and a subset $m = \{i_1 < i_2 < \dots < i_D\}$ of $\{2; \dots; n\}$ (with $m = \emptyset$ if $D = 0$) and setting $i_0 = 1$, $i_{D+1} = n + 1$, we consider the associated D_m -dimensional linear subspace of \mathbb{R}^n defined by

$$S_m = \left\{ \left(\sum_{j=0}^D \beta_j \mathbb{1}_{I_j}(x_i) \right)_{1 \leq i \leq n} \mid \boldsymbol{\beta} = (\beta_0, \dots, \beta_D)^t \in \mathbb{R}^D \right\} \quad \text{with } I_j = [x_{i_j}, x_{i_{j+1}-1}].$$

Hence $D_m = |m| + 1$. Defining by \mathcal{M} the set of all subsets m of $\{2; \dots; n\}$, we use the family $\{S_m\}_{m \in \mathcal{M}}$ to define a penalized least squares estimator of s . If we denote by \mathcal{M}_D the set of all $m \in \mathcal{M}$ such that $D_m = D + 1$, we get

$$\log |\mathcal{M}_D| = \log \binom{D}{n-1} \leq D[1 + \log(n-1) - \log D],$$

with the usual convention $0 \log 0 = 0$. Setting $L_\emptyset = 2$ and $L_m = 2 + \log(n - 1) - \log |m|$ for $m \in \mathcal{M}, m \neq \emptyset$, we get

$$\begin{aligned} \sum_{m \in \mathcal{M}} \exp(-L_m D_m) &= e^{-2} + \sum_{D=1}^{n-1} |\mathcal{M}_D| \exp[-(D + 1)(2 + \log(n - 1) - \log D)] \\ &< e^{-2} + \sum_{D=1}^{n-1} \exp[-D - 1] < e^{-2} \left[1 + (1 - e^{-1})^{-1} \right]. \end{aligned}$$

An application of Theorem 1 shows that, if the weights and penalty function are given by (16), the penalized least squares estimator \tilde{s} associated to the models S_m satisfies a risk bound of the form

$$\begin{aligned} \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] &\leq C \left[\inf_{m \in \mathcal{M} \setminus \emptyset} \left\{ d^2(s, S_m) + \varepsilon^2 |m| \left[1 + \log \frac{n - 1}{|m|} \right] \right\} \wedge \left(d^2(s, S_\emptyset) + \varepsilon^2 \right) \right]. \quad (44) \end{aligned}$$

The presence of the $\log((n - 1)/|m|)$ factor in the risk when $|m| \geq 1$ 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 [10] or Theorem 5 of Birgé and Massart [11].

4.5 A simulation study

In this context of multiple change points detection our data driven procedure with a calibration based on the comparison with (42) has been successfully implemented by Lebarbier [26] with a resulting risk indeed close to the benchmark. Let us now give below a brief account of her results. In a first step she performed many simulations on various signals in order to determine a proper shape for the F function involved in the data driven penalty choice described in Sect. 4.3. She ended with $F(D) = 2 \log(n/D) + 5$. With her kind permission we present here a toy example illustrating the performance of the method based on this value of F . The observation points $x_i \in [0; 1]$ are i/n with $n = 100$ and $1 \leq i \leq 100$. She had at hand one noisy test sample Y_1, \dots, Y_n with $Y_i = s(i/n) + \xi_i$ where the signal and the noise variance are given respectively by

$$s(x) = 0.5 \mathbf{1}_{(0.2, 0.5]} + 2 \mathbf{1}_{(0.5, 1]} \quad \text{and} \quad \sigma^2 = 1.$$

It is possible here to evaluate by Monte-Carlo the partition m_0 which minimizes the risk $\mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right]$ for these particular values of n, s and σ^2 . It has a unique change point at 0.5 and the resulting \hat{s}_{m_0} based on the test sample Y_1, \dots, Y_n is shown in Fig. 1 (black line) together with the corresponding noisy signal (the set of points). Note that s_{m_0} differs from s illustrating the fact that the “true” model (with two change points) does not necessarily minimize the risk. Our

