Finite Sample Improvement of Akaike's Information Criterion

Adrien Saumard, Fabien Navarro

Finite Sample Improvement of Akaike's Information Criterion

Adrien Saumard, and Fabien Navarro

***Abstract*—Considering the selection of frequency histograms, we propose a modification of Akaike's Information Criterion that avoids overfitting, even when the sample size is small. We call this correction an over-penalization procedure. We emphasize that the principle of unbiased risk estimation for model selection can indeed be improved by addressing excess risk deviations in the design of the penalization procedure. On the theoretical side, we prove sharp oracle inequalities for the Kullback-Leibler divergence. These inequalities are valid with positive probability for any sample size and include the estimation of unbounded log-densities. Along the proofs, we derive several analytical lemmas related to the Kullback-Leibler divergence, as well as concentration inequalities, that are of independent interest. In a simulation study, we also demonstrate state-of-the-art performance of our over-penalization criterion for bin size selection, in particular outperforming AICc procedure.**

***Index Terms*—model selection, bin size, AIC corrected, over-penalization, small sample size.**

I. Introduction

SINCE its introduction by Akaike in the early seventies [1], the celebrated Akaike's Information Criterion (AIC) has been an essential tool for the statistician and its use is almost systematic in problems of model selection and estimator selection for prediction. By choosing among estimators or models constructed from finite degrees of freedom, AIC recommends maximizing the log-likelihood of the estimators penalized by their corresponding degrees of freedom. This procedure has found pathbreaking applications in density estimation, regression, time series or neural network analysis, to name a few ([29]). Because of its simplicity and negligible computation cost—whenever the estimators are given—, it is also far from outdated and continues to serve as one of the most useful devices for model selection in high-dimensional statistics. For instance, it can be used to efficiently tune the Lasso ([54]).

Any substantial and principled improvement of AIC is likely to have a significant impact on the practice of model choices and we bring in this paper an efficient and theoretically grounded solution to the problem of overfitting that can occur when using AIC on small to medium sample sizes.

The fact that AIC tends to be unstable and therefore perfectible in the case of small sample sizes is well known to practitioners and has long been noted. Suguira [50] and Hurvich and Tsai [33] have proposed the so-called AICc (for AIC corrected), which tends to penalize more than AIC. However, the derivation of AICc comes from an asymptotic analysis where the dimension of the models are considered fixed relative to the sample size. In fact, such an assumption does not fit the usual practice of model selection, where the largest models are of dimensions close to the sample size.

Building on considerations from the general nonasymptotic theory of model selection developed during the nineties (see for instance [13] and [39]) and in particular on Castellan's analysis [27], Birgé and Rozenholc [20] have considered an AIC modification specifically designed for the selection of the bin size in histogram selection for density estimation. Indeed, results of [27]—and more generally results of [13]—advocate to take into account in the design of penalty the number of models to be selected. The importance of the cardinality of the collection of models for model selection is in fact a very general phenomenon and one of the main outcomes of the nonasymptotic model selection theory. In the bin size selection problem, this corresponds to adding a small amount to AIC. Unfortunately, the theory does not specify uniquely the term to be added to AIC. In order to choose a good one, intensive experiments were conducted in [20].

We propose an approach of optimal model selection that naturally leads to consider a quantile risk estimation rather than the well-known unbiased risk estimation principle. The latter principle is at the core of Akaike's model selection procedure and is more generally the main model selection principle, which underlies procedures such as Stein's Unbiased Risk Estimator ([48]) or cross-validation ([8]). We note that it is more efficient to estimate a quantile of the risk of the estimators - the level of the quantile depending on the size of the collection of models - rather than its mean. We call it an over-penalization procedure, because it systematically involves adding small terms to traditional penalties such as AIC. The term of over-penalization is indeed rather commonly used in the literature to describe the need to inflate criteria designed from the unbiased risk principle (see for instance [11, Section 8.4] and references therein).

We are interested in the present article by producing a sharp oracle inequality from a procedure of penalization of the empirical risk. But it should be mentioned that other kinds of procedures exist, also allowing to derive oracle inequalities for the model selection problem. Indeed, in the density estimation context for the Kullback-Leibler loss, [25], [53] propose to use an aggregation scheme to ensure an optimal oracle inequality. But there are two essential differences with our framework. Firstly, the above mentioned articles consider the estimators as fixed, a classical assumption in aggregation literature. Secondly, they work with a bounded setting, whereas our results are valid with only finite moment assumptions.

Another possible procedure allowing to obtain oracle in-

Adrien Saumard is with Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France (e-mail: adrien.saumard@ensai.fr).

Fabien Navarro is with the SAMM Laboratory, Paris 1 Panthéon-Sorbonne University, Paris, France (e-mail: fabien.navarro@math.cnrs.fr).

Manuscript received June 16, 2020; last revised July 1, 2021.

equalities would be Lepskii-type procedures ([37], [31]). While the rationale behind this kind of procedure is very general, obtaining sharp results in terms of constants in the oracle inequalities and performing a sharp calibration of the quantities involved in the procedure seem to be rather difficult problems, substantially different from empirical risk penalization, that have only been considered in a few, recent articles ([34], [35]).

Lets us now detail our contributions.

- Considering the problem of density estimation by selecting a histogram, we prove a sharp, nonasymptotic oracle inequality for our procedure. Indeed, we describe a control of Kullback-Leibler (KL) divergence - also called excess risk - of the selected histogram that is valid with positive probability for any sample size. We emphasize that this strong feature may not be possible when considering AIC. We also stress that up to our knowledge, our oracle inequality is the first nonasymptotic result comparing the KL divergence of the selected model to the KL divergence of the oracle in an unbounded setting. Indeed, oracle inequalities in density estimation are generally expressed in terms of Hellinger distance - which is easier to handle than the KL divergence, because it is bounded - for the selected model.
- In order to prove our oracle inequality, we improve upon the previously best known concentration inequality for the chi-square statistics ([27], [39]) and this allows us to gain an order of magnitude in the control of the deviations of the excess risks of the estimators. Our result on the chi-square statistics is general and of independent interest.
- We also prove new Bernstein-type concentration inequalities for log-densities that are unbounded. Again, these probabilistic results, which are naturally linked to information theory, are general and of independent interest.
- We generalize previous results of Barron and Sheu [14] regarding the existence of margin relations in maximum likelihood estimation (MLE). Indeed, related results of [14] where established under boundedness of the log-densities and we extend them to unbounded log-densities with moment conditions.
- Finally, from a practical point of view, we bring a nonasymptotic improvement of AIC that has, in its simplest form, the same computational cost as AIC. Our most efficient correction proceeds with a data-driven calibration of the over-penalization term. It appears in our experiments that the latter correction outperforms AIC on small and medium sample sizes, but also most often surpasses existing AIC corrections such as AICc or Birgé-Rozenholc's procedure.

Let us end this introduction by detailing the organization of the paper.

We present our over-penalization procedure in Section II. More precisely, we detail in Sections II-A and II-B our model selection framework related to MLE via histograms. Then in Section II-C we define formally over-penalization procedures. Section III is devoted to statistical guarantees related to over-penalization. In particular, as concentration properties of the excess risks are at the heart of the design of an over-penalization, we detail them in Section III-A. We then deduce a sharp oracle inequality in Section III-B and highlight the theoretical advantages compared to an AIC analysis. New mathematical tools of a probabilistic and analytical nature and of independent interest are presented in Section IV. Section V contains the experiments, with detailed practical procedures. We consider two different practical variations of over-penalization and compare them with existing penalization procedures. The proofs are gathered in a supplementary material [47], which also provides further theoretical developments that complement the description of our over-penalization procedure.

II. Statistical Framework and Notations

A. Maximum Likelihood Density Estimation

We are given n independent observations $(\xi_1, \ldots, \xi_n)$ with unknown common distribution P on a measurable space $(\mathcal{Z}, \mathcal{T})$. We assume that there exists a known probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that P admits a density f_* with respect to μ: $f_* = dP/d\mu$. Our goal is to estimate the density f_*.

For an integrable function f on $\mathcal{Z}$, we set $Pf = P(f) = \int_{\mathcal{Z}} f(z)\, dP(z)$ and $\mu f = \mu(f) = \int_{\mathcal{Z}} f(z)\, d\mu(z)$. If $P_n = 1/n \sum_{i=1}^n \delta_{\xi_i}$ denotes the empirical distribution associated to the sample $(\xi_1, \ldots, \xi_n)$, then we set $P_n f = P_n(f) = 1/n \sum_{i=1}^n f(\xi_i)$. Moreover, taking the conventions $\ln 0 = -\infty$, $0 \ln 0 = 0$ and defining $(x)_+ = x \vee 0$ and $(x)_- = -x \vee 0$, we set

$$\mathcal{S} = \left\{ f : \mathcal{Z} \longrightarrow \mathbb{R}_+;\ \int_{\mathcal{Z}} f d\mu = 1 \text{ and } P(\ln f)_+ < \infty \right\} .$$

We assume that the unknown density f_* belongs to $\mathcal{S}$.

Note that since $P(\ln f_*)_- = -\int f_* \ln f_* \mathbb{1}_{f_* \leq 1} d\mu < \infty$, the fact that f_* belongs to $\mathcal{S}$ is equivalent to $\ln(f_*) \in L_1(P)$, the space of integrable functions on $\mathcal{Z}$ with respect to P.

We consider the MLE of the density f_*. To do so, we define the so-called risk $P(-\ln f)$ of a function $f \in \mathcal{S}$ through the following formula,

$$P(-\ln f) = P(\ln f)_- - P(\ln f)_+ \in \mathbb{R} \cup \{+\infty\} .$$

Also, the excess risk of a function f with respect to the density f_*, that is the difference between the risk of f and the risk of f_*, is classically given in this context by the KL divergence of f with respect to f_*. Recall that for two probability distributions P_f and P_g on $(\mathcal{Z}, \mathcal{T})$ of respective densities f and g with respect to μ, the KL divergence of P_g with respect to P_f is defined to be

$$\mathcal{K}(P_f, P_g) = \begin{cases} \int_{\mathcal{Z}} \ln\left(\frac{dP_f}{dP_g}\right) dP_f = \int_{\mathcal{Z}} f \ln\left(\frac{f}{g}\right) d\mu & \text{if} \quad P_f \ll P_g \\ \infty & \text{otherwise.} \end{cases}$$

By a slight abuse of notation we denote $\mathcal{K}(f, g)$ rather than $\mathcal{K}(P_f, P_g)$ and by the Jensen inequality we notice that $\mathcal{K}(f, g)$ is a nonnegative quantity, equal to zero if and only if $f = g$

$\mu - a.s.$ Hence, for any $f \in \mathcal{S}$, the excess risk of a function f with respect to the density f_* satisfies

$$P(-\ln f) - P(-\ln f_*) = \int_{\mathcal{Z}} \ln\left(\frac{f_*}{f}\right) f_* d\mu = \mathcal{K}(f_*, f) \geq 0$$

and this nonnegative quantity is equal to zero if and only if $f_* = f$ $\mu - a.s.$ Consequently, the unknown density f_* is uniquely defined by

$$f_* = \arg\min_{f \in \mathcal{S}} \{P(-\ln f)\} \ .$$

For a model m, that is a subset $m \subset \mathcal{S}$, we define the maximum likelihood estimator on m, whenever it exists, by

$$\hat{f}_m \in \arg\min_{f \in m} \{P_n(-\ln f)\} = \arg\min_{f \in m} \left\{ \frac{1}{n} \sum_{i=1}^{n} -\ln f(\xi_i) \right\}. \tag{1}$$

B. Histogram Models

The models m that we consider here to define the maximum likelihood estimators as in (1) are made of histograms defined on a fixed partition of $\mathcal{Z}$. More precisely, for a finite partition Λ_m of $\mathcal{Z}$ of cardinality $|\Lambda_m| = D_m + 1$, $D_m \in \mathbb{N}$, we set

$$m = \left\{ f = \sum_{I \in \Lambda_m} \beta_I \mathbb{1}_I \ ; \ (\beta_I)_{I \in \Lambda_m} \in \mathbb{R}_+^{D_m+1}, \right.$$
$$\left. f \geq 0 \text{ and } \sum_{I \in \Lambda_m} \beta_I \mu(I) = 1 \right\} \ .$$

Note that the smallest affine space containing m is of dimension D_m. The quantity D_m can thus be interpreted as the number of degrees of freedom in the (parametric) model m. We assume that any element I of the partition Λ_m is of positive measure with respect to μ: for all $I \in \Lambda_m$, $\quad \mu(I) > 0$. As the partition Λ_m is finite, we have $P(\ln f)_+ < \infty$ for all $f \in m$ and so $m \subset \mathcal{S}$. We state in the next proposition some well-known properties that are satisfied by histogram models submitted to the procedure of MLE ([39, Section 7.3]).

