

Data-driven Calibration of Penalties for Least-Squares Regression

Sylvain Arlot

SYLVAIN.ARLOT@MATH.U-PSUD.FR

Pascal Massart

PASCAL.MASSART@MATH.U-PSUD.FR

Univ Paris-Sud, UMR 8628

Laboratoire de Mathématiques

Orsay, F-91405 ; CNRS, Orsay, F-91405 ;

INRIA Saclay, Projet Select

Editor: John Lafferty

Abstract

Penalization procedures often suffer from their dependence on multiplying factors, whose optimal values are either unknown or hard to estimate from data. We propose a completely data-driven calibration algorithm for these parameters in the least-squares regression framework, without assuming a particular shape for the penalty. Our algorithm relies on the concept of minimal penalty, recently introduced by Birgé and Massart (2007) in the context of penalized least squares for Gaussian homoscedastic regression. On the positive side, the minimal penalty can be evaluated from the data themselves, leading to a data-driven estimation of an optimal penalty which can be used in practice; on the negative side, their approach heavily relies on the homoscedastic Gaussian nature of their stochastic framework.

The purpose of this paper is twofold: stating a more general heuristics for designing a data-driven penalty (the *slope heuristics*) and proving that it works for penalized least-squares regression with a random design, even for heteroscedastic non-Gaussian data. For technical reasons, some exact mathematical results will be proved only for regressogram bin-width selection. This is at least a first step towards further results, since the approach and the method that we use are indeed general.

Keywords: Data-driven Calibration, Non-parametric Regression, Model Selection by Penalization, Heteroscedastic Data, Regressogram

1. Introduction

In the last decades, model selection has received much interest, commonly through penalization. In short, penalization chooses the model minimizing the sum of the empirical risk (how well the algorithm fits the data) and of some measure of complexity of the model (called penalty); see FPE (Akaike, 1970), AIC (Akaike, 1973), Mallows' C_p or C_L (Mallows, 1973). Many other penalization procedures have been proposed since, among which Rademacher complexities (Koltchinskii, 2001; Bartlett et al., 2002), local Rademacher complexities (Bartlett et al., 2005; Koltchinskii, 2006), bootstrap penalties (Efron, 1983), resampling and V -fold penalties (Arlot, 2008b,c).

Model selection can target two different goals. On the one hand, a procedure is *efficient* (or asymptotically optimal) when its quadratic risk is asymptotically equivalent to the risk

of the oracle. On the other hand, a procedure is *consistent* when it chooses the smallest true model asymptotically with probability one. This paper deals with *efficient* procedures, without assuming the existence of a true model.

A huge amount of literature exists about efficiency. First Mallows' C_p , Akaike's FPE and AIC are asymptotically optimal, as proved by Shibata (1981) for Gaussian errors, by Li (1987) under suitable moment assumptions on the errors, and by Polyak and Tsybakov (1990) under sharper moment conditions, in the Fourier case. Non-asymptotic oracle inequalities (with some leading constant $C > 1$) have been obtained by Barron et al. (1999) and by Birgé and Massart (2001) in the Gaussian case, and by Baraud (2000, 2002) under some moment assumptions on the errors. In the Gaussian case, non-asymptotic oracle inequalities with leading constant C_n tending to 1 when n tends to infinity have been obtained by Birgé and Massart (2007).

However, from the practical point of view, both AIC and Mallows' C_p still present serious drawbacks. On the one hand, AIC relies on a strong asymptotic assumption, so that for small sample sizes, the optimal multiplying factor can be quite different from one. Therefore, 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 , assumed to be constant. When real data are involved, σ^2 has to be estimated separately and independently from any model, which is a difficult task. Moreover, the best estimator of σ^2 (say, with respect to the quadratic error) quite unlikely leads to the most efficient model selection procedure. Contrary to Mallows' C_p , the data-dependent calibration rule defined in this article is not a "plug-in" method; it focuses directly on efficiency, which can improve significantly the performance of the model selection procedure.

Existing penalization procedures present similar or stronger drawbacks than AIC and Mallows' C_p , often because of a gap between theory and practice. For instance, oracle inequalities have only been proved for (global) Rademacher penalties multiplied by a factor two (Koltchinskii, 2001), while they are used without this factor (Lozano, 2000). As proved by Arlot (2007, Chapter 9), this factor is necessary in general. Therefore, the optimal calibration of these penalties is really an issue. The calibration problem is even harder for local Rademacher complexities: theoretical results hold only with large calibration constants, particularly the multiplying factor, and no optimal values are known. One of the purposes of this paper is to address the issue of optimizing the multiplying factor for general-shape penalties.

Few automatic calibration algorithms are available. The most popular ones are certainly cross-validation methods (Allen, 1974; Stone, 1974), in particular V -fold cross-validation (Geisser, 1975), because these are general-purpose methods, relying on a widely valid heuristics. However, their computational cost can be high. For instance, V -fold cross-validation requires the entire model selection procedure to be performed V times for each candidate value of the constant to be calibrated. For penalties proportional to the dimension of the models, such as Mallows' C_p , alternative calibration procedures have been proposed by George and Foster (2000) and by Shen and Ye (2002).

A completely different approach has been proposed by Birgé and Massart (2007) for calibrating dimensionality-based penalties. Since this article extends their approach to a

much wider range of applications, let us briefly recall their main results. In Gaussian homoscedastic regression with a fixed design, assume that each model is a finite-dimensional vector space. Consider the penalty $\text{pen}(m) = KD_m$, where D_m is the dimension of the model m and $K > 0$ is a positive constant, to be calibrated. First, there exists a *minimal* constant K_{\min} , such that the ratio between the quadratic risk of the chosen estimator and the quadratic risk of the oracle is asymptotically infinite if $K < K_{\min}$, and finite if $K > K_{\min}$. Second, when $K = K^* := 2K_{\min}$, the penalty KD_m yields an efficient model selection procedure. In other words, *the optimal penalty is twice the minimal penalty*. This relationship characterizes the “slope heuristics” of Birgé and Massart (2007).

A crucial fact is that the minimal constant K_{\min} can be estimated from the data, since large models are selected if and only if $K < K_{\min}$. This leads to the following strategy for choosing K from the data. For every $K \geq 0$, let $\hat{m}(K)$ be the model selected by minimizing the empirical risk penalized by $\text{pen}(D_m) = KD_m$. First, compute K_{\min} such that $D_{\hat{m}(K)}$ is “huge” for $K < K_{\min}$ and “reasonably small” when $K \geq K_{\min}$; explicit values for “huge” and “small” are proposed in Section 3.3. Second, define $\hat{m} := \hat{m}(2K_{\min})$. Such a method has been successfully applied for multiple change points detection by Lebarbier (2005).

From the theoretical point of view, the issue for understanding and validating this approach is the existence of a minimal penalty. This question has been addressed for Gaussian homoscedastic regression with a fixed design by Birgé and Massart (2001, 2007) when the variance is known, and by Baraud et al. (2007) when the variance is unknown. Non-Gaussian or heteroscedastic data have never been considered. This article contributes to fill this gap in the theoretical understanding of penalization procedures.

The calibration algorithm proposed in this article relies on a generalization of Birgé and Massart’s slope heuristics (Section 2.3). In Section 3, the algorithm is defined in the least-squares regression framework, for general-shape penalties. The shape of the penalty itself can be estimated from the data, as explained in Section 3.4.

The theoretical validation of the algorithm is provided in Section 4, from the *non-asymptotic point of view*. Non-asymptotic means in particular that the collection of models is allowed to depend on n : in practice, it is usual to allow the number of explanatory variables to increase with the number of observations. Considering models with a large number of parameters (for example 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 point of view allows us not to assume that the regression function is described with a small number of parameters.

The existence of minimal penalties for *heteroscedastic regression with a random design* (Theorem 2) is proved in Section 4.3. In Section 4.4, by proving that twice the minimal penalty has some optimality properties (Theorem 3), we extend the so-called slope heuristics to heteroscedastic regression with a random design. Moreover, neither Theorem 2 nor Theorem 3 assume the data to be Gaussian; only mild moment assumptions are required.

For proving Theorems 2 and 3, each model is assumed to be the vector space of piecewise constant functions on some partition of the feature space. This is indeed a restriction, but we conjecture that it is mainly technical, and that the slope heuristics remains valid at least in the general least-squares regression framework. We provide some evidence for this by proving two key concentration inequalities without the restriction to piecewise constant functions. Another argument supporting this conjecture is that recently several simulation

studies have shown that the slope heuristics can be used in several frameworks: mixture models (Maugis and Michel, 2008), clustering (Baudry, 2007), spatial statistics (Verzelen, 2008), estimation of oil reserves (Lepez, 2002) and genomics (Villers, 2007). Although the slope heuristics has not been formally validated in these frameworks, this article is a first step towards such a validation, by proving that the slope heuristics can be applied whatever the shape of the ideal penalty.

This paper is organized as follows. The framework and the slope heuristics are described in Section 2. The resulting algorithm is defined in Section 3. The main theoretical results are stated in Section 4. All the proofs are given in Appendix A.

2. Framework

In this section, we describe the framework and the general slope heuristics.

2.1 Least-squares regression

Suppose we observe some data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$, independent with common distribution P , where the feature space \mathcal{X} is typically a compact set of \mathbb{R}^d . The goal is to predict Y given X , where $(X, Y) \sim P$ is a new data point independent of $(X_i, Y_i)_{1 \leq i \leq n}$. Denoting by s the regression function, that is $s(x) = \mathbb{E}[Y | X = x]$ for every $x \in \mathcal{X}$, we can write

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \quad (1)$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise level and ϵ_i are i.i.d. centered noise terms, possibly dependent on X_i , but with mean 0 and variance 1 conditionally to X_i .