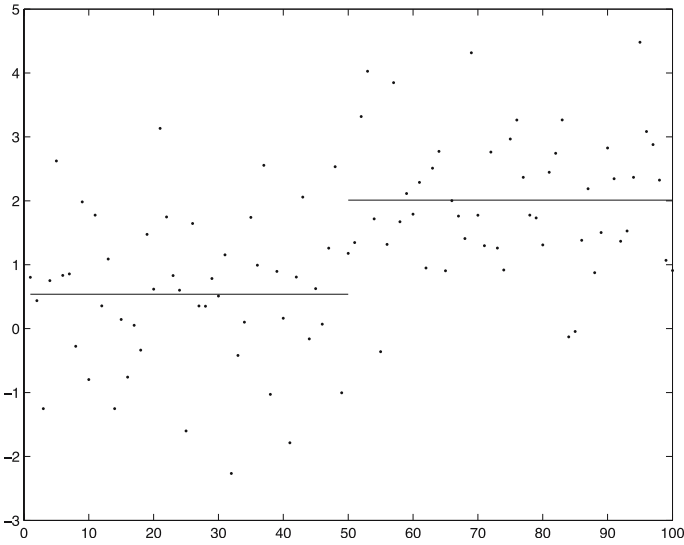


Fig. 1 Estimator based on the data driven penalty

estimator built according to the recipe of Sect. 4.3 actually leads, on this test sample, to $\hat{m} = m_0$, resulting in exactly the same figure. Finally Lebarbier has computed the estimator of s based on Mallows' C_p which is given in Fig. 2. It is clear that Mallows' C_p does not work, leading to a rather erratic behaviour of the estimator.

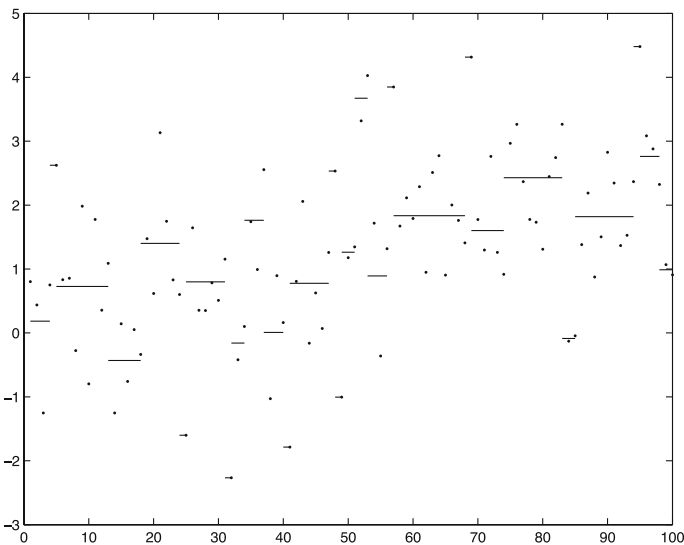


Fig. 2 Estimator based on Mallows' C_p

Sect. 3.1.3, the number of models with the same dimension is too large so that the selected dimension $D_{\hat{m}}$ blows up when the penalty is not heavy enough.

5 Proofs

5.1 Proving the existence of \tilde{s}

We recall that our observation is the process $Y(t)$ given by (1) where Z is a linear isonormal process on \mathbb{S} and s an unknown function in \mathbf{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

$$\zeta_m \sim \mathcal{N}(0, \mathbf{I}_m) \quad \text{and} \quad V_m = \|\zeta_m\|^2 \sim \chi^2(D_m), \tag{45}$$

where $\mathcal{N}(0, \mathbf{I}_m)$ 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 least squares estimator \hat{s}_m on S_m is the minimizer, with respect to $t \in S_m$ of

$$\begin{aligned} \gamma(t) &= \|t\|^2 - 2Y(t) = \|s - t\|^2 - \|s\|^2 - 2\varepsilon Z(t) \\ &= \|s - s_m\|^2 + \|t - s_m\|^2 - \|s\|^2 - 2\varepsilon \langle t, \zeta_m \rangle. \end{aligned} \tag{46}$$