Proposition II.1 *Let*

$$f_m = \sum_{I \in \Lambda_m} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ .$$

Then $f_m \in m$ and f_m is called the KL projection of f_* onto m. Moreover, it holds

$$f_m = \arg\min_{f \in m} P(-\ln f) \ .$$

The following Pythagorean-like identity for the KL divergence holds, for every $f \in m$,

$$\mathcal{K}(f_*, f) = \mathcal{K}(f_*, f_m) + \mathcal{K}(f_m, f) \ . \tag{2}$$

The maximum likelihood estimator on m is well-defined and corresponds to the so-called frequency histogram associated to the partition Λ_m. We have the following formulas,

$$\hat{f}_m = \sum_{I \in \Lambda_m} \frac{P_n(I)}{\mu(I)} \mathbb{1}_I \text{ and } P_n\left(\ln\left(\frac{\hat{f}_m}{f_m}\right)\right) = \mathcal{K}(\hat{f}_m, f_m) \ .$$

Remark II.1 *Histogram models are special cases of general exponential families exposed for example in Barron and Sheu [14] (see also Castellan [27] for the case of exponential models of piecewise polynomials). The projection property (2) can be generalized to exponential models (see [14, Lemma 3] and Csiszár [30]).*

C. Over-Penalization

We define in Section II-C1 below our model selection procedure. Then we provide in Section II-C2 a graphical insight on the benefits of over-penalization.

1) Over-Penalization as Estimation of the Ideal Penalty: We are given a collection of histogram models denoted $\mathcal{M}_n$, with finite cardinality depending on the sample size n, and its associated collection of maximum likelihood estimators $\left\{\hat{f}_m; m \in \mathcal{M}_n\right\}$. By taking a (nonnegative) penalty function pen on $\mathcal{M}_n$,

$$\text{pen} : m \in \mathcal{M}_n \longmapsto \text{pen}(m) \in \mathbb{R}^+ \ ,$$

the output of the penalization procedure (also called the selected model) is by definition any model satisfying,

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n(-\ln \hat{f}_m) + \text{pen}(m) \right\} \ . \tag{3}$$

We aim at selecting an estimator $\hat{f}_{\widehat{m}}$ with a KL divergence, pointed on the true density f_*, as small as possible. Hence, we want our selected model to have a performance as close as possible to the excess risk achieved by an oracle model (not necessarily unique), defined to be,

$$m_* \in \arg\min_{m \in \mathcal{M}_n} \left\{ \mathcal{K}(f_*, \hat{f}_m) \right\} \tag{4}$$
$$= \arg\min_{m \in \mathcal{M}_n} \left\{ P(-\ln \hat{f}_m) \right\} \ . \tag{5}$$

Recall that the celebrated AIC procedure corresponds to using a penalty $\text{pen}_{\text{AIC}}(m) = D_m/n$ in criterion (3). To understand further this choice and the possibility of an improvement, let us discuss the notion of an ideal penalty. From (5), it is seen that an ideal penalty in the optimization task (3) is given by

$$\text{pen}_{\text{id}}(m) = P(-\ln \hat{f}_m) - P_n(-\ln \hat{f}_m) \ ,$$

since in this case, the criterion $\text{crit}_{\text{id}}(m) = P_n(-\ln \hat{f}_m) + \text{pen}_{\text{id}}(m)$ is equal to the true risk $P(-\ln \hat{f}_m)$. However pen_{id} is unknown and, at some point, we need to give some estimate of it. In addition, pen_{id} is random, but we may not be able to provide a penalty, even random, whose fluctuations at a fixed model m would be positively correlated to the fluctuations of $\text{pen}_{\text{id}}(m)$. This means that we are rather searching for an estimate of a *deterministic functional* of pen_{id}. But which functional would be convenient? The answer to this question is essentially contained in the solution of the following problem.

Problem 1. *For any fixed $\beta \in (0,1)$ find the* deterministic penalty $\text{pen}_{\text{id},\beta} : \mathcal{M}_n \to R_+$, *that minimizes the value of C, among constants $C > 0$ which satisfy the following oracle inequality*,

$$\mathbb{P}\left(\mathcal{K}(f_*, \hat{f}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}(f_*, \hat{f}_m) \right\} \right) \geq 1 - \beta \ . \tag{6}$$

The solution - or even the existence of a solution - to the problem given in (6) is not easily accessible and depends on assumptions on the law P of data and on approximation properties of the models. In the following, we give a reasonable candidate for $\mathrm{pen}_{\mathrm{id},\beta}$. Indeed, let us set $\beta_{\mathcal{M}} = \beta/\mathrm{Card}(\mathcal{M}_n)$ and define

$$\mathrm{pen}_{\mathrm{opt},\beta}(m) = q_{1-\beta_{\mathcal{M}}}\left\{\mathcal{K}(f_m, \hat{f}_m) + \mathcal{K}(\hat{f}_m, f_m)\right\} , \quad (7)$$

where $q_\lambda\{Z\} = \inf\{q \in \mathbb{R}; \mathbb{P}(Z \le q) \ge \lambda\}$ is the quantile of level λ for the real random variable Z. Note that that the penalty $\mathrm{pen}_{\mathrm{opt},\beta}$ is unknown to the statistician. Our claim is that $\mathrm{pen}_{\mathrm{opt},\beta}$ has a theoretical interest since it gives in (6) a constant C which is close to one, under some general assumptions (see Section III for precise results). Let us explain now why $\mathrm{pen}_{\mathrm{opt},\beta}$ should lead to a nearly optimal model selection.

We set

$$\Omega_0 = \bigcap_{m\in\mathcal{M}_n}\left\{\mathcal{K}(f_m, \hat{f}_m) + \mathcal{K}(\hat{f}_m, f_m) \le \mathrm{pen}_{\mathrm{opt},\beta}(m)\right\} .$$

We see, by definition of $\mathrm{pen}_{\mathrm{opt},\beta}$ and by a simple union bound over the models $m \in \mathcal{M}_n$, that the event Ω_0 is of probability at least $1-\beta$. By definition of $\widehat{m}$ we have, for any $m \in \mathcal{M}_n$,

$$P_n(-\ln \hat{f}_{\widehat{m}}) + \mathrm{pen}_{\mathrm{opt},\beta}(\widehat{m}) \le P_n(-\ln \hat{f}_m) + \mathrm{pen}_{\mathrm{opt},\beta}(m) . \quad (8)$$

Now, by centering by $P(-\ln f_*)$, using simple algebra and using the fact that on Ω_0, we have $\mathrm{pen}_{\mathrm{opt},\beta}(\widehat{m}) - (\mathcal{K}(f_{\widehat{m}}, \hat{f}_{\widehat{m}}) + \mathcal{K}(\hat{f}_{\widehat{m}}, f_{\widehat{m}})) \ge 0$, Inequality (8) gives on Ω_0,

$$\begin{aligned}\mathcal{K}(f_*, \hat{f}_{\widehat{m}}) \le & \mathcal{K}(f_*, \hat{f}_m) \\ & + \underbrace{\left[\mathrm{pen}_{\mathrm{opt},\beta}(m) - (\mathcal{K}(f_m, \hat{f}_m) + \mathcal{K}(\hat{f}_m, f_m))\right]}_{(a)} \\ & + \underbrace{(P_n - P)\left(\ln(f_{\widehat{m}}/f_m)\right)}_{(b)} .\end{aligned}$$

In order to get an oracle inequality as in (6), it remains to control (a) and (b) in terms of the excess risks $\mathcal{K}(f_*, \hat{f}_m)$ and $\mathcal{K}(f_*, \hat{f}_{\widehat{m}})$. Quantity (a) is related to deviations bounds for the true and empirical excess risks of the M-estimators $\hat{f}_m$ and quantity (b) is related to fluctuations of empirical bias around the bias of the models. Suitable controls of these quantities will give sharp oracle inequalities.

We define an over-penalization procedure as follows.

Definition II.1 *A penalization procedure as defined in (3) is said to be an over-penalization procedure if the penalty* pen *that is used satisfies* $\mathrm{pen}(m) \ge \mathrm{pen}_{\mathrm{opt},\beta}(m)$ *for all* $m \in \mathcal{M}_n$ *and for some* $\beta \in (0, 1/2)$.

Based on concentration inequalities for the excess risks (see Section III-A) we propose the following over-penalization penalty for histogram selection,

$$\mathrm{pen}_+(m) = \left(1 + C\varepsilon_n^+(m)\right)\frac{D_m}{n} , \quad (9)$$

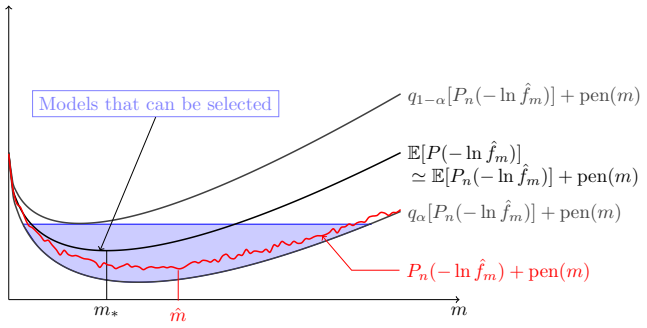


Fig. 1. A schematic view of the situation corresponding to a selection procedure based on the unbiased risk principle. The penalized empirical risk (in red) fluctuates around the expectation of the true risk. The size of the deviations typically increases with the model size, making the shape of the curves possibly flat for the largest models of the collection. Consequently, the chosen model can potentially be very large and lead to overfitting.

where C is a constant that must depend on the distribution of data and is thus unknown in general and $\varepsilon_n^+(m) = \max\left\{\sqrt{D_m \ln(n+1)/n}; \sqrt{\ln(n+1)/D_m}; \ln(n+1)/D_m\right\}$. Hence, C should be either fixed *a priori* ($C = 1$ or 2 are typical choices) or estimated using data (see Section V for further details about the choice of C). The logarithmic terms appearing in (9) are linked to our choice of β and to the cardinal of the collection of models, since in our proofs we take $\beta = (n+1)^{-2}$ and we consider a constant α such that $\ln \mathrm{Card}(\mathcal{M}_n) + \ln(\beta) \le \alpha \ln(n+1)$. The constant α then enters in the constant C of (9). We show below nonasymptotic accuracy of such procedure, both theoretically (assuming a good choice of C) and practically.

2) Graphical insights on over-penalization: Let us provide a graphical perspective on our over-penalization procedure.

