Data-driven calibration of penalties for least-squares regression

Sylvain Arlot, Pascal Massart

Journal of Machine Learning Research 0 (0000) 0 Submitted 3/08; Published 0/00

Data-driven calibration of penalties for least-squares regression

Sylvain Arlot SYLVAIN.ARLOT@MATH.U-PSUD.FR
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathematiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select

Pascal Massart PASCAL.MASSART@MATH.U-PSUD.FR
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathematiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select

Editor: Editor

Abstract

Penalization procedures often suffer from their dependence on multiplying factors, whose optimal values are either unknown or hard to estimate from the data. In this paper, we propose a completely data-driven calibration method for this parameter in the least-squares regression framework, without assuming a particular shape for the penalty. Our algorithm relies on the concept of minimal penalty, which has been introduced in a recent paper by Birgé and Massart (2007) in the context of penalized least squares for Gaussian homoscedastic regression. Interestingly, the minimal penalty can be evaluated from the data themselves, which leads to a data-driven estimation of an optimal penalty that one can use in practice. Unfortunately their approach heavily relies on the homoscedastic Gaussian nature of the stochastic framework that they consider.

Our purpose in this paper is twofold: stating a more general heuristics to design a data-driven penalty (the *slope heuristics*) and proving that it works for penalized least squares random design regression, even when the data is heteroscedastic. For some technical reasons which are explained in the paper, we could prove some precise mathematical results only for histogram bin-width selection. Even though we could not work at the level of generality that we were expecting, this is at least a first step towards further results. Our mathematical results hold in some specific framework, but the approach and the method that we use are indeed general.

Keywords: Data-driven calibration, Non-parametric regression, Model selection by penalization, Heteroscedastic data, Histogram

1. Introduction

Model selection has received much interest in the last decades. A very common approach is penalization. In a nutshell, it chooses the model which minimizes the sum of the empirical risk (how does the algorithm fit the data) and some complexity measure of the model (called the penalty). This is the case of FPE (Akaike, 1970), AIC (Akaike, 1973) and Mallows' C_p or

C_L (Mallows, 1973). Many other penalization procedures have been proposed since, among which we mention Rademacher complexities (Koltchinskii, 2001; Bartlett et al., 2002), local Rademacher complexities (Bartlett et al., 2005; Koltchinskii, 2006), bootstrap, resampling and V-fold penalties (Efron, 1983; Arlot, 2008b,a), to name but a few.

In this article, we consider the question of the *efficiency* of such penalization procedures, *i.e.* that their quadratic risk is asymptotically equivalent to the risk of the oracle. This property is often called asymptotic optimality. It does not mean that the procedure finds out a "true model" (which may not even exist), which would be the *consistency* problem. A procedure is efficient when it makes the best possible use of the data in terms of the quadratic risk of the final estimator.

There is a huge amount of literature about this question. Consider first Mallows' C_p and Akaike's FPE and AIC. Their asymptotic optimality has been proven by Shibata (1981) for Gaussian errors, Li (1987) under suitable moment assumptions on the errors, and Polyak and Tsybakov (1990) for sharper moment conditions in the Fourier case. Then, non-asymptotic oracle inequalities (with a constant $C > 1$) have been proven by Barron et al. (1999) and Birgé and Massart (2001) in the Gaussian case, and Baraud (2000, 2002) under some moment conditions on the errors. In the Gaussian case, non-asymptotic oracle inequalities with a constant C_n which goes to 1 when n goes to infinity have been obtained by Birgé and Massart (2007).

However, both AIC and Mallows' C_p still have serious drawbacks from the practical viewpoint. Indeed, AIC relies on a strong asymptotic assumption, so that the optimal multiplying factor may be quite different from one for small sample sizes. This is why corrected versions of AIC have been proposed (Sugiura, 1978; Hurvich and Tsai, 1989). On the other hand, the optimal calibration of Mallows' C_p requires the knowledge of the noise level σ^2, which is assumed to be constant. With real data, one has to estimate σ^2 separately, but it is hard to make it independently from any model. In addition, it is quite unlikely that the best estimator of σ^2 automatically leads to the most efficient model selection procedure. One of the purposes of this article is to provide a data-dependent calibration rule which directly aims at the efficiency of the final procedure. Focusing directly on efficiency may improve significantly the more classical "plug-in" method, in terms of the performance of the model selection procedure itself.

Actually, most of the penalization procedures have similar or even stronger drawbacks, often because of a gap between theoretical results and their practical use. For instance, their is a factor 2 between the (global) Rademacher complexities for which theoretical results have been proven, and the way they are used in practice (Lozano, 2000). Since this factor is unavoidable in some sense (Arlot (2007), Chap. 9), the optimal calibration of these penalties is a practical issue. The problem is tougher for local Rademacher complexities, since theoretical results are only valid with very large calibration constants (in particular the multiplying factor), and no one knows which are their optimal values. One of our goals is to address this question for such general-shape penalties (in particular data-dependent penalties), at least for the optimization of the multiplying factor.

There are not so many calibration algorithms available. Obviously, the most popular ones are cross-validation methods (Allen, 1974; Stone, 1974), in particular V-fold cross-validation (Geisser, 1975), in particular because these are general-purpose methods, relying

on a heuristics likely to be widely valid. However, their computational cost may be too heavy, because they require to perform V times the entire model selection procedure for each candidate value of the constant to be calibrated. For penalties based on the dimension of the models (assumed to be vector spaces), such as Mallows' C_p, an alternative calibration procedure has been proposed by George and Foster (2000).

A completely different approach is the one of Birgé and Massart (2007), who have also considered dimensionality based penalties. Since our purpose is to extend their approach to a much wider range of applications, let us recall briefly their main claims. In the Gaussian homoscedastic regression on a fixed-design framework, assume that each model is a finite-dimensional vector space. Then, consider the penalty $\mathrm{pen}(m) = KD_m$, where D_m is the dimension of the model m and $K > 0$ is a positive constant, to be calibrated. In several situations, it turns out that the *optimal* constant $K^\star$ (*i.e.* the one which leads to an asymptotically efficient procedure) is exactly twice the *minimal* constant $K_{\min}$ (defined as the one under which the ratio between the quadratic risk of the chosen estimator and the quadratic risk of the oracle goes to infinity with the sample size). In other words, *the optimal penalty is twice the minimal penalty*, which is called the "slope heuristics" by Birgé and Massart.

A crucial fact is that the minimal constant $K_{\min}$ can be estimated from the data, because very large models are selected if and only if $K < K_{\min}$. This leads them to the following strategy for choosing K from the data. Define $\widehat{m}(K)$ the model selected by $\mathrm{pen}(D_m) = KD_m$ as a function of K. First, compute $K_{\min}$ such that $D_{\widehat{m}(K)}$ is huge for $K < K_{\min}$ and reasonable when $K \geq K_{\min}$. Second, define $\widehat{m} := \widehat{m}(2K_{\min})$. Such a method has been successfully applied for multiple change points detection by Lebarbier (2005).

From the theoretical viewpoint, a crucial question to understand (and validate) this approach is the existence of a minimal penalty. In other words, how much should we penalize at least? In the framework of Gaussian regression on a fixed-design, this question has been addressed by Birgé and Massart (2001, 2007) and Baraud et al. (2007) (the latter considering the unknown variance case). However, nothing is known for non Gaussian or heteroscedastic data. One of our goals is thus to fill part of this gap in the theoretical understanding of penalization procedures.

In this paper, we use a similar link between minimal and optimal penalties, in order to calibrate any penalty (namely, the favorite penalty of the final user, including all the aforementioned penalties, and not necessarily dimensionality-based penalties), in a more general framework (*e.g.*, we allow the noise to be *heteroscedastic* and non-Gaussian, which is much more realistic). This leads us to Algorithm 1, which is defined in Sect. 3.1 in the least-squares regression framework, and relies on a generalization of the slope heuristics.

We then tackle the theoretical validation of this algorithm, from the *non-asymptotic viewpoint*. By non-asymptotic, we mean in particular that the collection of models is allowed to depend on n. This is quite natural since it is common in practice to introduce more explanatory variables (for instance) when one has more observations. Considering models with a large number of parameters (*e.g.* of the order of a power of the sample size n) is also necessary to approximate functions belonging to a general approximation space. Thus, the non-asymptotic viewpoint allows us not to assume that the regression function can be described with a very small number of parameters.

First, we prove the existence of minimal penalties for *heteroscedatic regression on a random-design* (Thm. 1). Then, we prove in the same framework that twice the minimal penalty has some optimality properties (Thm. 2), which means that we have extended the so-called slope heuristics to heteroscedatic least-squares regression on a random-design. For proving such a result, we have to assume that each model is the vector space of piecewise constant functions on some partition of the feature space. This is quite a restriction, but we conjecture that it is mainly technical, and that the slope heuristics stays valid at least in the general least-square regression framework. We provide some evidence for this by proving two key concentration inequalities without the restriction to histograms.

Another argument supporting this conjecture is that several simulation studies have shown recently that the slope heuristics could be used in several frameworks: mixture models (Maugis and Michel, 2007), clustering (Baudry, 2007), spatial statistics (Verzelen, 2007), estimation of oil reserves (Lepez, 2002) and genomics (Villers, 2007). Our results do not give a formal proof for these applications of the slope heuristics (*cf.* Sect. 3.2 for instances of completely data-driven penalties for which we have proven rigorously that our algorithm is working). However, they are a first step towards such a result, by proving that it can be applied when the ideal penalty has a general shape.

This paper is organized as follows. We describe the framework and our main heuristics in Sect. 2. The resulting algorithm is defined in Sect. 3. Our main theoretical results are stated in Sect. 4. Appendix A is devoted to computational issues. All the proofs are given in Appendix B.

2. Framework

2.1 Least-squares regression

We observe some data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, i.i.d. with common law P. Our goal is to predict Y given X, where $(X, Y) \sim P$ is independent from the data. Denoting by s the regression function, we can write

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{1}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with mean 0 and variance 1 conditionally to X_i. Typically, the feature space $\mathcal{X}$ is a compact set of $\mathbb{R}^d$.

Given a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$, its quality is measured by the (quadratic) prediction loss

$$\mathbb{E}_{(X,Y)\sim P}\left[\gamma(t,(X,Y))\right] =: P\gamma(t) \qquad \text{where} \quad \gamma(t,(x,y)) = (t(x)-y)^2$$

is the least-square contrast. Then, the Bayes predictor (*i.e.* the minimizer of $P\gamma(t)$ over the set of all predictors) is the regression function s, and we define the excess loss as

$$\ell(s,t) := P\gamma(t) - P\gamma(s) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2 \quad .$$

Given a particular set of predictors S_m (called a *model*), we define the best predictor over S_m

$$s_m := \arg\min_{t\in S_m} \left\{ P\gamma(t) \right\} \quad ,$$

and its empirical counterpart

$$\widehat{s}_m := \arg\min_{t \in S_m} \{ P_n \gamma(t) \}$$

(when it exists and is unique), where $P_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$. This estimator is the well-known *empirical risk minimizer*, also called least-square estimator since γ is the least-square contrast.

2.2 Ideal model selection

We now assume that we have a family of models $(S_m)_{m \in \mathcal{M}_n}$, hence a family of estimators $(\widehat{s}_m)_{m \in \mathcal{M}_n}$ (*via* empirical risk minimization). We are looking for some data-dependent $\widehat{m} \in \mathcal{M}_n$ such that $\ell(s, \widehat{s}_{\widehat{m}})$ is as small as possible. This is the model selection problem. For instance, we would like to prove some oracle inequality of the form

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \} + R_n$$

in expectation or on an event of large probability, with C close to 1 and $R_n = o(n^{-1})$.

General penalization procedures can be described as follows. Let $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$ be some penalty function, possibly data-dependent. Then, define

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ \mathrm{crit}(m) \} \qquad \text{with} \qquad \mathrm{crit}(m) := P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \ . \tag{2}$$

Since the ideal criterion crit is the true prediction error $P\gamma(\widehat{s}_m)$, the *ideal penalty* is

$$\mathrm{pen}_{\mathrm{id}}(m) := P\gamma(\widehat{s}_m) - P_n \gamma(\widehat{s}_m) \ .$$

Of course, this quantity is unknown because it depends on the true distribution P. A natural idea is to choose pen as close as possible to $\mathrm{pen}_{\mathrm{id}}$ for every model $m \in \mathcal{M}_n$. We show below, in a very general setting, that when pen estimates well the ideal penalty $\mathrm{pen}_{\mathrm{id}}$, $\widehat{m}$ satisfies an oracle inequality with a leading constant C close to 1.

By definition of $\widehat{m}$,

$$\forall m \in \mathcal{M}_n, \quad P_n \gamma(\widehat{s}_{\widehat{m}}) \leq P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}) \ .$$

For every $m \in \mathcal{M}_n$, we define

$$p_1(m) = P(\gamma(\widehat{s}_m) - \gamma(s_m)) \quad p_2(m) = P_n(\gamma(s_m) - \gamma(\widehat{s}_m)) \quad \delta(m) = (P_n - P)(\gamma(s_m))$$

so that

$$\ell(s, \widehat{s}_m) = P_n \gamma(\widehat{s}_m) + p_1(m) + p_2(m) - \delta(m) - P\gamma(s) \ .$$

We then have, for every $m \in \mathcal{M}_n$,

$$\ell(s, \widehat{s}_{\widehat{m}}) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(\widehat{m}) \leq \ell(s, \widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}_{\mathrm{id}})(m) \ . \tag{3}$$

So, in order to derive an oracle inequality from (3), we have to show that for every $m \in \mathcal{M}_n$, $\mathrm{pen}(m)$ is close to $\mathrm{pen}_{\mathrm{id}}(m)$.