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

$$\gamma(\hat{s}_m) = \|s - s_m\|^2 - \|s\|^2 - \varepsilon^2 V_m - 2\varepsilon Z(s_m) \tag{47}$$

and

$$\|\hat{s}_m - s\|^2 = \|s - s_m\|^2 + \varepsilon^2 V_m. \tag{48}$$

Since

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

it follows from (47) that

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

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

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

Under the assumption (10), we derive that, on some set Ω_ξ of probability larger than $1 - \Sigma e^{-\xi}$, for all m simultaneously,

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

Consequently, if (13) holds and η is small enough, depending on K and θ , one gets

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

for all $m \notin \overline{\mathcal{M}}$ such that $L_m D_m \geq \xi\eta^{-1}$. Since $\overline{\mathcal{M}}$ is finite, this implies that $\gamma_n(\hat{s}_m) + \text{pen}(m)$ tends to infinity with $L_m D_m$. By (10), there is only a finite number of m such that $L_m D_m \leq n$, whatever the integer n . One therefore concludes that (10) and (13) imply that there exists a minimizer \hat{m} of $\gamma(\hat{s}_m) + \text{pen}(m)$ on the set Ω_ξ and therefore a.s. since ξ is arbitrary.

5.2 Proof of Theorem 1

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

$$\|s\|^2 + 2\varepsilon Z(s) + \gamma(\hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) = \inf_{m \in \mathcal{M}} \left\{ \|s\|^2 + 2\varepsilon Z(s) + \gamma(\hat{s}_m) + \text{pen}(m) \right\} \tag{49}$$

and from (47) and (48) 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 (47) again to evaluate $\gamma(\hat{s}_{\hat{m}})$ we derive from (49) that

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

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 (48), we finally get

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

or equivalently,

$$\|s - \tilde{s}\|^2 = (1 - \theta)^{-1} \left(\Delta_{\hat{m}} + \inf_{m \in \mathcal{M}} R_m \right), \tag{50}$$

where

$$\Delta_m = (2 - \theta)\varepsilon^2 V_m + 2\varepsilon d_m U_m - \theta d_m^2 - \text{pen}(m) \tag{51}$$

and

$$R_m = d_m^2 + \text{pen}(m) - \varepsilon^2 V_m - 2\varepsilon d_m U_m. \tag{52}$$

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 $\mathbb{P} \left[\Omega_{\xi, m}^c \right] \leq \exp(-x_m)$ and therefore

$$\mathbb{P} \left[\Omega_{\xi}^c \right] \leq \sum_{m \in \mathcal{M}} \exp(-L_m D_m - \xi) = \Sigma \exp(-\xi). \tag{53}$$

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

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

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

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

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

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

It then follows from (51) and our definition of the penalty function that

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

Since this inequality holds whatever $m \in \mathcal{M}$ one can conclude from (13) that

$$\Delta_{\hat{m}} \mathbb{1}_{\Omega_\xi} \leq \left(\varepsilon^2 \xi \left[(2 - \theta)\alpha^{-1} + 2\theta^{-1} \right] + \sup_{m \in \overline{\mathcal{M}}} \{Q_m - \text{pen}(m)\} \right) \mathbb{1}_{\Omega_\xi} \quad (54)$$

and therefore, by (53), for all $\xi > 0$,

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

Integrating with respect to ξ , we get

$$\mathbb{E}_s[\Delta_{\hat{m}}] \leq \Sigma \varepsilon^2 \left[(2 - \theta)\alpha^{-1} + 2\theta^{-1} \right] + \sup_{m \in \overline{\mathcal{M}}} \{Q_m - \text{pen}(m)\}. \quad (55)$$

Since it follows from (52) that

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

we conclude from (50) and (55) that (15) holds.