If the penalty pen is chosen according to the unbiased risk estimation principle, then it should satisfy, for any model $m \in \mathcal{M}_n$,

$$\mathbb{E}\left[P_n(-\ln \hat{f}_m) + \mathrm{pen}(m)\right] \sim \mathbb{E}\left[P(-\ln \hat{f}_m)\right] .$$

In other words, the curve $\mathcal{C}_n : m \mapsto P_n(-\ln \hat{f}_m) + \mathrm{pen}(m)$ fluctuates around its mean, which is essentially the curve $\mathcal{C}_P : m \mapsto \mathbb{E}[P(-\ln \hat{f}_m))]$, see Figure 1. Asymptotically, the empirical risk $P_n(-\ln \hat{f}_m)$ behaves as a deterministic value (for a fixed model m), which consists to the theoretical bias of the model m, plus half of Akaike's penalty. Thus, asymptotically, the fluctuations of the empirical risk are indeed smaller than the penalty for models of reasonably small bias. But our point is that for small to moderate sample sizes, the fluctuations of the empirical risk may be non-negligible and should be compensated.

More precisely, the largest is the model m, the largest are the fluctuations of $P_n(-\ln \hat{f}_m) = \mathcal{K}(\hat{f}_m, f_m) + P_n(-\ln f_m)$. This is seen for instance through the concentration inequality (13) for the empirical excess risk $\mathcal{K}(\hat{f}_m, f_m)$, that is stated in Theorem III.1 below. Consequently, it can happen that the curve $\mathcal{C}_n$ is quite flat for the largest models and that the selected model is among the largest of the collection, see Figure 1.

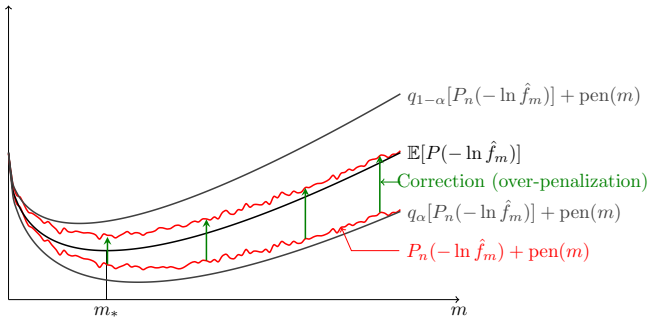


Fig. 2. The correction that should be applied to an unbiased risk estimation procedure would ideally be of the size of the deviations of the risk for each model of the collection.

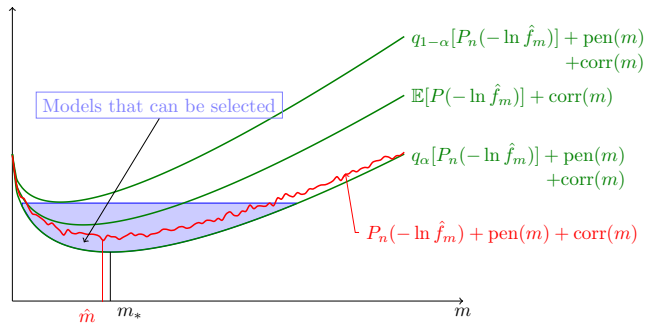


Fig. 3. After a suitable correction, the minimum of the red curve has a better shape. In addition, the region of models that can be possibly selected is substantially smaller and in particular avoids the largest models of the collection.

By using an over-penalization procedure instead of the unbiased risk estimation principle, we compensate the deviations for the largest models and thus obtain a thinner region of potential selected models, see Figures 2 and 3. In other words, we tend to avoid overfitting and by doing so, we ensure a reasonable performance of our over-penalization procedure in situations where unbiased risk estimation fails. As already discussed, this is particularly the case when the amount of data is small to moderate.

III. Theoretical Guarantees

We state here our theoretical results pertaining to the behavior of our over-penalization procedure. As explained in Section II-C, concentration inequalities for true and empirical excess risks are essential tools for understanding our model selection problem and we state them in Section III-A. In Section III-B, we give a sharp oracle inequality.

A. True and empirical excess risks' concentration

In this section, we fix the linear model m made of histograms and we are interested by concentration inequalities for the true excess risk $\mathcal{K}(f_m, \hat{f}_m)$ on m and for its empirical counterpart $\mathcal{K}(\hat{f}_m, f_m)$.

Theorem III.1 *Let $n \geq 1$ be a positive integer and let α, A_+, A_- and A_Λ be positive constants. Take m a model of histograms defined on a fixed partition Λ_m of $\mathcal{Z}$.We set $D_m = |\Lambda_m| - 1$. Assume that $1 < D_m \leq A_+ n/(\ln(n+1)) \leq n$ and*

$$0 < A_\Lambda \leq D_m \inf_{I \in \Lambda_m} \{P(I)\} \ . \tag{10}$$

If $(\alpha+1) A_+/A_\Lambda \leq \tau = \sqrt{\sqrt{6} - 3/\sqrt{2}} < 0.58$, then a positive constant A_0 exists, only depending on α, A_+ and A_Λ, such that by setting

$$\varepsilon_n^+(m) = \max\left\{ \sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}; \frac{\ln(n+1)}{D_m} \right\} \tag{11}$$

and

$$\varepsilon_n^-(m) = \max\left\{ \sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}} \right\} ,$$

we have, on an event of probability at least $1 - 4(n+1)^{-\alpha}$,

$$\left(1 - A_0 \varepsilon_n^-(m)\right) \frac{D_m}{2n} \leq \mathcal{K}\left(f_m, \hat{f}_m\right) \leq \left(1 + A_0 \varepsilon_n^+(m)\right) \frac{D_m}{2n} , \tag{12}$$

$$\left(1 - A_0 \varepsilon_n^-(m)\right) \frac{D_m}{2n} \leq \mathcal{K}\left(\hat{f}_m, f_m\right) \leq \left(1 + A_0 \varepsilon_n^+(m)\right) \frac{D_m}{2n} . \tag{13}$$

The proof of Theorem III.1, that can be found in the supplementary material [47, Section 2], is based on an improvement of independent interest of the previously best known concentration inequality for the chi-square statistics. See Section IV-A below for the precise result.

We obtain in Theorem III.1 sharp upper and lower bounds for the true and empirical excess risks on m. They are optimal at the first order since the leading constants are equal in the upper and lower bounds. They show the concentration of the true and empirical excess risks around the value $D_m/(2n)$. One should also notice that if $D_m > 1$, one always has $\mathbb{E}[\mathcal{K}(f_m, \hat{f}_m)] = +\infty$ since there is a positive (very small) probability that $\hat{f}_m$ vanishes on at least one element of the partition Λ_m.

Moreover, Theorem III.1 establishes equivalence with high probability of the true and empirical excess risks for models of reasonable dimension. This is in accordance with the celebrated Wilks's phenomenon, that ensures here that both $2n\mathcal{K}(f_m, \widehat{f}_m)$ and $2n\mathcal{K}(f_m, \widehat{f}_m)$ converge in distribution towards a chi-square distribution $\chi^2_{D_m}$ with D_m degrees of freedom, while their difference converges in probability to 0.

Concerning the control of the deviations in displays (12) and (13), we see more precisely that if $D_m \ll \sqrt{n}$, then the deviations are indeed of the order of a chi-square distribution with D_m degrees of freedom ([36, Lemma 1]). Indeed, the deviations at the right of $2n\mathcal{K}(f_m, \widehat{f}_m)$ and $2n\mathcal{K}(f_m, \widehat{f}_m)$ are smaller than the maximum between a sub-Gaussian term of order $\sqrt{D_m}$ and a sub-exponential term of order 1. The deviations at the left are of the order of a sub-Gaussian term proportional to $\sqrt{D_m}$. On the contrary, if $D_m \ll \sqrt{n}$, then the term reflecting the approximation of the scaled KL divergences to the chi-square statistics dominates over the previous sub-Gaussian term and is of order $D_m^{3/2}/\sqrt{n}$.

Another direction to get nonasymptotic bounds on the (rescaled) excess risks could be to look at the (Kolmogorov) distance to the $\chi^2_{D_m}$ distribution. The likelihood ratio is investigated in [2] in this perspective, using Stein's method for probability approximation. An open question would be in our case to determine precisely when the Kolgomorov distance between the rescaled excess risk $\chi^2_{D_m}$ distribution is competitive with the deviations of the latter. Does a transition occur around $D_m \approx \sqrt{n}$ as in our bounds?

Concentration inequalities for the excess risks as in Theorem III.1 is a new and exciting direction of research related to the theory of statistical learning and to high-dimensional statistics. Boucheron and Massart [22] obtained a pioneering result describing the concentration of the empirical excess risk around its mean, a property that they call a high-dimensional Wilks phenomenon. Then a few authors obtained results describing the concentration of the true excess risk around its mean [28], [42], [44] or around its median [16], [17] for (penalized) least square regression and in an abstract M-estimation framework [52]. In particular, recent results of [52] include the case of MLE on exponential models and as a matter of fact, on histograms. Nevertheless, we believe that Theorem III.1 is a valuable addition to the literature on this line of research since we obtain here not only concentration around a fixed point, but an explicit value $D_m/2n$ for this point. On the contrary, the concentration point is available in [52] only through an implicit formula involving local suprema of the underlying empirical process.

The principal assumption in Theorem III.1 is Inequality (10) of lower regularity of the partition with respect to P. It is ensured as soon as the density f_* is uniformly bounded from below and the partition is lower regular with respect to the reference measure μ (which will be the Lebesgue measure in our experiments). No restriction on the largest values of f_* are needed. In particular, we do not restrict to the bounded density estimation setting.

Castellan [26] proved inequalities that are related but weaker than those stated in Theorem III.1 above. She also asked for a lower regularity property of the partition, as in [26, Proposition 2.5], where she derived a sharp control of the KL divergence of the histogram estimator on a fixed model. More precisely, Castellan assumes that there exists a positive constant B such that

$$\inf_{I\in\Lambda_m} \mu(I) \geq B\frac{(\ln(n+1))^2}{n} \ . \tag{14}$$

This latter assumption is thus weaker than (10) - in the case where the target is uniformly bounded from below, as assumed by Castellan - for models of dimensions D_m that are smaller than the order $n(\ln(n+1))^{-2}$. We could assume (14) instead of (10) and restrict the dimensions D_m to be smaller than $A_+ n/(\ln(n+1))^2$ in order to derive Theorem III.1. This would lead to less precise results for second order terms in the deviations of the excess risks but the first order bounds would be preserved. More precisely, if we replace assumption (10) in Theorem III.1 by Castellan's assumption (14), a careful look at the proofs shows that the conclusions of Theorem III.1 are still valid for $\varepsilon_n^+(m) = \max\left\{(\ln(n+1))^{-1/2}; \sqrt{\ln(n+1)/D_m}; \ln(n+1)/D_m\right\}$ and $\varepsilon_n^-(m) = \max\left\{(\ln(n+1))^{-1/2}; \sqrt{\ln(n+1)/D_m}\right\}$. Thus assumption (10) is not a fundamental restriction in comparison to [26].

B. An Oracle Inequality

Let us state first the set of assumptions required to establish the nonasymptotic optimality of the over-penalization procedure. These assumptions will be discussed in more detail at the end of this section.

Set of assumptions (**SA**)

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

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $m \in \mathcal{M}_n$,

$$D_m \leq A_{\mathcal{M},+}\frac{n}{(\ln(n+1))^2} \leq n \ .$$

(P3) Richness of $\mathcal{M}_n$: there exist $c^-_{rich}, c^+_{rich} > 0$ such that for any $\lambda \in (0,1)$, there exists a model $m \in \mathcal{M}_n$ such that $D_m \in \left[\left\lceil c^-_{rich} n^\lambda \right\rceil, \left\lceil c^+_{rich} n^\lambda \right\rceil\right]$.