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

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

is the least-squares contrast. The minimizer of $P\gamma(t)$ over the set of all predictors, called Bayes predictor, is the regression function s . Therefore, the excess loss is defined as

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

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

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

with 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-squares estimator since γ is the least-squares contrast.

2.2 Ideal model selection

Let us assume that we are given a family of models $(S_m)_{m \in \mathcal{M}_n}$, hence a family of estimators $(\hat{s}_m)_{m \in \mathcal{M}_n}$ obtained by empirical risk minimization. The model selection problem consists in looking for some data-dependent $\hat{m} \in \mathcal{M}_n$ such that $\ell(s, \hat{s}_{\hat{m}})$ is as small as possible. For instance, it would be convenient to prove some oracle inequality of the form

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

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

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

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

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

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

This quantity is unknown because it depends on the true distribution P . A natural idea is to choose $\text{pen}(m)$ as close as possible to $\text{pen}_{\text{id}}(m)$ for every $m \in \mathcal{M}_n$. We will show below, in a general setting, that when pen is a good estimator of the ideal penalty pen_{id} , then \hat{m} satisfies an oracle inequality with leading constant C close to 1.

By definition of \hat{m} ,

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

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

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

so that

$$\begin{aligned} \text{pen}_{\text{id}}(m) &= p_1(m) + p_2(m) - \delta(m) \\ \text{and} \quad \ell(s, \hat{s}_m) &= P_n \gamma(\hat{s}_m) + p_1(m) + p_2(m) - \delta(m) - P\gamma(s) . \end{aligned}$$

Hence, for every $m \in \mathcal{M}_n$,

$$\ell(s, \hat{s}_{\hat{m}}) + (\text{pen} - \text{pen}_{\text{id}})(\hat{m}) \leq \ell(s, \hat{s}_m) + (\text{pen} - \text{pen}_{\text{id}})(m) . \quad (3)$$

Therefore, in order to derive an oracle inequality from (3), it is sufficient to show that for every $m \in \mathcal{M}_n$, $\text{pen}(m)$ is close to $\text{pen}_{\text{id}}(m)$.

2.3 The slope heuristics

If the penalty is too big, the left-hand side of (3) is larger than $\ell(s, \widehat{s}_{\widehat{m}})$ so that (3) implies an oracle inequality, possibly with large leading constant C . On the contrary, if the penalty is too small, the left-hand side of (3) may become negligible with respect to $\ell(s, \widehat{s}_{\widehat{m}})$ (which would make C explode) or—worse—may be nonpositive. In the latter case, no oracle inequality may be derived from (3). We shall see in the following that $\ell(s, \widehat{s}_{\widehat{m}})$ blows up if and only if the penalty is smaller than some “minimal penalty”.

Let us consider first the case $\text{pen}(m) = p_2(m)$ in (2). Then, $\mathbb{E}[\text{crit}(m)] = \mathbb{E}[P_n \gamma(s_m)] = P \gamma(s_m)$, so that \widehat{m} approximately minimizes its bias. Therefore, \widehat{m} is one of the more complex models, and the risk of $\widehat{s}_{\widehat{m}}$ is large. Let us assume now that $\text{pen}(m) = K p_2(m)$. If $0 < K < 1$, $\text{crit}(m)$ is a decreasing function of the complexity of m , so that \widehat{m} is again one of the more complex models. On the contrary, if $K > 1$, $\text{crit}(m)$ increases with the complexity of m (at least for the largest models), so that \widehat{m} has a small or medium complexity. This argument supports the conjecture that the “minimal amount of penalty” required for the model selection procedure to work is $p_2(m)$.

In many frameworks such as the one of Section 4.1, it turns out that

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

Hence, the ideal penalty $\text{pen}_{\text{id}}(m) \approx p_1(m) + p_2(m)$ is close to $2p_2(m)$. Since $p_2(m)$ is a “minimal penalty”, the optimal penalty is close to twice the minimal penalty:

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

This is the so-called “slope heuristics”, first introduced by Birgé and Massart (2007) in a Gaussian homoscedastic setting. Note that a formal proof of the validity of the slope heuristics has only been given for Gaussian homoscedastic least-squares regression with a fixed design (Birgé and Massart, 2007); up to the best of our knowledge, the present paper yields the second theoretical result on the slope heuristics.

This heuristics has some applications because the minimal penalty can be estimated from the data. Indeed, when the penalty smaller than pen_{min} , the selected model \widehat{m} is among the more complex. On the contrary, when the penalty is larger than pen_{min} , the complexity of \widehat{m} is much smaller. This leads to the algorithm described in the next section.

3. A data-driven calibration algorithm

Now, a data-driven calibration algorithm for penalization procedures can be defined, generalizing a method proposed by Birgé and Massart (2007) and implemented by Lebarbier (2005).

3.1 The general algorithm

Assume that the shape $\text{pen}_{\text{shape}} : \mathcal{M}_n \mapsto \mathbb{R}^+$ of the ideal penalty is known, from some prior knowledge or because it had first been estimated, see Section 3.4. Then, the penalty $K^* \text{pen}_{\text{shape}}$ provides an approximately optimal procedure, for some unknown constant $K^* > 0$. The goal is to find some \widehat{K} such that $\widehat{K} \text{pen}_{\text{shape}}$ is approximately optimal.

Let D_m be some known complexity measure of the 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 Section 2.3, the following algorithm provides an optimal calibration of the penalty $\text{pen}_{\text{shape}}$.

Algorithm 1 (Data-driven penalization with slope heuristics)

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

$$\hat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \{P_n \gamma(\hat{s}_m) + K \text{pen}_{\text{shape}}(m)\} .$$

2. Find $\hat{K}_{\min} > 0$ such that $D_{\hat{m}(K)}$ is “huge” for $K < \hat{K}_{\min}$ and “reasonably small” for $K > \hat{K}_{\min}$.
3. Select the model $\hat{m} := \hat{m}(2\hat{K}_{\min})$.

A computationally efficient way to perform the first step of Algorithm 1 is provided in Section 3.2. The accurate definition of \hat{K}_{\min} is discussed in Section 3.3, including explicit values for “huge” and “reasonably small”. Then, once $P_n \gamma(\hat{s}_m)$ and $\text{pen}_{\text{shape}}(m)$ are known for every $m \in \mathcal{M}_n$, the complexity of Algorithm 1 is $\mathcal{O}(\text{Card}(\mathcal{M}_n)^2)$ (see Algorithm 2 and Proposition 1). This can be a decisive advantage compared to cross-validation methods, as discussed in Section 4.6.

3.2 Computation of $(\hat{m}(K))_{K \geq 0}$

Step 1 of Algorithm 1 requires to compute $\hat{m}(K)$ for every $K \in (0, +\infty)$. A computationally efficient way to perform this step is described in this subsection.

We start with some notations:

$$\forall m \in \mathcal{M}_n, \quad f(m) = P_n \gamma(\hat{s}_m) \quad g(m) = \text{pen}_{\text{shape}}(m)$$

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

Since the latter definition can be ambiguous, let us choose any total ordering \preceq on \mathcal{M}_n such that g is non-decreasing, which is always possible if \mathcal{M}_n is at most countable. Then, $\hat{m}(K)$ 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 $(\hat{m}(K))_{K \geq 0}$ can be computed efficiently is its particular shape.

Indeed, the proof of Proposition 1 shows that $K \mapsto \hat{m}(K)$ is piecewise constant, and non-increasing for \preceq . Then, the whole trajectory $(\hat{m}(K))_{K \geq 0}$ can be summarized by

- the number of jumps $i_{\max} \in \{0, \dots, \text{Card}(\mathcal{M}_n) - 1\}$,
- the location of the jumps: an increasing sequence of nonnegative reals $(K_i)_{0 \leq i \leq i_{\max} + 1}$, with $K_0 = 0$ and $K_{i_{\max} + 1} = +\infty$,

- a non-increasing sequence of models $(m_i)_{0 \leq i \leq i_{\max}}$,

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

Algorithm 2 (Step 1 of Algorithm 1) For every $m \in \mathcal{M}_n$, define $f(m) = P_n \gamma(\widehat{s}_m)$ and $g(m) = \text{pen}_{\text{shape}}(m)$. 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 with respect to \preceq).
- *Step i , $i \geq 1$:* Let

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

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

$$K_i := \inf \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \text{ s.t. } m \in G(m_{i-1}) \right\} \quad (4)$$

$$\text{and } m_i := \min_{\preceq} F_i \quad \text{with } F_i := \arg \min_{m \in G(m_{i-1})} \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \right\} .$$

Proposition 1 (Correctness of Algorithm 2) If \mathcal{M}_n is finite, Algorithm 2 terminates and $i_{\max} \leq \text{Card}(\mathcal{M}_n) - 1$. With the notations of Algorithm 2, let $\widehat{m}(K)$ be the smallest element of

$$E(K) := \arg \min_{m \in \mathcal{M}_n} \{f(m) + Kg(m)\} \quad \text{with respect to } \preceq .$$

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

It is proved in Section A.2. In the change-point detection framework, a similar result has been proved by Lavielle (2005).

Proposition 1 also gives an upper bound on the computational complexity of Algorithm 2; since the complexity of each step is $\mathcal{O}(\text{Card } \mathcal{M}_n)$, Algorithm 2 requires less than $\mathcal{O}(i_{\max} \text{Card } \mathcal{M}_n) \leq \mathcal{O}((\text{Card } \mathcal{M}_n)^2)$ operations. In general, this upper bound is pessimistic since $i_{\max} \ll \text{Card } \mathcal{M}_n$.