5.3 Proof of Proposition 1

Let m be given in \mathcal{M} . It follows from (30) that

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

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

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

and by (75) (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,

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

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

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

By assumption, there exists some integer k depending on \bar{H}, θ 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

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

with constants C_1 and C_2 depending only on \bar{H}, θ and η . Therefore for N large enough (depending on \bar{H}, θ, η 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 (48) that

$$\|s - \tilde{s}\|^2 = \varepsilon^2 V_{\hat{m}} + \|s - s_{\hat{m}}\|^2 \geq \varepsilon^2 V_{\hat{m}} + \|s - s_N\|^2 \tag{56}$$

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, [21])

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

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 (56) since $\Phi(\sqrt{\tau}) = (\theta + 2)/4$ and therefore

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

5.4 Proof of Proposition 2

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 $W_{(1)} < \dots < W_{(n)}$ 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

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

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

$$\gamma(\hat{m}) = -\|\hat{s}_{\hat{m} \cap \Lambda_1}\|^2 - \varepsilon^2 \sum_{j=1}^k W_{(n+1-j)} \quad \text{with } k = |\hat{m} \cap \Lambda_2| \tag{57}$$

and that

$$\|s - \hat{s}_{\hat{m}}\|^2 = \|s - \hat{s}_{\hat{m} \cap \Lambda_1}\|^2 + \varepsilon^2 \sum_{j=1}^k W_{(n+1-j)}. \tag{58}$$

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

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

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

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

then from (57)

$$\sum_{j=1}^k W_{(n+1-j)} \geq \sum_{j=1}^J W_{(n+1-j)} - (2 - 2\alpha - \eta)(1 - \delta)|\bar{m}| \log N. \tag{59}$$

Since

$$n \geq N \left(1 - \delta AN^{\alpha-1}\right) \quad \text{and} \quad (1 - \delta)|\bar{m}| \leq J \leq |\bar{m}| \leq AN^\alpha, \tag{60}$$

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

$$\mathbb{P}[\Omega'] \geq 1 - \left[\exp\left(\frac{9}{8}\right) - 1\right]^{-1} > 1/2, \tag{61}$$

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 (59) and (60), when $N \rightarrow +\infty$,

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

It then follows from (58) and (61) that

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

which concludes the proof.

5.5 Proof of Proposition 3

For $D \geq 1$, the variables V_m , defined by (45), 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 $z > 0$,

$$\log \left(\mathbb{P} \left[\sup_{m \in \mathcal{M}_D} V_m < z \right] \right) = |\mathcal{M}_D| \log \left(1 - \mathbb{P} \left[\chi^2(D) \geq z \right] \right). \tag{62}$$

An application of (74) with $x = \alpha D + 2 \log(D + 3)$, $\rho = 0$ and $b = 2$ gives,

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

Setting

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

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

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

For $u \leq 1/16$, $u^{-1} \log(1 - u) \geq 16 \log(1 - 1/16) > -1.033$. It then follows from (62) and (35), that

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

and therefore,

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

Finally,

$$\begin{aligned} \mathbb{P} \left[\sup_{m \in \mathcal{M} \setminus \emptyset} \left\{ V_m - G(\alpha)D_m - 2 \left(2 + \alpha^{-1/2} \right) \log(D_m + 3) \right\} \geq 0 \right] \\ \leq 1.033 \sum_{D \geq 1} (D + 3)^{-2} < 0.3. \end{aligned} \tag{63}$$

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

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

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

$$\frac{\alpha n}{4} \geq \log n \geq \frac{1}{2} \log(4\pi) + \log(2a + \sqrt{2a}) + \frac{1}{n} \left[\frac{1}{6} + \frac{1}{2(a + \sqrt{a})^2} + \frac{1}{a + \sqrt{a}} \right], \tag{64}$$

$$\sum_{j \geq n} \exp\left(-\sqrt{j} \left(1 - e^{-\alpha j}\right)\right) \leq 0.2 \quad \text{and} \quad G(\alpha) \geq 8 \left(1 + (2/3)\alpha^{-1/2}\right) \frac{\log n}{n}. \tag{65}$$