(Asm) The unknown density f_* satisfies some moment condition and is uniformly bounded from below: there exist some constants $A_{\min} > 0$ and $p \in (1, +\infty]$ such that,

$$\int_{\mathcal{Z}} f_*^p \left[(\ln f_*)^2 \vee 1\right] d\mu < +\infty$$

and

$$\inf_{z\in\mathcal{Z}} f_*(z) \geq A_{\min} > 0 \ . \tag{15}$$

(Alr) Lower regularity of the partition with respect to μ: there exists a positive finite constant A_Λ such that, for all $m \in \mathcal{M}_n$,

$$D_m \inf_{I\in\Lambda_m} \mu(I) \geq A_\Lambda \geq A_{\mathcal{M},+}(\alpha_{\mathcal{M}}+6)/\tau \ ,$$

where $\tau = \sqrt{\sqrt{6} - 3/\sqrt{2}} > 0$.

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

$$C_- D_m^{-\beta_-} \leq \mathcal{K}(f_*, f_m) \leq C_+ D_m^{-\beta_+} \ .$$

We are now ready to state our main theorem related to the performance of over-penalization.

Theorem III.2 *Take an integer $n \geq 1$ and two real constants $p \in (1, +\infty]$ and $r \in (0, p-1)$. For some $\Delta > 0$, consider the following penalty,*

$$\mathrm{pen}(m) = \left(1 + \Delta\varepsilon_n^+(m)\right)\frac{D_m}{n} \ , \qquad \text{for all } m \in \mathcal{M}_n \ . \tag{16}$$

*Assume that the set of assumptions (**SA**) holds and that*

$$\beta_- < p(1+\beta_+)/(1+p+r) \text{ or } p/(1+r) > \beta_- + \beta_-/\beta_+ - 1 \ . \tag{17}$$

Then there exists an event Ω_n of probability at least $1-(n+1)^{-2}$ and some positive constant A_1 depending only on the

*constants defined in (**SA**) such that, if $\Delta \geq A_1 > 0$ then we have on Ω_n,*

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq (1+\delta_n) \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}\left(f_*, \hat{f}_m\right) \right\} , \qquad (18)$$

where $\delta_n = L_{(\mathbf{SA}),\Delta,r} \left(\ln(n+1)\right)^{-1/2}$ works.

The proof of Theorem III.2 and further descriptions of the behavior of the procedure can be found in the supplementary material [47, Section 2.2].

We derive in Theorem III.2 a pathwise oracle inequality for the KL excess risk of the selected estimator, with constant almost one. Our result thus establishes the nonasymptotic quasi-optimality of over-penalization with respect to the KL divergence. More precisely, the convergence rate $\delta_n \propto 1/\sqrt{\ln(n+1)}$ in Inequality (18) is sufficient to ensure the asymptotic efficiency of the procedure and the question of the optimality of this rate under the assumptions of Theorem III.2 remains open.

The convergence rate is better in the leading constant of Inequality (33) of Theorem 2.3 of the supplementary material [47], but at the price of adding a remainder term to the oracle inequality (33). The rate δ_n then comes from comparing the bounds on the excess risk of an oracle model with the remainder term of Inequality (33) and under Assumption (17) of Theorem 3.2, this is the best rate that we can get from our computations. However, if we have more precise relations between β_-, β_+ and p than in Assumption (17), then the rate δ_n may be better, typically polynomially decreasing in n. For instance, taking the special case where $p = +\infty$, Assumption (17) is automatically satisfied and if we assume further that $\beta_- = \beta_+ =: \beta$, then it is easy to check from the proof of Inequality (34) in the supplementary material that $\delta_n \propto (\ln(n+1))^{3/2}/n^{1/(1+\beta)}$ works.

Note that the lower bound A_1 on the constant Δ that is required for our over-penalization to ensure oracle inequality (18) is unknown in general, since it depends on the constants involved in the set of assumptions (**SA**). In section V-A below, we propose either to set an ad hoc value for Δ, such as $\Delta = 1$, or to provide a data-driven calibration of it, that is based on the estimation of the variability of the empirical risk. The latter procedure achieves the best performances in our simulations. However, obtaining theoretical statistical guarantees for the data-driven calibration of Δ seems unreachable at this point, as it is rather delicate and involves several steps of computations (see Section V-A for further details).

Note also that our choice of the lower bound $1-(n+1)^{-2}$ for the probability on which the oracle inequality (18) is achieved, is rather arbitrary but it is quite a classical choice in model selection (as for instance in [10], [43], [45]), because it allows to integrate - at least for bounded losses - the trajectorial oracle inequality, to obtain an oracle inequality in expectation. In our case, the Kullback-Leibler divergence taken on the estimators has an infinite expectation - as already discussed in Section III-A - but our choice is still sensible. Indeed, having a more general polynomial bound in n would not change the essence of our result.

We could work with more irregular partitions and grant Assumption (14) corresponding to [26]. This would give another form of over-penalization. But we have two remarks on this point. Firstly, despite working with Assumption (14), we would still need the assumption that the density f_* is uniformly bounded from below - as in [26] -, but in this case Assumption (**Alr**) of lower-regularity of the partitions is arguably the most natural, since one would typically consider regular partitions to estimate such density. Secondly, the form of the over-penalization (16) would be different using Assumption (14) but the algorithm that allows to calibrate empirically the over-penalization term - procedure AIC_a in our experiments - would give actually essentially the same penalty as the one deduced from Assumption (**Alr**), since it is only based on an estimation of the deviations of the empirical risk for large models and on the fact that the excess risks concentrate at an exponential rate (see Section V-A).

It is worth noting that three features related to oracle inequality (18) significantly improve upon the literature. Firstly, inequality (18) expresses the performance of the selected estimator through its KL divergence and compare it to the KL divergence of the oracle. Nonasymptotic results pertaining to (robust) maximum likelihood based density estimation usually control the Hellinger risk of the estimator [27], [39], [20], [19], [12]. The main reason is that the Hellinger risk is easier to handle than the KL divergence from a mathematical point of view. For instance, the Hellinger distance is bounded by one while the KL divergence can be infinite. However, from an M-estimation perspective, the natural excess risk associated with likelihood optimization is indeed the KL divergence and not the Hellinger distance. These two risks are provably close to each other in the bounded setting [39], but may behave very differently in general.

Second, nonasymptotic results describing the performance of procedures based on penalized likelihood, by comparing more precisely the (Hellinger) risk of the estimator to the KL divergence of the oracle, all deal with the case where the log-density to be estimated is bounded ([27], [39]). Here, we substantially extend the setting by considering only the existence of a finite polynomial moment for the large values of the density to be estimated.

Finally, the oracle inequality (18) is always valid with positive probability, larger than $3/4$. To our knowledge, any other oracle inequality describing penalization performance for maximum likelihood density estimation is valid with positive probability only when the sample size n is greater than an integer n_0 which depends on the constants defining the problem and that is thus unknown. For instance, the quantities are controlled in [26] only on an event Ω_m (see (2.8) in [26]), that is of probability bounded below by $1 - C/n$ (see (2.10) in [26]), for C an unknown constant that depends on the parameters of the problem. So it can happen for $n < C$ that $\mathbb{P}(\Omega_m) = 0$. In such case, Castellan's results, even for the Hellinger distance, are empty (it would give an upper-bound for the Hellinger distance that would be greater than one, which is trivial, see the reminder term in the oracle inequality of Theorem 3.2 in [26]).

We emphasize that we control the risk of the selected estimator *for any sample size* and that this property is highly valuable in practice when dealing with small to medium

sample sizes. Based on the arguments developed in Section II-C, we believe that such a feature of Theorem III.2 is accessible only through the use of over-penalization and we conjecture in particular that *it is impossible using AIC to achieve such a control of the KL divergence of the selected estimator for any sample size*.

Let us mention that we give in [47, Theorem 2.3] of the supplementary material a more general result than Theorem III.2 above, considering penalties of the form,

$$\mathrm{pen}(m) = \mathrm{pen}_\theta(m) = (\theta + \Delta\varepsilon_n^+(m))\frac{D_m}{n},$$

for $\theta > 1/2$. Taking $\theta = 1$ is actually the best theoretical choice since it allows to optimize the bound given in [47, Theorem 2.3], in such a way that an oracle inequality is achieved, with leading constant converging to one. This choice, that is made in Theorem III.2 above, also corresponds to penalizing more than AIC, since the penalty is then greater than Akaike's penalty. Our more general penalty of [47, Theorem 2.3], that depends on $\theta > 1/2$, can, however, be smaller than Akaike's penalty if $\Delta\varepsilon_n^+(m) < 1-\theta$, which is asymptotically true for $\theta < 1$. But taking $\theta \neq 1$ is asymptotically a bad choice, since AIC is asymptotically efficient (at least in good cases). On the contrary, if $\Delta\varepsilon_n^+(m) > 1-\theta$, which can happen for small values of n, then $\mathrm{pen}_\theta(m)$ is greater than Akaike's penalty and this is, to our understanding, precisely the reason why we obtain a non-trivial inequality for any sample size n in [47, Theorem 2.3].

The oracle inequality (18) is valid under conditions (17) relating the values of the bias decaying rates β_- and β_+ to the order p of finite moment of the density f_* and the parameter r. In order to understand these latter conditions, let us assume for simplicity that $\beta_- = \beta_+ =: \beta$. Then the conditions (17) both reduce to $\beta < p/(1+r)$. As r can be taken as close to zero as we want, the latter inequality reduces to $\beta < p$. In particular, if the density to be estimated is bounded ($p = +\infty$), then conditions (17) are automatically satisfied. If on the contrary the density f_* only has finite polynomial moment p, then the bias should not decrease too fast. In light of the following comments, if f_* is assumed to be α-Hölderian, $\alpha \in (0,1]$, then $\beta \leq 2\alpha \leq 2$ and the conditions (17) are satisfied, in the case where $\beta_- = \beta_+$, as soon as $p \geq 2$.

To conclude this section, let us comment on the set of assumptions (**SA**). Assumption (**P1**) indicates that the collection of models has increasing polynomial complexity. This is well suited to bin size selection because in this case we usually select among a number of models which is strictly bounded from above by the sample size. In the same manner, Assumption (**P2**) is legitimate and corresponds to practice, where we aim at considering bin sizes for which each element of the partition contains a few sample points. Assumption (**P3**) ensures that there are enough models, that are well spread over possible dimensions. It is satisfied, of course, if one takes one model per dimension. From a technical viewpoint, assumption (**P3**) allows to obtain an oracle inequality (18) without a remainder term. See [47, Section 2.2] for technical details about this latter point.

Assumption (**Asm**) imposes conditions on the moments density to be estimated. Assumption (15) stating that the unknown density is uniformly bounded from below is also granted in [26]. It is, moreover, assumed in [26, Theorem 3.4], when deriving an oracle inequality for the (weighted) KL excess risk of the histogram estimator, that the target is of finite sup-norm. This corresponds to the case where $p = +\infty$ in (**Asm**), but the condition where $p \in (1, +\infty)$ is, of course, more general. Furthermore, from a statistical perspective, the lower bound (15) is coherent since, by Assumption (**Alr**), we use models of lower-regular partitions with respect to the Lebesgue measure. In the case where Inequality (15) would not hold, one would typically have to consider exponentially many irregular histograms to take into account the possibly vanishing mass of some elements of the partitions (for more details on this aspect that goes beyond the scope of the present paper, see for instance [39]).