2.3 The slope heuristics

When the penalty pen is too large, the left-hand side of (3) stays larger than $\ell(s, \widehat{s}_{\widehat{m}})$ so that we can still obtain an oracle inequality (possibly with a large constant C). On the contrary, when pen is too small, the left-hand side of (3) can become negligible in front of $\ell(s, \widehat{s}_{\widehat{m}})$ (which makes C explode) or — worse — can be nonpositive (so that we can no longer derive an oracle inequality from (3)). We shall see in the following that this corresponds to the existence of a "minimal penalty".

Consider first the case $\mathrm{pen}(m) = p_2(m)$ in (2). Then, $\mathbb{E}[\mathrm{crit}(m)] = \mathbb{E}[P_n\gamma(s_m)] = P\gamma(s_m)$, so that $\widehat{m}$ tends to be the model with the smallest bias, hence the more complex one. As a consequence, the risk of $\widehat{s}_{\widehat{m}}$ is very large. When $\mathrm{pen}(m) = Kp_2(m)$, if $K < 1$, $\mathrm{crit}(m)$ is a decreasing function of the complexity of m, so that $\widehat{m}$ is still one of the more complex models. On the contrary, when $K > 1$, $\mathrm{crit}(m)$ starts to increase with the complexity of m (at least for the largest models), so that $\widehat{m}$ has a smallest complexity. This intuition supports the conjecture that the "minimal amount of penalty" required for the model selection procedure to work may be $p_2(m)$.

In several situations (such as the framework of Sect. 4.1, as we shall prove in the following), it turns out that

$$\forall m \in \mathcal{M}_n, \qquad p_1(m) \approx p_2(m) \ .$$

As a consequence, the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m) \approx p_1(m) + p_2(m)$ is close to $2p_2(m)$. On the other hand, $p_2(m)$ is actually a "minimal penalty". So, we deduce that the optimal penalty is close to twice the minimal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m) \approx 2\,\mathrm{pen}_{\mathrm{min}}(m) \ .$$

This is the so-called "slope heuristics", which was first introduced by Birgé and Massart (2007) in a Gaussian setting.

The practical interest of this heuristics is that the minimal penalty can be estimated from the data. Indeed, when the penalty is too small, the selected model $\widehat{m}$ is among the more complex. On the contrary, when the penalty is larger than the minimal one, the complexity of $\widehat{m}$ should be much smaller. This leads to the algorithm described in the next section.

3. A data-driven calibration algorithm

We are now in position to define a data-driven calibration algorithm for penalization procedures. It generalizes a method proposed by Birgé and Massart (2007) and implemented by Lebarbier (2005).

3.1 The general algorithm

Assume that we know the shape $\mathrm{pen}_{\mathrm{shape}} : \mathcal{M}_n \mapsto \mathbb{R}^+$ of the ideal penalty (because of some prior knowledge, or because we have been able to estimate it first, see Sect. 3.2). This means that the penalty $K^\star\,\mathrm{pen}_{\mathrm{shape}}$ provides an approximately optimal procedure, for some unknown constant $K^\star > 0$. Our goal is to find some $\widehat{K}$ such that $\widehat{K}\,\mathrm{pen}_{\mathrm{shape}}$ is approximately optimal.

We also assume that we know some complexity measure D_m for each model $m \in \mathcal{M}_n$. Typically, when the models are finite-dimensional vector spaces, D_m is the dimension of S_m. According to the "slope heuristics", detailed in Sect. 2.3, the following algorithm provides an optimal calibration of the penalty $\mathrm{pen}_{\mathrm{shape}}$.

Algorithm 1 (Data-driven penalization with slope heuristics)

1. *Compute the selected model* $\widehat{m}(K)$ *as a function of* $K > 0$

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma(\widehat{s}_m) + K \, \mathrm{pen}_{\mathrm{shape}}(m) \right\} \quad .$$

2. *Find* $\widehat{K}_{\min} > 0$ *such that* $D_{\widehat{m}(K)}$ *is "very large" for* $K < \widehat{K}_{\min}$ *and "reasonably small" for* $K > \widehat{K}_{\min}$.

3. *Select the model* $\widehat{m} = \widehat{m}\left(2\widehat{K}_{\min}\right)$.

Computational aspects of Algorithm 1 and the accurate definition of $\widehat{K}_{\min}$ are discussed in App. A. In particular, once $P_n\gamma\left(\widehat{s}_m\right)$ and $\mathrm{pen}_{\mathrm{shape}}(m)$ are known for every $m \in \mathcal{M}_n$, the first step of this algorithm can be performed with a complexity proportional to $\mathrm{Card}(\mathcal{M}_n)^2$ (*cf.* Algorithm 2 and Prop. 3). This is a crucial point compared to cross-validation methods, in particular when performing empirical risk minimization is computationally heavy.

3.2 Shape of the penalty

For using Algorithm 1 in practice, it is necessary to know *a priori*, or at least to estimate, the optimal shape $\mathrm{pen}_{\mathrm{shape}}$ of the penalty. We now explain how this can be done in several different situations.

At first reading, one can have in mind the simple example $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$. It is valid for homoscedastic least-squares regression on linear models, as shown by several papers mentioned in the introduction. Indeed, when $\mathrm{Card}(\mathcal{M}_n)$ is smaller than some power of n, it is well known that Mallows' C_p penalty — defined by $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma^2(X)\right] n^{-1} D_m$ — is asymptotically optimal. For larger collections $\mathcal{M}_n$, more elaborate results (Birgé and Massart, 2001, 2007) have shown that a penalty proportional to $\ln(n)\mathbb{E}\left[\sigma^2(X)\right] n^{-1} D_m$ (depending on the size of $\mathcal{M}_n$) is asymptotically optimal.

Algorithm 1 then provides an alternative to plugging an estimator of $\mathbb{E}\left[\sigma^2(X)\right]$ into the above penalties. We would like to underline two main advances with our approach. First, we avoid the difficult task of estimating $\mathbb{E}\left[\sigma^2(X)\right]$, which generally relies on the existence of a "large" model without bias. Our algorithm provides a model-free estimation of the multiplying factor in front of the penalty. Second, there is absolutely no reason that the best estimator $\widehat{\sigma^2}$ of $\mathbb{E}\left[\sigma^2(X)\right]$ (in terms of bias or quadratic risk, for instance) leads to the more efficient model selection procedure. For instance, it is well known that underpenalization (*i.e.* underestimating the multiplicative factor) leads to very poor performances, whereas overpenalization is generally less costly. Then, one can expect that minimizing the probability of underestimation of $\mathbb{E}\left[\sigma^2(X)\right]$ may lead to better performances than the bias. Adding that there are certainly several other important factors in order to optimize the choice of $\widehat{\sigma^2}$, some of them unknown, the plug-in approach seems quite tricky.

With Algorithm 1, we do not care about the bias or the quadratic risk of $2\widehat{K}_{\min}$ as an estimator of $2\mathbb{E}\left[\sigma^2(X)\right] n^{-1}$. Since we define $\widehat{K}_{\min}$ in terms of the output of the model selection procedure $\widehat{m}(K)$, we focus directly on the model selection problem. In particular, we guarantee that the selected model is not "too large", which solves part of the underpenalization issue.

In brief, we would like to emphasize that *Algorithm 1 with* $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$ *is quite different from a simple plug-in version of Mallows' C_p.* It leads to a really *data-dependent penalty, which may perform better in practice than the best deterministic penalty $K^\star D_m$.*

In a more general framework, Algorithm 1 allows to choose a different shape of penalty $\mathrm{pen}_{\mathrm{shape}}$.

For instance, in the heteroscedastic least-squares regression framework of Sect. 2.1, the optimal penalty is no longer proportional to the dimension D_m of the models. This can be shown from computations made by Arlot (2008b) when S_m is assumed to be the vector space of piecewise constant functions on a partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$:

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[(P - P_n)\gamma\left(\widehat{s}_m\right)\right] \approx \frac{1}{n}\sum_{\lambda\in\Lambda_m} \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X \in I_\lambda\right] \quad . \tag{4}$$

A more accurate result can even be found in Chap. 4 of (Arlot, 2007), where an example of model selection problem is given where no penalty proportional to D_m can be asymptotically optimal.

A first answer to this issue can be given when both the distribution of X and the shape of the noise level σ are known, which is simply to use (4) to compute $\mathrm{pen}_{\mathrm{shape}}$. This is of course unsatisfactory because one has seldom such a prior knowledge in practice.

Our suggest in this situation is the use of *resampling penalties* (Efron, 1983; Arlot, 2008a), or *V-fold penalties* (Arlot, 2008b) which have a much smaller computational cost. Indeed, up to a multiplicative factor (which is automatically estimated by Algorithm 1), these penalties should estimate well $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ in a general framework. In particular, their asymptotic optimality have been proven in the heteroscedastic least-squares regression framework by Arlot (2008b,a), in the framework of Sect. 4.1, and several theoretical results supports the conjecture of their validity much more generally.

3.3 The general prediction framework

In Sect. 2 and in the definition of Algorithm 1, we have restricted ourselves to the least-squares regression framework. This is actually not necessary at all to make Algorithm 1 well-defined, so that we can naturally extend it to the general prediction framework. More precisely, the (X_i, Y_i) can only be assumed to belong to $\mathcal{X}\times\mathcal{Y}$ for some general $\mathcal{Y}$, and $\gamma : S \times (\mathcal{X}\times\mathcal{Y}) \mapsto [0; +\infty)$ any contrast function. In particular, $\mathcal{Y} = \{0, 1\}$ leads to the binary classification problem, and a natural contrast function is the 0-1 loss $\gamma(t; (x, y)) = \mathbb{1}_{t(x)\neq y}$. In this case, the shape of the penalty $\mathrm{pen}_{\mathrm{shape}}$ can for instance be estimated with the global or local Rademacher complexities mentioned in introduction, as well as several other classical penalties.

However, one can wonder whether the slope heuristics of Sect. 2.3, upon which Algorithm 1 relies, can be extended to this general framework. We do not have a complete answer to these questions, but several preliminary evidence. First, in order to "prove" the

validity of the slope heuristics in the least-squares regression framework (with the theoretical results of Sect. 4), we use several concentration results which are valid in a very general setting, including binary classification. Even if the factor 2 (which comes from the closeness of $\mathbb{E}[p_1]$ and $\mathbb{E}[p_2]$, *cf.* Sect. 2.3) may not be universally valid, we conjecture that Algorithm 1 can be used in several settings outside the least-squares regression case. Second, as already mentioned at the end of the introduction, several empirical studies have shown that Algorithm 1 can be successfully applied for several problems, with several shapes for the penalty. A formal proof of this fact remains an interesting open problem, up to our knowledge.

4. Theoretical results

Algorithm 1 mainly relies on the "slope heuristics", which is developped in Sect. 2.2. The goal of this section is to provide a theoretical justification of this heuristics.

It is splitted into two main results. First, lower bounds on $D_{\widehat{m}}$ and the risk of $\widehat{s}_{\widehat{m}}$ when the penalty is smaller than $\mathrm{pen}_{\min}(m) := \mathbb{E}[p_2(m)]$ (Thm. 1). Second, an oracle inequality with constant almost one when $\mathrm{pen}(m) \approx 2\mathbb{E}[p_2(m)]$ (Thm. 2), relying on (3) and the comparison $p_1 \approx p_2$.

In order to prove these two theorems, we need two kinds of probabilistic results. First, both p_1, p_2 and δ concentrate around their expectations (which can be done in a quite general framework, at least for p_2 and δ, see App. B.5). Second, $\mathbb{E}[p_1(m)] \approx \mathbb{E}[p_2(m)]$ for every $m \in \mathcal{M}_n$. The latter point is quite hard in general, so that we must make a structural assumption on the models. This is why, in this section, we restrict ourselves to the histogram case, assuming that for every $m \in \mathcal{M}_n$, S_m is the set of piecewise constant functions on some fixed partition $(I_\lambda)_{\lambda\in\Lambda_m}$. We describe this framework in the next subsection.

Remember that we do not consider histograms as a final goal. We only make this assumption in order to prove some first theoretical results confirming that Algorithm 1 can be used in practical applications. Such theoretical results may also be quite interesting in order to understand better how to use this algorithm in practice.

4.1 Histograms

A "model of histograms" S_m is the the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda\in\Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for any probability measure on $\mathcal{X}$, computations are quite easy. This is the only reason why we assume that each S_m is a model of histograms in this section. In particular, we have:

$$s_m = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \quad \text{and} \quad \widehat{s}_m = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \ ,$$

where

$$\beta_\lambda := \mathbb{E}_P[Y \mid X \in I_\lambda] \qquad \widehat{\beta}_\lambda := \frac{1}{n\widehat{p}_\lambda} \sum_{X_i\in I_\lambda} Y_i \qquad \widehat{p}_\lambda := P_n(X \in I_\lambda) \ .$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i. Otherwise, we consider that the model m can not be chosen.

4.2 Main assumptions

For both our main results, we make the following assumptions.

First, $(S_m)_{m\in\mathcal{M}}$ is a family of histogram models satisfying

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}, c_{\mathrm{rich}}\sqrt{n}]$.