If $D \geq D(\alpha)$, then by (64) $y = g(\alpha)(\alpha D - 2 \log D) \geq aD$,

$$\sqrt{D} (a + \sqrt{a}) \leq (y/\sqrt{D}) + \sqrt{y} \leq \sqrt{D} (2a + \sqrt{2a})$$

and Corollary 2 below together with (64) imply that if $z = D + 2\sqrt{Dy} + 2y$,

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

It follows from Proposition 4 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 $y\theta(y/D) \leq \alpha D - 2 \log D$, $\log(\mathbb{P}[\chi^2(D) \geq z]) \geq -\alpha D + (\log D)/2$ and it follows from (62) and (35) that

$$\log\left(\mathbb{P}\left[\sup_{m \in \mathcal{M}_D} V_m < z\right]\right) \leq -|\mathcal{M}_D| \mathbb{P}[\chi^2(D) \geq z] \leq -\sqrt{D} (1 - e^{-\alpha D}). \tag{66}$$

Since $\sqrt{1-u} > 1 - 0.6u$ for $u \leq 1/2$, we derive from (64) that

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

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 (66) and (65) that

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

Together with (63), this means that $\mathbb{P}[\Omega] \geq 1/2$, if we denote by Ω the event defined by the set of inequalities

$$V_m < G(\alpha)D_m + 2 \left(2 + \alpha^{-1/2}\right) \log(D_m + 3), \quad \text{for all } m \in \mathcal{M}, \quad (67)$$

since $V_\emptyset = 0$ and

$$V_m \geq G(\alpha)F(\alpha)D_m - 4F(\alpha) \left(1 + (2/3)\alpha^{-1/2}\right) \log D_m \quad \text{if } D_m \geq D(\alpha). \quad (68)$$

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

$$\bar{D}[F(\alpha) - \lambda] \geq 4D(\alpha) \quad (69)$$

and

$$\bar{D}G(\alpha)[F(\alpha) - \lambda] \geq 2 \log(\bar{D} + 1) \left[4F(\alpha) \left(1 + (2/3)\alpha^{-1/2}\right) + \lambda\beta \left(\alpha^{-1/2} + 2\right) \right]. \quad (70)$$

For any $m \in \bar{\mathcal{M}}$, it follows from (69) that $D_m \geq \bar{D} > 4D(\alpha)$. Moreover, by (37) and (36),

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

and, since $s = s_m = 0$, $\gamma(\hat{s}_m) = -\varepsilon^2 V_m$ by (47). Therefore, by (68) and (70),

$$\begin{aligned} \gamma(\hat{s}_m) + \text{pen}(m) &\leq -\varepsilon^2 D_m G(\alpha)[F(\alpha) - \lambda] \\ &\quad + \varepsilon^2 \log(D_m + 1) \left[4F(\alpha) \left(1 + (2/3)\alpha^{-1/2} \right) + \lambda\beta \left(\alpha^{-1/2} + 2 \right) \right] \\ &\leq -(\varepsilon^2/2) D_m G(\alpha)[F(\alpha) - \lambda]. \end{aligned} \tag{71}$$

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

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

– If $\overline{\mathcal{M}}$ is finite, Theorem 1 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 $m \in \overline{\mathcal{M}}$, then $D' \geq D(\alpha) \geq 3$ by (69) and it follows from (67) and (65) that,

$$\begin{aligned} \inf_{D \leq D'} \inf_{m \in \mathcal{M}_D} (\gamma(\hat{s}_m) + \text{pen}(m)) &\geq -\varepsilon^2 \sup_{D \leq D'} \sup_{m \in \mathcal{M}_D} V_m \\ &> -\varepsilon^2 \left[G(\alpha) D' + 2 \left(2 + \alpha^{-1/2} \right) \log(D' + 3) \right] \\ &> -2\varepsilon^2 G(\alpha) D' \\ &> -\varepsilon^2 G(\alpha)[F(\alpha) - \lambda] D_m / 2. \end{aligned} \tag{72}$$

Comparing (71) with (72) and taking into account (69), one concludes, since m is arbitrary in $\overline{\mathcal{M}}$, that, on the set Ω ,

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

Since, by (48), $\|s - \tilde{s}\|^2 = \varepsilon^2 V_{\hat{m}}$, it follows from (68) and (65) 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 (38) follows since $\mathbb{P}[\Omega] \geq 1/2$.

