

Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression

Sylvain Arlot

Sylvain Arlot
CNRS ; Willow Project-Team
Laboratoire d'Informatique de l'Ecole Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France
e-mail: sylvain.arlot@ens.fr

Abstract: We consider the problem of choosing between several models in least-squares regression with heteroscedastic data. We prove that any penalization procedure is suboptimal when the penalty is proportional to the dimension of the model, at least for some typical heteroscedastic model selection problems. In particular, Mallows' C_p is suboptimal in this framework, as well as any "linear" penalty depending on both the data and their true distribution. On the contrary, optimal model selection is possible in this framework with data-driven penalties such as V -fold or resampling penalties (Arlot, 2008a,b). Therefore, estimating the "shape" of the penalty from the data is useful, even at the price of a higher computational cost.

AMS 2000 subject classifications: Primary 62G08; secondary 62G05, 62J05.

Keywords and phrases: non-parametric regression, model selection, penalization, heteroscedastic data, Mallows C_p , resampling penalties.

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 model fits 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, such as bootstrap penalties (Efron, 1983), resampling and V -fold penalties (Arlot, 2008a,b).

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 of "linear" penalties, that is penalties proportional to the dimension of the model. 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). The oracle inequalities of Birgé and Massart (2007) also apply to other data-driven linear penalties.

Nevertheless, Mallows' C_p and other linear penalties were only proved to be efficient for homoscedastic data, that is data such that the variance of the noise does not depend on the position in the feature space; such an assumption is unrealistic for many practical problems. This paper tackles the model selection problem when the noise-level varies over the feature space, that is *heteroscedastic data*. Mallows' C_p is empirically known to fail with heteroscedastic data, as showed for instance in a previous paper (Arlot, 2008a, Section 5).

Resampling-based model selection procedures are natural candidates for handling heteroscedastic data; among many examples, let us mention cross-validation (Allen, 1974; Stone, 1974), V -fold cross-validation (Geisser, 1975), resampling penalties (Efron, 1983; Arlot, 2008a) and V -fold penalties (Arlot, 2008b). In particular, resampling and V -fold penalties satisfy a non-asymptotic oracle inequality with leading constant C_n tending to 1 when n tends to infinity for selecting among regressogram estimators when data are heteroscedastic (Arlot, 2008a,b).

Compared to linear penalties, resampling methods can be computationally costly. The goal of this paper is to prove that the additional computational cost of resampling methods actually yields a better model selection efficiency than any linear penalization procedure.

More precisely, Theorem 1 shows a typical heteroscedastic model selection problem for which the excess loss of the estimator selected by any linear penalization procedure is larger than the excess loss of the oracle multiplied by an absolute constant $\kappa > 1$ with large probability (Section 3.1). In particular, *no linear penalty can be asymptotically optimal* for this model selection problem, even a penalty for which the multiplicative constant \hat{K} depends on the true distribution of the data.

Theorem 1 is a *non-asymptotic* result, meaning in particular that the collection of models is allowed to depend on the sample size 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 not to assume that the regression function is described with a small number of parameters.

The reason why such a strong negative result holds is that the ideal penalty is highly non-linear in the dimension of the models when data are heteroscedastic (Section 3.2). Hence, as soon as the collection of models is rich enough (in particular not restricted to regular histograms), any model that can be selected by a linear penalty yields an excess loss larger than the one of the oracle multiplied by $\kappa > 1$.

On the contrary, resampling and V -fold penalties satisfy a non-asymptotic oracle inequality with leading constant C_n tending to 1 when n tends to infinity for the model selection problem considered in Theorem 1 (Section 3.3). In particular, these penalties are more efficient than any linear penalty with large probability, at least when the sample size is large enough.

Note that the suboptimality of data-driven linear penalties for some heteroscedastic problems is highly intuitive and certainly known empirically by many practitioners. Indeed, when data are heteroscedastic, different parameters of the model are estimated with different uncertainties. Therefore, penalizing each parameter in the same way seems a poor strategy. Nevertheless, no theoretical result like Theorem 1 has ever been proved, up to the best of our knowledge, certainly because it requires precise non-asymptotic concentration inequalities which have only been proved recently (see Section 6.3).

Moreover, Theorem 1 is stronger than a minimax suboptimality result, at least from two aspects. First, Theorem 1 is not restricted to *data-driven* linear penalties: even linear penalization procedures using the knowledge of the true distribution are suboptimal. Second, since the proof of Theorem 1 applies to almost any heteroscedastic model selection problem, it proves that linear penalties are suboptimal for *each* of these problems with large probability, which is much stronger than being suboptimal in worst case. These are strong arguments against the practical use of linear penalties.

In Section 4, a simulation study shows that the negative result of Theorem 1 is not restricted to one particular model selection problem or to large sample sizes. Linear penalties probably fail to attain asymptotic optimality for almost any model selection problem in heteroscedastic regression, whereas resampling-based penalties generally perform better, even for small sample sizes.

Therefore, when data are heteroscedastic, whatever the sample size, the computational cost of resampling-based penalties is compensated by a significant improvement of model selection performance compared to linear penalties.

2. Framework

In this section, we describe the least-squares regression framework, model selection and the penalization approach.

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}^k . 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 $s(x) = \mathbb{E}[Y | X = x]$, 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 on 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*), the best predictor over S_m is defined by

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

The empirical counterpart of s_m is defined by

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

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

2.2. Model selection, penalization

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

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\} + R_n \quad (2)$$

in expectation or with 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

$$\hat{m} \in \arg \min_{m \in \mathcal{M}_n} \{ \text{crit}(m) \} \quad \text{with} \quad \text{crit}(m) := P_n \gamma(\hat{s}_m) + \text{pen}(m) . \quad (3)$$

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 $m \in \mathcal{M}_n$.

When $\text{Card}(\mathcal{M}_n) \leq Kn^\alpha$ for some $K, \alpha < \infty$ and $\text{pen}(m)$ is a good estimator of the ideal penalty $\text{pen}_{\text{id}}(m)$ for every $m \in \mathcal{M}_n$, then \hat{m} satisfies an oracle inequality (2) with large probability, with leading constant C close to 1 and $R_n \ll n^{-1}$ (see Arlot and Massart, 2008, for instance).

A classical penalization procedure in the least-squares regression framework is Mallows' C_p (Mallows, 1973). For every $m \in \mathcal{M}_n$, Mallows' penalty is defined as

$$\text{pen}(m) := \frac{2\sigma^2 D_m}{n} ,$$

where D_m is the dimension of the model S_m as a vector space and the noise-level $\sigma(\cdot)$ is assumed to be constant equal to σ . Various optimality results for Mallows' C_p have been proved, as noticed in Section 1, always assuming the noise-level to be constant. Note also that when σ is constant but unknown, it can be estimated from the data by

$$\hat{\sigma}^2 := \frac{d^2(Y_{1\dots n}, S_{\lfloor n/2 \rfloor})}{n - \lfloor n/2 \rfloor} , \quad (4)$$

where $Y_{1\dots n} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, d is the Euclidean distance on \mathbb{R}^n and $S_{\lfloor n/2 \rfloor}$ is a model (that is, a linear subspace of \mathbb{R}^n) of dimension $\lfloor n/2 \rfloor$. Baraud (2000, Section 6) proved an oracle inequality like (2) for Mallows' C_p when σ^2 is estimated by (4).

3. Main results

In this section, we first describe a typical heteroscedastic model selection problem for which any linear penalty—that is of the form $\text{pen}(m) = \hat{K}D_m$ where \hat{K} is allowed to depend on P and P_n —fails to attain asymptotic optimality (Section 3.1, Theorem 1). The main reason for this failure is that the ideal penalty is not a linear function of the dimension when data are heteroscedastic, as explained in Section 3.2. Finally, results proved in previous papers (Arlot, 2008a,b) are recalled in Section 3.3, showing that several resampling-based penalties satisfy with large probability a non-asymptotic oracle inequality (2) with leading constant $C = C_n$ tending to 1 when n tends to infinity in the framework of Theorem 1.

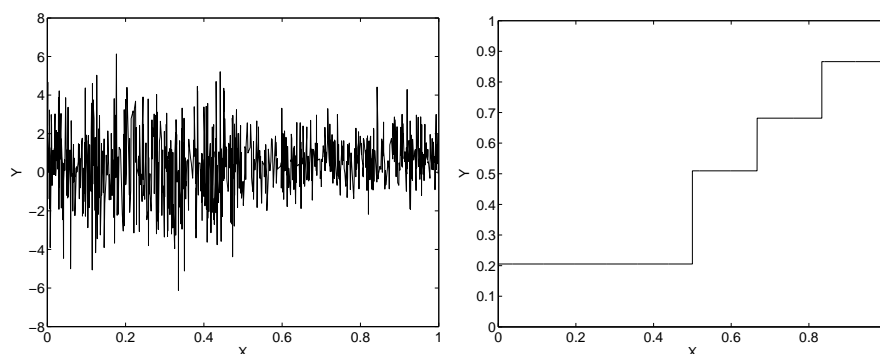


FIG 1. Framework of Theorem 1. Left: one data sample of size $n = 1000$, with $\epsilon_i \sim \mathcal{U}([- \sqrt{3}; \sqrt{3}])$. Right: The corresponding oracle estimator (the scales are different on the y-axis).

3.1. Suboptimality of linear penalization

Let us consider the framework of Section 2 with $\mathcal{X} = [0, 1]$ and assume the following. The data $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent and identically distributed (i.i.d.). For every i , X_i has a uniform distribution over \mathcal{X} and (X_i, Y_i) satisfies (1) with $s(x) = x$, that is $Y_i = X_i + \sigma(X_i)\epsilon_i$, where

$$\sigma(x) = \begin{cases} 2 & \text{if } x < \frac{1}{2} \\ 1 & \text{otherwise} \end{cases}$$

and $(\epsilon_i)_{1 \leq i \leq n}$ are i.i.d., independent of $(X_i)_{1 \leq i \leq n}$ and satisfy $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = 1$ and $\|\epsilon_i\|_\infty \leq 10$. For instance, ϵ_i uniformly distributed over $[-\sqrt{3}; \sqrt{3}]$ satisfies these properties. Such a data sample is represented on Figure 1 (left panel).