Assumption (**P1**) is quite classical when one aims at proving the asymptotic optimality of a model selection procedure (it is for instance implicitly assumed by Li (1987), in the homoscedastic fixed-design case).

For any penalty function $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$, we define the following model selection procedure:

$$\widehat{m} \in \arg \min_{m\in\mathcal{M}_n,\, \min_{\lambda\in\Lambda_m}\{\widehat{p}_\lambda\}>0} \left\{ P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \right\} \quad . \tag{5}$$

Moreover, we assume that the data $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d. and satisfy the following:

(**Ab**) The data is bounded: $\|Y_i\|_\infty \leq A < \infty$.

(**An**) Uniform lower-bound on the noise-level: $\sigma(X_i) \geq \sigma_{\min} > 0$ a.s.

($\mathbf{Ap_u}$) The bias decreases as a power of D_m: there exists $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell(s, s_m) \leq C_+ D_m^{-\beta_+} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) Lower regularity of the partitions for $\mathcal{L}(X)$: $D_m \min_{\lambda\in\Lambda_m}\{\mathbb{P}(X \in I_\lambda)\} \geq c_{\mathrm{r},\ell}^X$.

Further comments are made in the following about these assumptions, explaining in particular how to relax them.

4.3 Minimal penalties

Our first result is the existence of a minimal penalty.

Theorem 1 *Make all the assumptions of Sect. 4.2. Let $K \in [0;1)$, $L > 0$, and assume that there is an event of probability at least $1 - Ln^{-2}$ on which*

$$\forall m \in \mathcal{M}_n, \quad 0 \leq \mathrm{pen}(m) \leq K\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \quad . \tag{6}$$

Then, if $\widehat{m}$ is defined by (5), there exists two constants K_1, K_2 such that, with probability at least $1 - K_1 n^{-2}$,

$$D_{\widehat{m}} \geq K_2 n \ln(n)^{-1} \quad . \tag{7}$$

On the same event,

$$\ell(s, \widehat{s}_{\widehat{m}}) \geq \ln(n) \inf_{m\in\mathcal{M}_n} \left\{ \ell(s, \widehat{s}_m) \right\} \quad . \tag{8}$$

*The constants K_1 and K_2 may depend on K, L and constants in (**P1**), (**P2**), (**Ab**), (**An**), ($\mathbf{Ap_u}$) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$), but not on n.*

This theorem thus validates the first part of the heuristics of Sect. 2.3, proving that there is a minimal amount of penalization required, under which both the selected dimension $D_{\widehat{m}}$ and the quadratic risk of the final estimator $\ell\left(s, \widehat{s}_{\widehat{m}}\right)$ are blowing up. This coupling is quite interesting, since the dimension $D_{\widehat{m}}$ is known in practice, contrary to $\ell\left(s, \widehat{s}_{\widehat{m}}\right)$. It is then possible to detect from the data that the penalty is too small, as proposed in Algorithm. 1.

The main interest of this result is its coupling with Thm. 2 below. However, Thm. 1 is also of self-interest, since it helps to understand better the theoretical properties of penalization procedures. Indeed, it generalizes the results of Birgé and Massart (2007) on the existence of minimal penalties to heteroscedastic regression on a random design (even if we have to restrict to histogram models, as already explained). We then have a general formulation for the minimal penalty

$$\mathrm{pen}_{\min}(m) := \mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \quad ,$$

which includes situations where it is not proportional to the dimension D_m of the models (*cf.* Sect. 3.2 and references therein).

In addition, assumptions (**Ab**) and (**An**) on the data are much weaker than the Gaussian homoscedastic assumption. They are also much more realistic, and an important point is that they can be strongly relaxed. Roughly, the boundedness of the data can be replaced by some conditions on the moments of the noise, and the uniform lower bound of the data is no longer necessary when σ satisfies some mild regularity assumptions. We refer to (Arlot, 2008a) (in particular Sect. 4.3) for detailed statements of these assumptions, and explanations on how to adapt our proofs to these situations.

Finally, let us comment briefly ($\mathbf{Ap_u}$) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$). The upper bound ($\mathbf{Ap_u}$) on the bias occurs in most reasonable situations, for instance when $\mathcal{X} \subset \mathbb{R}^k$ is bounded, the partition $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and the regression function s is α-hölderian for some $\alpha > 0$ (β_+ depending on α and k). It ensures that large models have a significantly smaller bias than smaller ones (otherwise, the selected dimension would be allowed to be smaller with a significant probability). On the other hand, ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied at least for "almost regular" histograms, when X has a lower bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$.

The reason why we state Thm. 1 with a general formulation of ($\mathbf{Ap_u}$) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$) (instead of assuming that s is α-hölderian and X has a lower bounded density w.r.t Leb, for instance) is to point out the *generality* of the "minimal penalization" phenomenon. It occurs as soon as the models are not too pathological. In particular, we do not make any assumption on the distribution of X itself, but only that the models are not too badly chosen according to this distribution. Such a condition can be checked in practice if one has some prior knowledge on $\mathcal{L}(X)$, or if one has some unlabeled data (which is often the case).

4.4 Optimal penalties

Algorithm 1 relies on a link between the minimal penalty (pointed out by Thm. 1) and some optimal penalty. The following result is a formal proof of this link in our framework: penalties close to twice the minimal penalty satisfy an oracle inequality with a leading constant approximately equal to one.

Theorem 2 *Make all the assumptions of Sect. 4.2, and add the following:*

(**Ap**) *The bias decreases like a power of D_m: there exists $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_m^{-\beta_-} \leq \ell\left(s, s_m\right) \leq C_+ D_m^{-\beta_+} \quad .$$

Let $\delta \in (0,1)$, $L > 0$, and assume that there is an event of probability at least $1 - Ln^{-2}$ on which, for every $m \in \mathcal{M}_n$,

$$(2-\delta)\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \leq \mathrm{pen}(m) \leq (2+\delta)\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \quad . \tag{9}$$

Then, if $\widehat{m}$ is defined by (5) and $0 < \eta < \min\left\{\beta_+; 1\right\}/2$, there exists a constant K_3 and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_3 n^{-2}$, $D_{\widehat{m}} \leq n^{1-\eta}$ and

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq \left(\frac{1+\delta}{1-\delta} + \epsilon_n\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s, \widehat{s}_m\right)\right\} \quad . \tag{10}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq \left(\frac{1+\delta}{1-\delta} + \epsilon_n\right) \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{\ell\left(s, \widehat{s}_m\right)\right\}\right] + \frac{A^2 K_3}{n^2} \quad . \tag{11}$$

*The constant K_3 may depend on L, δ, η and the constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) and ($\mathbf{Ar}_\ell^{\mathbf{X}}$), but not on n. The small term ϵ_n is smaller than $\ln(n)^{-1/5}$; it can also be taken smaller than $n^{-\delta}$ for any $\delta \in (0; \delta_0(\beta_-, \beta_+))$ at the price of enlarging K_3.*

This theorem shows that twice the minimal penalty $\mathrm{pen}_{\min}$ pointed out by Thm. 1 satisfies an oracle inequality with a leading constant almost equal to one. It even stays valid when the penalty is only "close to" twice the minimal one, which means in particular that one can estimate the shape of the minimal penalty by resampling for instance (see Sect. 3.2). The rationale behind this theorem is that the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ is close to its expectation, which is itself close to $2\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right]$. Then, (3) directly implies an oracle inequality like (10), hence (11). In other words, we have proven the second part of the slope heuristics of Sect. 2.3.

Actually, Thm. 2 above is a corollary of a more general result (Thm. 5), that we state in App. B.2. In particular, if

$$\mathrm{pen}(m) \approx K\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \tag{12}$$

instead of (9), we can prove under the same assumptions that the same oracle inequality holds with a large probability, with a leading constant $C(K) + \epsilon_n$ instead of "almost one". When $K \in (1,2]$, we have $C(K) = (K-1)^{-1}$, and when $K > 2$, $C(K) = K - 1$. This means that for every $K > 1$, the penalty defined by (12) is efficient, up to a multiplicative constant. This is well known in the homoscedastic case (Birgé and Massart, 2001; Baraud, 2000, 2002), but new in the heteroscedastic one.

The most important consequences of this result follows from its *combination* with Thm. 1. We detail them in the next subsection. Let us first comment the additional

assumption (**Ap**), *i.e.* the lower bound on the bias. It means that s is not too well approximated by the models S_m, which may seem surprising. Notice that it is classical to assume that $\ell(s, s_m) > 0$ for every $m \in \mathcal{M}_n$, for proving the asymptotic optimality of Mallows' C_p (*cf.* Shibata (1981), Li (1987) and Birgé and Massart (2007)). Moreover, the stronger assumption (**Ap**) has already been made by Stone (1985) and Burman (2002) in the density estimation framework, for the same technical reasons as ours.

As detailed in (Arlot, 2008a) where a similar technique is used to derive an oracle inequality, when the lower bound in (**Ap**) is no longer assumed, (10) holds with two modifications in its right-hand side: the inf is restricted to models of dimension larger than $\ln(n)^{\gamma_1}$, and there is a remainder term $\ln(n)^{\gamma_2} n^{-1}$ (where γ_1 and γ_2 are numerical). This is essentially the same as (10), unless there is a model of small dimension with a very small bias, and the lower bound in (**Ap**) is sufficient to ensure that this do not happen. Notice that if there is such a very small model very close to s, it is hopeless to obtain an oracle inequality with a penalty which estimates $\mathrm{pen}_{\mathrm{id}}$, simply because deviations of $\mathrm{pen}_{\mathrm{id}}$ around its expectation would be much larger than the excess loss of the oracle. In such a situation, BIC-like methods are more appropriate.

Another argument in favour of (**Ap**) is that it is not too strong, because it is at least satisfied in the following case: $(I_\lambda)_{\lambda \in \Lambda_m}$ is "regular", X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$, and s is non-constant and α-hölderian (w.r.t. $\|\cdot\|_\infty$), with

$$\beta_1 = k^{-1} + \alpha^{-1} - (k-1)k^{-1}\alpha^{-1} \qquad \text{and} \qquad \beta_2 = 2\alpha k^{-1} \ .$$

We refer to Sect. 8.10 in (Arlot, 2007) for more details about this claim (including complete proofs).

We finally mention that this is not the only case where (**Ap**) holds, which is the reason why we use (**Ap**) as an assumption, and not these sufficient conditions (*cf.* the comments at the end of Sect. 4.3).

4.5 Main theoretical and practical consequences

Combining Thm. 1 and 2, we are now in position to "prove" the slope heuristics described in Sect. 2.3, as well as the validity of our Algorithm 1 (provided that $\mathrm{pen}_{\mathrm{shape}}$ is well chosen, for instance estimated by resampling).

4.5.1 Optimal penalty *vs.* minimal penalty

For the sake of simplicity, consider the penalty $K\mathbb{E}[p_2(m)]$ with any $K > 0$ (the same phenomenon occuring for a penalty approximately equal to this one). At first reading, one can think of the homoscedastic case where $\mathbb{E}[p_2(m)] \approx \sigma^2 D_m n^{-1}$, the general picture being quite similar (this generalization is one of the novelties of our results).

With Thm. 2, we have shown that it satisfies an oracle inequality with a leading constant $C_n(K)$ as soon as $K > 1$. Moreover, $C_n(2) \approx 1$. According to (Arlot, 2008b) (the proof of its Thm. 1, in particular Lemma 6), $C_n(K)$ stays away from 1 as soon as K is not close to 2. This means that $K = 2$ *is the optimal multiplying factor* in front of $\mathbb{E}[p_2(m)]$.

On the other hand, when $K < 1$, Thm. 1 shows that no oracle inequality can hold with a leading constant $C_n(K)$ smaller than $\ln(n)$ (and even much larger in most cases, according to the proof of Thm. 1). Since $C_n(K) \leq (K-1)^{-1} < \ln(n)$ as soon as $K > 1 + \ln(n)^{-1}$, this

means that $K = 1$ *is the minimal multiplying factor* in front of $\mathbb{E}[p_2(m)]$. More generally, we have proven that $\mathrm{pen}_{\min}(m) := \mathbb{E}[p_2(m)]$ is a *minimal penalty*.

In a nutshell, this is a formal proof of the heuristics of Sect. 2.3:

$$\textit{"optimal" penalty} \approx 2 \times \textit{"minimal" penalty} \ .$$

This has already been proposed by Birgé and Massart (2007), but their results were restricted to the Gaussian homoscedastic framework. In this paper, we extend them to a non-Gaussian and heteroscedastic setting.

4.5.2 Dimension jump

In addition, Thm. 1 and 2 prove the existence of a crucial phenomenon around the minimal penalty, which is the existence of a "dimension jump". This is the only reason why we can estimate the minimal penalty in practice (since the explosion of the prediction error can not be directly observed), so that Algorithm 1 strongly relies on it.

Indeed, consider again the penalty $K\mathbb{E}[p_2(m)]$, and define $\widehat{m}(K)$ the selected model as a function of K. For each $K > 0$, with a large probability, we have $D_{\widehat{m}(K)} \leq n^{1-\eta}$ if $K > 1$ and $D_{\widehat{m}(K)} \geq K_2 n(\ln(n))^{-1}$ if $K < 1$ (the constant K_2 depends on K). More precisely, a careful look at the proofs shows that *this holds simultaneously* in the following sense: there are constants $K_4, K_5 > 0$ and an event of probability $1 - K_4 n^{-2}$ on which

$$\begin{aligned} \forall K \in \left(0, 1 - \ln(n)^{-1}\right), \quad & D_{\widehat{m}(K)} \geq K_5 n(\ln(n))^{-2} \\ \text{and} \quad \forall K \in \left(1 + \ln(n)^{-1}, +\infty\right), \quad & D_{\widehat{m}(K)} \leq n^{1-\eta} \ . \end{aligned}$$

This means that there must be a *dimension jump* around $K = 1$, from dimensions of order at least $n(\ln(n))^{-2}$ to dimensions much smaller, of order at most $n^{1-\eta}$. Actually, there can be several jumps instead of only one, but they occur for very close values of K (at least when n is large).