5.6 Proof of Theorem 2

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

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

Since $\hat{s}_0 = 0$ and $\text{pen}(0) = 0$, it follows from (4) that $\hat{m} = 0$ if $\text{pen}(m) > \|\hat{s}_m\|^2$ 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 [9], 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

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

then

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

Putting everything together we can conclude that if

$$\varepsilon^{-2} \text{pen}(m) \geq D_m + A + 2\sqrt{(D_m + 2A)L_m D_m} + 2L_m D_m \quad \text{for all } m \in \mathcal{M}^*, \tag{73}$$

then

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

Since (73) is an immediate consequence of (39), (40) holds while the upper bound for the risk of \tilde{s} when $\text{pen}(m)$ is given by (16) follows from (17).

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*

$$\mathbb{P}\left[V + \rho U \geq D + \rho^2/(2b) + 2\sqrt{Dx} + bx\right] \leq \exp(-x) \quad \text{for any } b \geq 2 \tag{74}$$

and

$$\mathbb{P}\left[V + \rho U \leq D - 2\sqrt{(D + \rho^2/2)x}\right] \leq \exp(-x). \tag{75}$$

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

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

which implies by independence that

$$\log \mathbb{E}_s \left[e^{y(V-D+\rho U)} \right] \leq \frac{Dy^2}{1 - 2y} + \frac{y^2 \rho^2}{2}, \tag{76}$$

since $\mathbb{E}_s[tU] = \exp(t^2/2)$. If $b \geq 2$ the right-hand side of (76) can be bounded by $Dy^2/(1 - by) + y\rho^2/(2b)$ for $0 < y < b^{-1}$ which implies that

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

Inequality (74) then follows from Lemma 2 below with $a^2 = D$. Its proof is part of the proof of Lemma 8 of Birgé and Massart [10].

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

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

and (75) follows. □

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

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

where a and b are positive constants. Then

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

Lemma 3 *Let $W_{(1)} < \dots < W_{(n)}$ 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*

$$\mathbb{P} \left[W_{(n+1-j)} \leq \left[\Phi^{-1} \left(1 - \frac{j(1 + \theta)}{2n} \right) \right]^2 \right] \leq \exp \left[-\frac{j\theta^2}{2(1 + \theta)} \right], \tag{77}$$

and consequently if $\theta \geq 2.06$

$$W_{(n+1-j)} > \left[\Phi^{-1} \left(1 - \frac{j(1+\theta)}{2n} \right) \right]^2 \quad \text{for } 1 \leq j \leq \frac{n}{(1+\theta)}, \tag{78}$$

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 $0 < y \leq x$,

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

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 $W = [F^{-1}(U)]^2$ has the chi-square distribution with one degree of freedom. It follows that $W_{(j)}$ can be written as $[F^{-1}(U_{(j)})]^2$ where $U_{(1)} < \dots < U_{(n)}$ is an ordered sample of size n of the uniform distribution. Now set $x = j(1 + \theta)/n$. Since (77) 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

$$\begin{aligned} \mathbb{P}[U_{(n+1-j)} \leq 1-x] &= \mathbb{P}[U_{(n+1-j)} < 1-x] \\ &= \mathbb{P}[\mathcal{B}(n, x) < j] \\ &= \mathbb{P}[\mathcal{B}(n, x) < nx - j\theta]. \end{aligned} \tag{79}$$

Recalling from Massart ([30], Theorem 2) that, for $0 < y \leq p$,

$$\mathbb{P}[\mathcal{B}(n, p) - np < -ny] \leq \exp \left[-\frac{ny^2}{2(p-y/3)(1-p+y/3)} \right] < \exp \left[-\frac{ny^2}{2p} \right], \tag{80}$$

we derive from (79) that

$$\mathbb{P} \left[W_{(n+1-j)} \leq [F^{-1}(1-x)]^2 \right] = \mathbb{P} \left[U_{(n+1-j)} \leq 1-x \right] \leq \exp \left[-\frac{j\theta^2}{2(1+\theta)} \right] \tag{81}$$

and (77) follows since $F(t) = 2\Phi(t) - 1$. Summing the different probabilities gives (78). The last result follows from Feller ([16], Lemma 2, p. 175). \square