3.3 Definition of \widehat{K}_{\min}

Step 2 of Algorithm 1 estimates \widehat{K}_{\min} such that $\widehat{K}_{\min} \text{pen}_{\text{shape}}$ is the minimal penalty. The purpose of this subsection is to define properly \widehat{K}_{\min} as a function of $(\widehat{m}(K))_{K>0}$.

According to the slope heuristics described in Section 2.3, \widehat{K}_{\min} corresponds to a ‘‘complexity jump’’. If $K < \widehat{K}_{\min}$, $\widehat{m}(K)$ has a large complexity, whereas if $K > \widehat{K}_{\min}$, $\widehat{m}(K)$ has a small or medium complexity. Therefore, the two following definitions of \widehat{K}_{\min} are natural.

Let D_{thresh} be the largest ‘‘reasonably small’’ complexity, meaning the models with larger complexities should not be selected. When D_m is the dimension of S_m as a vector space,

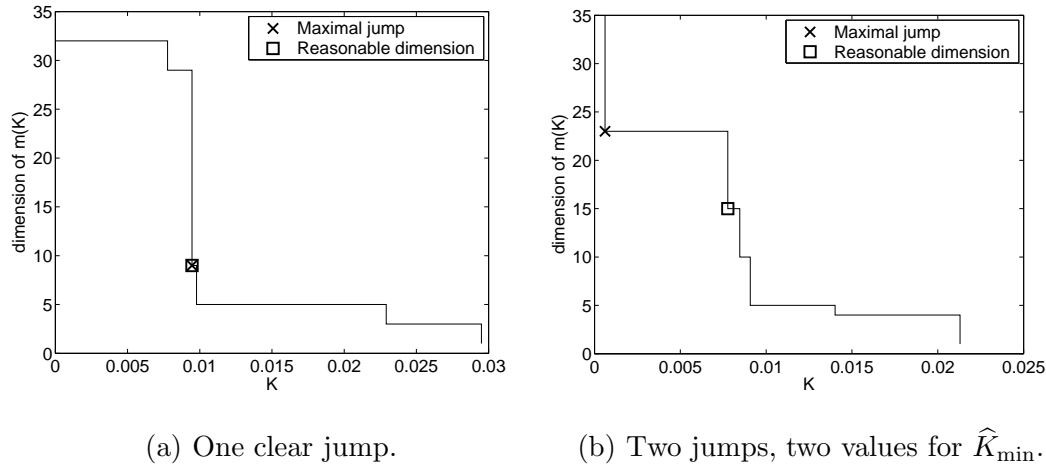


Figure 1: $D_{\hat{m}(K)}$ as a function of K for two different samples. Data are simulated according to (1) with $n = 200$, $X_i \sim \mathcal{U}([0, 1])$, $\epsilon_i \sim \mathcal{N}(0, 1)$, $s(x) = \sin(\pi x)$ and $\sigma \equiv 1$. The models $(S_m)_{m \in \mathcal{M}_n}$ are the sets of piecewise constant functions on regular partitions of $[0, 1]$, with dimensions between 1 and $n/(\ln(n))$. The penalty shape is $\text{pen}_{\text{shape}}(m) = D_m$ and the dimension threshold is $D_{\text{thresh}} = 19 \approx n/(2 \ln(n))$. See experiment S1 by Arlot (2008c, Section 6.1) for details.

$D_{\text{thresh}} \propto n/(\ln(n))$ or $n/(\ln(n))^2$ are natural choices since the dimension of the oracle is likely to be of order n^α for some $\alpha \in (0, 1)$. Then, define

$$\hat{K}_{\min} := \inf \{ K > 0 \text{ s.t. } D_{\hat{m}(K)} \leq D_{\text{thresh}} \} . \quad (\text{thresh})$$

With this definition, Algorithm 2 can be stopped as soon as the threshold is reached.

Another idea is that \hat{K}_{\min} should match with the largest complexity jump:

$$\hat{K}_{\min} := K_{i_{\text{jump}}} \quad \text{with} \quad i_{\text{jump}} = \arg \max_{i \in \{0, \dots, i_{\text{max}} - 1\}} \{ D_{m_{i+1}} - D_{m_i} \} . \quad (\text{max jump})$$

In order to ensure that there is a clear jump in the sequence $(D_{m_i})_{i \geq 0}$, it may be useful to add a few models of large complexity.

As an illustration, we compared the two definitions above (“threshold” and “maximal jump”) on 1000 simulated samples. The exact simulation framework is described below Figure 1. Three cases occurred:

1. There is one clear jump. Both definitions give the same value for \hat{K}_{\min} . This occurred for about 85% of the samples; an example is given on Figure 1a.
2. There are several jumps corresponding to close values of K . Definitions (thresh) and (max jump) give slightly different values for \hat{K}_{\min} , but the selected models $\hat{m}(2\hat{K}_{\min})$ are equal. This occurred for about 8.5% of the samples.

3. There are several jumps corresponding to distant values of K . Definitions (thresh) and (max jump) strongly disagree, giving different selected models $\widehat{m}(2\widehat{K}_{\min})$ at final. This occurred for about 6.5% of the samples; an example is given on Figure 1b.

The only problematic case is the third one, in which an arbitrary choice has to be made between definitions (thresh) and (max jump).

With the same simulated data, we have compared the prediction errors of the two methods by estimating the constant C_{or} that would appear in some oracle inequality,

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

With definition (thresh) $C_{\text{or}} \approx 1.88$; with definition (max jump) $C_{\text{or}} \approx 2.01$. For both methods, the standard error of the estimation is 0.04. As a comparison, Mallows' C_p with a classical estimator of the variance σ^2 has an estimated performance $C_{\text{or}} \approx 1.93$ on the same data.

The overall conclusion of this simulation experiment is that Algorithm 1 can be competitive with Mallows' C_p in a framework where Mallows' C_p is known to be optimal. Definition (thresh) for \widehat{K}_{\min} seems slightly more efficient than (max jump), but without convincing evidence. Indeed, both definitions depend on some arbitrary choices: the value of the threshold D_{thresh} in (thresh), the maximal complexity among the collection of models $(S_m)_{m \in \mathcal{M}_n}$ in (max jump). When n is small, say $n = 200$, choosing D_{thresh} is tricky since $n/(2 \ln(n))$ and \sqrt{n} are quite close. Then, the difference between (thresh) and (max jump) is likely to come mainly from the particular choice $D_{\text{thresh}} = 19$ than from basic differences between the two definitions.

In order to estimate \widehat{K}_{\min} as automatically as possible, we suggest to combine the two definitions; when the selected models $\widehat{m}(2\widehat{K}_{\min})$ differ, send a warning to the final user advising him to look at the curve $K \mapsto D_{\widehat{m}(K)}$ himself; otherwise, remain confident in the automatic choice of $\widehat{m}(2\widehat{K}_{\min})$.

3.4 Penalty shape

For using Algorithm 1 in practice, it is necessary to know *a priori*, or at least to estimate, the optimal shape $\text{pen}_{\text{shape}}$ of the penalty. Let us explain how this can be achieved in different frameworks.

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

Algorithm 1 then provides an alternative to plugging an estimator of $\mathbb{E}[\sigma^2(X)]$ into the above penalties. Let us detail two main advantages of our approach. First, we avoid the difficult task of estimating $\mathbb{E}[\sigma^2(X)]$ without knowing in advance some model to which the

true regression function belongs. Algorithm 1 provides a model-free estimation of the factor multiplying the penalty. Second, the estimator $\widehat{\sigma}^2$ of $\mathbb{E}[\sigma^2(X)]$ with the smallest quadratic risk is certainly far from being the optimal one for model selection. For instance, underestimating the multiplicative factor is well-known to lead to poor performances, whereas overestimating the multiplicative factor does not increase much the prediction error in general. Then, a good estimator of $\mathbb{E}[\sigma^2(X)]$ for model selection should overestimate it with a probability larger than $1/2$. Algorithm 1 satisfies this property automatically because \widehat{K}_{\min} so that the selected model cannot be too large.

In short, *Algorithm 1 with $\text{pen}_{\text{shape}}(m) = D_m$ 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^*D_m .*

In a more general framework, Algorithm 1 allows to choose a different shape of penalty $\text{pen}_{\text{shape}}$. For instance, in the heteroscedastic least-squares regression framework of Section 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, 2008c, Proposition 1) 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} :

$$\mathbb{E}[\text{pen}_{\text{id}}(m)] = \mathbb{E}[(P - P_n)\gamma(\widehat{s}_m)] \approx \frac{2}{n} \sum_{\lambda \in \Lambda_m} \mathbb{E}[\sigma(X)^2 \mid X \in I_\lambda] . \quad (5)$$

An exact result has been proved by Arlot (2008c, Proposition 1). Moreover, Arlot (2008a) gave an example of model selection problem in which no penalty proportional to D_m can be asymptotically optimal.

A first way to estimate the shape of the penalty is simply to use (5) to compute $\text{pen}_{\text{shape}}$, when both the distribution of X and the shape of the noise level σ are known. In practice, one has seldom such a prior knowledge.

We suggest in this situation to use *resampling penalties* (Efron, 1983; Arlot, 2008c), or *V-fold penalties* (Arlot, 2008b) which have much smaller computational costs. Up to a multiplicative factor (automatically estimated by Algorithm 1), these penalties should estimate correctly $\mathbb{E}[\text{pen}_{\text{id}}(m)]$ in any framework. In particular, resampling and V-fold penalties are asymptotically optimal in the heteroscedastic least-squares regression framework (Arlot, 2008b,c).

