Estimating and testing the order of a model

By Antoine CHAMBAZ

Université Paris-Sud
France Télécom R&D

We study in this paper some asymptotic properties of joint procedures of estimation and test of the order of a model. Given a collection of nested parametric models and a distribution in their union, the order of the latter is the order of the smallest model the distribution belongs to. It is a number that quantifies the complexity of the model regarding the whole collection. We define two estimators which yield two test procedures. We prove under mild assumptions their almost sure consistency, possibly without any prior bound on the order of interest $K^\star$. We also cope with their rates of under- and overestimation (*i.e.* when the estimated order is lower or larger than the true one, respectively). These rates respectively bound above the errors of second and first kind when testing $K^\star \leq K_0$ *vs.* $K^\star > K_0$. We get under mild assumptions a universal lower bound for the underestimation rate as well as upper bounds which all exponentially decrease with the sample size n. We prove that the best possible rate in exponential models is achieved. Finally, we show under mild assumptions that the overestimation rate is necessarily slower than exponential in n and get exponential rates in $n^{1-\delta}$ (any $\delta \in]0,1[$). The originality of our approach relies in the substitution of functional arguments to explicit calculus while dealing with maximum likelihood. This confers greater generality on our results. Some benchmark examples (mixture of Gaussians, abrupt changes and various regressions) are carefully addressed.

1. Introduction.

1.1. *Problem at stake.* Let $Z_1, \ldots, Z_n$ be n independent and identically distributed random variables (rv) taking their values in the measurable sample space $(\mathcal{Z}, \mathcal{F})$ with $\mathcal{Z}$ Polish. The observations are defined on a common probability space $(\Omega, \mathcal{A}, \mathrm{P})$ upon which all the rv will be defined in the sequel.

We assume that the distribution $P^\star$ of $Z_1, \ldots, Z_n$ may belong to the increasing family $\{\Pi_K\}_{K\geq 1}$ of models. Here, any Π_K is a parametric collection of probability distributions dominated by μ

$$\Pi_K = \{P_\theta = p_\theta \, \mathrm{d}\mu : \theta \in \Theta_K\} \subset \Pi_{K+1},$$

[1]Received

AMS 2000 subject classifications. Primary 60F10, 60G57, 62C99, 62F03, 62F05, 62F12

Key words and phrases. Empirical processes, Error exponents, Huber's trick, Hypothesis testing, Large deviations, Mixtures, Model selection, Moderate deviations, Order estimation, Penalization, Stein's lemma

where $\{(\Theta_K, d_K)\}_{K\geq 1}$ is an increasing family of metric parameter sets (since there is usually no confusion, we shall abbreviate in the sequel d_K to d). In the whole sequel, we shall suppose that the following assumption holds true:

A1 The parameter sets (Θ_K, d) are compact metric sets. The models Π_K are compact for the weak topology on the space $M_1(\mathcal{Z})$ of all the probability measures on $(\mathcal{Z}, \mathcal{F})$.

We call K the order of the model Π_K. It is also the order of any $P_\theta \in \Pi_K \setminus \Pi_{K-1}$ (by convention, $\Pi_0 = \varnothing$). Informally, K quantifies the complexity of Π_K among the whole collection. The order of $P^\star = P_{\theta^\star}$ is denoted $K^\star$, which is infinite whenever $P^\star$ does not belong to the collection.

In this paper, we mainly investigate two test procedures of

$$H_0: \ K^\star \leq K_0 \quad vs. \quad H_1: \ K^\star > K_0$$

with a particular view to the rates at which the first and second kind errors vanish as the sample size increases. To this end, we define two estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ of the order $K^\star$. Then, we reject H_0 in favour of H_1 whenever $\widehat{K}_n^{\mathrm{L}} > K_0$ or $\widehat{K}_n^{\mathrm{G}} > K_0$ according to the chosen test procedure. We must emphasize that the second and first kind errors are respectively bounded above by the rates of under- and overestimation, that is the rates at which underestimation (*i.e.* when the estimated order is lower than $K^\star$) and overestimation (*i.e.* when the estimated order is greater than $K^\star$) probabilities vanish.

$\widehat{K}_n^{\mathrm{G}}$ is defined by (3). It is a global (hence the G in its name) maximizer of a penalized maximum likelihood criterion (1). It always bounds above $\widehat{K}_n^{\mathrm{L}}$, the first local (hence the L) maximizer of the same criterion, *cf.* (2). Note that the computation of these estimators is a less demanding algorithmic task for $\widehat{K}_n^{\mathrm{L}}$ than for $\widehat{K}_n^{\mathrm{G}}$. We prove hereafter that both are almost surely consistent under mild assumptions, *without and with a prior bound* on $K^\star$, respectively. We get, still under mild assumptions, a universal lower bound for the underestimation rate as well as upper bounds which all exponentially decrease with n. We prove that the best possible rate in exponential model is achieved. Finally, we show under mild assumptions that the overestimation rate is necessarily slower than exponential in n and get exponential rates in $n^{1-\delta}$ (any $\delta \in]0,1[$). The originality of our approach relies in the substitution of functional arguments to explicit calculus while dealing with maximum likelihood. This confers greater generality on our results. Further comments follow.