Proposition 3, 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. Results of this type are definitely not new and one can find quite precise evaluations in Wallace [38] but in a form (comparison with Gaussian tails) which is not suitable for our needs. A more adequate approach appeared as Lemma 6.1 in Johnstone [23] 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 4 *Let X be a random variable with gamma distribution $\Gamma(t, 1)$. If $x > 0$ then*

$$\log (\mathbb{P} [X \geq t (1 + 2x + 2\sqrt{x})]) = -2tx\theta(x) - (1/2) \log(2\pi/\lambda) - \Phi, \tag{82}$$

with

$$\theta(x) = 1 + x^{-1/2} - (2x)^{-1} \log (1 + 2x + 2\sqrt{x}); \tag{83}$$

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

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

$$1 < \theta(x) < 1.196 \quad \text{and} \quad \lim_{x \rightarrow 0} \theta(x) = \lim_{x \rightarrow +\infty} \theta(x) = 1. \tag{84}$$

Remark Bound (82) is only useful for $\lambda < 2\pi$. Otherwise, since $2tx\theta(x) < 1.2/(2\lambda)$, (82) 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, [23], 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 $x^t e^{-x} (1 + t(x-t)^{-2})$, which implies, for $z > t$, that

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

□

Proof of Proposition 4 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 $1 > \Delta' > [1 + t(u + 1)^{-2}]^{-1}$. Since by Stirling's Formula (see Whittaker and Watson, [39] 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

$$\mathbb{P}[X \geq t + u] = \Delta \left(1 + ut^{-1}\right)^t e^{-u\sqrt{\delta/(2\pi)}}, \tag{85}$$

with

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

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

$$\log(\mathbb{P}[X \geq t(1 + 2x + 2\sqrt{x})]) = t[\log(1 + 2x + 2\sqrt{x}) - 2(x + \sqrt{x}) - (1/2)\log[2\pi/\lambda] - \Phi],$$

with $0 < \Phi < (12t)^{-1} + \log(1 + \lambda)$, which proves (82). As to (84), it can be derived from some elementary analytical considerations and numerical computations. \square

Corollary 2 *Let Y be a chi-square random variable with D degrees of freedom and $y > 0$. Then*

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

where the function θ defined by (83) satisfies (84) and

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

Proof Since Y has a distribution $\Gamma(D/2, 1/2)$, $X = Y/2$ has a distribution $\Gamma(D/2, 1)$. Applying Proposition 4 with $t = D/2$ and $x = y/D$, we get