The collection of models $(S_m)_{m \in \mathcal{M}_n}$ is defined by

$$\mathcal{M}_n = \left\{ (D_1, D_2) \text{ s.t. } 1 \leq D_1, D_2 \leq \frac{n}{2(\ln n)^2} \right\}$$

and for every $D_1, D_2 \in \mathbb{N} \setminus \{0\}$, $S_{(D_1, D_2)}$ is the model of piecewise-constant functions on the partition

$$\left\{ \left[\frac{k-1}{2D_1}; \frac{k}{2D_1} \right) \text{ s.t. } 1 \leq k \leq D_1 \right\} \cup \left\{ \left[\frac{D_2+k-1}{2D_2}; \frac{D_2+k}{2D_2} \right) \text{ s.t. } 1 \leq k \leq D_2 \right\}.$$

In other words, $(S_m)_{m \in \mathcal{M}_n}$ is the collection of histogram models which are regular on $[0, 1/2]$ and regular on $[1/2, 1]$ with two possibly different bin sizes. This collection is quite natural since it allows to adapt the bin size to the local noise-levels when data are heteroscedastic, which holds in this particular framework. For instance, on Figure 1 (right panel), the oracle estimator uses

a smaller bin size on $[1/2, 1]$ than on $[0, 1/2]$. Remark also that $\text{Card}(\mathcal{M}_n) \leq (n/(2(\ln n)^2))^2 \leq n^2$. Therefore, as explained in Section 2.2, penalization procedures using an estimator of $\text{pen}_{\text{id}}(m)$ for every $m \in \mathcal{M}_n$ as a penalty are relevant.

The main result of this paper is that any linear penalization procedure fails to attain asymptotic optimality for the above model selection problem.

Theorem 1. *There exist absolute constants $C, \eta > 0$ and an event of probability at least $1 - Cn^{-2}$ on which for every $K \geq 0$ and every*

$$\begin{aligned} \widehat{m}(K) &\in \arg \min_{m \in \mathcal{M}_n} \{P_n \gamma(\widehat{s}_m) + KD_m\} \ , \\ \ell\left(s, \widehat{s}_{\widehat{m}(K)}\right) &\geq (1 + \eta) \inf_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\} \ . \end{aligned} \quad (5)$$

Theorem 1 is proved in Section 6.4. Theorem 1 is a quite strong result: no linear penalization method can be asymptotically optimal for this model selection problem, even a data-dependent method using the knowledge of s and σ ! In particular, defining the *ideal linear penalization procedure* by

$$m_{\text{lin}}^* \in \widehat{m}(K^*) \quad \text{where} \quad K^* \in \arg \min_{K > 0} \left\{ \ell\left(s, \widehat{s}_{\widehat{m}(K)}\right) \right\} \ , \quad (6)$$

Theorem 1 proves that the choice m_{lin}^* is asymptotically suboptimal.

The proof of Theorem 1 relies on the fact that linear penalties can only select a small number of models, which was previously noticed by Breiman (1992, Section 5), who characterized the so-called ‘‘RSS-extreme submodels’’ $(\widehat{m}(K))_{K>0}$. Whereas Breiman stated that this limitation can be benefic from the computational point of view in some cases, Theorem 1 shows that it can also induce suboptimality when data are heteroscedastic.

Indeed, as illustrated by Figures 2 and 3, the oracle model

$$m^* \in \arg \min_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\} \quad (7)$$

is usually far from the path $(\widehat{m}(K))_{K>0}$ of models that can be selected with a linear penalty, hence from m_{lin}^* ; therefore, the excess loss of m_{lin}^* cannot be asymptotically equivalent to the one of the oracle model m^* . Note that the picture is even clearer in three other frameworks considered in Section 4 (see Figure 7).

Let us now add a few comments. First, the right-hand side of (5) is of order $n^{-2/3}$. Hence, no oracle inequality (2) can be proved with a constant C tending to one when n tends to infinity and a remainder term $R_n \ll n^{-2/3}$.

Second, results similar to Theorem 1 hold for several model selection problems with heteroscedastic data:

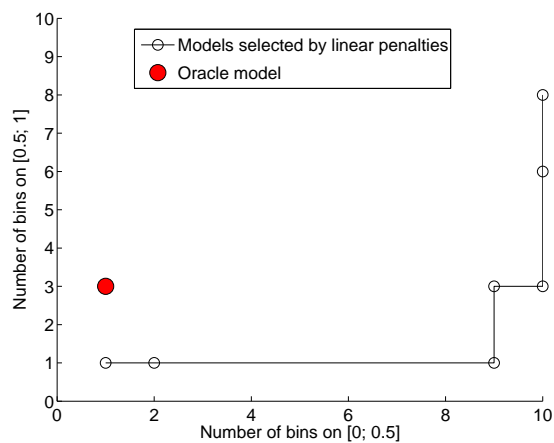


FIG 2. Framework of Theorem 1 with a sample size $n = 1000$ and $\epsilon_i \sim \mathcal{U}([- \sqrt{3}; \sqrt{3}])$: The oracle model m^* does not belong to the path $(\hat{m}(K))_{K>0}$ of models that can be selected with a linear penalty.

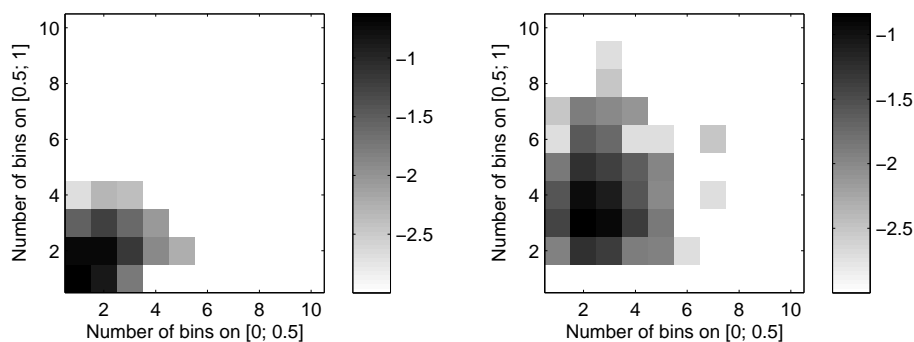


FIG 3. Framework of Theorem 1 with a sample size $n = 1000$ and $\epsilon_i \sim \mathcal{U}([- \sqrt{3}; \sqrt{3}])$. Left: $\log_{10} \mathbb{P}(m = m_{lin}^*)$ represented in \mathbb{R}^2 using $(D_{m,1}, D_{m,2})$ as coordinates, where m_{lin}^* is defined by (6); $N = 1000$ samples have been simulated for estimating the probabilities. Right: $\log_{10} \mathbb{P}(m = m^*)$ using the same representation and the same $N = 1000$ samples.

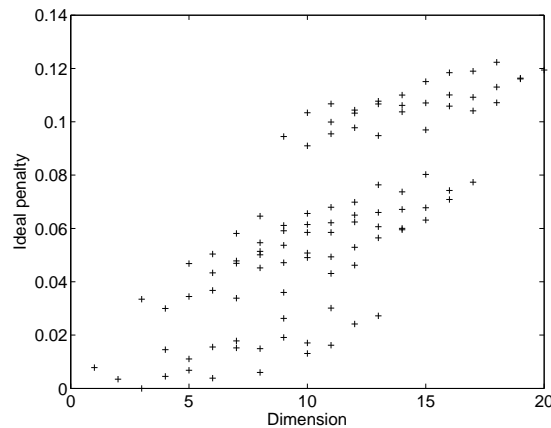


FIG 4. *Ideal penalty vs. D_m in the framework of Theorem 1, with a sample size $n = 1000$ and $\epsilon_i \sim \mathcal{U}([- \sqrt{3}; \sqrt{3}])$. A similar picture holds in expectation.*

- Theorem 1 is actually proved for any function σ satisfying $\int_0^{1/2} (\sigma(x))^2 dx \neq \int_{1/2}^1 (\sigma(x))^2 dx$ (the constant η depending only on the ratio between these two quantities), and $\|\epsilon_i\|_\infty$ can take any finite value. Therefore, it does not depend on any particular distribution of the errors.
- Concentration inequalities proved in a previous paper (Arlot, 2008a) show that the same result holds with unbounded noises ϵ (for instance Gaussian).
- The proof of Theorem 1 can be adapted to any regression function, provided that the bias $\ell(s, s_{D_1, D_2})$ remains close to $\alpha_1 D_1^{-2} + \alpha_2 D_2^{-2}$ when $\min(D_1, D_2)$ is large, which holds at least when s is continuously differentiable with

$$\alpha_1 = \frac{\|s'\|_{L^2([0,1/2])}^2}{96} \quad \text{and} \quad \alpha_2 = \frac{\|s'\|_{L^2([1/2,1])}^2}{96} .$$

In short, linear penalties are suboptimal for most heteroscedastic model selection problems, and we only chose this particular problem in the statement of Theorem 1 in order to keep the proof as simple as possible.

3.2. Non-linearity of the ideal penalty

The main argument used for proving Theorem 1 is that the ideal penalty is far from being a linear function of D_m when data are heteroscedastic. In the framework of Theorem 1, the non-linearity of $\text{pen}_{\text{id}}(m)$ as a function of D_m is illustrated by Figure 4. In order to understand better the drawbacks of linear penalties, we comment this point in this subsection.