We require in (**Ap**) that the quality of the approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of KL projections of the unknown density onto the models. For a density f_* uniformly bounded away from zero, the upper bound on the bias is satisfied when for example, $\mathcal{Z}$ is the unit interval, $\mu = \mathrm{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_m are regular and the density f_* belongs to the set $\mathcal{H}(H,\alpha)$ of α-hölderian functions for some $\alpha \in (0,1]$: if $f \in \mathcal{H}(H,\alpha)$, then for all $(x,y) \in \mathcal{Z}^2$

$$|f(x) - f(y)| \leq H|x-y|^\alpha \ .$$

In that case, $\beta_+ = 2\alpha$ is convenient and AIC-type procedures are adaptive to the parameters H and α, see [26].

In assumption (**Ap**) of Theorem III.2 we also assume that the bias $\mathcal{K}(f_*, f_m)$ is bounded from below by a power of the dimension D_m of the model m. This hypothesis is in fact quite classical as it has been used in [49], [24] for the estimation of density on histograms and also in [4], [5], [10] in the regression framework. Combining Lemmas 1 and 2 of Barron and Sheu [14] - see also Inequality (31) of Proposition IV.6 below - we can show that

$$\frac{1}{2}e^{-3\left\|\ln\left(\frac{f_*}{f_m}\right)\right\|_\infty}\int_{\mathcal{Z}}\frac{(f_m - f_*)^2}{f_*}d\mu \leq \mathcal{K}(f_*, f_m) \ .$$

Assuming for instance that the target is uniformly bounded, $\|f_*\|_\infty \leq A_*$, we get

$$\frac{A_{\min}^3}{2A_*^4}\int_{\mathcal{Z}}(f_m - f_*)^2 d\mu \leq \mathcal{K}(f_*, f_m) \ .$$

Now, since in the case of histograms the KL projection f_m is also the $L_2(\mu)$ projection of f_* onto m, we can apply Lemma 8.19 in Section 8.10 of Arlot [3] to show that assumption (**Ap**) is indeed satisfied for $\beta_- = 1 + \alpha^{-1}$, in the case where $\mathcal{Z}$ is the unit interval, $\mu = \mathrm{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_m are regular and the density f_* is a non-constant α-hölderian function.

IV. Probabilistic and Analytical Tools

In this section we set out some general results that are of independent interest and serve as tools for the mathematical

description of our statistical procedure. The first two sections contain new or improved concentration inequalities, for the chi-square statistics (Section IV-A) and for general log-densities (Section IV-B). We establish in Section IV-C some results that are related to the so-called margin relation in statistical learning and that are analytical in nature.

A. Chi-square Statistics' Concentration

The chi-square statistic plays an essential role in the proofs related to Section III-A. Let us recall its definition.

Definition IV.1 *Given some histogram model m, the chi-square statistics $\chi_n^2(m)$ is defined by*

$$\chi_n^2(m) = \int_{\mathcal{Z}} \frac{\left(\hat{f}_m - f_m\right)^2}{f_m} d\mu = \sum_{I \in m} \frac{(P_n(I) - P(I))^2}{P(I)} .$$

The following proposition provides an improvement upon the previously best known concentration inequality for the right tail of the chi-square statistics ([27], see also [39, Proposition 7.8] and [21, Theorem 12.13]).

Proposition IV.1 *For any $x, \theta > 0$, it holds*

$$\mathbb{P}\left(\chi_n(m)\,\mathbf{1}_{\Omega_m(\theta)} \geq \sqrt{\frac{D_m}{n}} + \left(1 + \sqrt{2\theta} + \frac{\theta}{6}\right)\sqrt{\frac{2x}{n}}\right) \leq \exp(-x) , \tag{19}$$

where we set $\Omega_m(\theta) = \bigcap_{I \in m} \{|P_n(I) - P(I)| \leq \theta P(I)\}$. More precisely, for any $x, \theta > 0$, it holds with probability at least $1 - e^{-x}$,

$$\chi_n(m)\,\mathbf{1}_{\Omega_m(\theta)} < \sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + 2\sqrt{\frac{\theta}{n}}\left(\sqrt{x} \wedge \left(\frac{xD_m}{2}\right)^{1/4}\right) + \frac{\theta}{3}\sqrt{\frac{x}{n}}\left(\sqrt{\frac{x}{D_m}} \wedge \frac{1}{\sqrt{2}}\right) . \tag{20}$$

The proof of Theorem IV.1 can be found in Section 1.1 of the supplementary material [47]. Essentially, we follow the same kind of arguments as those given in the proof of Castellan's inequality ([26, Inequality (4.27)]). In particular, the main tool is Bousquet's concentration inequality for the supremum of the empirical process at the right of its mean ([23]). However, we perform a slightly refined optimization of the quantities appearing in Bousquet's inequality.

Let us details the relationship of Proposition IV.1 with Castellan's inequality (in the form presented in [39, Proposition 7.8]), which is: for any $x, \varepsilon > 0$,

$$\mathbb{P}\left(\chi_n(m)\,\mathbf{1}_{\Omega_m(\varepsilon^2/(1+\varepsilon/3))} \geq (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}}\right)\right) \leq \exp(-x) . \tag{21}$$

By taking $\theta = \varepsilon^2/(1+\varepsilon/3) > 0$, we get $\varepsilon = \theta/6 + \sqrt{\theta^2/36 + \theta} > \theta/6 + \sqrt{\theta} > 0$. Assume that $D_m \geq 2x$. It is easy to check that Inequality (19) gives in this case a bound that is smaller than the one provided by Inequality (21). The essential improvement is that the constant in front of the term $\sqrt{D_m/n}$ is equal to one for our inequality instead of $1 + \epsilon$ for Castellan's.

To illustrate this improvement, let us mention that in our proofs we apply (19) with x proportional to $\ln(n+1)$ ([47, Section 2.1]). Hence, for most of the models of the collection, we have $x \ll D_m$ and as a result, the bounds that we obtain in Theorem III.1 by the use of Inequality (19) are substantially better than the bounds we would obtain by using Inequality (21) of [39]. More precisely, the deviation term $\sqrt{D_m \ln(n+1)/n}$ in (11) would be replaced by its square root $(D_m \ln(n+1)/n)^{1/4}$, thus degrading the order of magnitude for the deviations of the excess risks and changing the form of our over-penalization itself. Proposition IV.1 has thus a direct statistical impact in our study.

Finally, if $D_m \leq 2x$ then it is also easy to check that Inequality (20) improves upon Castellan's inequality (21).

The following result describes the concentration from the left of the chi-square statistics and is proved in the supplementary material [47, Section 1.1].

Proposition IV.2 *Let $\alpha,\ A_\Lambda > 0$. Assume $0 < A_\Lambda \leq D_m \inf_{I \in m}\{P(I)\}$. Then there exists a positive constant A_g depending only on A_Λ and α such that*

$$\mathbb{P}\left(\chi_n(m) \leq \left(1 - A_g\left(\sqrt{\frac{\ln(n+1)}{D_m}} \vee \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\right)\sqrt{\frac{D_m}{n}}\right) \leq (n+1)^{-\alpha} .$$

B. Bernstein type concentration inequalities for log-densities

The following propositions give concentration inequalities for the bias of log-densities. No structure is assumed for the densities, so these inequalities are general and may be of independent interest. These results are used in the proofs related to Theorem III.2 above by specifying the value of a density f to be equal to a projection f_m ([47, Section 2.1]).

Proposition IV.3 *Consider a density $f \in \mathcal{S}$. We have, for all $z \geq 0$,*

$$\mathbb{P}\left(P_n(\ln(f/f_*)) \geq \frac{z}{n}\right) \leq \exp(-z) . \tag{22}$$

Moreover, if we can take a finite quantity v which satisfies $v \geq \int (f \vee f_*)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu$, we have for all $z \geq 0$,

$$\mathbb{P}\left((P_n - P)(\ln(f/f_*)) \geq \sqrt{\frac{2vz}{n}} + \frac{2z}{n}\right) \leq \exp(-z) . \tag{23}$$

One can notice, with Inequality (22), that the empirical bias always satisfies some exponential deviations at the right of zero. In the Information Theory community, this inequality is also known as the "No Hyper-compression Inequality" ([32]).

Inequality (23) seems to be new and takes the form of a Bernstein-like inequality, even if the usual assumptions of Bernstein's inequality are not satisfied. In fact, we are able to

recover such a behavior by inflating the usual variance to the quantity v.

We now turn to concentration inequalities for the empirical bias at the left of its mean, where we also inflate the sub-Gaussian term to obtain a Bernstein-like inequality.

Proposition IV.4 *Let $r > 0$. For any density $f \in \mathcal{S}$ and for all $z \geq 0$, we have*

$$\mathbb{P}\left(P_n\left(\ln\left(f/f_*\right)\right) \leq -z/nr - (1/r)\ln\left(P\left[\left(f_*/f\right)^r\right]\right)\right) \leq \exp\left(-z\right) \ . \tag{24}$$

Moreover, if we can set a quantity w_r which satisfies $w_r \geq \int \left(\frac{f_*^{r+1}}{f^r} \vee f_*\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu$, then we get, for all $z \geq 0$,

$$\mathbb{P}\left(\left(P_n - P\right)\left(\ln\left(f/f_*\right)\right) \leq -\sqrt{\frac{2w_r z}{n}} - \frac{2z}{nr}\right) \leq \exp\left(-z\right) \ . \tag{25}$$

C. Margin-Like Relations

Our objective in this section is to control the variance terms v and w_r, appearing respectively in Propositions IV.3 and IV.4 above, in terms of the KL divergence pointed on the target f_*. This is done in Proposition IV.5 below under moment assumptions for f_*. Our inequalities generalize previous results of Barron and Sheu [14] obtained in the bounded setting (see also [39, Lemma 7.24]).

Proposition IV.5 *Let $p > 1$ and $c_+, c_- > 0$. Assume that the density f_* satisfies*

$$\begin{aligned} J &:= \int_{\mathcal{Z}} f_*^p \left(\left(\ln\left(f_*\right)\right)^2 \vee 1\right) d\mu < +\infty \\ Q &:= \int_{\mathcal{Z}} \frac{\left(\ln\left(f_*\right)\right)^2 \vee 1}{f_*^{p-1}} d\mu < +\infty \end{aligned} \tag{26}$$

Take a density f such that $0 < c_- \leq \inf_{z\in\mathcal{Z}}\left\{f\left(z\right)\right\} \leq \sup_{z\in\mathcal{Z}}\left\{f\left(z\right)\right\} \leq c_+ < +\infty$. Then, for some $A_{MR,d} > 0$ only depending on J, Q, p, c_+ and c_-, it holds

$$P\left[\left(\frac{f}{f_*} \vee 1\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right] \leq A_{MR,d}\mathcal{K}\left(f_*, f\right)^{1-\frac{1}{p}} \ . \tag{27}$$

More precisely,

$$A_{MR,d} = \left(4c_-^{1-p}\left(\left(\ln c_-\right)^2 \vee 1\right) J + 4c_+^p\left(\ln^2 c_+ \vee 1\right) Q\right)^{1/p}$$

holds. For any $0 < r \leq p-1$, we have the following inequality,

$$P\left[\left(\frac{f_*}{f} \vee 1\right)^r \left(\ln\left(\frac{f}{f_*}\right)\right)^2\right] \leq A_{MR,g}\mathcal{K}\left(f_*, f\right)^{1-\frac{r+1}{p}} \ , \tag{28}$$

available with

$$A_{MR,g} = \left(4c_-^{1-p}\left(\ln^2 c_- \vee 1\right) J + 2\left(\ln^2 c_+ + J + Q\right)\right)^{\frac{r+1}{p}} \ .$$

Proposition IV.5 states that the variance terms, appearing in the concentration inequalities of Section IV-B, are bounded from above, under moment restrictions on the density f_*, by a power less than one of the KL divergence pointed on f_*. The stronger are the moment assumptions, given in (26), the closer is the power to one. One can notice that J is a restriction on large values of f_*, whereas Q is related to values of f_* around zero.