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

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. \square

References

1. Abramovich, F., Benjamini, Y., Donoho, D.L., Johnstone, I.M.: Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, (2006)
2. Akaike, H.: Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217 (1969)
3. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, P.N., Csaki, F. (eds.) *Proceedings 2nd International Symposium on Information Theory*, Akademia Kiado, Budapest, pp. 267–281 (1973)
4. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974)
5. Akaike, H.: A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30**, Part A, 9–14 (1978)
6. Amemiya, T.: *Advanced Econometrics*. Basil Blackwell, Oxford (1985)
7. Barron, A.R., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301–415 (1999)
8. Barron, A.R., Cover, T.M.: Minimum complexity density estimation. *IEEE Trans. Inf. Theory* **37**, 1034–1054 (1991)
9. Birgé, L.: An alternative point of view on Lepski's method. In: de Gunst, M.C.M., Klaassen, C.A.J., van der Vaart, A.W. (eds.) *State of the Art in Probability and Statistics*, Festschrift for Willem R. van Zwet, Institute of Mathematical Statistics, Lecture Notes–Monograph Series, Vol. 36. 113–133 (2001)
10. Birgé, L., Massart, P.: Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375 (1998)
11. Birgé, L., Massart, P.: Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–268 (2001)
12. Birgé, L., Massart, P.: A generalized C_p criterion for Gaussian model selection. Technical Report No 647. Laboratoire de Probabilités, Université Paris VI (2001) <http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001>
13. Daniel, C., Wood, F.S.: *Fitting Equations to Data*. Wiley, New York (1971)
14. Draper, N.R., Smith, H.: *Applied Regression Analysis*, 2nd edn. Wiley, New York (1981)
15. Efron, B., Hastie, R., Johnstone, I.M., Tibshirani, R.: Least angle regression. *Ann. Statist.* **32**, 407–499 (2004)
16. Feller, W.: *An Introduction to Probability Theory and its Applications*, Vol. I (3rd edn.). Wiley, New York (1968)
17. George, E.I., Foster, D.P.: Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747 (2000)
18. Gey, S., Nédélec, E.: Model selection for CART regression trees. *IEEE Trans. Inf. Theory* **51**, 658–670 (2005)
19. Guyon, X., Yao, J.F.: On the underfitting and overfitting sets of models chosen by order selection criteria. *Jour. Multivar. Anal.* **70**, 221–249 (1999)
20. Hannan, E.J., Quinn, B.G.: The determination of the order of an autoregression. *J.R.S.S., B* **41**, 190–195 (1979)

21. Hoeffding, W.: Probability inequalities for sums of bounded random variables. *J.A.S.A.* **58**, 13–30 (1963)
22. Hurvich, K.L., Tsai, C.-L.: Regression and time series model selection in small samples. *Biometrika* **76**, 297–307 (1989)
23. Johnstone, I.: Chi-square oracle inequalities. In: de Gunst, M.C.M., Klaassen, C.A.J. van der Vaart, A.W. (eds.) *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet*, Institute of Mathematical Statistics, Lecture Notes–Monograph Series, Vol. 36. pp. 399–418 (2001)
24. Kneip, A.: Ordered linear smoothers. *Ann. Statist.* **22**, 835–866 (1994)
25. Lavielle, M., Moulines, E.: Least Squares estimation of an unknown number of shifts in a time series. *J. Time Series Anal.* **21**, 33–59 (2000)
26. Lebarbier, E.: Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.* **85**, 717–736 (2005)
27. Li, K.C.: Asymptotic optimality for C_p , C_L , cross-validation, and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958–975 (1987)
28. Loubes, J.-M., Massart, P.: Discussion of “Least angle regression” by Efron, B., Hastie, R., Johnstone, I., Tibshirani, R. *Ann. Statist.* **32**, 460–465 (2004)
29. Mallows, C.L.: Some comments on C_p . *Technometrics* **15**, 661–675 (1973)
30. Massart, P.: The tight constant in the D.K.W. inequality. *Ann. Probab.* **18**, 1269–1283 (1990)
31. McQuarrie, A.D.R., Tsai, C.-L.: *Regression and Time Series Model Selection*. World Scientific, Singapore (1998)
32. Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. *J.A.S.A.* **83**, 1023–1032 (1988)
33. Polyak, B.T., Tsybakov, A.B.: Asymptotic optimality of the C_p -test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, 293–306 (1990)
34. Rissanen, J.: Modeling by shortest data description. *Automatica* **14**, 465–471 (1978)
35. Schwarz, G.: Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464 (1978)
36. Shen, X., Ye, J.: Adaptive model selection. *J.A.S.A.* **97**, 210–221 (2002)
37. Shibata, R.: An optimal selection of regression variables. *Biometrika* **68**, 45–54 (1981)
38. Wallace, D.L.: Bounds on normal approximations to Student’s and the chi-square distributions. *Ann. Math. Stat.* **30**, 1121–1130 (1959)
39. Whittaker, E.T., Watson, G.N.: *A Course of Modern Analysis*. Cambridge University Press, London (1927)
40. Yang, Y.: Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950 (2005)
41. Yao, Y.C.: Estimating the number of change points via Schwarz criterion. *Stat. Probab. Lett.* **6**, 181–189 (1988)