Let S_m be the set of piecewise constant functions on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of \mathcal{X} . Then, the concentration inequalities of Section 6.3 show that for most of

the models, the ideal penalty is close to its expectation. Moreover, the expectation of the ideal penalty can be computed explicitly thanks to Proposition 1, first proved in a previous paper (Arlot, 2008b):

$$\mathbb{E}[\text{pen}_{\text{id}}(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} (2 + \delta_{n,p_\lambda}) \left((\sigma_\lambda^r)^2 + (\sigma_\lambda^d)^2 \right) \quad (8)$$

where

$$\begin{aligned} (\sigma_\lambda^r)^2 &:= \mathbb{E} \left[(Y - s(X))^2 \mid X \in I_\lambda \right] = \mathbb{E} \left[(\sigma(X))^2 \mid X \in I_\lambda \right] \\ (\sigma_\lambda^d)^2 &:= \mathbb{E} \left[(s(X) - s_m(X))^2 \mid X \in I_\lambda \right] \\ p_\lambda &:= \mathbb{P}(X \in I_\lambda) \quad \text{and} \quad |\delta_{n,p}| \leq \min \left\{ L_1, \frac{L_2}{(np)^{1/4}} \right\} \end{aligned}$$

for some absolute constants L_1, L_2 . The classical justification of Mallows' C_p is that when the X_i are deterministic and $\sigma(\cdot)$ is constant equal to σ ,

$$\mathbb{E}[\text{pen}_{\text{id}}(m)] = \frac{2\sigma^2 D_m}{n} = \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(D_m p_\lambda (\sigma_\lambda^r)^2 \right). \quad (9)$$

In general, there can be three differences between (9) and (8):

1. $\sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2 \neq \sum_{\lambda \in \Lambda_m} \left(D_m p_\lambda (\sigma_\lambda^r)^2 \right)$ except when either $\sigma(\cdot)$ is constant or $p_\lambda = D_m^{-1}$ for every $\lambda \in \Lambda_m$. Hence, when the noise-level is far from being constant, the ideal penalty is far from being proportional to the dimension of the models in general. In the framework of Theorem 1, the collection of models is rich enough so that $F(m) = D_m^{-1} \sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2$ varies over \mathcal{M}_n .
2. $(\sigma_\lambda^d)^2$ appears in (8) but not in (9); this term can be large if s is far from s_m , that is when s is highly non-smooth or D_m is small.
3. The term δ_{n,p_λ} is not exactly zero, especially when np_λ is small (see Section 6.2).

In Theorem 1, the only term making linear penalties fail is $\sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2$, which is not proportional to $D_m = D_{m,1} + D_{m,2}$ but to $2D_{m,1} + D_{m,2}$. It is not clear whether the two other differences between (9) and (8) can be sufficient to make linear penalties fail with homoscedastic data. Indeed, linear penalties have been proved to be asymptotically optimal for homoscedastic regression under mild assumptions, even when the design is random (see Baraud, 2002, and references given in Section 1). Therefore, $(\sigma_\lambda^d)^2$ and δ_{n,p_λ} may only have to be taken into account in the penalty for finite sample sizes.

3.3. Comparison with resampling-based penalties

In the framework of Theorem 1, both V -fold penalties (Arlot, 2008b) and “exchangeable” resampling penalties (Efron, 1983; Arlot, 2008a) are asymptotically

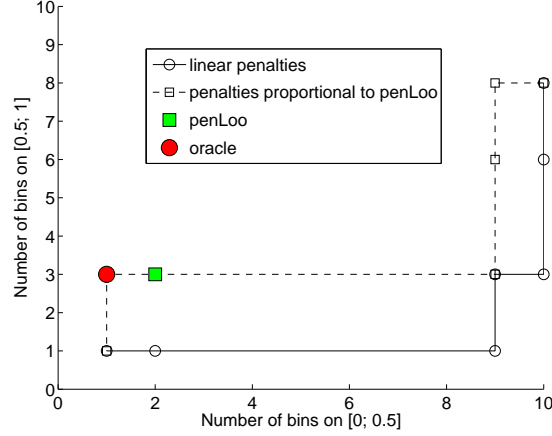


FIG 5. Framework of Theorem 1 with a sample size $n = 1000$ and $\epsilon_i \sim \mathcal{U}([- \sqrt{3}; \sqrt{3}])$: The path of models that can be selected with penalties proportional to penLoo is closest to the oracle than the path $(\hat{m}(K))_{K>0}$ of models that can be selected with a linear penalty.

optimal, despite the fact that they are general-purpose penalties. Results proved for V -fold penalties in a previous paper (Arlot, 2008b) are recalled in this subsection.

First, let us define V -fold penalties. Let $(B_j)_{1 \leq j \leq V}$ be a fixed partition of $\{1, \dots, n\}$ such that $|\text{Card}(B_j) - n/V| < 1$ for every j . For every j , define

$$P_n^{(-j)} = \frac{1}{n - \text{Card}(B_j)} \sum_{i \notin B_j} \delta_{(X_i, Y_i)}$$

and $\forall m \in \mathcal{M}_n, \hat{s}_m^{(-j)} \in \arg \min_{t \in S_m} \{P_n^{(-j)} \gamma(t)\}$.

Then, the V -fold penalty is defined by

$$\text{pen}_{\text{VF}}(m) := \frac{V-1}{V} \sum_{j=1}^V (P_n - P_n^{(-j)}) \gamma(\hat{s}_m^{(-j)}) . \quad (10)$$

As proved in a previous paper (Arlot, 2008b, Theorem 2), in the framework of Theorem 1, a constant C'_V (depending only on V) and an event of probability at least $1 - C'_V n^{-2}$ exist on which, for every

$$\hat{m}_{\text{penVF}} \in \arg \min_{m \in \mathcal{M}_n} \{P_n \gamma(\hat{s}_m) + \text{pen}_{\text{VF}}(m)\} ,$$

$$\ell(s, \hat{s}_{\hat{m}_{\text{penVF}}}) \leq \left(1 + (\ln n)^{-1/5}\right) \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\} .$$

In other words, provided that n is large enough, V -fold penalties perform strictly better than any linear penalty with large probability. Therefore, estimating the

shape of the ideal penalty¹ improves significantly the efficiency of the penalization procedure when data are heteroscedastic. Hence, the increased computational cost of V -fold penalties compared to Mallows' C_p is the price to pay for a versatile penalization procedure.

Moreover, the advantage of estimating the shape of the penalty by resampling compared to using the dimension of the model as a shape is clear from Figure 5. Indeed, the path of models $(\widehat{m}(K))_{K>0}$ that can be selected with linear penalties stays far from the oracle because when K decreases, the number of bins on $[0, 1/2]$ increases before the number of bins on $[1/2, 1]$. Intuitively, this happens because given the total number of bins, putting more bins on $[0, 1/2]$ where the noise-level is high decreases more the empirical risk.

On the contrary, penalties proportional to the Leave-one-out penalty (that is, the V -fold penalty with $V = n$ and $B_j = \{j\}$ for every j , denoted by penLoo) take into account the variations of the noise level over $[0, 1]$. Therefore, the path of models selected with penalties proportional to penLoo goes in the opposite direction: when K decreases, the number of bins on $[1/2, 1]$ increases first.

4. Simulation study

In this section, the suboptimality of linear penalties (Theorem 1) is illustrated by a short simulation study. In addition to the framework of Theorem 1 (called “Xu2-1”), we consider three model selection problems, called “X1-005”, “S0-1” and “HSd2”.

Data are generated according to (1), where n, s, σ and $\mathcal{L}(\epsilon_i)$ are given for each experiment in Table 1. The regression function and one particular sample for each experiment except Xu2-1 are plotted on Figure 6. The collection of models $(S_m)_{m \in \mathcal{M}_n}$ is defined as in Section 3.1 for X1-005 and S0-1, except that the maximal number of bins in each half of $[0, 1]$ is $n/(2 \ln n)$ instead of $n/(2(\ln n)^2)$, in order to keep a sufficiently large amount of models with $n = 200$. For reducing the computational cost, $(S_m)_{m \in \mathcal{M}_n}$ is restricted in HSd2 to “dyadic” partitions, that is

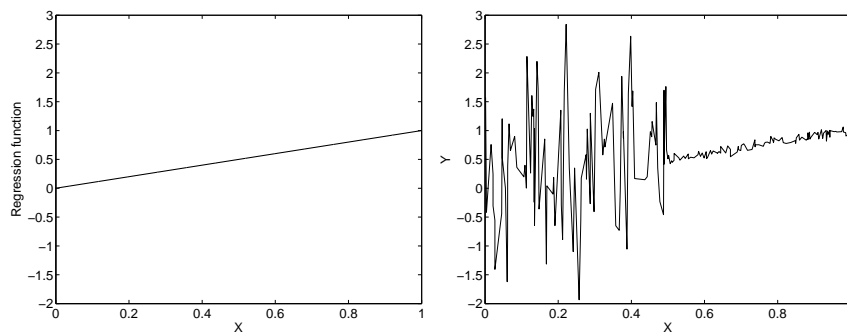
$$\mathcal{M}_n = \left\{ (2^{k_1}, 2^{k_2}) \text{ s.t. } 1 \leq 2^{k_1}, 2^{k_2} \leq \frac{n}{2} \right\}$$

with the notation of Section 3.1. In each experiment, the model of constant functions on $[0, 1]$ is also added to $(S_m)_{m \in \mathcal{M}_n}$.

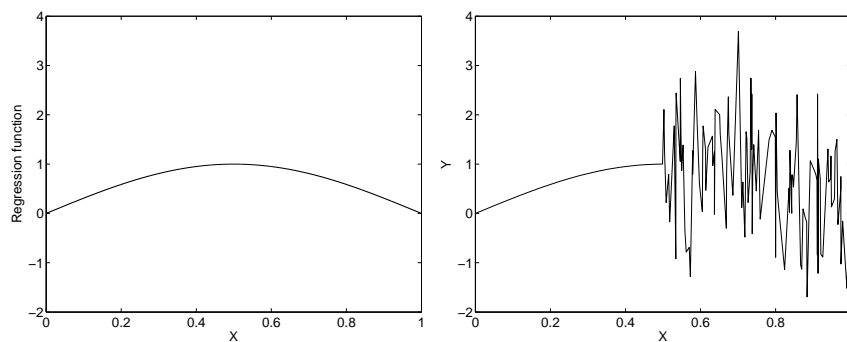
In each framework, $N = 1000$ independent data samples are generated; for each sample, the following model selection procedures are compared:

- Mallows' C_p procedure (Mal), defined as in Section 2.2, estimating the mean variance by (4) with $S_{n/2}$ defined as follows. First, determine the permutation τ of $\{1, \dots, n\}$ such that $(X_{\tau(i)})_{1 \leq i \leq n}$ is nondecreasing.