We call these inequalities "margin-like relations" because of their similarity with the margin relations known first in binary classification ([38], [51]) and then extended to empirical risk minimization (see [6], [40] for instance). Indeed, from a general point of view, margin relations relate the variance of contrasted functions (logarithm of densities here) pointed on the contrasted target to a function (in most cases, a power) of their excess risk.

Now we reinforce the restrictions on the values of f_* around zero. Indeed, we ask in the following proposition that the target is uniformly bounded away from zero.

Proposition IV.6 *Let $p > 1$ and $A_{\min}, c_+, c_- > 0$. Assume that the density f_* satisfies*

$$\begin{aligned} J &:= \int_{\mathcal{Z}} f_*^p \left(\left(\ln\left(f_*\right)\right)^2 \vee 1\right) d\mu < +\infty \\ \text{and } 0 < A_{\min} &\leq \inf_{z\in\mathcal{Z}} f_*\left(z\right) \ . \end{aligned}$$

Then there exists a positive constant $A_{MR,-}$ only depending on $A_{\min}, J, r$ and p such that, for any $m \in \mathcal{M}_n$,

$$P\left[\left(\frac{f_m}{f_*} \vee 1\right)\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq A_{MR,-}\mathcal{K}\left(f_*, f_m\right)^{1-1/p} \tag{29}$$

and for any $0 < r \leq p-1$,

$$P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r \left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq A_{MR,-}\mathcal{K}\left(f_*, f_m\right)^{1-\frac{r+1}{p}} \ . \tag{30}$$

If moreover $\ln\left(f_*\right) \in L_\infty\left(\mu\right)$, i.e. $0 < A_{\min} \leq \inf_{z\in\mathcal{Z}} f_*(z) \leq \left\|f_*\right\|_\infty < +\infty$, then there exists $\tilde{A} > 0$ only depending on $r, A_{\min}$ and $\left\|f_*\right\|_\infty$ such that, for any $m \in \mathcal{M}_n$,

$$\begin{aligned} &P\left[\left(\frac{f_m}{f_*} \vee 1\right)\ln^2\left(\frac{f_m}{f_*}\right)\right] \vee P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r \ln^2\left(\frac{f_m}{f_*}\right)\right] \\ &\quad \leq \tilde{A}\mathcal{K}\left(f_*, f_m\right) \ . \end{aligned} \tag{31}$$

Proposition IV.6 is stated only for projections f_m because we actually take advantage of their special form (as local means of the target) in the proof of the proposition. The benefit, compared to results of Proposition IV.5, is that Inequalities (29), (30) and (31) do not involve assumptions on the values of f_m.

V. Experiments

A simulation study is conducted to compare the numerical performance of the model selection procedures we discussed. We demonstrate the usefulness of our procedure on simulated data examples. The numerical experiments were performed using R.

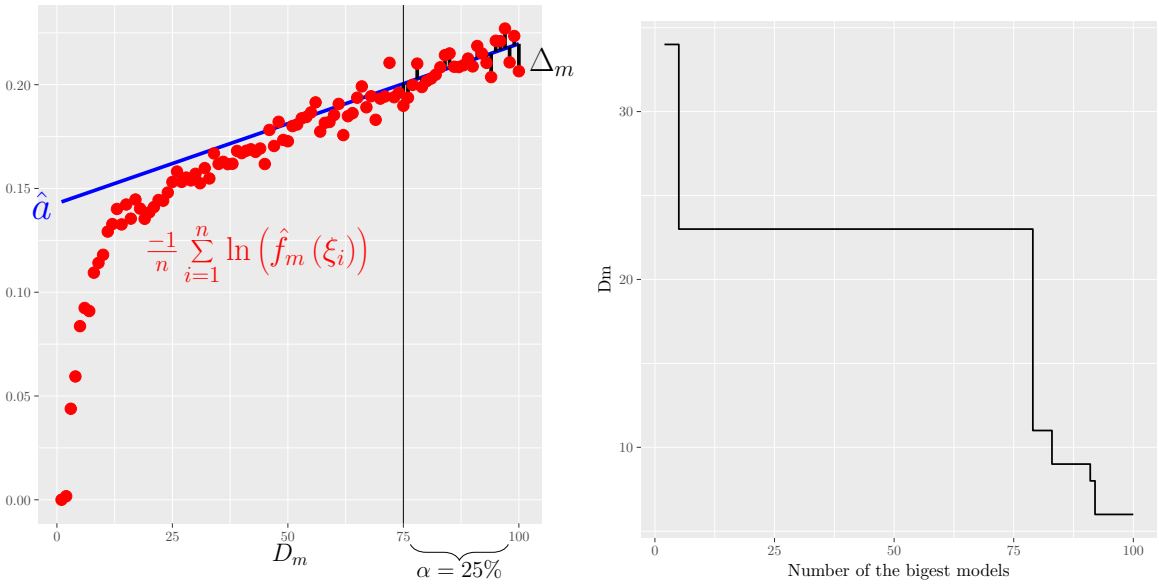


Fig. 4. Estimation of the over-penalization constant.

A. Experimental Setup

We have compared the numerical performance of our procedure with the classic methods of penalization of the literature on several densities. In particular, we consider the estimator of [20] and AICc ([33], [50]). We also report on AIC's behavior. In the following, we name the procedure of [20] by BR, and our criterion AIC_1 when the constant $C = 1$ in (9) and AIC_a for a fully adaptive, data-driven procedure which will be detailed below. More specifically, the performance of the following four model selection methods were compared:

1. AIC:

$$\widehat{m}_{\text{AIC}} \in \underset{m\in\mathcal{M}_n}{\arg\min}\left\{P_n(-\ln\hat{f}_m) + \frac{D_m}{n}\right\},$$

2. AICc:

$$\widehat{m}_{\text{AICc}} \in \underset{m\in\mathcal{M}_n}{\arg\min}\left\{P_n(-\ln\hat{f}_m) + \frac{D_m}{n - D_m - 1}\right\},$$

3. BR:

$$\widehat{m}_{\text{BR}} \in \underset{m\in\mathcal{M}_n}{\arg\min}\left\{P_n(-\ln\hat{f}_m) + \frac{D_m}{n} + \frac{\log^{2.5}(D_m+1)}{n}\right\},$$

4. AIC_1:

$$\widehat{m}_{\text{AIC}_1} \in \underset{m\in\mathcal{M}_n}{\arg\min}\left\{P_n(-\ln\hat{f}_m) + \text{pen}_{\text{AIC}_1}(m)\right\},$$

with

$$\text{pen}_{\text{AIC}_1}(m) = \left(1 + 1\times\varepsilon_n^+(m)\right)\frac{D_m}{n},$$

5. AIC_a:

$$\widehat{m}_{\text{AIC}_a} \in \underset{m\in\mathcal{M}_n}{\arg\min}\left\{P_n(-\ln\hat{f}_m) + \text{pen}_{\text{AIC}_a}(m)\right\},$$

$$\text{pen}_{\text{AIC}_a}(m) = \left(1 + \hat{C}\varepsilon_n^+(m)\right)\frac{D_m}{n},$$

where $\hat{C} = 6\times\text{median}_{\alpha\in\mathcal{P}}\hat{C}_\alpha$, with $\hat{C}_\alpha = \text{median}_{m\in\mathcal{M}_\alpha}|\hat{C}_m|$, where

$$\hat{C}_m = \frac{\Delta_m}{\max\left\{\sqrt{\frac{D_m}{n}};\sqrt{\frac{1}{D_m}}\right\}\frac{D_m}{2n}},$$

Δ_m is the least-squares distance between the opposite of the empirical risk $-P_n(\gamma(\widehat{f}_m))$ and a fitted line of equation $y = xD_m/(2n) + \hat{a}$ (Figure 4 at the left), $\mathcal{P}$ is the set of proportions α corresponding to the longest plateau of equal selected models when using penalty (9) with constant $C = \hat{C}_\alpha$ (Figure 4 at the right) and $\mathcal{M}_\alpha$ is the set of models in the collection associated to the proportion α of the largest dimensions.

The models that we used along the experiments are made of histogram densities defined on regular partitions of the interval $[0,1]$ (with the exception of the density Isosceles triangle which is supported on $[-1,1]$), from a cardinal equal to 1 to $\lceil n/\ln(n+1)\rceil$. Thus the cardinal of our collection of models is $\text{Card}(\mathcal{M}_n) = \lceil n/\ln(n+1)\rceil$.

We show the performance of the proposed method for a set of four test distributions (see Figure 5) and described in the *benchden*[1] R-package [41] which provides an implementation of the distributions introduced in [18].

Let us explain the ideas underlying the design of the procedure AIC_a given above. According to the definition of penalty $\text{pen}_{\text{opt},\beta}$ given in (7), the constant $\hat{C}$ in the penalty $\text{pen}_{\text{AIC}_a}$ should be computed so that the penalty provides an estimate of the quantile of order $1-\beta_{\mathcal{M}}$, where $\beta_{\mathcal{M}} = \beta/\text{Card}(\mathcal{M}_n)$, of the sum of excess risk and empirical excess risk on the models of the collection.

Based on Theorem III.1, we can also assume that the deviations of excess risk and excess empirical risk are of the same

[1]Available on the CRAN http://cran.r-project.org.

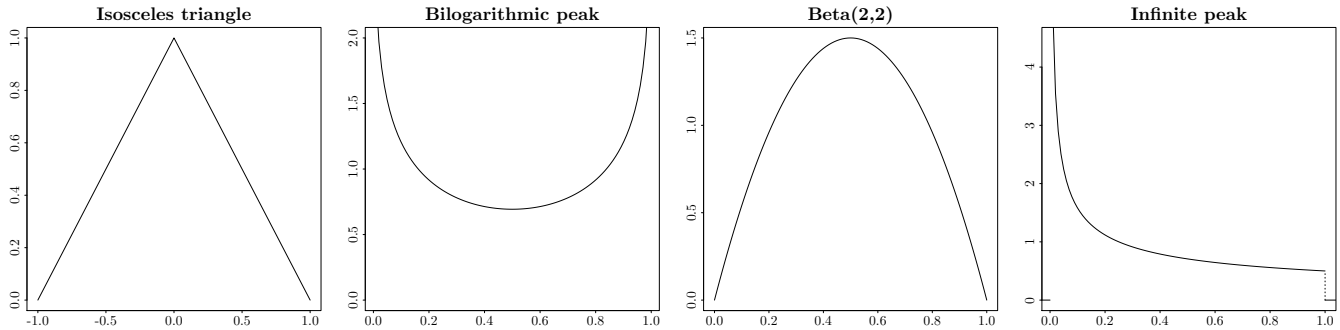


Fig. 5. Test densities f.

order. By choosing β of the order of $(n+1)^{-2}$ as in Theorem III.2 above, and considering that $\mathrm{Card}(\mathcal{M}_n) \simeq n \simeq n+1$, we arrive to a choice of $\beta_{\mathcal{M}} = (n+1)^{-3}$. The latter value impacts the over-penalization through a factor $3\ln(n+1)$ because the concentration of the excess risks is exponential. Putting things together, the over-penalization $\hat{C}$ should be given by $3 \times 2 = 6$ times the normalized deviations of the empirical excess risk.