Let us now come back to Algorithm 1. Defining a "reasonably small dimension" as any dimension smaller than $n(\ln(n))^{-3}$, we have proven that $\widehat{K}_{\min}$ must be close to the true "minimal" multiplying factor. When the penalty is $K\mathbb{E}[p_2(m)]$, we have

$$1 - \frac{1}{\ln(n)} \leq \widehat{K}_{\min} \leq 1 + \frac{1}{\ln(n)}$$

with a probability at least $1 - K_4 n^{-2}$. Notice that $n(\ln(n))^{-3}$ can be replaced by any dimension between $K_5 n(\ln(n))^{-2}$ and $n^{1-\eta}$, which are very far as soon as n is large enough. Hence, this dimension threshold does not have to be chosen accurately as soon as n is not small.

Combined with Thm. 2, this shows that *the model selection procedure of Algorithm 1 satisfies an oracle inequality with a leading constant smaller than* $1 + 2\ln(n)^{-1/5}$, on a large probability event. In addition, the same result holds when $\mathrm{pen}_{\mathrm{shape}}$ is only "close" to the ideal penalty shape, *e.g.* within a ratio $1 \pm \ln(n)^{-1}$. In particular, the resampling penalties of Efron (1983) and Arlot (2008b,a) satisfy this condition on a large probability event. We refer to Sect. 3.2 for further discussion on this question.

5. Conclusion

We have seen in this paper that it is possible to provide mathematical evidences that the method introduced by Birgé and Massart (2007) to design data-driven penalties remains efficient in a non Gaussian context. Our purpose in this conclusive section is to relate the heuristics that we have developped in Sect. 2 to the well known Mallows' C_p and Akaike's criteria and to the unbiased (or almost unbiased) estimation of the risk principle. To explain our idea which consists in guessing what is the right penalty to be used from the data themselves, let us come come back to Gaussian model selection. Towards this aim let us consider some empirical criterion γ_n (which can be the least squares criterion as in this paper but which could be the log-likelihood criterion as well). Let us also consider some collection of models $(S_m)_{m\in\mathcal{M}}$ and in each model S_m some minimizer s_m of $t \mapsto \mathbb{E}[\gamma_n(t)]$ over S_m (assuming that such a point does exist). Defining for every $m \in \mathcal{M}$,

$$\widehat{b}_m = \gamma_n(s_m) - \gamma_n(s) \text{ and } \widehat{v}_m = \gamma_n(s_m) - \gamma_n(\widehat{s}_m) \ ,$$

minimizing some penalized criterion

$$\gamma_n(\widehat{s}_m) + \mathrm{pen}(m)$$

over $\mathcal{M}$ amounts to minimize

$$\widehat{b}_m - \widehat{v}_m + \mathrm{pen}(m) \ .$$

The point is that $\widehat{b}_m$ is an unbiased estimator of the bias term $\ell(s, s_m)$. If we have in mind to use concentration arguments, one can hope that minimizing the quantity above will be approximately equivalent to minimize

$$\ell(s, s_m) - \mathbb{E}[\widehat{v}_m] + \mathrm{pen}(m) \ .$$

Since the purpose of the game is to minimize the risk $\mathbb{E}[\ell(s, \widehat{s}_m)]$, an ideal penalty would therefore be

$$\mathrm{pen}(m) = \mathbb{E}[\widehat{v}_m] + \mathbb{E}[\ell(s_m, \widehat{s}_m)] \ .$$

In the Mallows' C_p case (for Gaussian fixed design regression least squares), the models S_m are linear and $\mathbb{E}[\widehat{v}_m] = \mathbb{E}[\ell(s_m, \widehat{s}_m)]$ are explicitly computable (at least if the level of noise is assumed to be known). For Akaike's penalized log-likelihood criterion, this is similar, at least asymptotically. More precisely, one uses the fact that

$$\mathbb{E}[\widehat{v}_m] \approx \mathbb{E}[\ell(s_m, \widehat{s}_m)] \approx \frac{D_m}{2n} \ ,$$

where D_m stands for the number of parameters defining model S_m. The conclusion of these considerations is that Mallows' C_p as well as Akaike's criterion are indeed both based on the unbiased risk (or asymptotically unbiased) estimation principle.

The first idea that we are using in this paper is that one can go further in this direction and that the approximation $\mathbb{E}[\widehat{v}_m] \approx \mathbb{E}[\ell(s_m, \widehat{s}_m)]$ remains valid even in a non-asymptotic context. If one believes in it then a good penalty becomes $2\mathbb{E}[\widehat{v}_m]$ or equivalently (having still in mind concentration arguments) $2\widehat{v}_m$. This in some sense explains the rule of thumb which is given by Birgé and Massart (2007) and further studied in this paper, and connect

it to Mallows' C_p and Akaike's heuristics. Indeed, the minimal penalty is $\widehat{v}_m$ while the optimal penalty should be $\widehat{v}_m + \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right]$ and their ratio is approximately equal to 2.

The second idea that we are using in this paper is that one can guess the minimal penalty from the data. There are indeed several ways to perform the estimation of the minimal penalty. Here we are using the jump of dimension which occurs around the minimal penalty. When the shape of the minimal penalty is (at least approximately) of the form αD_m, this amounts to estimate the unknown value α by the *slope* of the graph of $\gamma_n\left(\widehat{s}_m\right)$ for large enough values of D_m. It is easy to extend this method to other shapes of penalties, simply by replacing D_m by some (known!) function $f\left(D_m\right)$.

It is even possible to combine resampling ideas with the slope heuristics by taking a random function f which is built from a randomized empirical criterion. As shown by Arlot (2007) this approach turns out to be much more efficient than the rougher choice $f\left(D_m\right) = D_m$ for highly heteroscedastic random regression frameworks. Of course, the question of the optimality of the slope heuristics remains widely open but we believe that on the one hand this heuristics can be helpful in practice and that on the other hand, proving its efficiency even on a toy model as we did in this paper is already something.

Let us finally mention that contrary to Birgé and Massart (2007), we have restricted our study to the situation where the collection of models $\mathcal{M}_n$ is "small", *i.e.* has a size growing at most like a power of n. For several problems, such that complete variable selection, this assumption does not hold, and it is known from the homoscedastic case that the minimal penalty is much larger than $\mathbb{E}\left[p_2(m)\right]$. For instance, using the results by Birgé and Massart (2007) in the Gaussian case, Émilie Lebarbier has used the slope heuristics with $f\left(D_m\right) = D_m\left(2.5 + \ln\left(\frac{n}{D_m}\right)\right)$ for multiple change points detection from n noisy data. Let us now explain how we expect to generalize their heuristics to the non-Gaussian heteroscedastic case.

First, group the models according to some complexity index C_m (for instance their dimensions, or the approximate value of their resampling penalty suitably normalized): for $C \in \left\{1, \ldots, n^k\right\}$, define $\widetilde{S_C} = \bigcup_{C_m = C} S_m$. Then, replace the model selection problem with the family $(S_m)_{m \in \mathcal{M}_n}$ by a "complexity selection problem", *i.e.* model selection with the family $\left(\widetilde{S_C}\right)_{1 \leq C \leq n^k}$. We conjecture that this grouping of the models is sufficient to take into account the richness of $\mathcal{M}_n$ for the optimal calibration of the penalty. A theoretical justification of this point may rely on the extension of our results to any kind of model, not only histogram ones (each $\widetilde{S_C}$ is not an "histogram model", since it is even not a vector space). As already mentioned, this remains an interesting open problem.

Appendix A. Computational aspects of the slope heuristics

With Algorithm 2 (possibly combined with resampling penalties for step 1), we have a completely data-driven and optimal model selection procedure. From the practical viewpoint, the last two problems may be steps 1 and 2. First, at step 1, how can we compute exactly $\widehat{m}(K)$ for every $K \in (0, +\infty)$, this latter set being uncountable? The answer is that the whole trajectory $\left(\widehat{m}(K)\right)_{K \geq 0}$ can be described with a small number of parameters, which can be computed fastly. This point is the object of Sect. A.1. Second, at step 2, how can the jump of dimension be detected automatically in practice? In other words, how should

$\widehat{K}_{\min}$ be defined exactly, as a function of $(\widehat{m}(K))_{K\geq 0}$? We try to answer this question in Sect. A.2.

A.1 Computation of $(\widehat{m}(K))_{K\geq 0}$

For every model $m \in \mathcal{M}_n$, define

$$f(m) = P_n\gamma(\widehat{s}_m) \qquad\qquad g(m) = \mathrm{pen}_{\mathrm{shape}}(m)$$

$$\text{and} \qquad \forall K \geq 0, \qquad \widehat{m}(K) \in \arg\min_{m\in\mathcal{M}_n} \{f(m) + Kg(m)\} \ .$$

Since the latter definition can be ambiguous, we choose any total ordering $\preceq$ on $\mathcal{M}_n$ such that g is non-decreasing. Then, $\widehat{m}(K)$ is defined as the smallest element of

$$E(K) := \arg\min_{m\in\mathcal{M}_n} \{f(m) + Kg(m)\}$$

for $\preceq$. The main reason why the whole trajectory $(\widehat{m}(K))_{K\geq 0}$ can be computed efficiently is its very particular shape.

Indeed, the results below (mostly Lemma 4) show that $K \mapsto \widehat{m}(K)$ is piecewise constant, and non-increasing for $\preceq$. We then have

$$\forall i \in \{0, \dots, i_{\max}\}, \quad \forall K \in [K_i, K_{i+1}), \qquad \widehat{m}(K) = m_i \ ,$$

and the whole trajectory $(\widehat{m}(K))_{K\geq 0}$ can be represented by:

- a non-negative integer $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n) - 1$ (the number of jumps),
- an increasing sequence of positive reals $(K_i)_{0\leq i\leq i_{\max}+1}$ (the location of the jumps, with $K_0 = 0$ and $K_{i_{\max}+1} = +\infty$)
- a non-increasing sequence of models $(m_i)_{0\leq i\leq i_{\max}}$.

We are now in position to give an efficient algorithm for step 1 in Algorithm 2. The point is that the K_i and the m_i can be computed sequentially, each step having a complexity proportional to $\mathrm{Card}(\mathcal{M}_n)$. This means that its overall complexity is lower than a constant times $i_{\max}\,\mathrm{Card}(\mathcal{M}_n) \leq \mathrm{Card}(\mathcal{M}_n)^2$ (and the latter bound is quite pessimistic in general). Notice also that Algorithm 2 can be stopped earlier if the only goal is to identify $\widehat{K}_{\min}$ (which may be done only with the first m_i).

Algorithm 2 (Step 1 of Algorithm 1) *For every $m \in \mathcal{M}_n$, define $f(m) = P_n\gamma(\widehat{s}_m)$ and $g(m) = \mathrm{pen}_{\mathrm{shape}}(m)$. Choose $\preceq$ any total ordering on $\mathcal{M}_n$ such that g is non-decreasing.*

- *Init: $K_0 = 0$, $m_0 = \arg\min_{m\in\mathcal{M}_n}\{f(m)\}$ (when this minimum is attained several times, m_0 is defined as the smallest one for $\preceq$).*
- *Step i, $i \geq 1$: Let*

$$G(m_{i-1}) := \{m \in \mathcal{M}_n \ \ s.t. \ \ f(m) > f(m_{i-1}) \quad \text{and} \quad g(m) < g(m_{i-1})\} \ .$$

If $G(m_{i-1}) = \emptyset$, then put $K_i = +\infty$, $i_{\max} = i-1$ and stop.
Otherwise, define

$$K_i := \inf \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \;\; s.t. \;\; m \in G(m_{i-1}) \right\} \tag{13}$$

and m_i the smallest element (for $\preceq$) of

$$F_i := \arg \min_{m \in G(m_{i-1})} \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \right\} \;\; .$$

The validity of Algorithm 2 is justified by the following proposition, showing that these K_i and m_i are the same as the ones describing $\left(\widehat{m}(K)\right)_{K \geq 0}$.

Proposition 3 *If $\mathcal{M}_n$ is finite, Algorithm 2 terminates and $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n) - 1$. Using the notations of Algorithm 2, and defining $\widehat{m}(K)$ as the smallest element (for $\preceq$) of*

$$E(K) := \arg \min_{m \in \mathcal{M}_n} \left\{ f(m) + K g(m) \right\} \;\; ,$$

$(K_i)_{0 \leq i \leq i_{\max}+1}$ is increasing and $\forall i \in \{0, \ldots, i_{\max} - 1\}$, $\forall K \in [K_i, K_{i+1})$, $\widehat{m}(K) = m_i$.

It is proven in Sect. A.3.

A.2 Definition of $\widehat{K}_{\min}$

We now come to the question of defining $\widehat{K}_{\min}$ as a function of $\left(\widehat{m}(K)\right)_{K>0}$. As we have mentioned in Sect. 4.5.2, it corresponds to a "dimension jump", which should be observable since the whole trajectory of $\left(D_{\widehat{m}(K)}\right)_{K \geq 0}$ is known.