3.5 The general prediction framework

Section 2 and definition of Algorithm 1 have restricted ourselves to the least-squares regression framework. Actually, this is not necessary at all to make Algorithm 1 well-defined, so that it can naturally be extended to the general prediction framework. More precisely, the (X_i, Y_i) can 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, for which a natural contrast function is the 0–1 loss $\gamma(t; (x, y)) = \mathbf{1}_{t(x) \neq y}$. In this case, the shape of the penalty $\text{pen}_{\text{shape}}$ can for instance be estimated with the global or local Rademacher complexities mentioned in Section 1.

However, a natural question is whether the slope heuristics of Section 2.3, upon which Algorithm 1 relies, can be extended to the general framework. Several concentration results used to prove the validity of the slope heuristics in the least-squares regression framework in

this article are valid in a general setting including binary classification. Even if the factor 2 coming from the closeness of $\mathbb{E}[p_1]$ and $\mathbb{E}[p_2]$ (see Section 2.3) may not be universally valid, we conjecture that Algorithm 1 can be used in other settings than least-squares regression. Moreover, as already mentioned at the end of Section 1, empirical studies have shown that Algorithm 1 can be successfully applied to several problems, with different shapes for the penalty. To our knowledge, to give a formal proof of this fact remains an interesting open problem.

4. Theoretical results

Algorithm 1 mainly relies on the “slope heuristics”, developed in Section 2.2. The goal of this section is to provide a theoretical justification of this heuristics.

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

In order to prove both theorems, two probabilistic results are necessary. First, p_1 , p_2 and δ concentrate around their expectations; for p_2 and δ , it is proved in a general framework in Appendix A.6. 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 to prove in general, so that we must make an assumption on the models. Therefore, in this section, we restrict ourselves to the regressogram case, assuming that for every $m \in \mathcal{M}_n$, S_m is the set of piecewise constant functions on some fixed partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of \mathcal{X} . This framework is described precisely in the next subsection. Although we do not consider regressograms as a final goal, the theoretical results proved for regressograms help to understand better how to use Algorithm 1 in practice.

4.1 Regressograms

Let S_m be the the set of piecewise constant functions on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of \mathcal{X} . The empirical risk minimizer \widehat{s}_m on S_m is called a *regressogram*. S_m is a vector space of dimension $D_m = \text{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda \in \Lambda_m}$. Since this basis is orthogonal in $L^2(\mu)$ for any probability measure μ on \mathcal{X} , computations are quite easy. 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

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

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

4.2 Main assumptions

In this section, we make the following assumptions. First, each model S_m is a set of piecewise constants functions on some fixed partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of X . Second, the family $(S_m)_{m \in \mathcal{M}_n}$ satisfies:

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

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

Assumption **(P1)** is quite classical for proving the asymptotic optimality of a model selection procedure; it is for instance implicitly assumed by Li (1987) in the homoscedastic fixed-design case. Assumption **(P2)** is merely technical and can be changed if necessary; it only ensures that $(S_m)_{m \in \mathcal{M}_n}$ does not contain only models which are either too small or too large.

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

$$\hat{m} \in \arg \min_{m \in \mathcal{M}_n, \min_{\lambda \in \Lambda_m} \{\hat{p}_\lambda\} > 0} \{P_n \gamma(\hat{s}_m) + \text{pen}(m)\} . \quad (6)$$

Moreover, the data $(X_i, Y_i)_{1 \leq i \leq n}$ are assumed to be i.i.d. and to satisfy:

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

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

(Apu) The bias decreases as a power of D_m : there exist some $\beta_+, C_+ > 0$ such that

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

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

Further comments are made in Sections 4.3 and 4.4 about these assumptions, in particular about their possible weakening.

4.3 Minimal penalties

Our first result concerns the existence of a minimal penalty. In this subsection, **(P2)** is replaced by the following strongest assumption:

(P2+) $\exists c_0, c_{\text{rich}} > 0$ s.t. $\forall l \in [\sqrt{n}, c_0 n / (c_{\text{rich}} \ln(n))]$, $\exists m \in \mathcal{M}_n$ s.t. $D_m \in [l, c_{\text{rich}} l]$.

The reason why **(P2)** is not sufficient to prove Theorem 2 below is that at least one model of dimension of order $n / \ln(n)$ should belong to the family $(S_m)_{m \in \mathcal{M}_n}$; otherwise, it may not be possible to prove that such models are selected by penalization procedures beyond the minimal penalty.

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

$$\forall m \in \mathcal{M}_n, \quad 0 \leq \text{pen}(m) \leq K \mathbb{E} [P_n (\gamma(s_m) - \gamma(\hat{s}_m))] . \quad (7)$$

Then, there exist two positive constants K_1, K_2 such that, with probability at least $1 - K_1 n^{-2}$,

$$D_{\hat{m}} \geq K_2 n \ln(n)^{-1} ,$$

where \hat{m} is defined by (6). On the same event,

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

The constants K_1 and K_2 may depend on K, L and constants in **(P1)**, **(P2+)**, **(Ab)**, **(An)**, **(Ap_u)** and **(Ar_ℓ^X)**, but do not depend on n .

This theorem thus validates the first part of the heuristics of Section 2.3, proving that a minimal amount of penalization is required; when the penalty is smaller, the selected dimension $D_{\hat{m}}$ and the quadratic risk of the final estimator $\ell(s, \hat{s}_{\hat{m}})$ blow up. This coupling is quite interesting, since the dimension $D_{\hat{m}}$ is known in practice, contrary to $\ell(s, \hat{s}_{\hat{m}})$. It is then possible to detect from the data whether the penalty is too small, as proposed in Algorithm 1.

The main interest of this result is its combination with Theorem 3 below. Nevertheless Theorem 2 is also interesting by itself for understanding 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 with a random design, even if we have to restrict to regressograms. Moreover, we have a general formulation for the minimal penalty

$$\text{pen}_{\min}(m) := \mathbb{E}[P_n(\gamma(s_m) - \gamma(\hat{s}_m))] = \mathbb{E}[p_2(m)] ,$$

which can be used in frameworks situations where it is not proportional to the dimension D_m of the models (see Section 3.4 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 moreover can be strongly relaxed. Roughly speaking, boundedness of data can be replaced by conditions on moments of the noise, and the uniform lower bound σ_{\min} is no longer necessary when σ satisfies some mild regularity assumptions. We refer to Arlot (2008c, Section 4.3) for detailed statements of these assumptions, and explanations on how to adapt proofs to these situations.

Finally, let us comment on conditions **(Ap_u)** and **(Ar_ℓ^X)**. The upper bound **(Ap_u)** on the bias occurs in the 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 medium and large models have a significantly smaller bias than smaller ones; otherwise, the selected dimension would be allowed to be too small with significant probability. On the other hand, **(Ar_ℓ^X)** is satisfied at least for “almost regular” partitions $(I_\lambda)_{\lambda \in \Lambda_m}$, when X has a lower bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$.

Theorem 2 is stated with a general formulation of **(Ap_u)** and **(Ar_ℓ^X)**, instead of assuming for instance that s is α -Hölderian and X has a lower bounded density w.r.t. Leb, in order to point out the *generality* of the “minimal penalization” phenomenon. It occurs as soon as the models are not too much 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 some prior knowledge on $\mathcal{L}(X)$ is available; if part of the data are unlabeled—a usual case—, classical density estimation procedures can be applied for estimating $\mathcal{L}(X)$ from unlabeled data (Devroye and Lugosi, 2001).

4.4 Optimal penalties

Algorithm 1 relies on a link between the minimal penalty pointed out by Theorem 2 and some optimal penalty. The following result is a formal proof of this link in the framework we consider: penalties close to twice the minimal penalty satisfy an oracle inequality with leading constant approximately equal to one.

Theorem 3 *Suppose all the assumptions of Section 4.2 are satisfied together with*

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

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

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

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

Then, for every $0 < \eta < \min\{\beta_+, 1\}/2$, there exist a constant K_3 and a sequence ϵ_n tending to zero at infinity such that, with probability at least $1 - K_3 n^{-2}$,

$$D_{\widehat{m}} \leq n^{1-\eta} \quad \text{and} \quad \ell(s, \widehat{s}_{\widehat{m}}) \leq \left(\frac{1 + \delta}{1 - \delta} + \epsilon_n \right) \inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \} , \quad (10)$$

where \widehat{m} is defined by (6). Moreover, we have the oracle inequality

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

*The constant K_3 may depend on L, δ, η and the constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) and (**Ar $_{\ell}^X$**), but not on n . The term ϵ_n is smaller than $\ln(n)^{-1/5}$; it can be made 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 pen_{\min} pointed out by Theorem 2 satisfies an oracle inequality with leading constant almost one. In other words, the slope heuristics of Section 2.3 is valid. The consequences of the combination of Theorems 2 and 3 are detailed in Section 4.5.

The oracle inequality (10) remains valid when the penalty is only close to twice the minimal one. In particular, *the shape of the penalty can be estimated by resampling as suggested in Section 3.4.*

Actually, Theorem 3 above is a corollary of a more general result stated in Appendix A.3, Theorem 5. If

$$\text{pen}(m) \approx K \mathbb{E} [P_n (\gamma(s_m) - \gamma(\hat{s}_m))] \quad (11)$$

instead of (9), under the same assumptions, an oracle inequality with leading constant $C(K) + \epsilon_n$ instead of $1 + \epsilon_n$ holds with large probability. The constant $C(K)$ is equal to $(K - 1)^{-1}$ when $K \in (1, 2]$ and to $C(K) = K - 1$ when $K > 2$. Therefore, for every $K > 1$, the penalty defined by (11) is efficient up to a multiplicative constant. This result is new in the heteroscedastic framework.