Moreover, considering the largest models in the collection neglects questions of bias and, therefore, the median of the normalized deviations of the empirical risk around its mean for the largest models should be a reasonable estimator of the constant C.

Finally, the remaining problem is to give a tractable definition to the "largest models" in the collection. To do this, we choose a proportion α of the largest dimensions of the models at hand and calculate using these models an estimator $\hat{C}_\alpha$ of the constant C in (9). We then proceed for each α in a grid of values between 0 and 1 to a model selection step by over-penalization using the constant $C = \hat{C}_\alpha$. This gives us a graph of the selected dimensions with respect to the proportions (Figure 4 at the right). Finally, we define our over-penalization constant $\hat{C}$ as the median of the values of the constants $\hat{C}_\alpha$, $\alpha \in \mathcal{P}$ where $\mathcal{P}$ is the largest plateau in the graph of the selected dimensions with respect to proportions α.

Note that we make use of the plot of the empirical risk as a function of the dimension D_m. This is a common point with the slope estimation procedure in the so-called slope heuristics [7], [15], but our use of the plot of the empirical risk substantially differs from the slope estimation, in that we consider that the slope is known and is given by Akaike's penalty and we estimate the order of deviations of the empirical risk around this slope, for large enough models.

B. Results

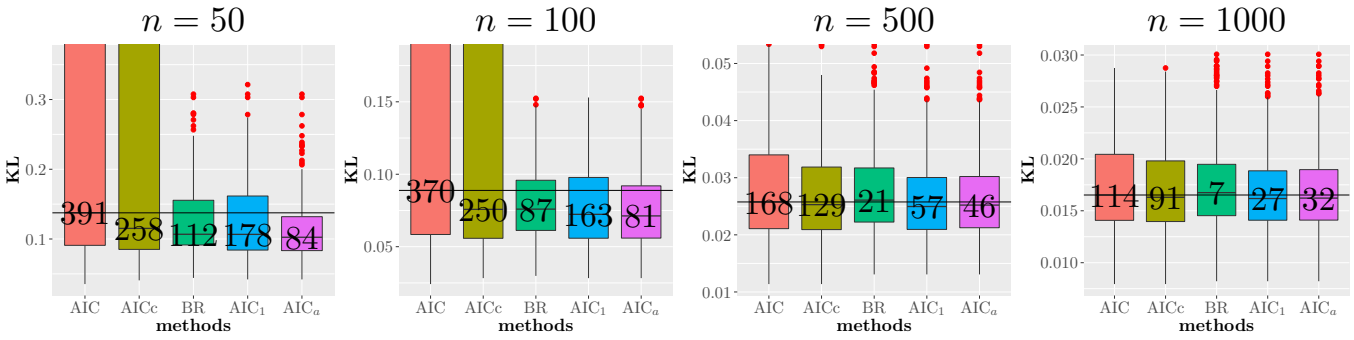
We compared procedures on $N = 1000$ independent data sets of size n ranging from 50 to 1000. We estimate the quality of the model selection strategies using the median KL divergence, on the one hand, and the median squared Hellinger distance, on the other hand. Boxplots were made of the KL risk - resp. the Hellinger distance - over the N trials. The horizontal lines of the boxplots indicate the $5\%, 25\%, 50\%, 75\%$, and 95% quantiles of the error distribution. The median value of AIC (horizontal black line) is also superimposed for visualization purposes. It can be seen from Figure 6 (resp. Figure 7) that, as expected, for each method and in all cases, the KL divergence (resp. the squared Hellinger distance) decreases as the sample size increases. We also see clearly that there is generally a substantial advantage in modifying AIC for sample sizes smaller than 1000.

We see from Figure 6 pertaining to KL divergence, that AIC_a is quite clearly the most advisable procedure in practice for small to moderate sample sizes, since it is the most stable while being one of the most efficient procedures. It indeed outperforms all the other procedures for a very small sample size (50 or 100) and is as good as AIC_1 (and comparable or better than the other procedures) for a moderate sample size. The picture is quite the same when looking at the Hellinger risk (Figure 7), except that now AIC_a and AIC_1 have comparable performances in all settings.

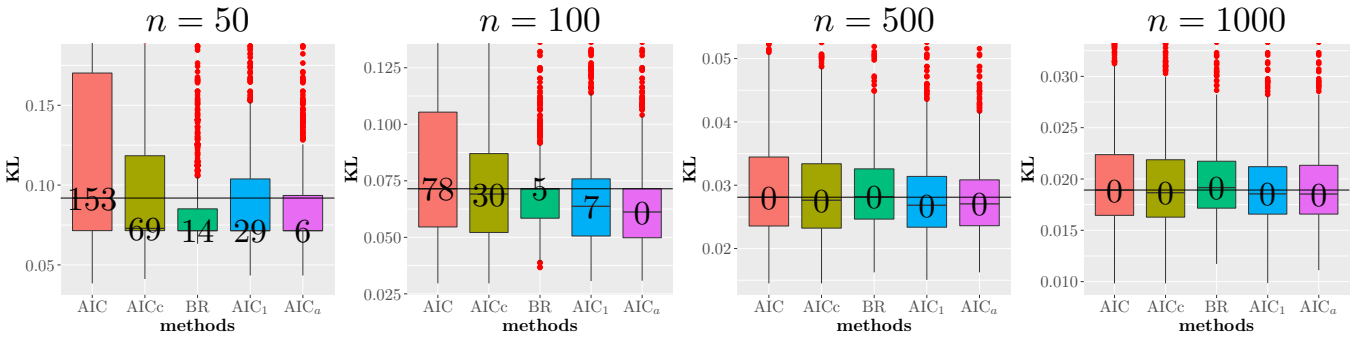
But AIC_a comes at a price of more computations that the other considered procedures. If a computational simplicity equivalent to AIC is required, then we recommend using AIC_1 rather than AICc or BR. Indeed, compared to AIC_1, it seems that AICc is not penalizing enough, which translates into a worse performance for samples equal to 50 and 100. On the contrary, it seems that the BR criterion penalizes too much. As a result, its performance deteriorates relative to other methods as the sample size increases.

VI. Conclusion

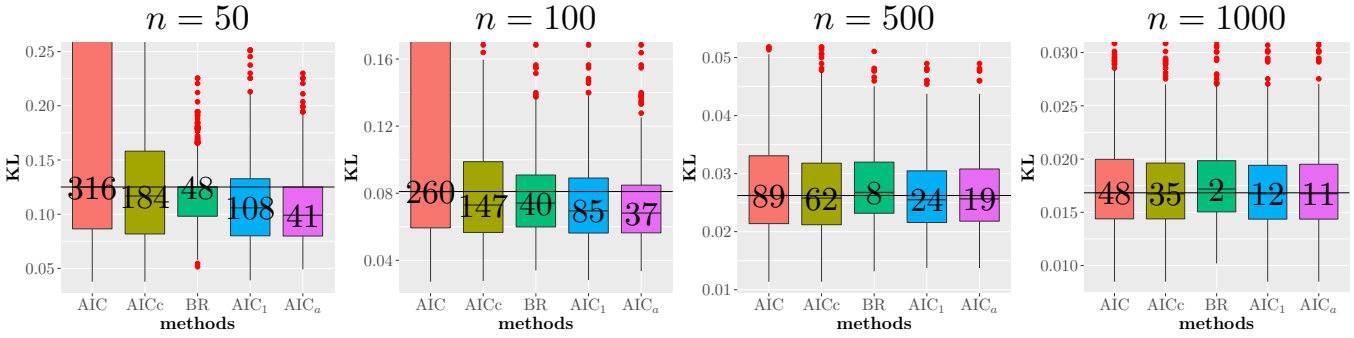
In this work, we tackled the delicate, but well-known question of the lack of efficiency of AIC for small to moderate sample sizes. Several modifications of AIC have been already proposed, such as AICc ([33]) or the correction due to Birgé and Rozhenholc ([20]). We introduced a new correction of AIC that is based on estimating the quantiles - at the right order - of the true and empirical excess risks of the estimators at hand. By focusing on histograms, we were able to give sharp concentration bounds for the excess risks and to discuss the quality of our model selection procedure in an unbounded setting. We provided more precisely an oracle inequality that holds with positive probability without any remainder term and for any sample size. We also provided an algorithm of data-driven calibration of our correction term, that seems to be most often in our experiments the most accurate procedure.



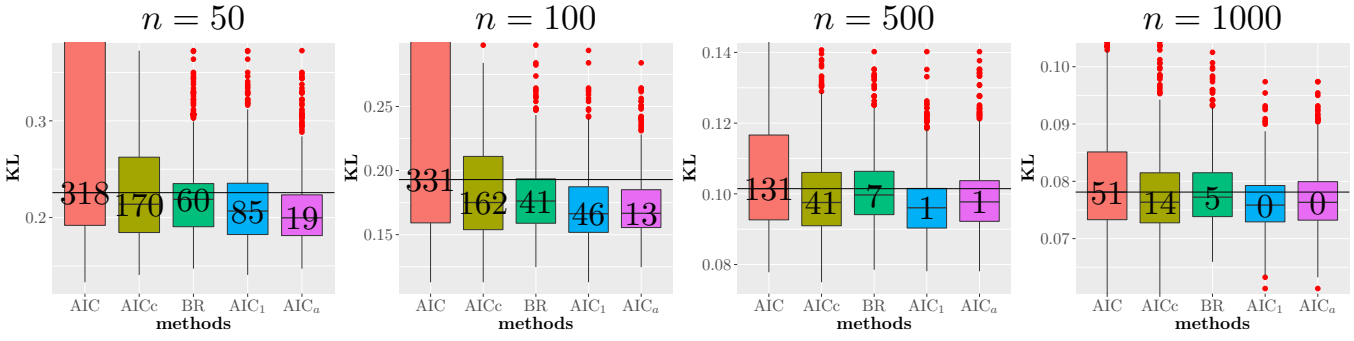
(a) Isosceles triangle



(b) Bilogarithmic peak

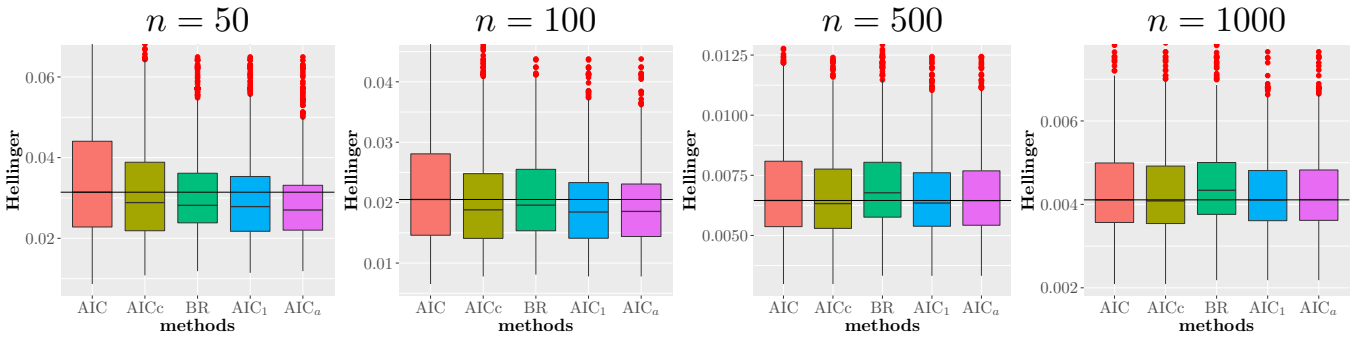


(c) Beta (2,2)