1.2. *Two maximum penalized likelihood estimators.* Let us denote $\ell_\theta = \log p_\theta$ (for any $\theta \in \Theta_\infty = \cup_{K\geq 1}\Theta_K$, with $\ell_{\theta^\star} = \ell^\star$) and define the log-likelihood ℓ_n of the observations

$$\ell_n(\theta) = \sum_{i=1}^n \ell_\theta(Z_i) \quad (\text{any } \theta \in \Theta_\infty).$$

The penalized maximum likelihood criterion for the model Π_K is

$$\operatorname{crit}(n, K) = \sup_{\theta\in\Theta_K} \ell_n(\theta) - \operatorname{pen}(n, K), \tag{1}$$

where the positive penalty function pen satisfies the following assumption throughout this paper:

A2 On the one hand, $\mathrm{pen}(n, \cdot)$ is an increasing function for any $n \geq 1$. On the other hand, $\mathrm{pen}(n, K) \to \infty$ as $n \to \infty$ and $\mathrm{pen}(n, K) = o(n)$ for any $K \geq 1$.

We shall focus on two estimators of the order $K^\star$ of $P^\star$, namely

$$\widehat{K}_n^{\mathrm{L}} = \inf \{K \geq 1 : \mathrm{crit}(n, K) \geq \mathrm{crit}(n, K+1)\}, \tag{2}$$

$$\widehat{K}_n^{\mathrm{G}} = \arg\sup_{K \geq 1} \{\mathrm{crit}(n, K)\} \geq \widehat{K}_n^{\mathrm{L}}. \tag{3}$$

The need for a penalization term while estimating by minimization of an empirical criterion is well known since Mallows (1973), Akaike (1974), Rissanen (1978) and Schwarz (1978) published their seminal papers. The literature about penalization is abundant and it would be illusory to try to present here a comprehensive overview on that subject. We nonetheless refer the reader to (Barron et al., 1999; Massart, 2000) for an insight on model selection *via* penalization with a view to risk bounds. We must stress on that occasion that our point of view diverges from the latter risk bound model selection. We indeed choose the loss function $K \mapsto 1\!\!1\{K \neq K^\star\}$. We also refer the reader to the collection of papers concerned with estimation of an order that we shall cite hereafter: indeed, all of them use some kind of penalization.

1.3. *Forthcoming results and methods.* In the sequel, we shall use linear functional notations for expectations and integrals, *i.e.* write λf for the integral $\int f \mathrm{d}\lambda$ of a function f wrt a measure λ.

We shall assume that all the expressions involving extrema and empirical processes are measurable. It suffices for instance that each of them equals an extremum over a countable set, almost surely.

Almost sure consistency. We shall first prove that, under mild assumptions, $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ are consistent almost surely, respectively *without and with a prior bound* $K_{\max}$ *on* $K^\star$ (*cf.* Theorems 1, 2, 3).

The result for $\widehat{K}_n^{\mathrm{L}}$ does not require a prior bound on $K^\star$, but a peculiar additional assumption is needed: For all $K < K^\star$, the inclusion $\Pi_K \subset \Pi_{K+1}$ is strengthened by the condition $H(\Pi_{K+1} \,|\, P^\star) < H(\Pi_K \,|P^\star)$.

Here, $H(P\,|Q) = P \log \frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Kullback-Leibler divergence of P wrt Q when $P << Q$ (it is infinite otherwise) and $H(\Pi\,|Q) = \inf\{H(P\,|Q) : P \in \Pi\}$, $H(P\,|\Pi) = \inf\{H(P\,|Q) : Q \in \Pi\}$. We refer the reader to (Dupuis & Ellis, 1997) for an overview of some useful properties the Kullback-Leibler divergence satisfies.

The proofs of consistency are twofold, since we have to consider separately the cases of underestimation and overestimation. They indeed correspond to different phenomena, *i.e.* second and first kind errors respectively, as already stressed before. From a more technical point of view:

- when underestimating, one has to control the bias term $\sup_{\theta \in \Theta_K} \ell_n(\theta)$.

Now, if we set $K_1 \geq 1$ and K_2 equals to $K_1 + 1 \leq K^\star$ or $K^\star$, the bias term satisfies almost surely

$$\limsup_{n\to\infty} n^{-1}\left(\sup_{\theta\in\Theta_{K_1}} \ell_n(\theta) - \sup_{\theta\in\Theta_{K_2}} \ell_n(\theta)\right) < 0$$

(*cf.* Lemma 1) and particularly, this motivates the condition $\mathrm{pen}(n, K) = o(n)$.

- when overestimating, one has to control the variance term $\mathrm{pen}(n, K)$ thanks to the bias term.

  Now, if we set $K_2 > K^\star$ and K_1 equals to $K^\star$ or $K_2 - 1$, the likelihood ratio statistic

$$\sup_{\theta\in\Theta_{K_2}} \ell_n(\theta) - \sup_{\theta\in\Theta_{K_1}} \ell_n(\theta)$$

  should be thought as bounded in probability (in the classical framework of a test of a null hypothesis versus its alternative, a Taylor expansion yields under mild assumptions that the likelihood ratio statistic is asymptotically halved-chi-squared distributed). This particularly motivates the condition $\mathrm{pen}(n, K) \to \infty$ as $n \to \infty$.

The proofs of consistency heavily rely on a Bounded Law of the Iterated Logarithm (bounded LIL) due to Dudley and Philipp (1983). It yields some conditions on the minimum rate of divergence of $\mathrm{pen}(n, K)$. We postpone further comments on the conditions needed for the penalization function in Section 1.6.