Let us comment the additional assumption **(Ap)**, that is the lower bound on the bias. Assuming $\ell(s, s_m) > 0$ for every $m \in \mathcal{M}_n$ is classical for proving the asymptotic optimality of Mallows' C_p (Shibata, 1981; Li, 1987; Birgé and Massart, 2007). **(Ap)** has been made by Stone (1985) and Burman (2002) in the density estimation framework, for the same technical reasons as ours. Assumption **(Ap)** is satisfied in several frameworks, such as the following: $(I_\lambda)_{\lambda \in \Lambda_m}$ 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

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

We refer to Arlot (2007, Section 8.10) for a complete proof.

When the lower bound in **(Ap)** is no longer assumed, (10) holds with two modifications in its right-hand side (for details, see Arlot, 2008c, Remark 9): 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 $\gamma_1, \gamma_2 > 0$ are numerical constants. This is equivalent to (10), unless there is a model of small dimension with a small bias. The lower bound in **(Ap)** ensures that it cannot happen. Note that if there is a small model close to s , it is hopeless to obtain an oracle inequality with a penalty which estimates pen_{id} , simply because deviations of pen_{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; for instance, Csiszár (2002) and Csiszár and Shields (2000) showed that BIC penalties are minimal penalties for estimating the order of a Markov chain.

4.5 Main theoretical and practical consequences

The slope heuristics and the correctness of Algorithm 1 follow from the combination of Theorems 2 and 3.

4.5.1 OPTIMAL AND MINIMAL PENALTIES

For the sake of simplicity, let us consider the penalty $K \mathbb{E} [p_2(m)]$ with any $K > 0$; any penalty close to this one satisfies similar properties. At first reading, one can think of the homoscedastic case where $\mathbb{E} [p_2(m)] \approx \sigma^2 D_m n^{-1}$; one of the novelties of our results is that the general picture is quite similar.

According to Theorem 3, the penalization procedure associated with $K \mathbb{E} [p_2(m)]$ satisfies an oracle inequality with leading constant $C_n(K)$ as soon as $K > 1$, and $C_n(2) \approx 1$. Moreover, results proved by Arlot (2008b) imply that $C_n(K) \geq C(K) > 1$ as soon as K is not close to 2. Therefore, $K = 2$ is the optimal multiplying factor in front of $\mathbb{E} [p_2(m)]$.

When $K < 1$, Theorem 2 shows that no oracle inequality can hold with leading constant $C_n(K) < \ln(n)$. Since $C_n(K) \leq (K-1)^{-1} < \ln(n)$ as soon as $K > 1 + \ln(n)^{-1}$, $K = 1$ is the *minimal multiplying factor* in front of $\mathbb{E}[p_2(m)]$. More generally, $\text{pen}_{\min}(m) := \mathbb{E}[p_2(m)]$ is proved to be a *minimal penalty*.

In short, Theorems 2 and 3 prove the slope heuristics described in Section 2.3:

$$\text{“optimal” penalty} \approx 2 \times \text{“minimal” penalty} .$$

Birgé and Massart (2007) have proved the validity of the slope heuristics in the Gaussian homoscedastic framework. This paper extends their result to a non-Gaussian and heteroscedastic setting.

4.5.2 DIMENSION JUMP

In addition, Theorems 2 and 3 prove the existence of a crucial phenomenon: there exists a “dimension jump”—complexity jump in the general framework—around the minimal penalty. Let us consider again the penalty $K\mathbb{E}[p_2(m)]$. As in Algorithm 1, let us define

$$\hat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \{P_n \gamma(\hat{s}_m) + K\mathbb{E}[p_2(m)]\} .$$

A careful look at the proofs of Theorems 2 and 3 shows that there exist constants $K_4, K_5 > 0$ and an event of probability $1 - K_4 n^{-2}$ on which

$$\forall 0 < K < 1 - \frac{1}{\ln(n)}, D_{\hat{m}(K)} \geq \frac{K_5 n}{(\ln(n))^2} \quad \text{and} \quad \forall K > 1 + \frac{1}{\ln(n)}, D_{\hat{m}(K)} \leq n^{1-\eta} . \quad (12)$$

Therefore, the dimension $D_{\hat{m}(K)}$ of the selected model jumps around the minimal value $K = 1$, from values of order $n(\ln(n))^{-2}$ to $n^{1-\eta}$.

Let us now explain why Algorithm 1 is correct, assuming that $\text{pen}_{\text{shape}}(m)$ is close to $\mathbb{E}[p_2(m)]$. With definition (thresh) of \hat{K}_{\min} and a threshold $D_{\text{thresh}} \propto n(\ln(n))^{-3}$, (12) ensures that

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

with a large probability. Then, according to Theorem 3, the output of Algorithm 1 satisfies an oracle inequality with leading constant C_n tending to one as n tends to infinity.

4.6 Comparison with data-splitting methods

Tuning parameters are often chosen by cross-validation or by another data-splitting method, which suffer from some drawbacks compared to Algorithm 1.

First, V -fold cross-validation, leave- p -out and repeated learning-testing methods require a larger computation time. Indeed, they need to perform the empirical risk minimization process for each model several times, whereas Algorithm 1 only needs to perform it once.

Second, V -fold cross-validation is asymptotically suboptimal when V is fixed, as shown by (Arlot, 2008b, Theorem 1). The same suboptimality result is valid for the hold-out, when the size of the training set is not asymptotically equivalent to the sample size n . On the contrary, Theorems 2 and 3 prove that Algorithm 1 is asymptotically optimal in a framework

including the one used by (Arlot, 2008b, Theorem 1) for proving the suboptimality of V -fold cross-validation. Hence, the quadratic risk of Algorithm 1 should be smaller, within a factor $\kappa > 1$.

Third, hold-out with a training set of size $n_t \sim n$, for instance $n_t = n - \sqrt{n}$ or $n_t = n(1 - \ln(n)^{-1})$, is known to be unstable. The final output \hat{m} strongly depends on the choice of a particular split of the data. According to the simulation study of Section 3.3, Algorithm 1 is far more stable.

To conclude, compared to data splitting methods, Algorithm 1 is either faster to compute, more efficient in terms of quadratic risk, or more stable. Then, Algorithm 1 should be preferred each time it can be used. Another approach is to use aggregation techniques, instead of selecting one model. As shown by several results (see for instance Tsybakov, 2004; Lecué, 2007), aggregating estimators built upon a training simple of size $n_t \sim n$ can have an optimal quadratic risk. Moreover, aggregation requires approximately the same computation time as Algorithm 1, and is much more stable than the hold-out. Hence, it can be an alternative to model selection with Algorithm 1.

5. Conclusion

This paper provides mathematical evidence that the method introduced by Birgé and Massart (2007) for designing data-driven penalties remains efficient in a non-Gaussian framework. The purpose of this conclusion is to relate the slope heuristics developed in Section 2 to the well known Mallows' C_p and Akaike's criteria and to the unbiased estimation of the risk principle.

Let us come back to Gaussian model selection in order to explain how to guess what is the right penalty from the data themselves. Let γ_n be some empirical criterion (for instance the least-squares criterion as in this paper, or the log-likelihood criterion), $(S_m)_{m \in \mathcal{M}_n}$ be a collection of models and for every $m \in \mathcal{M}_n$ s_m be some minimizer of $t \mapsto \mathbb{E}[\gamma_n(t)]$ over S_m (assuming that such a point exists). Minimizing some penalized criterion

$$\gamma_n(\hat{s}_m) + \text{pen}(m)$$

over \mathcal{M}_n amounts to minimize

$$\hat{b}_m - \hat{v}_m + \text{pen}(m) ,$$

$$\text{where } \forall m \in \mathcal{M}_n, \hat{b}_m = \gamma_n(s_m) - \gamma_n(s) \text{ and } \hat{v}_m = \gamma_n(s_m) - \gamma_n(\hat{s}_m) .$$

The point is that \hat{b}_m is an unbiased estimator of the bias term $\ell(s, s_m)$. Having concentration arguments in mind, minimizing $\hat{b}_m - \hat{v}_m + \text{pen}(m)$ can be conjectured approximately equivalent to minimize

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

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

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

In Gaussian least-squares regression with a fixed design, the models S_m are linear and $\mathbb{E}[\hat{v}_m] = \mathbb{E}[\ell(s, \hat{s}_m)]$ is explicitly computable if the noise level is constant and known;

this leads to Mallows' C_p penalty. When γ_n is the log-likelihood,

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

asymptotically, where D_m stands for the number of parameters defining model S_m ; this leads to Akaike's Information Criterion (AIC). Therefore, both Mallows' C_p and Akaike's criterion are based on the unbiased (or asymptotically unbiased) risk estimation principle.

This paper goes further in this direction, using that $\mathbb{E}[\hat{v}_m] \approx \mathbb{E}[\ell(s_m, \hat{s}_m)]$ remains a valid approximation in a non-asymptotic framework. Then, a good penalty becomes $2\mathbb{E}[\hat{v}_m]$ or $2\hat{v}_m$, having in mind concentration arguments. Since \hat{v}_m is the minimal penalty, this explains the slope heuristics (Birgé and Massart, 2007) and connects it to Mallows' C_p and Akaike's heuristics.

The second main idea developed in this paper is that the minimal penalty can be estimated from the data; Algorithm 1 uses the jump of complexity which occurs around the minimal penalty, as shown in Sections 3.3 and 4.5.2. Another way to estimate the minimal penalty when it is (at least approximately) of the form αD_m is to estimate α by the *slope* of the graph of $\gamma_n(\hat{s}_m)$ for large enough values of D_m ; this method can be extended to other shapes of penalties, simply by replacing D_m by some (known!) function $f(D_m)$.