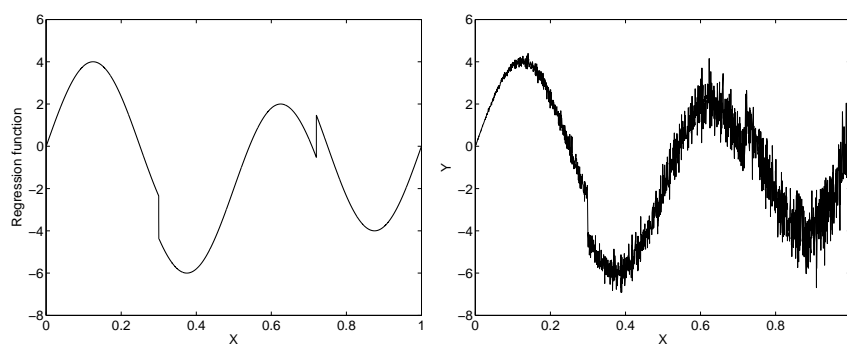
¹The shape of the ideal penalty is defined as the way it depends on m up to an increasing linear transformation.



Experiment X1-005



Experiment S0-1



Experiment HSd2 (see Donoho and Johnstone, 1995, for a definition of the HeaviSine function)

FIG 6. *Left: Regression functions. Right: One particular data sample.*

Second, let $S_{n/2}$ be generated by the family $(e_{\tau(2i-1)} + e_{\tau(2i)})_{1 \leq i \leq n/2}$ where e_j is the j -th vector of the canonical basis of \mathbb{R}^n . In other words,

$$\hat{\sigma} := \frac{1}{n} \sum_{i=1}^{n/2} (Y_{\tau(2i)} - Y_{\tau(2i-1)})^2 .$$

- Mal with a penalty multiplied by a factor $C_{\text{ov}} \in \{1.25; 2\}$, in order to test for overpenalization.
- The ideal linear penalization procedure (IdLin) defined by (6).
- Leave-one-out penalization (penLoo), that is the V -fold penalty defined by (10) with $V = n$ and $B_j = \{j\}$ for every j .
- penLoo with a penalty multiplied by a factor $C_{\text{ov}} \in \{1.25; 2\}$, in order to test for overpenalization.

In addition, before performing each procedure, the models S_m with less than 2 data points in one piece of their associated partition are removed from the family $(S_m)_{m \in \mathcal{M}_n}$; this intends to make the comparison between leave-one-out penalties (which require this preliminary step) and other procedures clearer.

The signal-to-noise ratio is rather small in the four experimental settings considered here, and the collection of models is quite large. Therefore, we can expect overpenalization to be necessary, especially for Xu2-1, X1-005 and S0-1 (see Arlot, 2008a, Section 7.3.2, for more details on overpenalization).

Then, the model selection performance of each procedure is evaluated by the following benchmark:

$$C_{\text{or}} := \frac{\mathbb{E} [\ell(s, \hat{s}_m)]}{\mathbb{E} [\inf_{m \in \mathcal{M}_n} \ell(s, \hat{s}_m)]} . \quad (11)$$

Basically, C_{or} is the constant that should appear in an oracle inequality (2) holding in expectation with $R_n = 0$. The values of C_{or} evaluated for each procedure in each experiment are reported in Table 1.

The conclusion of this simulation study is that the failure of linear penalties actually occurs for finite sample sizes, and not only in the framework of Theorem 1. Indeed, in experiments X1-005, S0-1 and HSd2, even the ideal linear penalization procedure (IdLin) performs significantly worse than the Leave-one-out penalty (penLoo) multiplied by the right (deterministic) overpenalization factor C_{ov} . Estimating C_{ov} from data may not be easy in general, but some proposals have been made for instance in the author's Ph.D. thesis (Arlot, 2007, Section 11.3.3).

The reason why linear penalties fail can be visualized on Figure 7, which is the equivalent of Figure 3 for experiments X1-005, S0-1 and HSd2. The distribution of the oracle model m^* is almost disjoint from the distribution of m_{lin}^* . Importantly, this phenomenon happens for various regression functions (even non-smooth ones) and various kinds of heteroscedastic noises. Therefore,

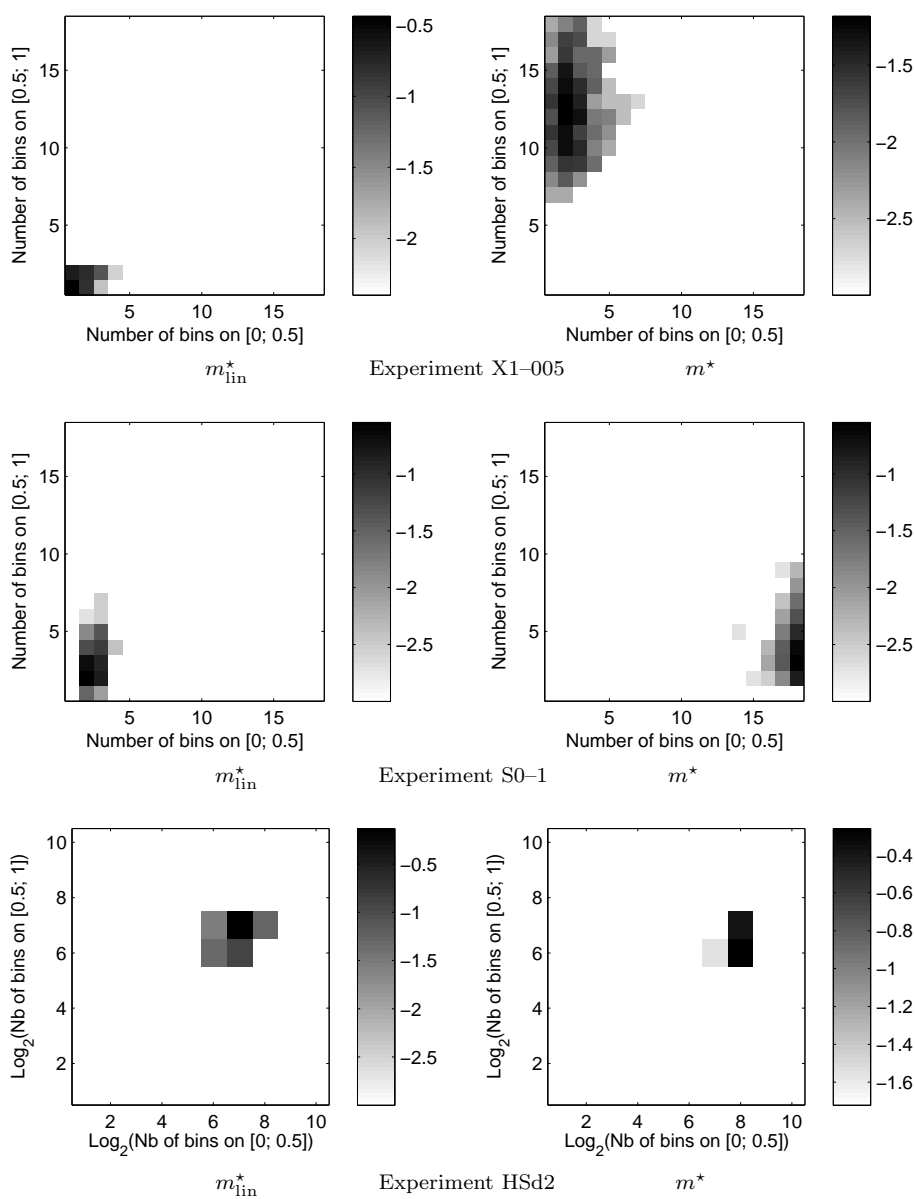


FIG 7. Same as Figure 3 for the three other experiments. Left: $\log_{10} \mathbb{P}(m = m_{lin}^*)$ represented in \mathbb{R}^2 using $(D_{m,1}, D_{m,2})$ as coordinates, where m_{lin}^* is defined by (6); $N = 1000$ samples have been simulated for estimating the probabilities. Right: $\log_{10} \mathbb{P}(m = m^*)$ using the same representation and the same samples. The distributions of m^* and m_{lin}^* are almost disjoint.

TABLE 1
Accuracy indices C_{or} for each algorithm in two experiments, \pm a rough estimate of uncertainty of the value reported (that is the empirical standard deviation divided by \sqrt{N}). In each column, the more accurate algorithms (taking the uncertainty into account) are bolded.

Experiment	Xu2-1	X1-005	S0-1	HSd2
n	1 000	200	200	2048
$s(x)$	x	x	$\sin(\pi x)$	HeaviSine(x)
$\sigma(x)$	$1 + \mathbf{1}_{x < 1/2}$	$\frac{1}{20} + \frac{19}{20} \mathbf{1}_{x < 1/2}$	$\mathbf{1}_{x \geq 1/2}$	x
$\mathcal{L}(\epsilon_i)$	$\mathcal{U}([- \sqrt{3}; \sqrt{3}])$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$
Mal	3.419 \pm 0.066	9.571 \pm 0.199	6.500 \pm 0.124	1.525 \pm 0.012
Mal \times 1.25	3.051 \pm 0.064	7.898 \pm 0.204	5.411 \pm 0.121	1.373 \pm 0.010
Mal \times 2	2.435 \pm 0.048	4.504 \pm 0.142	3.363 \pm 0.076	1.527 \pm 0.004
IdLin	1.650 \pm 0.023	2.039 \pm 0.032	2.132 \pm 0.030	1.207 \pm 0.005
penLoo	2.523 \pm 0.056	3.207 \pm 0.115	2.439 \pm 0.065	1.171 \pm 0.004
penLoo \times 1.25	2.164 \pm 0.047	2.496 \pm 0.087	2.063 \pm 0.048	1.158 \pm 0.003
penLoo \times 2	1.882 \pm 0.031	1.803 \pm 0.047	1.986 \pm 0.037	1.157 \pm 0.003

the suboptimality result of Theorem 1 is certainly valid for a wide range of model selection problems with heteroscedastic data.