As an illustration to this question, we represented on Fig. 1 $D_{\widehat{m}(K)}$ as a function of K, for two simulated samples. On the left (a), the dimension jump is quite clear, and we expect a formal definition of $\widehat{K}_{\min}$ to find this jump. The same picture holds for approximately 85% of the data sets. On the right (b), there seems to be several jumps, and a proper definition of $\widehat{K}_{\min}$ is problematic. What is sure is the necessity to find some automatic choice for $\widehat{K}_{\min}$, that is defining it properly.

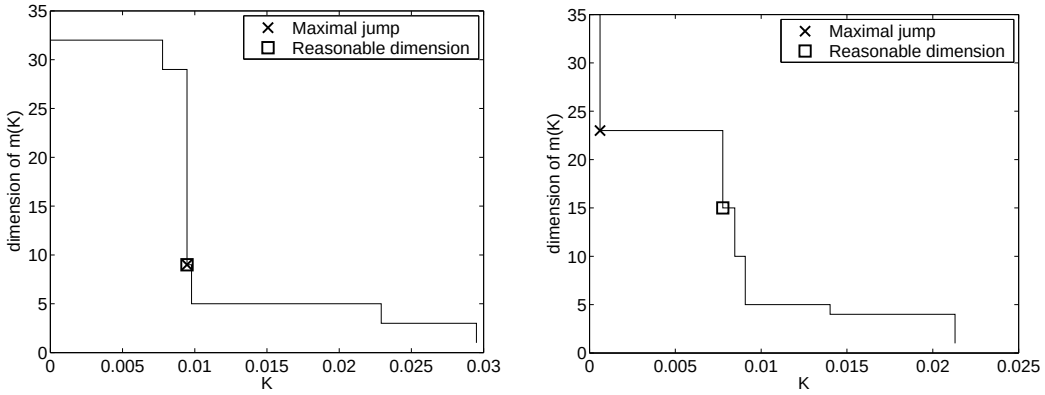
We now propose two definitions that seem reasonable to us. For the first one, choose a threshold $D_{\text{reas.}}$, of order $n/(\ln(n))$, corresponding to the largest "reasonable" dimension for the selected model. Then, define

$$\widehat{K}_{\min} := \inf \left\{ K > 0 \;\text{ s.t. }\; D_{\widehat{m}(K)} \leq D_{\text{reas.}} \right\} \;\; .$$

With this definition, one can stop Algorithm 2 as soon as the threshold is reached. However, $\widehat{K}_{\min}$ may depend strongly on the choice of the threshold, which may not be quite obvious in the non-asymptotic situation (where $n/\ln(n)$ is not so far from n).

Our second idea is that $\widehat{K}_{\min}$ should match with the largest dimension jump, *i.e.*

$$\widehat{K}_{\min} := K_{i_{\text{max.jump}}} \qquad \text{with} \qquad i_{\text{max.jump}} = \arg \max_{i \in \{0, \ldots, i_{\max} - 1\}} \left\{ D_{m_{i+1}} - D_{m_i} \right\} \;\; .$$



(a) One clear jump. (b) Two jumps, two values for $\widehat{K}_{\min}$.

Figure 1: $D_{\widehat{m}(K)}$ as a function of K for two different samples. Data are simulated with $X \sim \mathcal{U}([0,1])$, $\epsilon \sim \mathcal{N}(0,1)$, $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$. $(S_m)_{m \in \mathcal{M}_n}$ is the collection of regular histogram models with dimension between 1 and $n/(\ln(n))$. $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$. "Reasonable dimensions" are below $n/(2\ln(n)) \approx 19$. See (Arlot, 2008b) for details (experiment S1).

Although this definition may seem less arbitrary than the previous one, it still depends strongly on $\mathcal{M}_n$, which may not contain so many large models for computational reasons. In order to ensure that there is a clear jump, an idea may be to add a few models of dimension $\approx n/2$, so that at least one has a well-defined empirical risk minimizer $\widehat{s}_m$. In practice, several huge models with a well-defined $\widehat{s}_m$ may be necessary, in order to decrease the variability of $\widehat{K}_{\min}$. This modification has the default of being quite arbitrary.

As an illustration, we compared the two definitions above ("reasonable dimension" *vs.* "maximal jump") on one thousand simulated samples similar to the one of Fig. 1. Three cases occured:

1. The values of $\widehat{K}_{\min}$ do not differ (about 85% of the data sets; this is the (a) situation).
2. The values of $\widehat{K}_{\min}$ differ, but the selected models $\widehat{m}\left(2\widehat{K}_{\min}\right)$ are still equal (about 8.5% of the data sets).
3. The finally selected models are different (about 6.5% of the data sets; this is the (b) situation).

Hence, in this non-asymptotic framework, the formal definition of $\widehat{K}_{\min}$ does not matter in general, but stays problematic in a few cases.

In terms of prediction error, we have compared the two methods by estimating the constant C_{or} that would appear in some oracle inequality:

$$C_{\mathrm{or}} := \frac{\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n}\left\{\ell\left(s, \widehat{s}_m\right)\right\}\right]} \cdot$$

With the "reasonable dimension" definition, $C_{\mathrm{or}} \approx 1.88$. With the "maximal jump" definition, $C_{\mathrm{or}} \approx 2.01$. As a comparison, Mallows' C_p (with a classical estimator of the variance σ^2) has a performance of $C_{\mathrm{or}} \approx 1.93$ on the same data. For the three procedures, the standard deviation of the estimator of C_{or} is about 0.04. See Chap. 4 of (Arlot, 2007) for more details. This preliminary simulation study shows that Algorithm 1 works efficiently (it is competitive with Mallows' C_p in a situation where this one is also optimal). It also suggests that the "reasonable dimension" definition may be better, but without very convincing evidence.

In order to make the choice of $\widehat{K}_{\min}$ as automatic as possible, we suggest to use simultaneously the two methods. When the selected models are not the same, then, send a warning to the final user, advising him to look at the curve $K \mapsto D_{\widehat{m}(K)}$ himself. Otherwise, stay confident in the automatic choice of $\widehat{m}(2\widehat{K}_{\min})$.

A.3 Proof of Prop. 3

First of all, since $\mathcal{M}_n$ is finite, the infimum in (13) is attained as soon as $G(m_{i-1}) \neq \emptyset$, so that m_i is well defined for every $i \leq i_{\max}$. Moreover, by construction, $g(m_i)$ decreases with i, so that all the $m_i \in \mathcal{M}_n$ are distinct. Hence, Algorithm 2 terminates and $i_{\max}+1 \leq \mathrm{Card}(\mathcal{M}_n)$. We now prove by induction the following property for every $i \in \{0, \ldots, i_{\max}\}$:

$$\mathcal{P}_i : \qquad K_i < K_{i+1} \quad \text{and} \quad \forall K \in [K_i, K_{i+1}), \quad \widehat{m}(K) = m_i \ .$$

Notice also that K_i can always be defined by (13) with the convention $\inf \emptyset = +\infty$.

$\mathcal{P}_0$ holds true

By definition of K_1, it is clear that $K_1 > 0$ (it may be equal to $+\infty$ if $G(m_0) = \emptyset$). For $K = K_0 = 0$, the definition of m_0 is the one of $\widehat{m}(0)$, so that $\widehat{m}(K) = m_0$. For $K \in (0, K_1)$, Lemma 4 shows that either $\widehat{m}(K) = \widehat{m}(0) = m_0$ or $\widehat{m}(K) \in G(0)$. In the latter case, by definition of K_1,

$$\frac{f(\widehat{m}(K)) - f(m_0)}{g(m_0) - g(\widehat{m}(K))} \geq K_1 > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_0) + Kg(m_0)$$

which is contradictory with the definition of $\widehat{m}(K)$. Hence, $\mathcal{P}_0$ holds true.

$\mathcal{P}_i \Rightarrow \mathcal{P}_{i+1}$ for every $i \in \{0, \ldots, i_{\max} - 1\}$

Assume that $\mathcal{P}_i$ holds true. First, we have to prove that $K_{i+2} > K_{i+1}$. Since $K_{i_{\max}+1} = +\infty$, this is clear if $i = i_{\max} - 1$. Otherwise, $K_{i+2} < +\infty$ and m_{i+2} exists. Then, by definition of m_{i+2} and K_{i+2} (resp. m_{i+1} and K_{i+1}), we have

$$f(m_{i+2}) - f(m_{i+1}) = K_{i+2}(g(m_{i+1}) - g(m_{i+2})) \tag{14}$$

$$f(m_{i+1}) - f(m_i) = K_{i+1}(g(m_i) - g(m_{i+1})) \ . \tag{15}$$

Moreover, $m_{i+2} \in G(m_{i+1}) \subset G(m_i)$, and $m_{i+2} \prec m_{i+1}$ (because g is non-decreasing). Using again the definition of K_{i+1}, we have

$$f(m_{i+2}) - f(m_i) > K_{i+1}(g(m_i) - g(m_{i+2})) \tag{16}$$

(otherwise, we would have $m_{i+2} \in F_{i+1}$ and $m_{i+2} \prec m_{i+1}$, which is not possible). Combining the difference of (16) and (15) with (14), we have

$$K_{i+2}(g(m_{i+1}) - g(m_{i+2})) > K_{i+1}(g(m_{i+1}) - g(m_{i+2})) \ ,$$

so that $K_{i+2} > K_{i+1}$ (since $g(m_{i+1}) > g(m_{i+2})$).

Second, we prove that $\widehat{m}(K_{i+1}) = m_{i+1}$. From $\mathcal{P}_i$, we know that for every $m \in \mathcal{M}_n$, for every $K \in [K_i, K_{i+1})$, $f(m_i) + Kg(m_i) \leq f(m) + Kg(m)$. Taking the limit when K goes to K_{i+1}, we obtain that $m_i \in E(K_{i+1})$. By (15), we then have $m_{i+1} \in E(K_{i+1})$. On the other hand, if $m \in E(K_{i+1})$, Lemma 4 shows that either $f(m) = f(m_i)$ and $g(m) = g(m_i)$ or $m \in G(m_i)$. In the first case, $m_{i+1} \prec m$ (because g is non-decreasing). In the second one, $m \in F_{i+1}$, so $m_{i+1} \preceq m$. Since $\widehat{m}(K_{i+1})$ is the smallest element of $E(K_{i+1})$, we have proven that $m_{i+1} = \widehat{m}(K_{i+1})$.

Last, we have to prove that $\widehat{m}(K) = m_{i+1}$ for every $K \in (K_1, K_2)$. From the last statement of Lemma 4, we have either $\widehat{m}(K) = \widehat{m}(K_1)$ or $\widehat{m}(K_1) \in G(\widehat{m}(K))$. In the latter case (which is only possible if $K_{i+2} < \infty$), by definition of K_{i+2},

$$\frac{f(\widehat{m}(K)) - f(m_{i+1})}{g(m_{i+1}) - g(\widehat{m}(K))} \geq K_{i+2} > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_{i+1}) + Kg(m_{i+1})$$

which is contradictory with the definition of $\widehat{m}(K)$. ■

Lemma 4 *Use the notations of Prop. 3 and its proof. If $0 \leq K < K'$, $m \in E(K)$ and $m' \in E(K')$, then we have either*

(a) $f(m) = f(m')$ and $g(m) = g(m')$.

(b) $f(m) < f(m')$ and $g(m) > g(m')$.

In particular, we have either $\widehat{m}(K) = \widehat{m}(K')$ or $\widehat{m}(K') \in G(\widehat{m}(K))$.

Proof By definition of $E(K)$ and $E(K')$, we have

$$f(m) + Kg(m) \leq f(m') + Kg(m') \tag{17}$$
$$f(m') + K'g(m') \leq f(m) + K'g(m) \ . \tag{18}$$

Summing (17) and (18) gives $(K' - K)g(m') \leq (K' - K)g(m)$ so that

$$g(m') \leq g(m) \ . \tag{19}$$

Since $K \geq 0$, (17) and (19) give $f(m) + Kg(m) \leq f(m') + Kg(m)$, *i.e.*

$$f(m) \leq f(m') \ . \tag{20}$$

Moreover, using (18), $g(m) = g(m')$, implies $f(m') \leq f(m)$, *i.e.* $f(m) = f(m')$ by (20). In the same way, (17) and (19) show that $f(m) = f(m')$ imply $g(m) = g(m')$. In both cases, (a) is satisfied. Otherwise, $f(m) < f(m')$ and $g(m) > g(m')$, *i.e.* (b) is satisfied.

The last statement follows by taking $m = \widehat{m}(K)$ and $m' = \widehat{m}(K')$, because g is non-decreasing, so that the minimum of g in $E(K)$ is attained by $\widehat{m}(K)$. ■

Appendix B. Proofs

B.1 Conventions and notations

In the following, when we do not want to write explicitly some constants, we use the letter L. It means "some absolute constant, possibly different from a line to another, or even within the same line". When L is not numerical, but depends on some parameters $p_1, \ldots, p_k$, it is written $L_{p_1,\ldots,p_k}$. $L_{(\mathbf{SH1})}$ (resp. $L_{(\mathbf{SH5})}$) denotes a constant that depends only on the set of assumptions of Thm. 1 (resp. Thm. 5), including (**P1**) and (**P2**).