Rate of underestimation. We shall then focus on the underestimation probabilities. First, we shall obtain an optimal lower bound for any estimator of the order that does not almost surely overestimate the true order (*cf.* Proposition 4). It is shown under mild assumptions that such an estimator $\widetilde{K}_n$ of $K^\star$ satisfies

$$\liminf_{n\to\infty} \log \mathrm{P}(\widetilde{K}_n < K^\star) \geq -\inf_{K<K^\star} H(\Pi_K | P^\star) < 0.$$

The proof relies on the application of a general version of Stein's lemma (Bahadur et al., 1980), see Lemma 3.

Second, we shall derive upper bounds on the underestimation probabilities for our estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ (*cf.* Theorems 5, 6).

The main feature of our approach should be emphasized now: the analysis of underestimation and overestimation events is based on functional arguments, *i.e.* are described as events concerning the empirical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{Z_i}$. Furthermore, we shall avoid (as far as possible) using the properties of the maximum likelihood estimator of the parameter $\theta^\star$ associated with the true distribution $P^\star$, though our estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ are maximum likelihood estimators.

Thus for instance, the event $[\widehat{K}_n^{\mathrm{L}} = K < K^\star]$ is obviously included in the event $[\mathrm{crit}(n, K) \geq \mathrm{crit}(n, K + 1)]$ which in turn is included (for n large enough and any $\alpha > 0$) in the events $[\mathbb{P}_n \in \Lambda_{\alpha,K}]$, where $\Lambda_{\alpha,K} = \{Q : \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \geq -\alpha\}$.

The crucial tool is therefore a Sanov's Large Deviations Principle (LDP) for the empirical measure $\mathbb{P}_n$. We actually use the generalized Sanov's Theorem 4 from (Léonard & Najim, 2000), regarding $\mathbb{P}_n$ as a linear functional on the set of functions f which admit some exponential moments wrt $P^\star$, *i.e.* equivalently f such that there exists $a > 0$ ensuring $P^\star \exp(a|f|) < \infty$. The latter particularly yields

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} \wedge \widehat{K}_n^{\mathrm{G}} < K^\star) < 0.$$

Third, we shall prove that $\widehat{K}_n^{\mathrm{G}}$ achieves the best underestimation rate whenever the models we consider are exponential (*cf.* Proposition 5). The proof involves H-projections as defined and studied by Csiszár (1975) and Čencov (1982) and convexity arguments.

Rate of overestimation. Finally, we consider the overestimation probabilities. Stein's lemma is used again and implies that the rate of convergence is slower than exponential in n for any estimator $\widetilde{K}_n$ which does not almost surely underestimate the true order $K^\star$, *i.e.* (*cf.* Proposition 6)

$$\lim_{n\to\infty} n^{-1} \log \mathrm{P}(\widetilde{K}_n > K^\star) = 0.$$

We derive some rates of overestimation from a Moderate Deviations Principle (MDP) due to Wu (1994), again with a linear functional approach. It is shown under mild assumptions (*cf.* Theorems 7, 8) that

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} \vee \widehat{K}_n^{\mathrm{G}} > K^\star) < 0$$

whenever $\mathrm{pen}(n, K) = v_n\, D(K)$ with $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ also increasing such that $n^{1/2} v_n^{-1} = o(1)$ (a secondary assumption on $\{v_n\}$ is also necessary).

1.4. *Companion examples: Mixtures, Abrupt changes and Various regressions.* Let us now introduce four companion examples that fit in our framework – they will be carefully addressed in Section 5.

Mixture of distributions. Let $X_1, \ldots, X_n$ be n independent *hidden* rv that take their values in an unknown finite subset $\{u_1^\star, \ldots, u_{K^\star}^\star\}$ of the compact set $\mathcal{U}$ of $\mathbb{R}^q$ (we do not know $K^\star$ either), with unknown distribution $\mathrm{P}(X_i = u_j^\star) = \pi_j^\star$ $(j = 1, \ldots, K^\star)$. We observe

- **Location mixture of Gaussian distributions example (MGM):**

  $$Z_i = X_i + \sigma\, e_i \quad (i = 1, \ldots, n),$$

  where $e_1, \ldots, e_n$ are independent, drawn along a centered Gaussian distribution of variance 1 independently of $X_1, \ldots, X_n$ and σ is *a priori* known.

- **Location-scale mixture of Gaussian distributions example (RKO):**

  $$Z_i = X_{i,1} + X_{i,2}\, e_i \quad (i = 1, \ldots, n),$$

  where $e_1, \ldots, e_n$ are independent, drawn along a centered Gaussian distribution of variance 1 independently of $X_1 = (X_{1,1}, X_{1,2}), \ldots, X_n = (X_{n,1}, X_{n,2})$.

Abrupt changes. Let $(\mathcal{X}, \mathcal{B}, P)$ be an open subset of $\mathbb{R}^q$ equipped with the trace of the Borel σ-field and a probability measure P dominated by the Lebesgue measure $\mu^{\otimes q}$, whose density wrt $\mu^{\otimes q}$ is denoted p. Let $\tau^\star$ be an unknown finite partition of $\mathcal{X}$ of (unknown, too) cardinality $K^\star$, *i.e.* a finite collection $(\tau_j^\star)_{1\le j\le K^\star}$ of subsets $\tau_j^\star$ of $\mathcal{X}$ such that $P(\mathcal{X}\setminus \cup_j \tau_j^\star) = 0$, $P(\tau_j^\star) > 0$ and $P(\tau_j \cap \tau_{j'}) = 0$ for any $j, j' \neq j$. Finally, let $\mathcal{M}$ be a compact subset of $\mathbb{R}$, $m_j^\star$ $(j = 1, \dots, K^\star)$ be unknown elements of $\mathcal{M}$ and $f^\star(x) = \sum_j m_j^\star 1\!\!1\{x \in \tau_j^\star\}$. We observe

- **Abrupt changes in the mean example (AC):**
$$Z_i = (X_i, Y_i), \text{ with } \qquad Y_i = f^\star(X_i) + \sigma\, e_i \quad (i = 1, \dots, n),$$
where $X_1, \dots, X_n$ are independent and P-distributed; $e_1, \dots, e_n$ are independent of $X_1, \dots, X_n$, with centered Gaussian distribution of variance 1; σ is *a priori* known.