The slope heuristics can even be combined with resampling ideas, by taking a function f built from a randomized empirical criterion. As shown by Arlot (2008a), this approach is much more efficient than the rougher choice $f(D_m) = D_m$ for heteroscedastic regression frameworks. The question of the optimality of the slope heuristics in general remains an open problem; nevertheless, we believe that this heuristics can be useful in practice, and that proving its efficiency in this paper helps to understand it better.

Let us finally mention that contrary to Birgé and Massart (2007), we assume in this paper that the collection of models \mathcal{M}_n is "small", that is $\text{Card}(\mathcal{M}_n)$ grows at most like a power of n . For several problems, such that complete variable selection, larger collections of models have to be considered; then, it is known from the homoscedastic case that the minimal penalty is much larger than $\mathbb{E}[p_2(m)]$. Nevertheless, Émilie Lebarbier has used the slope heuristics with $f(D_m) = D_m \left(2.5 + \ln \left(\frac{n}{D_m} \right) \right)$ for multiple change-points detection from n noisy data, using the results by Birgé and Massart (2007) in the Gaussian case.

Let us now explain how we expect to generalize the slope heuristics to the non-Gaussian heteroscedastic case when \mathcal{M}_n is large. First, group the models according to some complexity index C_m such as their dimensions D_m ; for $C \in \{1, \dots, n^k\}$, 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", that is model selection with the family $(\widetilde{S}_C)_{1 \leq C \leq n^k}$. 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 could rely on the extension of our results to any kind of model, since \widetilde{S}_C is not a vector space in general.

Acknowledgments

The authors gratefully acknowledge the anonymous referees for several suggestions and references.

Appendix A. Proofs

This appendix is devoted to the proofs of the results stated in the paper. Proposition 1 is proved in Section A.2; Theorem 3 is proved in Sections A.3 and A.4; Theorem 2 is proved in Section A.5; the remaining sections are devoted to probabilistic results used in the main proofs and technical proofs.

A.1 Conventions and notations

In the rest of the paper, L denotes a universal constant, not necessarily the same at each occurrence. When L is not universal, but depends on p_1, \dots, p_k , it is written L_{p_1, \dots, p_k} . Similarly, $L_{(\mathbf{SH2})}$ (resp. $L_{(\mathbf{SH5})}$) denotes a constant allowed to depend on the parameters of the assumptions made in Theorem 2 (resp. Theorem 5), including **(P1)** and **(P2)**. We also make use of the following notations:

- $\forall a, b \in \mathbb{R}$, $a \wedge b$ 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.
- $\forall I_\lambda \subset \mathcal{X}$, $p_\lambda := P(X \in I_\lambda)$ and $\sigma_\lambda^2 := \mathbb{E} \left[(Y - s_m(X))^2 \mid X \in I_\lambda \right]$.
- Since $\mathbb{E}[p_1(m)]$ is not well-defined (because of the event $\{\min_{\lambda \in \Lambda_m} \{\widehat{p}_\lambda\} = 0\}$), we have to take the following convention

$$p_1(m) = \widetilde{p}_1(m) := \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda > 0} p_\lambda \left(\beta_\lambda - \widehat{\beta}_\lambda \right)^2 + \sum_{\lambda \in \Lambda_m \text{ s.t. } \widehat{p}_\lambda = 0} p_\lambda \sigma_\lambda^2 .$$

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 (Theorems 2 and 5).

A.2 Proof of Proposition 1

First, since \mathcal{M}_n is finite, the infimum in (4) 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 different; hence, Algorithm 2 terminates and $i_{\max} + 1 \leq \text{Card}(\mathcal{M}_n)$. We now prove by induction the following property for every $i \in \{0, \dots, i_{\max}\}$:

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

Notice also that K_i can always be defined by (4) 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 ,

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

hence

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

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

$\mathcal{P}_i \Rightarrow \mathcal{P}_{i+1}$ FOR EVERY $i \in \{0, \dots, i_{\max} - 1\}$

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

$$f(m_{i+2}) - f(m_{i+1}) = K_{i+2}(g(m_{i+1}) - g(m_{i+2})) \quad (13)$$

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

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

$$f(m_{i+2}) - f(m_i) > K_{i+1}(g(m_i) - g(m_{i+2})) \quad (15)$$

(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 (15) and (14) with (13), we have

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

hence $K_{i+2} > K_{i+1}$, 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 tends to K_{i+1} , it follows that $m_i \in E(K_{i+1})$. By (14), 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 proved 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 *With the notations of Proposition 1 and its proof, if $0 \leq K < K'$, $m \in E(K)$ and $m' \in E(K')$, then one of the two following statements holds true:*

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

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

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

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

$$f(m) + Kg(m) \leq f(m') + Kg(m') \quad (16)$$

$$f(m') + K'g(m') \leq f(m) + K'g(m) \quad (17)$$

Summing (16) and (17) gives $(K' - K)g(m') \leq (K' - K)g(m)$ so that

$$g(m') \leq g(m) \quad (18)$$

Since $K \geq 0$, (16) and (18) give $f(m) + Kg(m) \leq f(m') + Kg(m)$, that is

$$f(m) \leq f(m') \quad (19)$$

Moreover, (19) and (17) imply $g(m) = g(m')$, hence $f(m') \leq f(m)$, that is $f(m) = f(m')$ by (19). Similarly, (16) and (18) show that $f(m) = f(m')$ imply $g(m) = g(m')$. In both cases, (a) is satisfied. Otherwise, $f(m) < f(m')$ and $g(m) > g(m')$, that is the (b) statement.

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 $\widehat{m}(K)$. \blacksquare

A.3 A general oracle inequality

First of all, let us state a general theorem, from which Theorem 3 is an obvious corollary.

Theorem 5 *Suppose all the assumptions of Section 4.2 are satisfied together with*

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

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

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

$$\begin{aligned} & \mathbb{E} [c_1 P(\gamma(\widehat{s}_m) - \gamma(s_m)) + c_2 P_n(\gamma(s_m) - \gamma(\widehat{s}_m))] \\ & \leq \text{pen}(m) \leq \mathbb{E} [C_1 P(\gamma(\widehat{s}_m) - \gamma(s_m)) + C_2 P_n(\gamma(s_m) - \gamma(\widehat{s}_m))] \quad . \end{aligned} \quad (20)$$

Then, for every $0 < \eta < \min\{\beta_+; 1\}/2$, there exist a constant K_3 and a sequence ϵ_n tending to zero at infinity such that, with probability at least $1 - K_3 n^{-2}$,

$$\begin{aligned} & D_{\widehat{m}} \leq n^{1-\eta} \\ \text{and} \quad & \ell(s, \widehat{s}_{\widehat{m}}) \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n \right] \inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \} \end{aligned} \quad (21)$$

where \widehat{m} is defined by (6). Moreover, we have the oracle inequality

$$\mathbb{E}[\ell(s, \widehat{s}_{\widehat{m}})] \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n \right] \mathbb{E} \left[\inf_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\} \right] + \frac{A^2 K_3}{n^2}. \quad (22)$$

The constant K_3 may depend on $L, \eta, \xi, c_1, c_2, C_1, C_2$ and constants in **(P1)**, **(P2)**, **(Ab)**, **(An)**, **(Ap)** and **(Ar X)**, but not on n . The term ϵ_n is smaller than $\ln(n)^{-1/5}$; it can be made 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 (20) on the penalty is motivated by the fact that the ideal shape of penalty $\mathbb{E}[\text{pen}_{\text{id}}(m)]$ (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 Theorem 5, Arlot (2008b,c) has proved that resampling and V -fold penalties satisfy condition (20) with constants $c_1 + c_2 = 2 - \delta_n$, $C_1 + C_2 = 2 + \delta_n$ (for some absolute sequence δ_n tending to zero at infinity), and some numerical constant $\xi > 0$. Then, Theorem 5 shows that such a penalization procedure satisfies an oracle inequality with leading constant tending to 1 asymptotically.

The rationale behind Theorem 5 is that if $\text{pen}(m)$ is close to $c_1 p_1(m) + c_2 p_2(m)$, then $\text{crit}(m) \approx \ell(s, s_m) + c_1 p_1(m) + (c_2 - 1)p_2(m)$. When $c_1 = c_2 = 1$, this is exactly the ideal criterion $\ell(s, \widehat{s}_m)$. When $c_1 + c_2 = 2$ 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 enough. The 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 Theorem 3, as K_3), one can replace the condition $c_2 > 1$ by $c_1 + c_2 > 1$ and $c_1, c_2 \geq 0$.

A.4 Proof of Theorem 5

This proof is similar to the one of Arlot (2008c, Theorem 1). We give it for the sake of completeness.