Let us emphasize that performing as well as IdLin is far from being easy for a data-driven procedure. Therefore, penLoo should mainly be compared with Mal in terms of model selection performance, leaving the choice of the overpenalization factor free for both procedures. Table 1 clearly shows that Mal fails to select a correct model for heteroscedastic model selection problems, that is heteroscedastic data and a collection $(S_m)_{m \in \mathcal{M}_n}$ allowing to take into account the variations of the noise level. Moreover, the performance gap between Mal and penLoo is large in the four experiments and not only due to a lack of overpenalization. Indeed, for a wide range of overpenalization factors, the estimated model selection performance C_{or} of Mal is uniformly much worse than the one of penLoo.

Finally, another illustration of the suboptimality of linear penalties for heteroscedastic model selection problems is provided by Figure 8.

When the sample size increase, the model selection performance of IdLin remains approximately constant (close to 2) while the model selection performance of penLoo constantly decreases (with $C_{\text{ov}} = 1.25$ because overpenalization is still needed for $n = 3\,000$ and we could not consider larger sample sizes for computational reasons). This illustrates perfectly, with finite sample sizes, the difference between IdLin and penLoo which emerges from the comparison between Theorem 1 and the optimality results on resampling-based penalties recalled in Section 3.3.

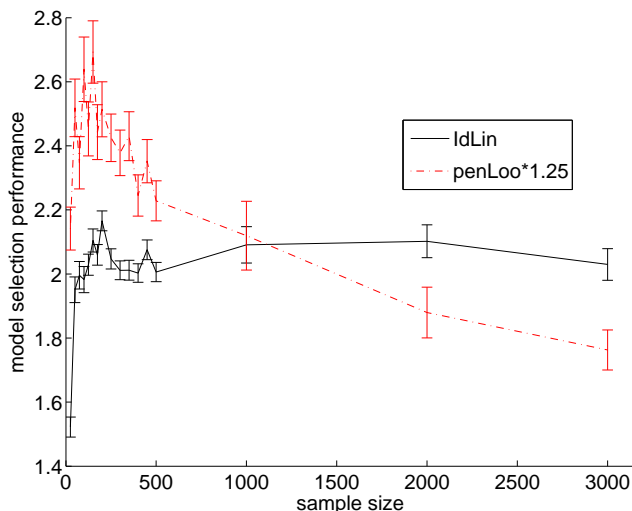


FIG 8. C_{or} as a function of n for experiment X1-005. Each estimated value of C_{or} is obtained with $N = 1000$ data samples for $n < 1000$ and $N = 200$ data samples for $n \geq 1000$ for computational reasons. The error bars represent the empirical standard deviation divided by \sqrt{N} .

5. Conclusion

Both theoretical and experimental results from this paper show that penalties proportional to the dimension—especially Mallows' C_p —should not be used for heteroscedastic model selection. Indeed, assuming that models should only be penalized proportionally to their dimension is convenient from the computational point of view but definitely not adapted to heteroscedasticity. As soon as the collection of models is rich enough to take into account heteroscedasticity (the collections considered in the paper being typical examples), using a linear penalty restricts the choice among models which are all far from the oracle. Therefore, linear penalties have a fundamental drawback which can only be solved by estimating properly the shape of the ideal penalty.

As mentioned in Section 3.3, resampling penalties are natural candidates for model selection with heteroscedastic data, in particular because they are optimal in frameworks where linear penalties are suboptimal, as proved in previous papers (Arlot, 2008a,b). The simulation experiment of Section 4 confirms this advice by showing that the leave-one-out penalty strongly outperforms Mallows' C_p ; a properly calibrated leave-one-out penalty is even unbeatable by any linear penalty, even using the knowledge of the true distribution.

Several other estimators have been proposed for regression with heteroscedastic data, for instance by Galtchouk and Pergamenschikov (2008) and by Gendre (2008) using model selection. Nevertheless, these approaches mostly use model selection as a tool for estimation since they consider very particular collections

of models from which they select an estimator of the regression function. When the model selection problem is given *a priori* and heteroscedasticity of data makes it difficult, resampling penalization (or cross-validation) are certainly the only model selection methods that can be optimal with no information on the variations of the noise-level, up to the best of our knowledge.

To conclude, solving a model selection problem which is difficult because of heteroscedasticity requires at least the computational cost of V -fold penalties.

Once the shape of the penalty is properly estimated, the next question is how to calibrate the constant in front of the penalty. In the histogram case, the constant is known for V -fold and resampling penalties, at least when the sample size is large enough, but it may not be optimal for other kinds of models. A completely data-driven procedure has been proposed by Birgé and Massart (2007) with dimensionality-based penalties, and extended to general shapes by Arlot and Massart (2008). Theoretical results have been proved in these papers for either homoscedastic Gaussian data and general linear models or heteroscedastic data and histogram models.

Nevertheless, a calibration procedure proved to be asymptotically optimal for general models and heteroscedastic data is still needed. In addition, the optimal calibration of penalties for a fixed sample size really matters, as showed in the simulation study of Section 4; though, up to the best of our knowledge, this problem is still widely open.

6. Proof of Theorem 1

Before proving Theorem 1, let us define some notation and recall probabilistic results from other papers (Arlot, 2008a,b; Arlot and Massart, 2008) that are used in the proof.

6.1. Notation

In the rest of the paper, L denotes an absolute 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} .

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

$$\begin{aligned} p_1(m) &:= P\gamma(\widehat{s}_m) - P\gamma(s_m) & p_2(m) &:= P_n\gamma(s_m) - P_n\gamma(\widehat{s}_m) \\ \bar{\delta}(m) &:= (P_n - P)(\gamma(s_m) - \gamma(s)) \\ A_n(m) &:= \min_{\lambda \in \Lambda_m} \{ \text{Card} \{ i \text{ s.t. } X_i \in I_\lambda \} \} & B_n(m) &:= \min_{\lambda \in \Lambda_m} \{ n\mathbb{P}(X_i \in I_\lambda) \} . \end{aligned}$$

6.2. Probabilistic tools: expectations

Proposition 1 (Proposition 1 and Lemma 7, Arlot, 2008b). *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda \in \Lambda_m}$. Then,*

$$\mathbb{E}[p_1(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} (1 + \delta_{n,p_\lambda}) \sigma_\lambda^2 \quad (12)$$

$$\mathbb{E}[p_2(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \sigma_\lambda^2 \quad (13)$$

$$\text{where } \sigma_\lambda^2 := \mathbb{E} \left[(Y - s(X))^2 \mid X \in I_\lambda \right] = (\sigma_\lambda^d)^2 + (\sigma_\lambda^r)^2 ,$$

$p_\lambda = \mathbb{P}(X \in I_\lambda)$ and $\delta_{n,p}$ only depends on (n, p) . Moreover, $\delta_{n,p}$ is small when the product np is large:

$$|\delta_{n,p}| \leq \min \left\{ L_1, L_2(np)^{-1/4} \right\} ,$$

where L_1 and L_2 are absolute constants.

Note that $\delta_{n,p}$ can be made explicit:

$$\delta_{n,p} = np \mathbb{E} [Z^{-1} \mathbf{1}_{Z>0}] - 1$$

where Z is a binomial random variable with parameters (n, p) .

Remark 1. Since we deal with histograms, \hat{s}_m is not defined when $\min_{\lambda \in \Lambda_m} \hat{p}_\lambda = 0$, which occurs with positive probability. Therefore, a convention for $p_1(m)$ as to be chosen on the event $A_n(m) = 0$ (which has a small probability) so that $p_1(m)$ has a finite expectation (see Arlot, 2008b, for details). This convention is purely formal, since the statement of Theorem 1 does not involve the expectation of $p_1(m)$. The important point is that the same convention is used in Proposition 2 below.

6.3. Probabilistic tools: concentration inequalities

We state in this section some concentration results on the components of the ideal penalty, using for $p_1(m)$ the same convention as in Proposition 1.

Proposition 2 (Proposition 12, Arlot, 2008a). *Let $\gamma > 0$. Assume that $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n \geq 1$, $\|Y\|_\infty \leq A < \infty$ and*

$$D_m^{-1} \sum_{\lambda \in \Lambda_m} \mathbb{E} [\sigma(X)^2 \mid X \in I_\lambda] \geq Q > 0 .$$

Then, an event of probability at least $1 - Ln^{-\gamma}$ exists on which

$$p_1(m) \geq \mathbb{E}[p_1(m)] - L_{A,Q,\gamma} \left[(\ln n)^2 D_m^{-1/2} + e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad (14)$$

$$p_1(m) \leq \mathbb{E}[p_1(m)] + L_{A,Q,\gamma} \left[(\ln n)^2 D_m^{-1/2} + \sqrt{D_m} e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad (15)$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{A,Q,\gamma} D_m^{-1/2} \ln(n) \mathbb{E}[p_2(m)] . \quad (16)$$

Lemma 3 (Proposition 8, Arlot and Massart, 2008). *Assume that $\|Y\|_\infty \leq A < \infty$. Then for any $x \geq 0$, an event of probability at least $1 - 2e^{-x}$ exists on which*

$$\forall \eta > 0, \quad |\bar{\delta}(m)| \leq \eta \ell(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3}\right) \frac{A^2 x}{n}. \quad (17)$$

Lemma 4 (Lemma 12, Arlot, 2008b). *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 $(n; (p_\lambda)_{\lambda \in \Lambda_m})$, $\gamma > 0$. Assume that $\text{Card}(\Lambda_m) \leq n$ and $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n > 0$. Then, an event of probability at least $1 - Ln^{-\gamma}$ exists on which*

$$\min_{\lambda \in \Lambda_m} \{n\hat{p}_\lambda\} \geq \frac{\min_{\lambda \in \Lambda_m} \{np_\lambda\}}{2} - 2(\gamma + 1) \ln n. \quad (18)$$

6.4. Proof of Theorem 1

We actually prove a more general result, assuming only that ϵ and σ satisfy the following:

$$\|\epsilon\|_\infty \leq E < \infty \quad \text{and} \quad (\sigma_a)^2 \geq (1 + \epsilon)(\sigma_b)^2 \quad \text{for some } \epsilon > 0$$

where $(\sigma_a)^2 := \int_0^{1/2} (\sigma(x))^2 dx$ and $(\sigma_b)^2 := \int_{1/2}^1 (\sigma(x))^2 dx$.