Various regressions. Let $\{t_K\}_{K\ge 1}$ be a uniformly bounded, free system of continuous functions on $[0, 1]$. Let $\mathcal{U}$ be a compact subset of $\mathbb{R}$ that contains 0, $\theta_j^\star \in \mathcal{U}$ $(j = 1, \dots, K^\star)$ be unknown elements of $\mathcal{U}$ (we do not know $K^\star$ either) such that $\theta_{K^\star}^\star \neq 0$ and $f^\star = \sum_j \theta_j^\star t_j$. We observe

- **Various regressions example (VR):**
$$Z_i = (X_i, Y_i), \text{ with } \qquad Y_i = f^\star(X_i) + \sigma\, e_i \quad (i = 1, \dots, n),$$
where $X_1, \dots, X_n$ are independent, uniformly distributed on $[0, 1]$; $e_1, \dots, e_n$ are independent of $X_1, \dots, X_n$, with centered Gaussian distribution of variance 1; σ is *a priori* known.

1.5. *Earlier references for comparison.* Let us briefly survey the history of order estimation. The selection is of course again far from being comprehensive. We tried to pick up some important papers regarding the encompassed results and enhancements, methods and bibliographies. The interested reader should refer to the papers themselves for more precise overviews for each of them.

In most of earlier works on order estimation, the choice of the framework is contingent on the need for tractable explicit calculus. This remark applies to (Henna, 1985), (Senoussi, 1990), (Leroux, 1992); (Antoniadis & Berruyer, 1986), (Lindsay, 1989), (Dacunha-Castelle & Gassiat, 1997); (Haughton, 1989), (Dacunha-Castelle & Gassiat, 1999), (Keribin, 2000), (Haughton & Keribin, 2001); (Chambaz, 2002). In (Merhav, 1989), the author focuses in exponential families on the underestimation rates for estimators whose overestimation rates are assumed exponential in n. This is an important restriction, *cf.* Proposition 6. Besides, the example of (Guyon & Yao, 1999) is particularly relevant. Here, the framework allows dependence of the observations and nonasymptotic evaluations of under- and overestimation probabilities are derived (with a prior bound on $K^\star$). Their assumptions (standard conditions of identifiability and smoothness) are satisfied in various models (including regression models with least squares estimation, time series and random fields) as soon as the empirical criterion whose minimization yields the order estimator is *factorized*, *i.e.* whenever there exist a deterministic function U and a *finite dimensional* statistic

T_n of the observations $Z_1, \dots, Z_n$ such that the criterion is of the form $U(\theta, T_n)$. We emphasize that this condition is not fulfilled *e.g.* for mixture models.

In contrast, no explicit calculus are needed in (James et al., 2001), where the order estimation relies on kernel density estimation and an iterative scheme. The same distinction holds true for (Gassiat, 2002), where the author notably addresses the estimation of the order of a mixture with Markov regime thanks to two novel, general, simple yet powerful inequalities on likelihood ratios.

In conclusion, there is still another competitive field of research concerned with order estimation into which our framework does not fit. It is the field of the related (but different) estimations of the order of a Markov chain on a finite alphabet or a Hidden Markov Model. There too, one can distinguish the Markov chain framework from the Hidden Markov Model one on the basis of the tractability of calculus. Thus, subtle arguments based on martingales allow Csiszár and Shields to prove that the Bayesian Information Criterion (BIC) leads to an almost surely consistent estimator of the order, *without any prior bound on* $K^\star$ (*cf.* Section 1.6 for further important remarks). On the contrary, the point of view in (Gassiat & Boucheron, 2001) and in the current one have much in common, and particularly the "functional approach". Gassiat and Boucheron prove the almost sure consistency of estimators related to the BIC estimator *without any prior bound on* $K^\star$. They also derive tight upper bounds on the rate of underestimation by resort to large deviations techniques for empirical processes (more comments in the sequel). From our point of view, this illustrates the generality of our approach.

1.6. *Conditions on the penalization. Huber's trick.* We have argued informally in Section 1.3 why the penalty function has to satisfy **A2**. The condition $\mathrm{pen}(n, K) = o(n)$ will be sufficient for dealing with the underestimation. On the contrary, the condition $\mathrm{pen}(n, K) \to \infty$ as $n \to \infty$, which is necessary for dealing with the overestimation, will be made precise *via* some lower bounds on this rate of divergence.

Besides, general considerations on penalization (particularly in a framework of model selection with a view to risk bounds) yields that the lower is the penalization $\mathrm{pen}(n, K)$, the better. So, we are interested in exhibiting as small lower bounds on the latter rate of convergence as possible.