We also make use of the following notations:

- for every $a, b \in \mathbb{R}$, $a \wedge b$ is the minimum of a and b, $a \vee b$ is the maximum of a and b, $a_+ = a \vee 0$ is the positive part of a and $a_- = a \wedge 0$ is its negative part.
- for every $I_\lambda \subset \mathcal{X}$, $p_\lambda := P(X \in I_\lambda)$ and $\sigma_\lambda^2 := \mathbb{E}\left[\left(Y - s_m(X) \right)^2 \middle| \ X \in I_\lambda \right]$.
- Since $\mathbb{E}\left[p_1(m) \right]$ is not well-defined (because of the event $\{ \min_{\lambda \in \Lambda_m} \{ \widehat{p}_\lambda \} = 0 \}$), we have to take the following convention

$$p_1(m) = \widetilde{p_1}(m) := \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda > 0} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda \right)^2 + \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda = 0} p_\lambda \sigma_\lambda^2 \ . \tag{21}$$

  Remark that $p_1(m) = \widetilde{p_1}(m)$ when $\min_{\lambda \in \Lambda_m} \{ \widehat{p}_\lambda \} > 0$), so that this convention has no consequences on the final results (Thm. 1 and 5).

B.2 A general oracle inequality

First of all, let us state a general theorem, from which Thm. 2 is an obvious corollary.

Theorem 5 *Make all the assumptions of Sect. 4.2, and add the following:*

(**Ap**) *The bias decreases like a power of D_m: there exists $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_m^{-\beta_-} \leq \ell\left(s, s_m \right) \leq C_+ D_m^{-\beta_+} \ .$$

Let $L, \xi, c_1, c_2, C_1, C_2 \geq 0$ such that $c_2 > 1$ and assume that there is an event of probability at least $1 - Ln^{-2}$ on which, for every $m \in \mathcal{M}_n$ such that $D_m \geq \ln(n)^\xi$,

$$\begin{aligned} &\mathbb{E}\left[c_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m) \right) + c_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m) \right) \right] \\ &\quad \leq \operatorname{pen}(m) \leq \mathbb{E}\left[C_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m) \right) + C_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m) \right) \right] \ . \end{aligned} \tag{22}$$

Then, if $\widehat{m}$ is defined by (5) *and $0 < \eta < \min\{\beta_+; 1\}/2$, there exists a constant K_3 and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_3 n^{-2}$, $D_{\widehat{m}} \leq n^{1-\eta}$ and*

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n \right] \inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \} \quad . \tag{23}$$

Moreover, we have the oracle inequality

$$\mathbb{E}[\ell(s, \widehat{s}_{\widehat{m}})] \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n \right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \} \right] + \frac{A^2 K_3}{n^2} \quad . \tag{24}$$

The constant K_3 may depend on L, η, ξ, c_1, c_2, C_1, C_2 and constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) *and* ($\mathbf{Ar}_\ell^{\mathbf{X}}$), *but not on n. The small term ϵ_n is smaller than $\ln(n)^{-1/5}$; it can also be taken smaller than $n^{-\delta}$ for any $\delta \in (0; \delta_0(\beta_-, \beta_+))$ at the price of enlarging K_3.*

The particular form of condition (22) on the penalty is motivated by the fact that the ideal shape of penalty $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ (or equivalently $\mathbb{E}[2p_2(m)]$) is unknown in general. Then, it has to be estimated from the data, for instance by resampling. Under the assumptions of Thm. 5, it has been proven by Arlot (2008b,a) that resampling penalties satisfy condition (22) with constants $c_1 + c_2 = 2 - \delta_n$ and $C_1 + C_2 = 2 + \delta_n$ (for some absolute sequence δ_n converging to zero at infinity), at least for models of dimension larger than $\ln(n)^\xi$ (where ξ depends on the constants in the assumptions on the data).

In such a situation (obtained by resampling or not), (23) shows that we have an *asymptotically optimal* model selection procedure.

The rationale behind this theorem is that if pen is close to $c_1 p_1 + c_2 p_2$, then $\mathrm{crit}(m) = \ell(s, s_m) + c_1 p_1(m) + (c_2 - 1) p_2(m)$. If $c_1 = c_2 = 1$, this is exactly the ideal criterion $\ell(s, \widehat{s}_m)$. If $c_1 + c_2 = 2$ with $c_1 \geq 0$ and $c_2 > 1$, we obtain the same result because $p_1(m)$ and $p_2(m)$ are quite close (at least when D_m is large). This closeness between p_1 and p_2 is the keystone of the slope heuristics. Notice that if $\max_{m \in \mathcal{M}_n} D_m \leq K_3' (\ln(n))^{-1} n$ (for some constant K_3' depending only on the assumptions of Thm. 2, as K_3), one can replace the condition $c_2 > 1$ by $c_1 + c_2 > 1$ and $c_1, c_2 \geq 0$.

B.3 Proof of Thm. 5

This proof is very similar to the one of Thm. 2 of (Arlot, 2007). We give it for the sake of completeness.

From (3), we have for each $m \in \mathcal{M}_n$ such that $A_n(m) := \min_{\lambda \in \Lambda_m} \{ n \widehat{p}_\lambda \} > 0$

$$\ell(s, \widehat{s}_{\widehat{m}}) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m}) \right) \leq \ell(s, \widehat{s}_m) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m) \right) \quad . \tag{25}$$

with $\mathrm{pen}'_{\mathrm{id}}(m) := p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}(m) + (P - P_n)\gamma(s)$ and $\overline{\delta}(m) := (P_n - P)(\gamma(s_m) - \gamma(s))$. It is sufficient to control $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ for every $m \in \mathcal{M}_n$.

We will thus use the concentration inequalities of Sect. B.5 with $x = \gamma \ln(n)$ and $\gamma = 2 + \alpha_{\mathcal{M}}$. Define $B_n(m) = \min_{\lambda \in \Lambda_m} \{ n p_\lambda \}$. Let Ω_n be the event on which

- for every $m \in \mathcal{M}_n$, (22) holds

- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$:

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{(\mathbf{SH5})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n(m)}\right]\mathbb{E}\left[p_2(m)\right] \tag{34}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{(\mathbf{SH5})}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n(m)}\right]\mathbb{E}\left[p_2(m)\right] \tag{35}$$

- for every $m \in \mathcal{M}_n$ such that $B_n(m) > 0$:

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n(m)^{-1}\ln(n)} - \frac{L_{(\mathbf{SH5})}\ln(n)^2}{\sqrt{D_m}}\right)\mathbb{E}\left[p_2(m)\right] \ . \tag{36}$$

$$\left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right| \leq \frac{L_{(\mathbf{SH5})}\ln(n)}{\sqrt{D_m}}\left[\ell\left(s, s_m\right) + \mathbb{E}\left[p_2(m)\right]\right] \tag{33}$$

$$\left|\overline{\delta}(m)\right| \leq \frac{\ell\left(s, s_m\right)}{\sqrt{D_m}} + L_{(\mathbf{SH5})}\frac{\ln(n)}{\sqrt{D_m}}\mathbb{E}\left[p_2(m)\right] \tag{31}$$

From Prop. 9 (for $\widetilde{p_1}$), Prop. 8 (for p_2), Prop. 6 (for $\overline{\delta}(m)$), we have

$$\mathbb{P}\left(\Omega_n\right) \geq 1 - L\sum_{m \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{c_{\mathcal{M}}}n^{-2} \ .$$

For every $m \in \mathcal{M}_n$ such that $D_m \leq L_{c^X_{\mathrm{r},\ell}}\, n\ln(n)^{-1}$, $(\mathbf{Ar}^{\mathbf{X}}_{\ell})$ implies that $B_n(m) \geq L^{-1}\ln(n) \geq 1$. As a consequence, on Ω_n, if $\ln(n)^7 \leq D_m \leq L_{c^X_{\mathrm{r},\ell}}\, n\ln(n)^{-1}$:

$$\begin{aligned}\max\left\{\left|\widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m)\right]\right|, \left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right|, \left|\overline{\delta}(m)\right|\right\} \\ \leq \frac{L_{(\mathbf{SH5})}\mathbb{E}\left[\ell\left(s, s_m\right) + p_2(m)\right]}{\ln(n)}\end{aligned}$$

Using (37) (in Prop. 10) and the fact that $B_n(m) \geq L^{-1}\ln(n)$,

$$\frac{(c_1+c_2)\left(1-\widetilde{\delta_n}\right)}{2} \leq \mathbb{E}\left[\mathrm{pen}(m)\right] \leq \frac{(C_1+C_2)\left(1+\widetilde{\delta_n}\right)}{2}\mathbb{E}\left[\widetilde{p_1}(m) + p_2(m)\right]$$

with $0 \leq \widetilde{\delta_n} \leq L\ln(n)^{-1/4}$. We deduce: if $n \geq L_{(\mathbf{SH5})}$, for every $m \in \mathcal{M}_n$ such that $\ln(n)^7 \leq D_m \leq L_{c^X_{\mathrm{r},\ell}}\, n\ln(n)^{-1}$, on Ω_n,

$$\begin{aligned}\left[\left(c_1+c_2-2\right)_- - \frac{L_{(\mathbf{SH5})}}{\ln(n)^{1/4}}\right]p_1(m) \leq \left(\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}\right)(m) \\ \leq \left[\left(C_1+C_2-2\right)_+ + \frac{L_{(\mathbf{SH5})}}{\ln(n)^{1/4}}\right]p_1(m) \ .\end{aligned}$$

We need to assume that n is large enough in order to upper bound $\mathbb{E}\left[p_2(m)\right]$ in terms of $p_1(m)$, since we only have

$$p_1(m) \geq \left[1 - \frac{L_{(\mathbf{SH5})}}{\ln(n)^{1/4}}\right]_+ \mathbb{E}\left[p_2(m)\right]$$

in general.

Combined with (25), this gives: if $n \geq L_{(\mathbf{SH5})}$,

$$\ell\,(s,\widehat{s}_{\widehat{m}})\,\mathbb{1}_{\ln(n)^5 \leq D_{\widehat{m}} \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \leq \left[\frac{1+(C_1+C_2-2)_+}{(c_1+c_2-1)\wedge 1} + \frac{L_{(\mathbf{SH5})}}{\ln(n)^{1/4}}\right] \times \inf_{m\in\mathcal{M}_n \text{ s.t. } \ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}},c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \{\ell\,(s,\widehat{s}_m)\}\ . \tag{26}$$

Define the oracle model $m^\star \in \arg\min\{\ell\,(s,\widehat{s}_m)\}$. We prove below that for any $c>0$ and $\alpha > (1-\beta_+)_+/2$, if $n \geq L_{(\mathbf{SH5}),c,\alpha}$, then, on Ω_n:

$$\ln(n)^7 \leq D_{\widehat{m}} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1} \tag{27}$$

$$\ln(n)^7 \leq D_{m^\star} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1}\ . \tag{28}$$

The result follows since $L_{(\mathbf{SH5})}\ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ for $n \geq L_{(\mathbf{SH5})}$. We finally remove the condition $n \geq n_0 = L_{(\mathbf{SH5})}$ by choosing $K_3 = L_{(\mathbf{SH5})}$ such that $K_3 n_0^{-2} \geq 1$.

Proof of (27) By definition, $\widehat{m}$ minimizes $\mathrm{crit}(m)$ over $\mathcal{M}_n$. It thus also minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = \ell\,(s,s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $\mathcal{M}_n$.

1. Lower bound on $\mathrm{crit}'(m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. We then have

$$\begin{aligned}
\ell\,(s,s_m) &\geq C_-\,(\ln(n))^{-7\beta_-} \qquad \text{from } (\mathbf{Ap})\\
\mathrm{pen}(m) &\geq 0\\
p_2(m) &\leq L_{(\mathbf{SH5})}\sqrt{\frac{\ln(n)}{n}} + L_{(\mathbf{SH5})}\frac{D_m}{n} \leq L_{(\mathbf{SH5})}\sqrt{\frac{\ln(n)}{n}} \qquad \text{from (32)}
\end{aligned}$$

   and from (31) (in Prop. 6),

$$\overline{\delta}(m) \geq -L_A\sqrt{\frac{\ell\,(s,s_m)\ln(n)}{n}} + L_A\frac{\ln(n)}{n} \geq -L_A\sqrt{\frac{\ln(n)}{n}}\ .$$

   We then have

$$\mathrm{crit}'(m) \geq L_{(\mathbf{SH5})}\,(\ln(n))^{-L_{\beta_-}}\ .$$

2. Lower bound for large models: let $m \in \mathcal{M}_n$ such that $D_m \geq n^{1/2+\alpha}$. From (22) and (32) (in Prop. 8),

$$\begin{aligned}
\mathrm{pen}(m) - p_2(m) &\geq (c_2-1)\,\mathbb{E}\,[p_2(m)] - L_A\sqrt{\frac{\ln(n)}{n}}\\
&\geq \frac{(c_2-1)\sigma_{\min}^2 D_m}{n} - L_A\sqrt{\frac{\ln(n)}{n}}
\end{aligned}$$

and from (29),

$$\overline{\delta}(m) \geq -L_{(\mathbf{SH5})}\sqrt{\frac{\ln(n)}{n}} \ .$$

Hence, if $D_m \geq n^{1/2+\alpha}$ and $n \geq L_{(\mathbf{SH5}),\alpha}$

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) + \overline{\delta}(m) - p_2(m) \geq L_{(\mathbf{SH5}),\alpha} n^{-1/2+\alpha} \ .$$

3. There exists a better model for $\mathrm{crit}(m)$: from (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n}$. If moreover $n \geq L_{c_{\mathrm{rich}},\alpha}$, then

$$\ln(n)^7 \leq \sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}}\sqrt{n} \leq n^{1/2+\alpha} \ .$$

By (38) in Lemma 11, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.
Using (**Ap**),

$$\ell\left(s, s_{m_0}\right) \leq C_+ c_{\mathrm{rich}}^{\beta_+} n^{-\beta_+/2}$$

so that, when $n \geq L_{(\mathbf{SH5})}$,

$$\begin{aligned}\mathrm{crit}'(m_0) &\leq \ell\left(s, s_{m_0}\right) + \left|\overline{\delta}(m)\right| + \mathrm{pen}(m) \\ &\leq L_{(\mathbf{SH5})}\left(n^{-\beta_+/2} + n^{-1/2}\right) \ .\end{aligned}$$

If $n \geq L_{(\mathbf{SH5}),\alpha}$, this upper bound is smaller than the previous lower bounds for small and large models.