Theorem 1 thus corresponds to the case

$$E = 10 \quad (\sigma_a)^2 = 2 \quad (\sigma_b)^2 = \frac{1}{2}.$$

In the following, $L_{(\mathbf{H})} = L_{E, \sigma_a, \sigma_b}$ denotes any constant depending on the above parameters only.

The above condition on $\sigma(\cdot)$ imply that the last assumption of Proposition 2 holds since

$$\begin{aligned} D_m^{-1} \sum_{\lambda \in \Lambda_m} \mathbb{E} [\sigma(X)^2 \mid X \in I_\lambda] &= \frac{2D_{m,1}(\sigma_a)^2 + 2D_{m,2}(\sigma_b)^2}{D_{m,1} + D_{m,2}} \\ &\geq 2 \min \left\{ (\sigma_a)^2, (\sigma_b)^2 \right\} =: Q > 0. \end{aligned}$$

Let us consider the ideal criterion

$$\text{crit}_{\text{id}}(m) := \ell(s, \hat{s}_m) + \infty \mathbf{1}_{A_n(m)=0} = \ell(s, s_m) + p_1(m) + \infty \mathbf{1}_{A_n(m)=0},$$

and for every $K \geq 0$ the linearly penalized empirical criterion

$$\begin{aligned} \text{crit}_K(m) &:= P_n \gamma(\hat{s}_m) + KD_m - P_n \gamma(s) + \infty \mathbf{1}_{A_n(m)=0} \\ &= \ell(s, s_m) - p_2(m) + KD_m + \bar{\delta}(m) + \infty \mathbf{1}_{A_n(m)=0}. \end{aligned}$$

The goal is to prove that whatever $K \geq 0$, any

$$\widehat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \{ \text{crit}_K(m) \}$$

satisfies

$$\text{crit}_{\text{id}}(\widehat{m}(K)) \geq \left(1 + \frac{1}{230}\right) \inf_{m \in \mathcal{M}_n} \{ \text{crit}_{\text{id}}(m) \}$$

with large probability.

Decomposition of the criteria For every $m \in \mathcal{M}_n$,

$$\begin{aligned} \text{crit}_{\text{id}}(m) &= \ell(s, s_m) + p_1(m) + \infty \mathbf{1}_{A_n(m)=0} \\ &= \ell(s, s_m) + \mathbb{E}[p_1(m)] + (p_1(m) - \mathbb{E}[p_1(m)]) + \infty \mathbf{1}_{A_n(m)=0} \end{aligned} \quad (19)$$

and for every $K \geq 0$ and $m \in \mathcal{M}_n$,

$$\begin{aligned} \text{crit}_K(m) &= \ell(s, s_m) - p_2(m) + K D_m + \bar{\delta}(m) + \infty \mathbf{1}_{A_n(m)=0} \\ &= \ell(s, s_m) - \mathbb{E}[p_2(m)] + K D_m + \infty \mathbf{1}_{A_n(m)=0} \end{aligned} \quad (20)$$

$$+ (\mathbb{E}[p_2(m)] - p_2(m) + \bar{\delta}(m)) . \quad (21)$$

Bounds on $A_n(m)$ and $B_n(m)$ For every $D_1, D_2 \geq 1$,

$$B_n(D_1, D_2) = \min \left\{ \frac{n}{D_1}, \frac{n}{D_2} \right\} .$$

Using Lemma 4, for every D_1, D_2 , an event of probability at least $1 - Ln^{-4}$ exists such that

$$A_n(D_1, D_2) \geq \frac{B_n(D_1, D_2)}{2} - 10 \ln n \geq \min \left\{ \frac{n}{2D_1}, \frac{n}{2D_2} \right\} - 10 \ln n .$$

This lower bound is positive provided that $\max\{D_1, D_2\} \leq \frac{n}{11 \ln n}$, which holds if $(D_1, D_2) \in \mathcal{M}_n$ and $n \geq L$. The intersection Ω_1 of these $\text{Card}(\mathcal{M}_n) \leq n^2$ events has a probability larger than $1 - Ln^{-2}$.

Computation of the main terms In the decompositions (19) and (21), we have splitted both criteria into deterministic terms and random remainder terms. We start here by computing explicitly the deterministic terms, which are the most important ones. First, the bias is:

$$\begin{aligned} \ell(s, s_{(D_1, D_2)}) &= \mathbb{E} \left[(s_{(D_1, D_2)}(X) - s(X))^2 \right] \\ &= \int_0^{1/2} (s_{(D_1, D_2)}(x) - s(x))^2 dx + \int_{1/2}^1 (s_{(D_1, D_2)}(x) - s(x))^2 dx \\ &= D_1 \int_0^{1/(2D_1)} \left(\frac{1}{4D_1} - x \right)^2 dx + D_2 \int_0^{1/(2D_2)} \left(\frac{1}{4D_2} - x \right)^2 dx \\ &= \frac{1}{96D_1^2} + \frac{1}{96D_2^2} . \end{aligned}$$

In order to compute $\mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)]$, we use Proposition 1, and the fact that

$$\begin{aligned}\sigma_\lambda^2 &= \mathbb{E} \left[(Y - s(X))^2 \mid X \in I_\lambda \right] \\ &= \mathbb{E} \left[(s(X) - s_m(X))^2 \mid X \in I_\lambda \right] + \mathbb{E} \left[(\sigma(X))^2 \mid X \in I_\lambda \right] \\ &= \frac{1}{\text{Leb}(I_\lambda)} \mathbb{E} \left[(s(X) - s_m(X))^2 \mathbf{1}_{X \in I_\lambda} \right] + \frac{1}{\text{Leb}(I_\lambda)} \mathbb{E} \left[(\sigma(X))^2 \mathbf{1}_{X \in I_\lambda} \right] .\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}[p_1(D_1, D_2)] &= \frac{1}{n} \sum_{\lambda \in \Lambda_m} (1 + \delta_{n, p_\lambda}) \sigma_\lambda^2 \\ &= \frac{2D_1 (1 + \delta_{n, (2D_1)^{-1}})}{n} \int_0^{1/2} \left[(s(X) - s_{(D_1, D_2)}(x))^2 + (\sigma(x))^2 \right] dx \\ &\quad + \frac{2D_2 (1 + \delta_{n, (2D_2)^{-1}})}{n} \int_{1/2}^1 \left[(s(X) - s_{(D_1, D_2)}(x))^2 + (\sigma(x))^2 \right] dx \\ &= \frac{2D_1 (1 + \delta_{n, (2D_1)^{-1}})}{n} \left[\frac{1}{96D_1^2} + (\sigma_a)^2 \right] \\ &\quad + \frac{2D_2 (1 + \delta_{n, (2D_2)^{-1}})}{n} \left[\frac{1}{96D_2^2} + (\sigma_b)^2 \right] .\end{aligned}$$

Moreover,

$$|\delta_{n, (2D_1)^{-1}}| \leq \min \left\{ L_1, L_2(n/D_1)^{-1/4} \right\} \leq L (\ln n)^{-1/2} ,$$

and the same bound holds for $|\delta_{n, (2D_2)^{-1}}|$. Similarly,

$$\mathbb{E}[p_2(D_1, D_2)] = \frac{D_1}{n} \left[\frac{1}{48D_1^2} + 2(\sigma_a)^2 \right] + \frac{D_2}{n} \left[\frac{1}{48D_2^2} + 2(\sigma_b)^2 \right] .$$

Remark that in both cases, the terms of order $D_i^{-1}n^{-1}$ are negligible in front of the bias, because $D_i/n \leq 1/(2(\ln n)^2)$.

Control of the remainder terms: large models We now have to prove that $p_1(m) - \mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)] - p_2(m) + \bar{\delta}(m)$ are close to zero on a large probability event.

Let us first consider a model $m = (D_{m,1}, D_{m,2})$ of dimension $D_m \geq (\ln n)^6$. Since $\max\{D_{m,1}, D_{m,2}\} \leq n/(2(\ln n)^2)$, we also have on Ω_1

$$B_n(m) = \min \left\{ \frac{n}{2D_{m,1}}, \frac{n}{2D_{m,2}} \right\} \geq (\ln n)^2 \geq 1 .$$

From Proposition 2 (with $\gamma = 4$), an event of probability at least $1 - Ln^{-4}$ exists on which:

$$|p_1(m) - \mathbb{E}[p_1(m)]| \leq L_{(\mathbf{H})} (\ln n)^{-1} \mathbb{E}[p_2(m)] \quad (22)$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{(\mathbf{H})} (\ln n)^{-2} \mathbb{E}[p_2(m)] . \quad (23)$$

Moreover, from Lemma 3 (with $x = 4 \ln n$ and $\eta = (\ln n)^{-1}$), an event of probability at least $1 - 2n^{-4}$ exists on which

$$|\bar{\delta}(m)| \leq \frac{\ell(s, s_m)}{\ln n} + \frac{L(\mathbf{H})(\ln n)^2}{n} . \quad (24)$$

We now combine these controls with our previous computations, in order to make (19) and (21) more explicit. Let us consider the event Ω_2 on which (22)–(24) hold for every $m \in \mathcal{M}_n$ and $\Omega = \Omega_1 \cap \Omega_2$. The probability of Ω is larger than $1 - Ln^{-2}$.