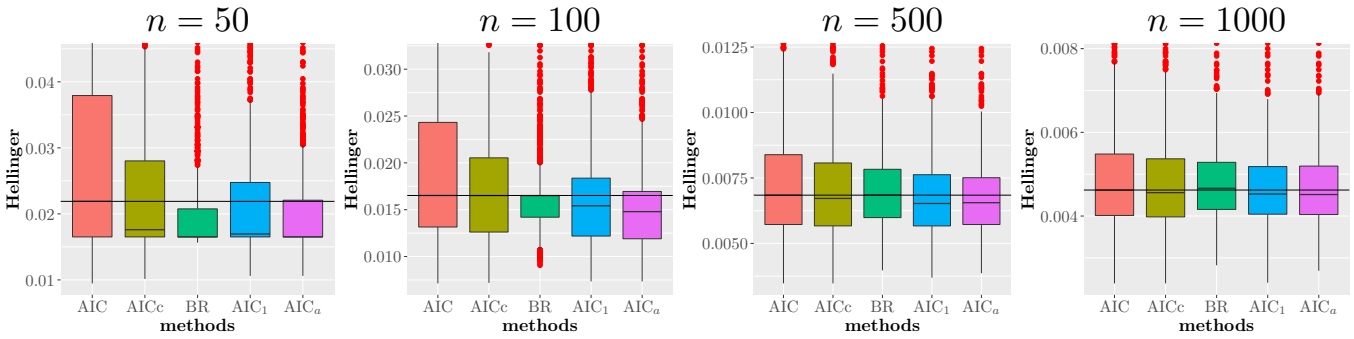


(d) Infinite peak

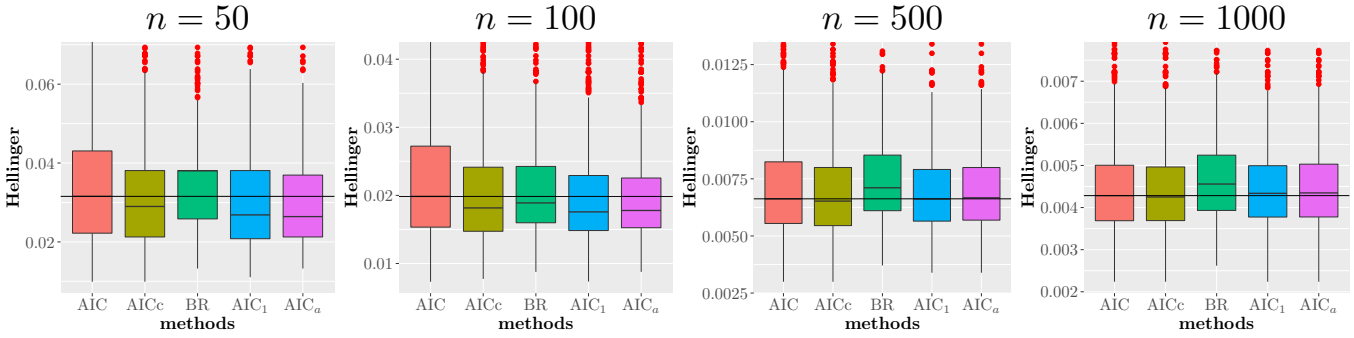
Fig. 6. KL divergence results. Box plots of the KL divergence to the true distribution for the estimated distribution. The solid black line corresponds to the AIC KL divergence median. The term inside the box is the number of times the KL divergence equals ∞ out of 1000.



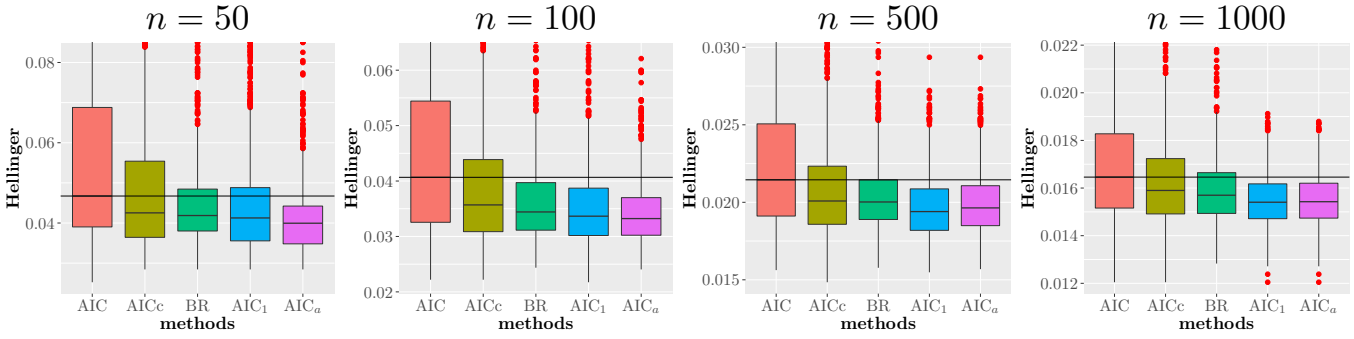
(a) Isosceles triangle



(b) Bilogarithmic peak



(c) Beta (2,2)



(d) Infinite peak

Fig. 7. Hellinger distance results. Box plots of the Hellinger distance to the true distribution for the estimated distribution. The solid black line corresponds to the AIC Hellinger distance median.

Many directions of research for extending this work are open. Indeed, one can notice that the rationale behind our over-penalization procedure is not based on the particular value of the MLE contrast or the specific choice of the models and that other M-estimation context could be tackled. The crucial point to understand is indeed the excess risk's concentration and so, available results constitute a good basis for future work [43], [46], [52]. Our over-penalization strategy could thus be investigated for more general exponential models in MLE estimation ([52]), or for other contrasts, such as the least-squares density contrast ([9], [52]) or the least-squares regression contrast (with projection estimators [46]) and even for regularized estimators ([52]). We could also tackle the correction of other model selection criteria than the theoretically designed penalties and in our opinion, the correction of V-fold penalties ([4], [9], [43]) and its comparison to the classical V-fold cross-validation is a particularly attractive direction of research.

VII. Supplementary material

The supplement [47] to "Finite sample improvement of Akaike's Information Criterion" contains in Sections 1 and 2 the proofs of the results described in this article as well as some theoretical extensions that complement the description of the over-penalization procedure.

Acknowledgment

The first author warmly thanks Matthieu Lerasle for instructive discussions on the topic of estimation by tests - which appeared to be useful in the process of this work - and Alain Célisse for a nice discussion at a early stage of this work. He is also grateful to Pascal Massart for having pushed him towards obtaining better oracle inequalities than in a previous version of this study. We owe thanks to Sylvain Arlot and Amandine Dubois for a careful reading that helped to correct some mistakes and improve the presentation of the paper. Finally, we deeply thank the associate editor and two anonymous referees for their painstaking and insightful comments that have led to an improvement of the article.

References

[1] H. 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.

[2] A. Anastasiou and G. Reinert. Bounds for the asymptotic distribution of the likelihood ratio. *Ann. Appl. Probab.*, 30(2):608–643, 2020.

[3] S. Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, Dec. 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[4] S. Arlot. $V-$fold cross-validation improved: $V-$fold penalization. arXiv:0802.0566v2, 2008.

[5] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[6] S. Arlot and P. L. Bartlett. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 05 2011.

[7] S. Arlot. Minimal penalties and the slope heuristics: a survey *J. SFdS*, 160(3):1–106, 2019.

[8] S. Arlot and A. Célisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[9] S. Arlot and M. Lerasle. Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation. *J. Mach. Learn. Res.*, Paper No. 208, 50 pp., 2016.

[10] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[11] S. Arlot. Minimal penalties and the slope heuristics: a survey. *J. SFdS*, 160(3):1–106, 2019.

[12] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

[13] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[14] A. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[15] J.-P. Baudry and C. Maugis and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[16] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Towards the study of least squares estimators with convex penalty. *Actes du 1er Congrès National de la SMF—Tours, 2016,* 109–136, Sémin. Congr., 31, *Soc. Math. France, Paris,* 2017.

[17] P. Bellec and A. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. *Modern Problems of Stochastic Analysis and Statistics*, 315–333, Springer Proc. Math. Stat., 208, *Springer, Cham,* 2017.

[18] A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38(3):3–59, 1994.

[19] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[20] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10:24–45 (electronic), 2006.

[21] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[22] S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011.

[23] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[24] P. Burman. Estimation of equifrequency histogram. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

[25] C. Butucea, J. F. Delmas, A. Dutfoy and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electron. J. Stat.*, 11(1):2258–2294, 2017.

[26] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université Paris-Sud*, 1999.

[27] G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.

[28] S. Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 12 2014.

[29] G. Claeskens and N. L. Hjort. *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2008.

[30] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[31] A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

[32] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[33] C. M. Hurvich and C.-L. Tsai. Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.*, 9(3):259–265, 1990.

[34] C. Lacour and P. Massart. Minimal penalty for the Goldenshluger-Lepski method. *Stochastic Process. Appl.*, 126(12):3774–3789, 2016.

[35] C. Lacour, P. Massart and V. Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335, 2017.

[36] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[37] O. V. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1991.

[38] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann.Statist.*, 27:1808–1829, 1999.

[39] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[40] P. Massart and E. Nédélec. Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366, 2006.

[41] T. Mildenberger and H. Weinert. The benchden package: Benchmark densities for nonparametric density estimation. *J. Stat. Softw.*, 46(14):1–14, 2012.
[42] A. Muro and S. van de Geer. Concentration behavior of the penalized least squares estimator. *Stat. Neerl.* 72 (2018), no. 2, 109–125.
[43] F. Navarro and A. Saumard. Slope heuristics and V-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM Probab. Stat.*, 21:412–451, 2017.
[44] A. Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6(1-2):579–655, 2012.
[45] A. Saumard. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Statist.*, 7:1184–1223, 2013.
[46] A. Saumard. A concentration inequality for the excess risk in least-squares regression with random design and heteroscedastic noise. arXiv preprint arXiv:1702.05063, 2017.
[47] A. Saumard and F.Navarro. Supplement to "Finite sample improvement of Akaike's Information Criterion". 2021.
[48] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.
[49] C. Stone. An asymptotically optimal histogram selection rule. *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*. Vol. 2. Wadsworth, 1984.
[50] N. Sugiura. Further analysts of the data by Akaike's information criterion and the finite corrections. *Commun. Stat. Theory Methods*, 7(1):13–26, 1978.
[51] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32:135–166, 2004.
[52] S. van de Geer and M. J. Wainwright. On concentration for (regularized) empirical risk minimization. *Sankhya A*, 79(2):159–200, Aug 2017.
[53] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.
[54] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007.

Adrien Saumard received a Ph.D. degree from Université de Rennes 1, France, in 2010, and is currently Associate Professor in the Center for Research in Economics and Statistics at Ecole Nationale de la Statistique et de l'Analyse de l'Information, France. His main research interests are in model selection, statistical learning and Stein's method in probability and statistics.

Fabien Navarro received the B.Sc., M.Sc. and Ph.D. degrees in Applied Mathematics from the University of Caen, Caen, France, in 2008, 2010 and 2013, respectively. From 2014 to 2015, he was a Research Assistant Professor with the department of Mathematics and Statistics, Concordia University, Montreal, Canada. From 2015 to 2021, he was an Assistant Professor with the Center for Research in Economics and Statistic, Ecole Nationale de la Statistique et de l'Analyse de l'Information, Bruz, France. He is currently an Associate Professor with the University of Paris 1 Panthéon-Sorbonne, Paris, France. His research interests include nonparametric statistics, inverse problems, computational harmonic analysis, sparse representations, machine learning and statistical approaches in graph signal processing.