Proof of (28) Recall that $m^\star$ minimizes $\ell\left(s, \widehat{s}_m\right) = \ell\left(s, s_m\right) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $\ell\left(s, \widehat{s}_m\right) = \infty$ if $A_n(m) = 0$.

1. Lower bound on $\ell\left(s, \widehat{s}_m\right)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < \left(\ln(n)\right)^7$. From (**Ap**), we have

$$\ell\left(s, \widehat{s}_m\right) \geq \ell\left(s, s_m\right) \geq C_- \left(\ln(n)\right)^{-7\beta_-} \ .$$

2. Lower bound on $\ell\left(s, \widehat{s}_m\right)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > n^{1/2+\alpha}$. From (36), for $n \geq L_{(\mathbf{SH5}),\alpha}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)\left(c_{\mathrm{r},\ell}^X\right)^{-1}\ln(n)} - \frac{L_{(\mathbf{SH5}),\alpha}}{n^{1/4}}\right)\mathbb{E}\left[\widetilde{p_2}(m)\right]$$
$$\text{so that} \quad \ell\left(s, \widehat{s}_m\right) \geq \widetilde{p_1}(m) \geq L_{(\mathbf{SH5}),\alpha} n^{-1/2+\alpha} \ .$$

3. There exists a better model for $\ell\left(s, \widehat{s}_m\right)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of (27) and assume that $n \geq L_{c_{\mathrm{rich}},\alpha}$. Then,

$$p_1(m_0) \leq L_{(\mathbf{SH5})}\mathbb{E}\left[p_2(m)\right] \leq L_{(\mathbf{SH5})} n^{-1/2}$$

and the arguments of the previous proof show that

$$\ell\left(s, \widehat{s}_{m_0}\right) \leq L_{(\mathbf{SH5})}\left(n^{-\beta_+/2} + n^{-1/2}\right)$$

which is smaller than the previous upper bounds for $n \geq L_{(\mathbf{SH5}),\alpha}$.

Classical oracle inequality Let Ω_n be the event on which (23) holds true. Then,

$$\begin{aligned}\mathbb{E}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)\right] &= \mathbb{E}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)\mathbb{1}_{\Omega_n^c}\right] \\ &\leq \left[2\eta - 1 + \epsilon_n\right]\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s,\widehat{s}_m\right)\right\}\right] + A^2K_3\mathbb{P}\left(\Omega_n^c\right)\end{aligned}$$

which proves (24). ■

B.4 Proof of Thm. 1

Similarly to the proof of Thm. 5, we consider the event Ω_n', of probability at least $1-L_{c_{\mathcal{M}}}n^{-2}$, on which:

- for every $m \in \mathcal{M}_n$, (6) (for pen), (36) (for $\widetilde{p_1}$), (32)–(33) (for p_2, with $x = \gamma\ln(n)$ and $\theta = \sqrt{\ln(n)/n}$) and (29)–(31) (for $\overline{\delta}$, with $x = \gamma\ln(n)$ and $\eta = \sqrt{\ln(n)/n}$) hold true.
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$, (34) and (35) hold (for $\widetilde{p_1}$).

Lower bound on $D_{\widehat{m}}$ By definition, $\widehat{m}$ minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = \ell\left(s,s_m\right) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$. As in the proof of Thm. 5, we define $c = L_{c_{\mathrm{r},\ell}^X} > 0$ such that for every model of dimension $D_m \leq cn\ln(n)^{-1}$, $B_n(m) \geq L^{-1}\ln(n) \geq 1$. Let $d < 1$ to be chosen later.

1. Lower bound on $\mathrm{crit}'(m)$ for "small" models: assume that $m \in \mathcal{M}_n$ and $D_m \leq dcn\ln(n)^{-1}$. Then, $\ell\left(s,s_m\right) + \mathrm{pen}(m) \geq 0$ and from (29),

   $$\overline{\delta}(m) \geq -L_A\sqrt{\frac{\ln(n)}{n}} \ .$$

   If $D_m \geq \ln(n)^4$, (33) implies that

   $$p_2(m) \leq \left(1 + \frac{L_{(\mathbf{SH1})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m)\right] \leq \frac{L_{(\mathbf{SH1})}D_m}{n} \leq \frac{cdL_{(\mathbf{SH1})}}{\ln(n)} \ .$$

   On the other hand, if $D_m < \ln(n)^4$, (32) implies that

   $$p_2(m) \leq L_{(\mathbf{SH1})}\sqrt{\frac{\ln(n)}{n}} \ .$$

   We then have

   $$\mathrm{crit}'(m) \geq -dL_{(\mathbf{SH1})}\left(\ln(n)\right)^{-1} \ .$$

2. There exists a better model for $\mathrm{crit}(m)$: let $m_1 \in \mathcal{M}_n$ such that

   $$\ln(n)^4 \leq \frac{cdn}{c_{\mathrm{rich}}\ln(n)} \leq D_{m_1} \leq \frac{cn}{\ln(n)} \leq n \ .$$

From (**P2**), this is possible as soon as $n \geq L_{c_{\text{rich}},c,d}$. By (38) in Lemma 11, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.
We then have

$$\begin{aligned}
\ell(s, s_{m_1}) &\leq L_{(\mathbf{SH1}),c} \ln(n)^{\beta_+} n^{-\beta_+} \qquad \text{by } (\mathbf{Ap}) \\
p_2(m_1) &\geq \left(1 - \frac{L_{(\mathbf{SH1})}}{\ln(n)}\right) \mathbb{E}\left[p_2(m_1)\right] \qquad \text{by (33)} \\
\operatorname{pen}(m_1) &\leq K \mathbb{E}\left[p_2(m_1)\right] \qquad \text{by (6)} \\
\left|\overline{\delta}(m_1)\right| &\leq L_A \sqrt{\frac{\ln(n)}{n}} \qquad \text{by (29)}
\end{aligned}$$

so that

$$\begin{aligned}
\operatorname{crit}'(m_1) &\leq L_{(\mathbf{SH1}),c} \ln(n)^{\beta_+} n^{-\beta_+} + \left(K - 1 + \frac{L_{(\mathbf{SH1})}}{\ln(n)}\right) \mathbb{E}\left[p_2(m_1)\right] + L_A \sqrt{\frac{\ln(n)}{n}} \\
&\leq \frac{(K - 1 + L_{(\mathbf{SH1})}(\ln(n))^{-1}) \sigma_{\min}^2 c}{2 \ln(n)}
\end{aligned}$$

if $n \geq L_{(\mathbf{SH1}),c}$.

We now choose d such that the constant $dL_{(\mathbf{SH1})}$ appearing in the lower bound on $\operatorname{crit}'(m)$ for "small" models is smaller than $(1 - K - L_{(\mathbf{SH1})}(\ln(n))^{-1})\sigma_{\min}^2 c/2$, i.e. $d \leq L_{(\mathbf{SH1}),c}$. Then, we assume that $n \geq n_0 = L_{(\mathbf{SH1}),c,d} = L_{(\mathbf{SH1})}$. Finally, we remove this condition as before by enlarging K_1.

Risk of $D_{\widehat{m}}$ The proof of (8) is quite similar to the one of (28). First, for every model $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$ and $D_m \geq K_2 n \ln(n)^{-1}$, we have

$$\ell(s, \widehat{s}_m) \geq \widetilde{p_1}(m) \geq L_{(\mathbf{SH1})} K_2 \ln(n)^{-2} \qquad \text{by (36)} \ .$$

Then, the model $m_0 \in \mathcal{M}_n$ defined previously satisfies $A_n(m) \geq 1$, and

$$\ell(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{SH1})} \left(n^{-\beta_+/2} + n^{-1/2}\right) \ .$$

If $n \geq L_{(\mathbf{SH1})}$, the ratio between these two bounds is larger than $\ln(n)$, so that (8) holds. ■

B.5 Concentration inequalities used in the main proofs

We do not always assume in this section that models are made of histograms, but only that they are bounded by some finite A. First, we can control $\overline{\delta}(m)$ with general models and bounded data.

Proposition 6 *Assume that $\|Y\|_\infty \leq A < \infty$. Then for all $x \geq 0$, on an event of probability at least $1 - 2e^{-x}$:*

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta \ell(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right) \frac{A^2 x}{n} \ . \tag{29}$$

If moreover

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[p_2(m)\right]}{D_m} > 0 \ , \tag{30}$$

on the same event,

$$\left|\overline{\delta}(m)\right| \leq \frac{\ell\left(s, s_m\right)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q_m^{(p)}}\frac{\mathbb{E}[p_2(m)]}{\sqrt{D_m}}x \ . \tag{31}$$

Remark 7 *In the histogram case,*

$$Q_m^{(p)} = \frac{1}{D_m}\sum_{\lambda\in\Lambda_m}\sigma_\lambda^2 \geq \left(\sigma_{\min}\right)^2 > 0 \ .$$

Then, we derive a concentration inequality for $p_2(m)$ in the histogram case from a general result of Boucheron and Massart (2008) (Thm. 2.2 in a preliminary version).

Proposition 8 *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\|Y\|_\infty \leq A$ and define $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$.*

Then, for every $x \geq 0$, there exists an event of probability at least $1 - e^{1-x}$ on which for every $\theta \in (0;1)$,

$$\left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right| \leq L\left[\theta\ell\left(s, s_m\right) + \frac{A^2\sqrt{D_m}\sqrt{x}}{n} + \frac{A^2 x}{\theta n}\right] \tag{32}$$

for some absolute constant L. If moreover $\sigma(X) \geq \sigma_{\min} > 0$ a.s., we have on the same event:

$$\left|p_2(m) - \mathbb{E}\left[p_2(m)\right]\right| \leq \frac{L}{\sqrt{D_m}}\left[\ell\left(s, s_m\right) + \frac{A^2\mathbb{E}\left[p_2(m)\right]}{\sigma_{\min}^2}\left(\sqrt{x} + x\right)\right] \ . \tag{33}$$

Finally, we recall a concentration inequality for $p_1(m)$ that comes from (Arlot, 2008b). Its proof is particular to the histogram case.

Proposition 9 (Prop. 9, Arlot (2008b)) *Let $\gamma > 0$ and S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\|Y\|_\infty \leq A < \infty$, $\sigma(X) \geq \sigma_{\min} > 0$ a.s. and $\min_{\lambda\in\Lambda_m}\left\{np_\lambda\right\} \geq B_n > 0$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{A,\sigma_{\min},\gamma}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{34}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{A,\sigma_{\min},\gamma}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \ . \tag{35}$$

If we only have a lower bound $B_n > 0$, then, with probability at least $1 - Ln^{-\gamma}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n^{-1}\ln(n)} - \frac{L_{A,\sigma_{\min},\gamma}\ln(n)^2}{\sqrt{D_m}}\right)\mathbb{E}\left[p_2(m)\right] \ . \tag{36}$$

B.6 Additional results needed

A crucial result in the proofs of Thm. 5 and 1 is that $p_1(m)$ and $p_2(m)$ are close in expectation. This comes from (Arlot, 2007) (Sect. 5.7.2).

Proposition 10 (Lemma 7, Arlot (2008b)) *Let S_m be a model of histograms adapted to some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\min_{\lambda\in\Lambda_m}\{np_\lambda\}\geq B>0$. Then,*

$$\begin{aligned}\left(1-e^{-B}\right)^2\mathbb{E}\left[p_2(m)\right]&\leq\mathbb{E}\left[\widetilde{p_1}(m)\right]\\&\leq\left[2\wedge\left(1+5.1\times B^{-1/4}\right)+\left(B\vee 1\right)e^{-\left(B\vee 1\right)}\right]\mathbb{E}\left[p_2(m)\right]\quad.\end{aligned}\tag{37}$$

Finally, we need the following technical lemma in the proof of the main theorems.

Lemma 11 *Let $(p_\lambda)_{\lambda\in\Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ a multinomial vector of parameters $(n;(p_\lambda)_{\lambda\in\Lambda_m})$. Then, for all $\gamma>0$,*

$$\min_{\lambda\in\Lambda_m}\left\{n\widehat{p}_\lambda\right\}\geq\frac{\min_{\lambda\in\Lambda_m}\left\{np_\lambda\right\}}{2}-2(\gamma+1)\ln(n)\tag{38}$$

with probability at least $1-2n^{-\gamma}$.