Now, it must be emphasized that our asymptotic study of the problem intrinsically does not allow to obtain conditions on the dependence of $\mathrm{pen}(n, K)$ wrt K. Such conditions may be derived in a nonasymptotic framework, for instance *via* concentration inequalities as in (Barron et al., 1999; Massart, 2000), but certainly under restrictive assumptions. It must be stressed that Csiszár and Shields (2000) precisely identify the penalty function as $\mathrm{pen}(n, K) = |A|^K(|A|-1)/2 \cdot \log n$ (where A is the finite alphabet of the Markov chain). It seems that this is the minimal penalty that can be used in the Markov (on a finite alphabet) order estimation problem (*cf.* (Gassiat & Boucheron, 2001)). In (Gassiat & Boucheron, 2001), the behaviour in K of the penalty function is also made more precise.

In this paper, we shall give an example of what we call (perhaps familiarly) an Huber's trick in order to derive better lower bounds on pen, *i.e.* smaller ones as explained above. Our illustration of Huber's trick is given in Proposition 1 below, whose proof is postponed to Appendix A.1.

It applies to the proofs of Theorems 1, 2 (*cf.* Proposition 3) and Theorems 7, 8.

PROPOSITION 1. *Set $K_2 > K_1 \geq K^\star$, the order of $P^\star$. Then, both inequalities below hold true, the second one providing an example of Huber's trick:*

$$\sup_{\theta\in\Theta_{K_2}} |(\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star)| \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n \ell_\theta \quad \text{and} \tag{4}$$

$$\left(\sup_{\theta\in\Theta_{K_2}} \left|(\mathbb{P}_n - P^\star)\frac{\ell_\theta - \ell^\star}{H(P^\star|P_\theta)^{1/2}}\right|\right)^2 \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n \ell_\theta. \tag{5}$$

Thus for instance, when studying the rate of overestimation for $\mathrm{pen}(n, K) = v_n D(K)$, the main condition $n^{1/2}v_n^{-1} = o(1)$ obtained thanks to (4) of Proposition 1 can be replaced by the much slighter one $(\log n)v_n^{-1} = o(1)$ whenever (5) is invoked.

The general idea behind Huber's trick is that rescaling often enhances "performances" when empirical processes methods and techniques are involved. Such a trick *e.g.* allowed Pollard (1985) to propose some simple proofs of uniform central limit theorems, accordingly to the original idea of Huber as presented in its 1967 paper. One can find more recent examples in (Barron et al., 1999) (see Proposition 7 therein and the attached remark) and (Massart, 2000) in a framework of model selection. Or in (Gassiat, 2002), where the author derives the consistency of the maximum penalized marginal likelihood estimator of the order of a mixture with Markov regime (*i.e.* for instance as in **MGM** or **RKO**, with a finite state space Markov chain $\{X_i\}$ and conditionally independent Y_i's).

1.7. *Organization of the paper.* Sections 2, 3 and 4 are respectively devoted to the results about consistency, underestimation and overestimation. Section 5 is dedicated to the careful presentation of the examples sketched in the Introduction. We have postponed in Appendix some proofs of individual interest or not essential for the comprehension.

2. Consistency.

2.1. *Statement of the results.* We denote $p^\star$ and $\ell^\star$ the density and the log-density of $P^\star << \mu$ (forgetting the index $\theta^\star$) in order to encompass the case of $P^\star \notin \Pi_\infty = \cup_{K\geq 1}\Pi_K$.

From now on, Log denotes the truncated log, *i.e.* $\mathrm{Log}(x) = \log(x \vee e)$ $(x \in \mathbb{R})$. The function φ is defined by $\varphi(x) = x^2/\mathrm{Log}\,\mathrm{Log}(x)$ $(x \in \mathbb{R})$.

Let us state some assumptions (**C** stands for Consistency, **P** for Penalty) before our first result about consistency.

C1 $H(P^\star|\Pi_1)$ is finite.

C2 If $P^\star \notin \Pi_K$, then $H(P^\star|\Pi_{K+1}) < H(P^\star|\Pi_K)$.

C3 There exist $l, u \in \mathbb{R}^{\mathcal{Z}}$ such that $(u-l) \in L^1(P^\star)$ with

$$l \leq \ell^\star \leq u \quad \text{and} \quad l \leq \ell_\theta \leq u \quad (\text{any } \theta \in \Theta_\infty).$$

C4 The parameterization $\theta \mapsto \ell_\theta(z)$ from Θ_K to $\mathbb{R}$ is continuous for any $z \in \mathcal{Z}$ and $K \geq 1$.

C5a Let $K > K^\star$ be an integer to be chosen later. The class of functions

$$\mathcal{G}_K^{\mathrm{a}} = \left\{ g_\theta = (\,\ell_\theta - \ell^\star) : \theta \in \Theta_K \right\}$$

is $P^\star$-Donsker and $\varphi(u-l) \in L^1(P^\star)$.

PCa The penalty function satisfies, for any $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{and} \quad \limsup_{n\to\infty} \frac{(n \log\log n)^{1/2}}{\mathrm{pen}(n, K)} = 0.$$

For a recent presentation of Donsker classes, refer to (van der Vaart, 1998).

Let $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ be the two estimators we defined in (2) and (3).

THEOREM 1. *Assume* **C1**, **C3**, **C4** *and* **PCa**. *Suppose moreover that* $P^\star \in \Pi_\infty$, *precisely* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star - 1}$.