Then, for every $m = (D_{m,1}, D_{m,2}) \in \mathcal{M}_n$ such that $\min\{D_{m,1}, D_{m,2}\} \geq (\ln n)^6$, and every $K > 0$, some $\epsilon_{1,m}$ and $\epsilon_{2,m}$ exist such that the following holds. First,

$$\begin{aligned} \text{crit}_{\text{id}}(D_{m,1}, D_{m,2}) &= \frac{1}{96D_1^2} + \frac{2(\sigma_a)^2(1 + \epsilon_{1,m})D_1}{n} \\ &+ \frac{1}{96D_2^2} + \frac{2(\sigma_b)^2(1 + \epsilon_{1,m})D_2}{n} . \end{aligned} \quad (25)$$

Second,

$$\begin{aligned} \text{crit}_K(D_{m,1}, D_{m,2}) &= KD_m + \frac{1}{96D_{m,1}^2} - \frac{2(\sigma_a)^2(1 + \epsilon_{2,m})D_{m,1}}{n} \\ &+ \frac{1}{96D_{m,2}^2} - \frac{2(\sigma_b)^2(1 + \epsilon_{2,m})D_{m,2}}{n} . \end{aligned} \quad (26)$$

Third,

$$\max\{|\epsilon_{1,m}|, |\epsilon_{2,m}|\} \leq L(\ln n)^{-1/2} .$$

Small models When the dimension D_m of m is small, we have to control crit_{id} and crit_K in a different way. Roughly, the bias term will be much larger than the other ones, because $D_{m,1} \leq D_m$ and $D_{m,2} \leq D_m$.

More precisely, on Ω , for every $m \in \mathcal{M}_n$ such that $D_m \leq (\ln n)^6$,

$$\begin{aligned} \ell(s, s_m) &= \frac{1}{96D_{m,1}^2} + \frac{1}{96D_{m,2}^2} \geq \frac{1}{48(\ln n)^{12}} \\ \mathbb{E}[p_1(m)] &\geq 0 \\ \mathbb{E}[p_2(m)] &\leq \frac{L(\mathbf{H})D_m}{n} \leq \frac{L(\mathbf{H})(\ln n)^6}{n} \\ |\bar{\delta}(m)| &\leq \frac{\ell(s, s_m)}{\ln n} + \frac{L(\mathbf{H})(\ln n)^2}{n} \\ |p_1(m) - \mathbb{E}[p_1(m)]| &\leq L(\mathbf{H})(\ln n)^3 \mathbb{E}[p_2(m)] \leq \frac{L(\mathbf{H})(\ln n)^9}{n} \\ |p_2(m) - \mathbb{E}[p_2(m)]| &\leq L(\mathbf{H}) \ln(n) \mathbb{E}[p_2(m)] \leq \frac{L(\mathbf{H})(\ln n)^7}{n} . \end{aligned}$$

Hence, for every $K \geq 0$,

$$\text{crit}_{\text{id}}(D_{m,1}, D_{m,2}) \geq \frac{1}{48 (\ln n)^{12}} - \frac{L_{(\mathbf{H})} (\ln n)^9}{n} \quad (27)$$

$$\text{crit}_K(D_{m,1}, D_{m,2}) \geq K D_m + \left(1 - (\ln n)^{-1}\right) \frac{1}{48 (\ln n)^{12}} - \frac{L_{(\mathbf{H})} (\ln n)^7}{n} . \quad (28)$$

Conclusion: a deterministic lemma On the event previously defined, (26), (27) and (28) show that we can apply Lemma 5 below, with

$$\begin{aligned} \text{crit}_1 &= \text{crit}_{\text{id}} & \text{crit}_{2,K} &= \text{crit}_K \\ \alpha &= \frac{1}{96} & \beta_1 &= 2(\sigma_a)^2 & \beta_2 &= 2(\sigma_b)^2 \leq \beta_1/(1+\varepsilon) \\ \kappa_1 &= 6 & \kappa_2 &= \frac{1}{2} & \kappa_3 &= 12 & c_2 &= L & c_3 &= \frac{1}{50} , \end{aligned}$$

at least for $n \geq L_{(\mathbf{H})}$. This proves the result provided $n \geq n_0 = L_{(\mathbf{H})}$. To remove the latter condition, we enlarge the constant K in the probability bound so that $K \geq n_0^2$, hence $1 - K n^{-2} \leq 0$ when n is not large enough. \square

Lemma 5. Let $\alpha, (\beta_i)_{i=1,2}, (c_i)_{i=2,3}, (\kappa_i)_{1 \leq i \leq 3}$ be some positive constants, $n \in \mathbb{N}$ and $\mathcal{M}_n = \left\{1, \dots, n/(2(\ln n)^2)\right\} \times \left\{1, \dots, n/(2(\ln n)^2)\right\}$. Assume that $(1+\varepsilon)\beta_2 \leq \beta_1$ for some $\varepsilon > 0$. For every $m = (D_1, D_2) \in \mathcal{M}_n$, we define $D_{m,1} = D_1$ and $D_{m,2} = D_2$. Let crit_1 be some function $\mathcal{M}_n \mapsto \mathbb{R}$, and $(\text{crit}_{2,K})_{K \in [0, +\infty)}$ be a family of functions $\mathcal{M}_n \mapsto \mathbb{R}$ satisfying the following conditions:

(i) for every $m \in \mathcal{M}_n$,

$$\text{crit}_1(m) = \left(\frac{\alpha}{D_{m,1}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\alpha}{D_{m,2}^2} + \frac{\beta_2 D_{m,2}}{n} \right) (1 + \epsilon_{1,m}) \quad (29)$$

$$\begin{aligned} \text{crit}_{2,K}(m) &= K D_{m,1} + \left(\frac{\alpha}{D_{m,1}^2} - \frac{\beta_1 D_{m,1}}{n} \right) (1 + \epsilon_{2,m}) \\ &+ K D_{m,2} + \left(\frac{\alpha}{D_{m,2}^2} - \frac{\beta_2 D_{m,2}}{n} \right) (1 + \epsilon_{2,m}) \end{aligned} \quad (30)$$

with $\max_{i=1,2} \sup_{m \in \mathcal{M}_n} \text{s.t. } (\ln n)^{\kappa_1 \leq D_m} |\epsilon_{i,m}| \leq c_2 (\ln n)^{-\kappa_2}$.

(ii) for every $m \in \mathcal{M}_n$ such that $D_m := D_{m,1} + D_{m,2} < (\ln n)^{\kappa_1}$,

$$\text{crit}_1(m) \geq c_3 (\ln n)^{-\kappa_3} \quad (31)$$

$$\text{crit}_{2,K}(m) \geq c_3 (\ln n)^{-\kappa_3} . \quad (32)$$

Then, there exists some constants $\eta > 0$ (depending only on ε) and $n_0 > 0$ (depending on $\alpha, (\beta_i)_{i=1,2}, (c_i)_{i=2,3}, (\kappa_i)_{1 \leq i \leq 3}$) such that, if $n \geq n_0$, for every $K > 0$ and $\widehat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \text{crit}_{2,K}(m)$,

$$\text{crit}_1(\widehat{m}(K)) \geq (1 + \eta) \inf_{m \in \mathcal{M}_n} \{ \text{crit}_1(m) \} . \quad (33)$$

6.5. Proof of Lemma 5

The proof is in two steps. First, we control the dimensions of “good” models m for crit_1 (points A–B). Second, we show that no model in $(\widehat{m}(K))_{K>0}$ can be “good” (point C–E). In the following, $L_{(\mathbf{HL})}$ denotes any constant depending only on $\alpha, \varepsilon, (\beta_i)_{i=1,2}, (c_i)_{i=2,3}, (\kappa_i)_{1 \leq i \leq 3}$.

A. Upper bound on crit_1 for a “good” model Let $m^* \in \mathcal{M}_n$ be any model such that

$$\left| D_{m^*,1} - \left(\frac{2\alpha n}{\beta_1} \right)^{1/3} \right| \leq 1 \quad \text{and} \quad \left| D_{m^*,2} - \left(\frac{2\alpha n}{\beta_2} \right)^{1/3} \right| \leq 1 .$$

As soon as $n \geq L_{(\mathbf{HL})}$, such an m^* exists and satisfies $\min \{ D_{m^*,1}, D_{m^*,2} \} > (\ln n)^{\kappa_1}$. We then deduce from (29) that

$$\text{crit}_1(m^*) \leq \frac{3\alpha^{1/3}}{2^{2/3}\eta^{2/3}} \left(\beta_1^{2/3} + \beta_2^{2/3} \right) \left(1 + L_{c_2} (\ln n)^{-\min\{\kappa_1, \kappa_2\}} \right) . \quad (34)$$

B. Dimension of “good” models for crit_1 Let us fix some $\eta \in (0, 1/2)$, and let $m \in \mathcal{M}_n$ be such that $\text{crit}_1(m) \leq (1 + \eta) \text{crit}_1(m^*)$. The goal is to prove that m must have dimensions “close” to the ones of m^* .

First, the upper bound (34) is smaller than $c_3 (\ln n)^{-\kappa_3}$ (at least when $n \geq L_{(\mathbf{HL})}$). Therefore, we must have $D_m \geq (\ln n)^{\kappa_1}$, hence (29) holds with $|\varepsilon_{1,m}| \leq c_2 (\ln n)^{-\kappa_2}$.