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

$$\ell(s, \widehat{s}_{\widehat{m}}) - (\text{pen}'_{\text{id}}(\widehat{m}) - \text{pen}(\widehat{m})) \leq \ell(s, \widehat{s}_m) + (\text{pen}(m) - \text{pen}'_{\text{id}}(m)) \quad (23)$$

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

We will thus use the concentration inequalities of Section A.6 with $x = \gamma \ln(n)$ and $\gamma = 2 + \alpha_{\mathcal{M}}$. Define $B_n(m) = \min_{\lambda \in \Lambda_m} \{np_\lambda\}$, and Ω_n the event on which

- for every $m \in \mathcal{M}_n$, (20) holds
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$, (29) and (30) hold:

$$\begin{aligned} \widetilde{p}_1(m) &\geq \mathbb{E}[\widetilde{p}_1(m)] - L_{(\text{SH5})} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n(m)} \right] \mathbb{E}[p_2(m)] \\ \widetilde{p}_1(m) &\leq \mathbb{E}[\widetilde{p}_1(m)] + L_{(\text{SH5})} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m} e^{-LB_n(m)} \right] \mathbb{E}[p_2(m)] \end{aligned}$$

- for every $m \in \mathcal{M}_n$ such that $B_n(m) > 0$, (31), (28) and 26 hold:

$$\begin{aligned}\tilde{p}_1(m) &\geq \left(\frac{1}{2 + (\gamma + 1)B_n(m)^{-1} \ln(n)} - \frac{L_{(\mathbf{SH5})} \ln(n)^2}{\sqrt{D_m}} \right) \mathbb{E}[p_2(m)] \\ |p_2(m) - \mathbb{E}[p_2(m)]| &\leq \frac{L_{(\mathbf{SH5})} \ln(n)}{\sqrt{D_m}} [\ell(s, s_m) + \mathbb{E}[p_2(m)]] \\ |\bar{\delta}(m)| &\leq \frac{\ell(s, s_m)}{\sqrt{D_m}} + L_{(\mathbf{SH5})} \frac{\ln(n)}{\sqrt{D_m}} \mathbb{E}[p_2(m)]\end{aligned}$$

From Proposition 11 (for \tilde{p}_1), Proposition 10 (for p_2) and Proposition 8 (for $\bar{\delta}(m)$),

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

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

$$\max \{ |\tilde{p}_1(m) - \mathbb{E}[\tilde{p}_1(m)]|, |p_2(m) - \mathbb{E}[p_2(m)]|, |\bar{\delta}(m)| \} \leq \frac{L_{(\mathbf{SH5})} \mathbb{E}[\ell(s, s_m) + p_2(m)]}{\ln(n)}$$

Using (32) (in Proposition 12) and the fact that $B_n(m) \geq L^{-1} \ln(n)$,

$$\frac{(c_1 + c_2) (1 - \tilde{\delta}_n)}{2} \leq \mathbb{E}[\text{pen}(m)] \leq \frac{(C_1 + C_2) (1 + \tilde{\delta}_n)}{2} \mathbb{E}[\tilde{p}_1(m) + p_2(m)]$$

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

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

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

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

in general. Combined with (23), this gives: if $n \geq L_{(\mathbf{SH5})}$,

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

We now use Lemmas 6 and 7 below to control on Ω_n the dimensions of the selected model \hat{m} and the oracle model $m^* \in \arg \min_{m \in \mathcal{M}_n} \{ \ell(s, \hat{s}_m) \}$.

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 $K_3 n_0^{-2} \geq 1$.

Classical oracle inequality Since (21) holds true on Ω_n ,

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

which proves (22). ■

Lemma 6 (Control on the dimension of the selected model) *Let $c > 0$ and $\alpha > (1 - \beta_+)_+ / 2$. Then, if $n \geq L_{(\text{SH5}),c,\alpha}$, on the event Ω_n defined in the proof of Theorem 5,*

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

Lemma 7 (Control on the dimension of the oracle model) *Define the oracle model $m^* \in \arg \min_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\}$. Let $c > 0$ and $\alpha > (1 - \beta_+)_+ / 2$. Then, if $n \geq L_{(\text{SH5}),c,\alpha}$, on the event Ω_n defined in the proof of Theorem 5,*

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

Proof of Lemma 6 By definition, \widehat{m} minimizes $\text{crit}(m)$ over \mathcal{M}_n . It thus also minimizes

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

over \mathcal{M}_n .

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

$$\begin{aligned} \ell(s, s_m) &\geq C_- (\ln(n))^{-7\beta_-} \quad \text{from (Ap)} \\ \text{pen}(m) &\geq 0 \end{aligned}$$

$$p_2(m) \leq L_{(\text{SH5})} \sqrt{\frac{\ln(n)}{n}} + L_{(\text{SH5})} \frac{D_m}{n} \leq L_{(\text{SH5})} \sqrt{\frac{\ln(n)}{n}} \quad \text{from (27)}$$

and from (26) (in Proposition 8),

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

We then have

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

2. Lower bound for large models: let $m \in \mathcal{M}_n$ such that $D_m \geq n^{1/2+\alpha}$. From (20) and (27) (in Proposition 10),

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

and from (24),

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

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

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

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

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

By (33) in Lemma 13, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.

Using **(Ap)**,

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

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

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

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

Proof of Lemma 7 Recall that m^* minimizes $\ell(s, \hat{s}_m) = \ell(s, s_m) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $\ell(s, \hat{s}_m) = \infty$ if $A_n(m) = 0$.

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

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

2. Lower bound on $\ell(s, \hat{s}_m)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > n^{1/2+\alpha}$. From (31), for $n \geq L_{(\mathbf{SH5}),\alpha}$,

$$\tilde{p}_1(m) \geq \left(\frac{1}{2 + (\gamma + 1) \left(c_{r,\ell}^X \right)^{-1} \ln(n)} - \frac{L_{(\mathbf{SH5}),\alpha}}{n^{1/4}} \right) \mathbb{E}[\tilde{p}_2(m)]$$

$$\text{so that } \ell(s, \hat{s}_m) \geq \tilde{p}_1(m) \geq L_{(\mathbf{SH5}),\alpha} n^{-1/2+\alpha} .$$

3. There exists a better model for $\ell(s, \hat{s}_m)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of Lemma 6 and assume that $n \geq L_{c_{\text{rich}},\alpha}$. Then,

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

and the arguments of the previous proof show that

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

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

A.5 Proof of Theorem 2

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

- for every $m \in \mathcal{M}_n$, (7) (for pen), (31) (for \tilde{p}_1), (27)–(28) (for p_2 , with $x = \gamma \ln(n)$ and $\theta = \sqrt{\ln(n)/n}$) and (24)–(26) (for $\bar{\delta}$, with $x = \gamma \ln(n)$ and $\eta = \sqrt{\ln(n)/n}$) hold true.
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$, (29) and (30) hold (for \tilde{p}_1).

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

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

over $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$. As in the proof of Theorem 5, we define $c = L_{c_{r,\ell}^x} > 0$ 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 $c' = \min(c, c_0)$ and $d \in (0, 1)$ a constant to be chosen later.

1. Lower bound on $\text{crit}'(m)$ for “small” models: assume that $m \in \mathcal{M}_n$ and $D_m \leq dc' n \ln(n)^{-1}$. Then, $\ell(s, s_m) + \text{pen}(m) \geq 0$ and from (24),

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

If $D_m \geq \ln(n)^4$, (28) implies that

$$p_2(m) \leq \left(1 + \frac{L(\mathbf{SH2})}{\ln(n)}\right) \mathbb{E}[p_2(m)] \leq \frac{L(\mathbf{SH2})D_m}{n} \leq \frac{c'dL(\mathbf{SH2})}{\ln(n)}.$$

On the other hand, if $D_m < \ln(n)^4$, (27) implies that

$$p_2(m) \leq L(\mathbf{SH2}) \sqrt{\frac{\ln(n)}{n}}.$$

We then have

$$\text{crit}'(m) \geq -dL(\mathbf{SH2}) (\ln(n))^{-1}.$$

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

$$\ln(n)^4 \leq \frac{c'dn}{c_{\text{rich}} \ln(n)} \leq D_{m_1} \leq \frac{c'n}{\ln(n)} \leq n.$$

From **(P2+)**, this is possible as soon as $n \geq L_{c_{\text{rich}}, c', d}$. By (33) in Lemma 13, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.

We then have

$$\begin{aligned} \ell(s, s_{m_1}) &\leq L(\mathbf{SH2}), c' \ln(n)^{\beta_+} n^{-\beta_+} && \text{by (Ap)} \\ p_2(m_1) &\geq \left(1 - \frac{L(\mathbf{SH2})}{\ln(n)}\right) \mathbb{E}[p_2(m_1)] && \text{by (28)} \\ \text{pen}(m_1) &\leq K \mathbb{E}[p_2(m_1)] && \text{by (7)} \\ |\bar{\delta}(m_1)| &\leq L_A \sqrt{\frac{\ln(n)}{n}} && \text{by (24)} \end{aligned}$$

so that

$$\begin{aligned} \text{crit}'(m_1) &\leq L_{(\mathbf{SH2}),c'} \ln(n)^{\beta_+} n^{-\beta_+} + \left(K - 1 + \frac{L_{(\mathbf{SH2})}}{\ln(n)} \right) \mathbb{E}[p_2(m_1)] + L_A \sqrt{\frac{\ln(n)}{n}} \\ &\leq \frac{(K - 1 + L_{(\mathbf{SH2})}(\ln(n))^{-1}) \sigma_{\min}^2 c'}{2 \ln(n)} \end{aligned}$$

if $n \geq L_{(\mathbf{SH2}),c'}$.

We now choose d such that the constant $dL_{(\mathbf{SH2})}$ appearing in the lower bound on $\text{crit}'(m)$ for “small” models is smaller than $(1 - K - L_{(\mathbf{SH2})}(\ln(n))^{-1}) \sigma_{\min}^2 c' / 2$, that is $d \leq L_{(\mathbf{SH2}),c'}$. Then, we assume that $n \geq n_0 = L_{(\mathbf{SH2}),c',d} = L_{(\mathbf{SH2})}$. Finally, we remove this condition as before by enlarging K_1 .