• *If* **C5a** *is fulfilled for* $K = K^\star + 1$, if **C2** is satisfied, *then* without any prior bound on $K^\star$,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = K^\star \right) = 1.$$

• *Furthermore,* if we know a prior bound $K_{\max}$ on $K^\star$, *if moreover* **C5a** *holds true for* $K = K_{\max}$, *then* without invoking **C2**,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = K^\star \right) = 1,$$

We prove in Section 5 that the assumptions of the theorem are satisfied in **MGM** for $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ and for $\widehat{K}_n^{\mathrm{G}}$ in **VR**. In **AC**, only almost never asymptotic under-estimation holds true (*cf.* Proposition 2).

Indeed, all the assumptions above are mild (instead of **C1**, one actually needs the existence of a minimal $K_{\min} \geq 1$ such that $H(P^\star | \Pi_{K_{\min}}) < \infty$: one then considers the models Π_K for $K \geq K_{\min}$) but perhaps **C5a**, whose goal is to guarantee the application of the bounded LIL of (Dudley & Philipp, 1983).

Now, a moderate strengthening of **C5a** (*cf.* **C5b** below) in conjunction with Huber's trick inequality (5) yield the stark relaxed assumption **PCb** on the penalty function needed in Theorem 2. We prove in Section 5 that the assumptions of this alternative theorem about consistency are satisfied in the **MGM**.

Let $H(\theta)$ (any $\theta \in \Theta_\infty$) denote throughout the quantity $H(P_\theta | P^\star)$:

C5b Let $K > K^\star$ be an integer to be chosen later. The class of functions

$$\mathcal{G}_K^{\mathrm{b}} = \left\{ g_\theta = \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}} : \theta \in \Theta_K,\ H(\theta) > 0 \right\}$$

is $P^\star$-Donsker. Moreover, there exists an envelope function G for the class $\mathcal{G}_K^{\mathrm{b}}$ such that $\varphi(G) \in L^1(P^\star)$.

PCb The penalty function also satisfies, for any $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{and} \quad \limsup_{n\to\infty} \frac{\log\log n}{\mathrm{pen}(n, K)} = 0.$$

Theorem 2. *Assume* **C1, C3, C4** *and* **PCb**. *Suppose moreover that* $P^\star \in \Pi_\infty$, *precisely* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star - 1}$.
• *If* **C5b** *is fulfilled for* $K = K^\star + 1$, if **C2** is satisfied, *then* without any prior bound on $K^\star$,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = K^\star\right) = 1.$$

• *Besides,* if we know a prior bound $K_{\max}$ on $K^\star$, *if moreover* **C5b** *holds true for* $K = K_{\max}$, *then* without invoking **C2**,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = K^\star\right) = 1.$$

Finally, what does happen when the true distribution $P^\star$ does not belong to the collection of nested models ?

Theorem 3. *Assume* **C2, C3, C4** *hold true. Suppose moreover that* $P^\star \notin \Pi_\infty$. *Then*

$$\mathrm{P}\left(\liminf_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = \infty\right) = \mathrm{P}\left(\liminf_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = \infty\right) = 1.$$

2.2. *Proofs.*

The two estimators almost surely asymptotically do not underestimate. Lemma 1 casts some light on the behaviour of the bias term. Its proof essentially relies on the achievement of $H(P^\star|\Pi_K)$, the Strong law of large numbers and the Borel-Lebesgue property. It is the key of the proofs of Theorem 3 and Proposition 2 which copes with the underestimation case and gives the first halves of Theorems 1 and 2.

Lemma 1. *Assume* **C3** *and* **C4**. *Then almost surely, for any* $K \geq 1$,

$$\sup_{\theta\in\Theta_K} n^{-1}\Big(\ell_n(\theta) - \ell_n(\theta^\star)\Big) \underset{n\to\infty}{\longrightarrow} -H(P^\star|\Pi_K).$$

Proof of Theorem 3. It suffices to prove the result for $\widehat{K}_n^{\mathrm{L}}$ because it bounds below $\widehat{K}_n^{\mathrm{G}}$. On the event of probability 1 of Lemma 1 and for all $K \geq 1$, one has

$$\sup_{\theta\in\Theta_{K+1}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta \underset{n\to\infty}{\longrightarrow} H(P^\star|\Pi_K) - H(P^\star|\Pi_{K+1}) > 0$$

in virtue of **C2**, because $P^\star \notin \Pi_K$. Therefore, **A2** yields that

$$\mathrm{crit}(n, K+1) - \mathrm{crit}(n, K) \underset{n\to\infty}{\longrightarrow} \infty,$$

hence $\widehat{K}_n^{\mathrm{L}} > K$ for n large enough. This is true for any $K \geq 1$, so the proof is complete.

The next proposition is concerned with underestimation. Its assumptions are fulfilled in **MGM**, **RKO** and **AC** for $\widehat{K}_n^{\mathrm{L}}$. They hold true in the latter and also in **VR** for $\widehat{K}_n^{\mathrm{G}}$, see Section 5.

PROPOSITION 2. *Assume* **C3** *and* **C4**. *For* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$*:*
- *under* **C2**, $\widehat{K}_n^{\mathrm{L}}$ *almost surely asymptotically does not underestimate* $K^\star$.
- $\widehat{K}_n^{\mathrm{G}}$ *almost surely asymptotically does not underestimate* $K^\star$.