Define

$$C_{m,1} := D_{m,1} \left(\frac{\beta_1}{2\alpha n} \right)^{1/3} > 0 \quad C_{m,2} := D_{m,2} \left(\frac{\beta_2}{2\alpha n} \right)^{1/3} > 0$$

and for every $x > -1$, $f(x) = 2^{-2/3}(1+x)^{-2} + 2^{1/3}(1+x)$. Then, (i) and Lemma 6 below yield

$$\begin{aligned} \text{crit}_1(m) &\geq \frac{\alpha^{1/3}}{n^{2/3}} \left(\beta_1^{2/3} f(C_{m,1} - 1) + \beta_2^{2/3} f(C_{m,2} - 1) \right) \left(1 - c_2 (\ln n)^{-\kappa_2} \right) \\ &\geq \frac{3\alpha^{1/3}}{2^{2/3}\eta^{2/3}} \left(\beta_1^{2/3} + \beta_2^{2/3} \right) + \frac{3\alpha^{1/3}}{2^{14/3}\eta^{2/3}} \left(1 - c_2 (\ln n)^{-\kappa_2} \right) \\ &\quad \times \left(\beta_1^{2/3} \min \{ 1, (C_{m,1} - 1)^2 \} + \beta_2^{2/3} \min \{ 1, (C_{m,2} - 1)^2 \} \right) . \end{aligned}$$

Hence, using (34), $(1 + \eta) \text{crit}_1(m^*) \geq \text{crit}_1(m)$ implies

$$\begin{aligned} & 16 \left(\beta_1^{2/3} + \beta_2^{2/3} \right) \left(\eta + L_{c_2} (\ln n)^{-\min\{\kappa_1, \kappa_2\}} \right) \\ & \geq \left(\beta_1^{2/3} \min\{1, (C_{m,1} - 1)^2\} + \beta_2^{2/3} \min\{1, (C_{m,2} - 1)^2\} \right) \end{aligned}$$

In particular,

$$\begin{aligned} \min\{1, (C_{m,1} - 1)^2\} & \leq 16 \left(1 + \left(\frac{\beta_1}{\beta_2} \right)^{2/3} \right) \left(\eta + \frac{L_{c_2}}{(\ln n)^{\min\{\kappa_1, \kappa_2\}}} \right) \\ \min\{1, (C_{m,2} - 1)^2\} & \leq 16 \left(1 + \left(\frac{\beta_2}{\beta_1} \right)^{2/3} \right) \left(\eta + \frac{L_{c_2}}{(\ln n)^{\min\{\kappa_1, \kappa_2\}}} \right) . \end{aligned}$$

Let $\Delta \in (0, 1]$ and take

$$0 < \eta \leq \eta_\Delta := \frac{\Delta^2}{17 \left(1 + \left(\frac{\beta_1}{\beta_2} \right)^{2/3} \right)} < \frac{1}{2} ,$$

so that when $n \geq L_{(\mathbf{HL}), \Delta}$, both upper bounds are smaller than $16\Delta^2/17 < 1$, hence

$$\max\{|C_{m,1} - 1|, |C_{m,2} - 1|\} \leq \Delta . \quad (35)$$

We now start the second part of the proof: can any $\widehat{m}(K)$ be good for crit_1 ? Assume that $\Delta \in (0, 1]$, $K \geq 0$ and $\widehat{m}(K) \in \arg \min_{m \in \mathcal{M}_n} \{\text{crit}_{2,K}(m)\}$ exist such that

$$(1 + \eta_\Delta) \text{crit}_1(m^*) \geq \text{crit}_1(\widehat{m}(K)) . \quad (36)$$

In particular, $D_{\widehat{m}(K)} \geq (\ln n)^{\kappa_1}$ and (35) holds for $m = \widehat{m}(K)$. Considering separately different ranges of values of K , the idea of the proof is to show this implies a contradiction.

C. Small values of K Let us first assume that $Kn \leq \beta_1 + (\ln n)^{-1}$. Then,

$$\begin{aligned} 0 & \leq \text{crit}_{2,K} \left(n^{1/3} (\ln n)^{1/2}, D_{\widehat{m}(K),2} \right) - \text{crit}_{2,K}(\widehat{m}(K)) \\ & \leq (Kn - \beta_1) \left(n^{-2/3} (\ln n)^{1/2} - n^{-1} D_{\widehat{m}(K),1} \right) \\ & \quad + \left(\frac{\alpha}{n^{2/3} \ln n} - \frac{\alpha}{D_{\widehat{m}(K),1}^2} \right) + \frac{L_{(\mathbf{HL})}}{n^{2/3} (\ln n)^{\kappa_2}} . \end{aligned} \quad (37)$$

On the one hand, if $Kn - \beta_1 \leq 0$, (37) cannot hold for $n \geq L_{(\mathbf{HL})}$ because the first terms are negative and $L_{(\mathbf{HL})} n^{-2/3} (\ln n)^{-\kappa_2}$ is negligible in front of them. On the other hand, if $Kn - \beta_1 \geq 0$,

$$(Kn - \beta_1) \left(n^{-2/3} (\ln n)^{1/2} - n^{-1} D_{\widehat{m}(K),1} \right) \leq L_{(\mathbf{HL})} (\ln n)^{-1/2} n^{-2/3} ,$$

so that the (negative) term $-\alpha D_{\widehat{m}(K),1}^{-2}$ is the dominant one in the right-hand side of (37). Hence, (36) cannot hold for $n \geq L_{(\mathbf{HL})}$ if $Kn \leq \beta_1 + (\ln n)^{-1}$.

D. Intermediate values of K Assume now that $\beta_1 + (\ln n)^{-1} < Kn \leq \beta_2 + n(\ln n)^{-\max\{3\kappa_1, 2\kappa_3\}}$. Then, for $i = 1, 2$, $(2\alpha n / (Kn - \beta_i))^{1/3}$ is between $(\ln n)^{\kappa_1}$ and $n / (2((\ln n)^2))$, at least for $n \geq L_{(\mathbf{HL})}$, and $(Kn - \beta_i)^{2/3} n^{-2/3} \ll (\ln n)^{-\kappa_3}$. Therefore, a proof similar to the one of (35) (that is, points A–B of the current proof), with $(Kn - \beta_1, Kn - \beta_2)$ instead of (β_1, β_2) , leads to the following. For every $\Delta \in (0, 1]$, if $n \geq L_{(\mathbf{HL}), \Delta}$,

$$\left| D_{\widehat{m}(K), 1} \left(\frac{Kn - \beta_1}{2\alpha n} \right)^{1/3} - 1 \right| \leq \Delta \quad (38)$$

$$\left| D_{\widehat{m}(K), 2} \left(\frac{Kn - \beta_2}{2\alpha n} \right)^{1/3} - 1 \right| \leq \Delta . \quad (39)$$

If we combine (35) with (38) (with the same Δ), we obtain that if $n \geq L_{(\mathbf{HL}), \Delta}$,

$$\left(\frac{1 - \Delta}{1 + \Delta} \right)^3 + 1 \leq \frac{Kn}{\beta_1} \leq \left(\frac{1 + \Delta}{1 - \Delta} \right)^3 + 1 . \quad (40)$$

Similarly, the combination of (35) and (39) (with the same Δ) shows that if $n \geq L_{(\mathbf{HL}), \Delta}$,

$$\left(\frac{1 - \Delta}{1 + \Delta} \right)^3 + 1 \leq \frac{Kn}{\beta_2} \leq \left(\frac{1 + \Delta}{1 - \Delta} \right)^3 + 1 . \quad (41)$$

Choosing Δ small enough so that

$$\beta_2 \left[\left(\frac{1 + \Delta}{1 - \Delta} \right)^3 + 1 \right] < \beta_1 \left[\left(\frac{1 - \Delta}{1 + \Delta} \right)^3 + 1 \right] , \quad (42)$$

it follows from (40) and (41) that (36) cannot hold for $n \geq L_{(\mathbf{HL})}$ if $\beta_1 + (\ln n)^{-1} < Kn \leq \beta_2 + n(\ln n)^{-\max\{3\kappa_1, 2\kappa_3\}}$.

E. Large values of K Assume now that $Kn > \beta_2 + n(\ln n)^{-\max\{3\kappa_1, 2\kappa_3\}}$. Then,

$$\begin{aligned} 0 &\leq \text{crit}_{2, K} \left(D_{\widehat{m}(K), 1}, (\ln n)^{\kappa_1} \right) - \text{crit}_{2, K}(\widehat{m}(K)) \\ &\leq (Kn - \beta_2) \left(\frac{(\ln n)^{\kappa_1} - D_{\widehat{m}(K), 2}}{n} \right) + \frac{(Kn - \beta_2)n^{1/3}L_{(\mathbf{HL})}}{(\ln n)^{\kappa_2}} , \end{aligned} \quad (43)$$

which is not possible for $n \geq L_{(\mathbf{HL})}$ since $D_{\widehat{m}(K), 2}$ is of order $n^{1/3}$.

To conclude, whatever $K \geq 0$, (36) cannot hold when $\Delta = \Delta^*$ is the largest element of $(0, 1]$ satisfying (42) and $n \geq L_{(\mathbf{HL}), \Delta^*} = L_{(\mathbf{HL})}$, which is the desired result. \square

Lemma 6. Let $f : (-1, +\infty) \mapsto \mathbb{R}$ be defined by $f(x) = 2^{-2/3}(1+x)^{-2} + 2^{1/3}(1+x)$. Then, for every $x > -1$,

$$f(x) \geq 3 \times 2^{-2/3} + 3 \times 2^{-14/3} (x^2 \wedge 1) .$$

proof of Lemma 6. We apply the Taylor-Lagrange theorem to f (which is infinitely differentiable) at order two, between 0 and x . The result follows since $f(0) = 3 \times 2^{-2/3}$, $f'(0) = 0$ and $f''(t) = 6 \times 2^{-2/3} \times (1+t)^{-4} \geq 3 \times 2^{1/3-4}$ if $t \leq 1$. If $t > 1$, the result follows from the fact that $f' \geq 0$ on $[0, +\infty)$. \square

Acknowledgments

The author would like to thank gratefully Pascal Massart for several fruitful discussions.

References

- Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadors, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.
- Arlot, S. (2007). *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11. oai:tel.archives-ouvertes.fr:tel-00198803_v1.
- Arlot, S. (2008a). Model selection by resampling penalization. oai:hal.archives-ouvertes.fr:hal-00262478_v1.
- Arlot, S. (2008b). V -fold cross-validation improved: V -fold penalization. arXiv:0802.0566v2.
- Arlot, S. and Massart, P. (2008). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* To appear. arXiv:0802.0837v3.
- Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493.
- Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic).
- Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.
- Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.
- Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X -fixed prediction error. *J. Amer. Statist. Assoc.*, 87(419):738–754.

- Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331.
- Galtchouk, L. and Pergamenshchikov, S. (2008). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression via model selection. arXiv:0810.1173v1.
- Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328.
- Gendre, X. (2008). Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. arXiv:0807.2547v1.
- Li, K.-C. (1987). Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, 15:661–675.
- Polyak, B. T. and Tsybakov, A. B. (1990). Asymptotic optimality of the C_p -test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, 35(2):305–317.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68(1):45–54.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Geisser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.