Risk of $D_{\hat{m}}$ The proof of (8) is quite similar to the one of Lemma 7. 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

$$\ell(s, \hat{s}_m) \geq \tilde{p}_1(m) \geq L_{(\mathbf{SH2})} K_2 \ln(n)^{-2} \quad \text{by (31)} .$$

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

$$\ell(s, \hat{s}_{m_0}) \leq L_{(\mathbf{SH2})} \left(n^{-\beta_+/2} + n^{-1/2} \right) .$$

If $n \geq L_{(\mathbf{SH2})}$, the ratio between these two bounds is larger than $\ln(n)$, so that (8) holds. ■

A.6 Concentration inequalities used in the main proofs

In this section, we no longer assume that each model is the set of piecewise constant functions on some partition of \mathcal{X} . First, we control $\bar{\delta}(m)$ with general models and bounded data.

Proposition 8 *Assume that $\|Y\|_\infty \leq A < \infty$. Then for all $x \geq 0$, on an event of probability at least $1 - 2e^{-x}$:*

$$\forall \eta > 0, \quad |\bar{\delta}(m)| \leq \eta \ell(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3} \right) \frac{A^2 x}{n} . \quad (24)$$

If moreover

$$Q_m^{(p)} := \frac{n \mathbb{E}[p_2(m)]}{D_m} > 0 , \quad (25)$$

on the same event,

$$|\bar{\delta}(m)| \leq \frac{\ell(s, s_m)}{\sqrt{D_m}} + \frac{20}{3} \frac{A^2}{Q_m^{(p)}} \frac{\mathbb{E}[p_2(m)]}{\sqrt{D_m}} x . \quad (26)$$

Remark 9 (Regressogram case) *If S_m is the set of piecewise constant functions on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of \mathcal{X} ,*

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

Then, we derive a concentration inequality for $p_2(m)$ in the regressogram case from a general result by Boucheron and Massart (2008).

Proposition 10 *Let S_m be the model of piecewise constant functions associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$. Assume that $\|Y\|_\infty \leq A$ and define $p_2(m) = P_n(\gamma(s_m) - \gamma(\hat{s}_m))$.*

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

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L \left[\theta \ell(s, s_m) + \frac{A^2 \sqrt{D_m} \sqrt{x}}{n} + \frac{A^2 x}{\theta n} \right] \quad (27)$$

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

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq \frac{L}{\sqrt{D_m}} \left[\ell(s, s_m) + \frac{A^2 \mathbb{E}[p_2(m)]}{\sigma_{\min}^2} (\sqrt{x} + x) \right]. \quad (28)$$

Finally, we recall a concentration inequality for $p_1(m)$ proved by (Arlot, 2008b, Proposition 9). Its proof is particular to the regressogram case.

Proposition 11 (Proposition 9, Arlot (2008b)) *Let $\gamma > 0$ and S_m be the model of piecewise constant functions 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 $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n > 0$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\tilde{p}_1(m) \geq \mathbb{E}[\tilde{p}_1(m)] - L_{A, \sigma_{\min}, \gamma} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad (29)$$

$$\tilde{p}_1(m) \leq \mathbb{E}[\tilde{p}_1(m)] + L_{A, \sigma_{\min}, \gamma} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m} e^{-LB_n} \right] \mathbb{E}[p_2(m)]. \quad (30)$$

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

$$\tilde{p}_1(m) \geq \left(\frac{1}{2 + (\gamma + 1)B_n^{-1} \ln(n)} - \frac{L_{A, \sigma_{\min}, \gamma} \ln(n)^2}{\sqrt{D_m}} \right) \mathbb{E}[p_2(m)]. \quad (31)$$

A.7 Additional results needed

A crucial result in the proofs of Theorems 5 and 2 is that $p_1(m)$ and $p_2(m)$ are close in expectation; the following proposition was proved by Arlot (2008b, Lemma 7).

Proposition 12 (Lemma 7, Arlot (2008b)) *Let S_m be a model of piecewise constant functions adapted to some partition $(I_\lambda)_{\lambda \in \Lambda_m}$. Assume that $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B > 0$. Then,*

$$\begin{aligned} (1 - e^{-B})^2 \mathbb{E}[p_2(m)] &\leq \mathbb{E}[\tilde{p}_1(m)] \\ &\leq \left[2 \wedge \left(1 + 5.1 \times B^{-1/4} \right) + (B \vee 1) e^{-(B \vee 1)} \right] \mathbb{E}[p_2(m)]. \end{aligned} \quad (32)$$

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

Lemma 13 *Let $(p_\lambda)_{\lambda \in \Lambda_m}$ be non-negative real numbers of sum 1, $(n\hat{p}_\lambda)_{\lambda \in \Lambda_m}$ a multinomial vector of parameters $(n; (p_\lambda)_{\lambda \in \Lambda_m})$. Then, for all $\gamma > 0$,*

$$\min_{\lambda \in \Lambda_m} \{n\hat{p}_\lambda\} \geq \frac{\min_{\lambda \in \Lambda_m} \{np_\lambda\}}{2} - 2(\gamma + 1) \ln(n) \quad (33)$$

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

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

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

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

A.8 Proof of Proposition 8

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

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

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

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

since

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

and

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

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

$$\begin{aligned} (\gamma(t, \cdot) - \gamma(s, \cdot))^2 &= (t(X) - s(X))^2 (2(Y - s(X)) - t(X) + s(X))^2 \\ \text{and } \mathbb{E}[(Y - s(X))^2 | X] &\leq \frac{(2A)^2}{4} = A. \end{aligned}$$

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

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

and (24) 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 (25), we deduce (26). ■

A.9 Proof of Proposition 10

We apply here a result by Boucheron and Massart (2008, Theorem 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-squares regression framework. For every $t : \mathcal{X} \mapsto \mathbb{R}$ and $\epsilon > 0$, we define

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

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 $\phi_m(\sigma) \leq \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E} \left[\sup_{t \in S_m, d(u, t) \leq \sigma} |\bar{\gamma}_n(u) - \bar{\gamma}_n(t)| \right] \leq \phi_m(\sigma) . \quad (34)$$

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

$$\|p_2(m) - \mathbb{E}[p_2(m)]\|_q \leq \frac{L}{\sqrt{n}} \left[\sqrt{2q} \left(\sqrt{\ell(s, s_m)} \vee \varepsilon_{\star, m} \right) + q \frac{2}{\sqrt{n}} \right] . \quad (35)$$

Using now that S_m is the set of piecewise constant functions on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of \mathcal{X} , we can take

$$\phi_m(\sigma) = 3\sqrt{2}\sqrt{D_m} \times \sigma \quad \text{in (34)} . \quad (36)$$

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

Combining (35) with the classical link between moments and concentration (see for instance Arlot, 2007, Lemma 8.9), the first result follows. The second result is obtained by taking $\theta = D_m^{-1/2}$, as in Proposition 8. ■

Proof of (36) 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 $u \in S_m$,

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

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

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

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

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

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

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

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

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 $\sigma > 0$, that is

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

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

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

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 :

$$\psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2n}} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E} \left[\frac{(\sum_{i=1}^n (\mathbf{1}_{X_i \in I_\lambda} - p_\lambda))^2}{p_\lambda} \right]} + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E} \left[\frac{(\sum_{i=1}^n \mathbf{1}_{X_i \in I_\lambda} \eta_i)^2}{p_\lambda} \right]} \quad (37)$$

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

$$\mathbb{E} \left(\sum_{i=1}^n (\mathbf{1}_{X_i \in I_\lambda} - p_\lambda) \right)^2 = \sum_{i=1}^n \mathbb{E} (\mathbf{1}_{X_i \in I_\lambda} - p_\lambda)^2 = np_\lambda (1 - p_\lambda) \quad (38)$$

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

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

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

$$\mathbb{E} \left(\sum_{i=1}^n \mathbf{1}_{X_i \in I_\lambda} \eta_i \right)^2 = \sum_{i=1}^n \mathbb{E} [\mathbf{1}_{X_i \in I_\lambda} \eta_i^2] \leq np_\lambda \|\eta\|_\infty^2 \leq np_\lambda (2A)^2 . \quad (39)$$

Combining (37) with (38) and (39), 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

- Hirotsugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.
- Hirotsugu 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.

- Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.
- Sylvain Arlot. Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression, December 2008a. arXiv:0812.3141.
- Sylvain Arlot. *V-fold cross-validation improved: V-fold penalization*, February 2008b. arXiv:0802.0566v2.
- Sylvain Arlot. Model selection by resampling penalization, March 2008c. oai:hal.archives-ouvertes.fr:hal-00262478_v1.
- Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.
- Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.
- Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with unknown variance. To appear in *The Annals of Statistics*. arXiv:math.ST/0701250, 2007.
- Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
- 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.
- 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.
- Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- Stéphane Boucheron and Pascal Massart. A poor man’s wilks phenomenon. Personal communication, March 2008.
- Prabir Burman. Estimation of equipfrequency histograms. *Statist. Probab. Lett.*, 56(3):227–238, 2002.
- Imre Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48(6):1616–1628, 2002.
- Imre Csiszár and Paul C. Shields. The consistency of the BIC Markov order estimator. *Ann. Statist.*, 28(6):1601–1619, 2000.

- Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.
- Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.
- 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.
- Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.
- Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001.
- Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.
- Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8):1501–1510, 2005.
- Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.
- Guillaume Lecué. *Méthodes d'agrégation : optimalité et vitesses rapides*. PhD thesis, LPMA, University Paris VII, May 2007.
- 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.
- 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.
- Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. Technical Report 6549, INRIA, 2008.
- Boris T. Polyak and Alexandre 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.

- Xiaotong Shen and Jianming Ye. Adaptive model selection. *J. Amer. Statist. Assoc.*, 97 (457):210–221, 2002.
- Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.
- 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.
- Nariaki 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.
- Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.
- Nicolas Verzelen. *Gaussian graphical models and Model selection*. PhD thesis, University Paris XI, December 2008.
- Fanny Villers. *Tests et sélection de modèles pour l’analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris XI, December 2007.