PROOF OF PROPOSITION 2. Let us abbreviate "infinitely often" to io and prove that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star \text{ io}) = 0$ (minor changes allow to cope with $\widehat{K}_n^{\mathrm{G}}$). Since

$$\left[\widehat{K}_n^{\mathrm{L}} < K^\star \text{ io}\right] \subset \bigcup_{K=1}^{K^\star-1} \left[\widehat{K}_n^{\mathrm{L}} = K \text{ io}\right],$$

it suffices to prove that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K \text{ io}) = 0$ for any $1 \leq K < K^\star$. Let Ω_1 be the event of probability 1 of Lemma 1 and let us denote $\delta = H(P^\star|\Pi_{K+1}) - H(P^\star|\Pi_K) < 0$ (in virtue of **C2**).

$$\begin{aligned}\left[\widehat{K}_n^{\mathrm{L}} = K \text{ io}\right] \cap \Omega_1 &\subset \left[\sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K+1}} \mathbb{P}_n \ell_\theta \geq -\delta/2 \text{ io}\right] \cap \Omega_1 \\ &\subset \left[\liminf_{n\to\infty} \left\{\sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K+1}} \mathbb{P}_n \ell_\theta\right\} \geq -\delta/2\right] \cap \Omega_1,\end{aligned}$$

where the first inclusion stems from **A2** and is satisfied for n large enough and the second one is immediate. Now, Lemma 1 yields that the last event has probability zero, and this concludes the proof.

The two estimators almost surely asymptotically do not overestimate. The next proposition is concerned with the overestimation case and gives the second halves of Theorems 1 and 2.

PROPOSITION 3. *Assume* **C3** *and* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$.
- $\widehat{K}_n^{\mathrm{L}}$ *almost surely asymptotically does not overestimate* $K^\star$ *whenever* **C5a** *(resp.* **C5b***) holds true for* $K = K^\star + 1$ *and* pen *satisfies* **PCa** *(resp.* **PCb***);*
- $\widehat{K}_n^{\mathrm{G}}$ *almost surely asymptotically does not overestimate* $K^\star$ *whenever* **C5a** *(resp.* **C5b***) holds true for* $K = K_{\max}$*, where* $K_{\max}$ *is a prior bound on* $K^\star$*, and* pen *satisfies* **PCa** *(resp.* **PCb***).*

We prove in Section 5 that the assumptions of the proposition are satisfied for both alternatives **a** and **b** in **MGM**, and for the alternative **a** in **VR**.

We shall need the following lemma which is a simple consequence of the bounded LIL of (Dudley & Philipp, 1983) (*cf.* their Theorem 4.1).

LEMMA 2. *Let us assume that* **C5a** *(resp.* **C5b***) holds true for some* $K > K^\star$ *and denote* $\mathcal{G} = \mathcal{G}_K^{\mathrm{a}}$ *(resp.* $\mathcal{G} = \mathcal{G}_K^{\mathrm{b}}$*). Then, there exists a positive constant* C_K *such* *that, almost surely,*

$$\limsup_{n\to\infty} \frac{n^{1/2}\sup_{g\in\mathcal{G}} |(\mathbb{P}_n - P^\star)g|}{(\log\log n)^{1/2}} \le C_K.$$

PROOF OF PROPOSITION 3. The scheme of proof is simple: *primo* apply Proposition 1, and precisely (4) for the alternative **a**, (5) for the alternative **b** in order to, *secundo*, apply Lemma 2 and conclude.

• Let us prove that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star \text{ io}) = 0$. The event $[\widehat{K}_n^{\mathrm{L}} > K^\star \text{ io}]$ is included in

$$\left[\sup_{\theta\in\Theta_{K^\star+1}} \mathbb{P}_n\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n\,\ell_\theta \ge n^{-1}\{\mathrm{pen}(n,K^\star+1) - \mathrm{pen}(n,K^\star)\} \text{ io}\right]$$

which in turn is included in (we set $K = K^\star + 1$)

$$\left[\left\{\frac{(n\log\log n)^{1/2}}{\mathrm{pen}(n,K^\star)}\right\}\left\{\frac{n^{1/2}\sup_{g\in\mathcal{G}_K^{\mathrm{a}}} |(\mathbb{P}_n - P^\star)g|}{(\log\log n)^{1/2}}\right\} \ge \frac{\mathrm{pen}(n,K^\star+1)}{\mathrm{pen}(n,K^\star)} - 1 \text{ io}\right] \text{ or} \tag{6}$$

$$\left[\left\{\frac{\log\log n}{\mathrm{pen}(n,K^\star)}\right\}\left\{\frac{n^{1/2}\sup_{g\in\mathcal{G}_K^{\mathrm{b}}} |(\mathbb{P}_n - P^\star)g|}{(\log\log n)^{1/2}}\right\}^2 \ge \frac{\mathrm{pen}(n,K^\star+1)}{\mathrm{pen}(n,K^\star)} - 1 \text{ io}\right]. \tag{7}$$

Here, (6) is obtained by application of (4) of Proposition 1 when **C5a** holds true for $K = K^\star + 1$, while (7) is a consequence of Huber's trick (5) of the same proposition when **C5b** holds true for $K = K^\star + 1$. Now, **PCa** (resp. **PCb**) and Lemma 2 yield the result.