Proof By Bernstein inequality (Massart (2007), Prop. 2.9), for all $\lambda\in\Lambda_m$,

$$\mathbb{P}\left(n\widehat{p}_\lambda\geq(1-\theta)np_\lambda-\sqrt{2npx}-\frac{x}{3}\right)\geq 1-e^{-x}\quad.$$

Take $x=(\gamma+1)\ln(n)$ above, and remark that $\sqrt{2npx}\leq\frac{np}{2}+x$. The union bound gives the result since $\mathrm{Card}(\Lambda_m)\leq n$. ∎

B.7 Proof of Prop. 6

Since $\|Y\|_\infty\leq A$, we have $\|s\|_\infty\leq A$ and $\|s_m\|_\infty\leq A$. In fact, everything happens as if $S_m\cup\{s\}$ was bounded by A in L^∞.

We have

$$\overline{\delta}(m)=\frac{1}{n}\sum_{i=1}^n\left(\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))-\mathbb{E}\left[\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))\right]\right)$$

and assumptions of Bernstein inequality (Massart (2007), Prop. 2.9) are fulfilled with

$$c=\frac{8A^2}{3n}\quad\text{and}\quad v=\frac{8A^2\ell\left(s,s_m\right)}{n}$$

since

$$\left\|\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))-\mathbb{E}\left[\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))\right]\right\|_\infty\leq 8A^2$$

and

$$\begin{aligned}\mathrm{var}\left(\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))\right)&\leq\mathbb{E}\left[\left(\gamma(s_m,(X_i,Y_i))-\gamma(s,(X_i,Y_i))\right)^2\right]\\&\leq 8A^2\ell\left(s,s_m\right)\end{aligned}\tag{39}$$

because $\|s_m - s\|_\infty \leq 2A$ and

$$(\gamma(t,\cdot) - \gamma(s,\cdot))^2 = (t(X) - s(X))^2 \, (2(Y - s(X)) - t(X) + s(X))^2$$
$$\text{and} \quad \mathbb{E}\left[(Y - s(X))^2 \,\middle|\, X\right] \leq \frac{(2A)^2}{4} = A \ .$$

We obtain that, with probability at least $1 - 2e^{-x}$,

$$\left|\overline{\delta}(m)\right| \leq \sqrt{2vx} + c = \sqrt{\frac{16A^2 \ell\,(s, s_m)\, x}{n}} + \frac{8A^2 x}{3n}$$

and (29) follows since $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$. Taking $\eta = D_m^{-1/2} \leq 1$ and using $Q_m^{(p)}$ defined by (30), we deduce (31). ■

B.8 Proof of Prop. 8

We apply here a result by Boucheron and Massart (2008) (Thm. 2.2 in a preliminary version), in which it is only assumed that γ takes its values in $[0;1]$. This is satisfied when $\|Y\|_\infty \leq A = 1/2$. When $A \neq 1/2$, we apply this result to $(2A)^{-1}Y$ and recover the general result by homogeneity.

First, we recall this result in the bounded least-square regression framework. For every $t : \mathcal{X} \mapsto \mathbb{R}$ and $\epsilon > 0$, we define

$$d^2(s,t) = 2\ell\,(s,t) \qquad \text{and} \qquad w(\epsilon) = \sqrt{2}\epsilon \ .$$

Let ϕ_m belong to the class of nondecreasing and continuous functions $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $x \mapsto f(x)/x$ is nonincreasing on $(0; +\infty)$ and $f(1) \geq 1$. Assume that for every $u \in S_m$ and $\sigma > 0$ such that $\phi_m(\sigma) \leq \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E}\left[\sup_{t \in S_m,\, d(u,t) \leq \sigma} |\overline{\gamma}_n(u) - \overline{\gamma}_n(t)|\right] \leq \phi_m(\sigma) \ . \tag{40}$$

Let $\varepsilon_{\star,m}$ be the unique positive solution of the equation

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ .$$

Then, there exists some absolute constant L such that for every real number $q \geq 2$ one has

$$\|p_2(m) - \mathbb{E}[p_2(m)]\|_q \leq \frac{L}{\sqrt{n}}\left[\sqrt{2q}\left(\sqrt{\ell\,(s, s_m)} \vee \varepsilon_{\star,m}\right) + q\frac{2}{\sqrt{n}}\right] \ . \tag{41}$$

For every model S_m of histograms, of dimension D_m as a vector space, we can take

$$\phi_m(\sigma) = 3\sqrt{2}\sqrt{D_m} \times \sigma \qquad \text{in (40).} \tag{42}$$

The proof of this statement is made below. Then, $\varepsilon_{\star,m} = 6\sqrt{D_m}n^{-1/2}$.

Combining (41) with the classical link between moments and concentration (for instance Lemma 8.9 of Arlot (2007)), the first result follows. The second result is obtained by taking $\theta = D_m^{-1/2}$, as in Prop. 6.

Proof of (42) Let $u \in S_m$ and $d(u,t) = \sqrt{2}\,\|u(X) - t(X)\|_2$ for every $t : \mathcal{X} \mapsto \mathbb{R}$. Define $\psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ by

$$\psi(\sigma) = \mathbb{E}\left[\sup_{d(u,t)\leq\sigma,\, t\in S_m} |(P_n - P)(\gamma(u,\cdot) - \gamma(t,\cdot))|\right] \ .$$

We are looking for some nondecreasing and continuous function $\phi_m : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\phi_m(x)/x$ is nonincreasing, $\phi_m(1) \geq 1$ and for every $u \in S_m$,

$$\forall \sigma > 0 \quad \text{such that} \quad \phi_m(\sigma) \leq \sqrt{n}\sigma^2 \ , \qquad \phi_m(\sigma) \geq \sqrt{n}\psi(\sigma) \ .$$

We first look at a general upperbound on ψ.

Assume that $u = s_m$. If this is not the case, the triangular inequality shows that $\psi_{\text{general } u} \leq 2\psi_{u=s_m}$. Let us write

$$t = \sum_{\lambda\in\Lambda_m} t_\lambda \mathbb{1}_{I_\lambda} \qquad u = s_m = \sum_{\lambda\in\Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \ .$$

Computation of $P(\gamma(t,\cdot) - \gamma(s_m,\cdot))$ for some general $t \in S_m$:

$$\begin{aligned}
P(\gamma(t,\cdot) - \gamma(s_m,\cdot)) &= \mathbb{E}\left[(t(X) - Y)^2 - (s_m(X) - Y)^2\right] \\
&= \mathbb{E}\left[(t(X) - s_m(X))^2\right] + 2\mathbb{E}\left[(t(X) - s_m(X))(s_m(X) - s(X))\right] \\
&= \mathbb{E}\left[(t(X) - s_m(X))^2\right] \\
&= \sum_{\lambda\in\Lambda_m} p_\lambda (t_\lambda - \beta_\lambda)^2
\end{aligned}$$

since for every $\lambda \in \Lambda_m$, $\mathbb{E}\left[s(X) \mid \ X \in I_\lambda\right] = \beta_\lambda$.

Computation of $P_n(\gamma(t,\cdot) - \gamma(s_m,\cdot))$ for some general $t \in S_m$: with $\eta_i = Y_i - s_m(X_i)$, we have

$$\begin{aligned}
P(\gamma(t,\cdot) - \gamma(s_m,\cdot)) &= \frac{1}{n}\sum_{i=1}^n \left[(t(X_i) - Y_i)^2 - (u(X_i) - Y_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n (t(X_i) - u(X_i))^2 - \frac{2}{n}\sum_{i=1}^n \left[(t(X_i) - u(X_i))\eta_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)^2 \mathbb{1}_{X_i\in I_\lambda} - \frac{2}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda - u_\lambda)\mathbb{1}_{X_i\in I_\lambda}\eta_i \ .
\end{aligned}$$

Back to $(P_n - P)$ We sum the two inequalities above and use the triangular inequality:

$$\begin{aligned}|(P_n - P)(\gamma(t,\cdot) - \gamma(u,\cdot))| &\leq \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{\lambda \in \Lambda_m} (t_\lambda - u_\lambda)^2 (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda) \right| \\ &\quad + \left| \frac{2}{n} \sum_{i=1}^{n} \sum_{\lambda \in \Lambda_m} (t_\lambda - u_\lambda) \mathbb{1}_{X_i \in I_\lambda} \eta_i \right| \\ &\leq \frac{2A}{n} \sum_{\lambda \in \Lambda_m} \left[(\sqrt{p_\lambda}\, |t_\lambda - u_\lambda|) \, \frac{|\sum_{i=1}^{n} (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)|}{\sqrt{p_\lambda}} \right] \\ &\quad + \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left[(\sqrt{p_\lambda}\, |t_\lambda - u_\lambda|) \, \frac{|\sum_{i=1}^{n} \mathbb{1}_{X_i \in I_\lambda} \eta_i|}{\sqrt{p_\lambda}} \right]\end{aligned}$$

since $|t_\lambda - u_\lambda| \leq 2A$ for every $t \in S_m$.

We now assume that $d(u,t) \leq \sigma$ for some $\sigma > 0$, *i.e.*

$$d(u,t)^2 = 2 \sum_{\lambda \in \Lambda_m} p_\lambda (t_\lambda - u_\lambda)^2 \leq \sigma^2 \ .$$

From Cauchy-Schwarz inequality, we obtain for every $t \in S_m$ such that $d(u,t) \leq \sigma$

$$\begin{aligned}|(P_n - P)(\gamma(t,\cdot) - \gamma(u,\cdot))| &\leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^{n} (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda}} \\ &\quad + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda}}\end{aligned}$$

Back to ψ The upper bound above does not depend on t, so that the left-hand side of the inequality can be replaced by a supremum over $\{\, t \in S_m \text{ s.t. } d(u,t) \leq \sigma \,\}$. Taking expectations and using Jensen's inequality ($\sqrt{\cdot}$ being concave), we obtain an upper bound on ψ:

$$\psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^{n} (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda} \right]} + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda} \right]} \tag{43}$$

For every $\lambda \in \Lambda_m$, we have

$$\mathbb{E}\left(\sum_{i=1}^{n} (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda) \right)^2 = \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}_{X_i \in I_\lambda} - p_\lambda \right)^2 = n p_\lambda \left(1 - p_\lambda \right) \tag{44}$$

which simplifies the first term. For the second term, notice that

$$\begin{aligned}&\forall i \neq j, \qquad \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{1}_{X_j \in I_\lambda} \eta_i \eta_j \right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i \right] \mathbb{E}\left[\mathbb{1}_{X_j \in I_\lambda} \eta_j \right] \\ &\quad \text{and} \quad \forall i, \qquad \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i \right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{E}\left[\eta_i \mid \mathbb{1}_{X_i \in I_\lambda} \right] \right] = 0\end{aligned}$$

since η_i is centered conditionally to $\mathbb{1}_{X_i \in I_\lambda}$. Then,

$$\mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2 = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i^2\right] \leq n p_\lambda \left\|\eta\right\|_\infty^2 \leq n p_\lambda (2A)^2 \ . \tag{45}$$

Combining (43) with (44) and (45), we deduce that

$$\psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2}\sqrt{n}} \sqrt{D_m - 1} + \frac{2\sqrt{2}A\sigma}{\sqrt{n}} \sqrt{D_m} \leq 3A\sqrt{2} \frac{\sqrt{D_m}}{\sqrt{n}} \times \sigma \ .$$

As already noticed, we have to multiply this bound by 2 so that it is valid for every $u \in S_m$ and not only $u = s_m$.

The resulting upper bound (multiplied by $\sqrt{n}$) has all the desired properties for ϕ_m since $6A\sqrt{2}\sqrt{D_m} = 3\sqrt{2D_m} \geq 1$. The result follows. ■

References

Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970. ISSN 0020-3157.

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974. ISSN 0040-1706.

Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

Sylvain Arlot. Model selection by resampling penalization, March 2008a. hal-00262478.

Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008b. arXiv:0802.0566.

Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000. ISSN 0178-8051.

Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002. ISSN 1292-8100.

Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with unknown variance. Preprint. Arxiv:math.ST/0701250, January 2007.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051.

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005. ISSN 0090-5364.

Jean-Patrick Baudry. Clustering through model selection criteria. Poster session at One Day Statistical Workshop in Lisieux. `http://www.math.u-psud.fr/~baudry`, June 2007.

Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. ISSN 1435-9855.

Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. ISSN 0178-8051.

Stéphane Boucheron and Pascal Massart. A poor man's wilks phenomenon. Personal communication, March 2008.

Prabir Burman. Estimation of equifrequency histograms. *Statist. Probab. Lett.*, 56(3): 227–238, 2002. ISSN 0167-7152.

Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983. ISSN 0162-1459.

Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

Edward I. George and Dean P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000. ISSN 0006-3444.

Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989. ISSN 0006-3444.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001. ISSN 0018-9448.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006. ISSN 0090-5364.

Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

Vincent Lepez. *Some estimation problems related to oil reserves*. PhD thesis, University Paris XI, 2002.

Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987. ISSN 0090-5364.

Fernando Lozano. Model selection using rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany*. ICSC Academic Press, 2000.

Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Cathy Maugis and Bertrand Michel. A nonasymptotic penalized criterion for gaussian mixture model selection. a variable selection and clustering problems. In preparation, September 2007.

B. T. Polyak and A. B. Tsybakov. Asymptotic optimality of the C_p-test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, 35(2):305–317, 1990. ISSN 0040-361X.

Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981. ISSN 0006-3444.

Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. ISSN 0035-9246. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

N. Sugiura. Further analysis of the data by akaike's information criterion and the finite corrections. *Comm. Statist. A—Theory Methods*, 7(1):13–26, 1978.

Nicolas Verzelen. Model selection for graphical models. In preparation, September 2007.

Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris XI, December 2007.