• Let us consider now $\widehat{K}_n^{\mathrm{G}}$. The existence of the prior bound $K_{\max}$ allows to write

$$\left[\widehat{K}_n^{\mathrm{G}} > K^\star \text{ io}\right] = \bigcup_{K=K^\star+1}^{K_{\max}} \left[\widehat{K}_n^{\mathrm{G}} = K \text{ io}\right], \text{ with}$$

$$\left[\widehat{K}_n^{\mathrm{G}} = K \text{ io}\right] \subset \left[\sup_{\theta\in\Theta_K} \mathbb{P}_n\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n\,\ell_\theta \ge n^{-1}\{\mathrm{pen}(n,K) - \mathrm{pen}(n,K^\star)\} \text{ io}\right].$$

Thus, it suffices to prove that the above events have probability zero for all $K^\star+1 \le K \le K_{\max}$. Minor changes in the previous study of $\widehat{K}_n^{\mathrm{L}}$ yield the result.

3. Underestimation.

3.1. *Lower bound for the rate of underestimation* via *Stein's lemma.*

PROPOSITION 4. *Let $\widetilde{K}_n$ be any estimator of the order of the common distribution of $Z_1, \dots, Z_n$. One assumes that, for all $K_0 \ge 1$ and $P_0 \in \Pi_{K_0} \setminus \Pi_{K_0-1}$, $\widetilde{K}_n$ does not asymptotically almost surely overestimate K_0,* i.e.

$$\limsup_{n\to\infty} \mathrm{P}_{P_0}(\widetilde{K}_n > K_0) < 1, \tag{8}$$

whenever the Z_i's are P_0-distributed.

Let us denote again $P^\star$ the actual distribution of the Z_i's. We suppose that for all $\theta \in \Theta_{K^\star-1}$, $\ell_\theta, \ell_{\theta^\star} \in L^1(P_\theta)$. Then

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) \geq -\inf_{K<K^\star} H(\Pi_K \,|P^\star).$$

The proof of Proposition 4 involves Stein's Lemma 3 below in its most concise and therefore practical version due to Bahadur et al. (1980):

Lemma 3 (Bahadur et al.). *Let P, Q be two probability measures on $(\Omega, \mathcal{A})$ and a sequence $\{A_n\}$ of measurable sets. Then*

$$\liminf_{n\to\infty} Q(A_n) > 0 \quad \Longrightarrow \quad \liminf_{n\to\infty} n^{-1} \log P(A_n) \geq -H(Q\,|P). \tag{9}$$

Stein's lemma is quite famous in Statistics. It is as a versatile, powerful yet simple result due to Stein, though he never published it. Chernoff (1956) first mentions it in a framework of hypothesis testing. The corresponding version of the lemma is stated and proved *e.g.* in (Bahadur, 1967; Bahadur, 1971). It is intimately related to the notions of Bahadur efficiency and Hodges-Lehmann optimality, see *e.g.* (Kourouklis, 1991; Kallenberg & Kourouklis, 1992; van der Vaart, 1998). There, Stein's lemma provides a benchmark (*via* a minoration) and efficiency or optimality roughly consists, for a test, of achieving the lower bound. There is another example which illustrates both the versatility and the power of the discussed lemma: Gassiat & Boucheron (2001) invoke it in order to characterize the best underestimation exponent when estimating the order of a Hidden Markov Model.

Lemma 3 relies on the core of Stein's original proof (which is a change of probability argument). It is, in most cases, the key of its various versions.

Proof of Proposition 4. Let us temporarily denote P_{P_0} (resp. $\mathrm{P}_{P^\star}$) for P in order to emphasize that $Z_1, \ldots, Z_n$ are P_0-distributed (resp. $P^\star$-distributed). Set $K_0 < K^\star$, $\theta_0 \in \Theta_{K_O}$ and $\varepsilon > 0$. Choose $P = \mathrm{P}_{P^\star}$, $Q = \mathrm{P}_{P_{\theta_0}}$ and $A_n = [\widetilde{K}_n \leq K] \cap B_n$, with $B_n = \{|\mathbb{P}_n \ell_{\theta^\star} - \mathbb{P}_n \ell_{\theta_0} + H(\mathrm{P}_{\theta_0}\,|P^\star)| < \varepsilon\}$. Then the left hand condition of (9) is fulfilled because the Law of large numbers applies and thanks to (8). So, the right hand side of (9) holds true and this completes the proof, because $\mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) \geq \mathrm{P}_{P^\star}(\widetilde{K}_n \leq K_0)$ and K_0, θ_0 and ε are arbitrary.

3.2. *A few prerequisites.* We need to introduce some useful objects in order to state our main results on upper bounds for the underestimation rates in the next Section 3.3. We refer the reader to (Rao & Ren, 1991; Léonard & Najim, 2000) for comprehensive presentations of the following material.

The spaces $\mathcal{L}_\tau$ and $\mathcal{M}_\tau$. First, let $\tau(s) = \exp(|s|) - |s| - 1$ $(s \in \mathbb{R})$ be an even convex nonnegative function. Let us define

$$\mathcal{L}_\tau = \left\{ f \in \mathbb{R}^{\mathcal{Z}} : \exists\, a > 0,\ P^\star \tau(f/a) < \infty \right\}, \tag{10}$$

$$\mathcal{M}_\tau = \left\{ f \in \mathbb{R}^{\mathcal{Z}} : \forall\, a > 0,\ P^\star \tau(f/a) < \infty \right\} \subset \mathcal{L}_\tau. \tag{11}$$

The set $\mathcal{L}_\tau$ is a Banach space (called *Orlicz space*) when equipped with the norm

$$||f||_\tau = \inf\{a > 0 : P^\star \tau(f/a) \leq 1\} \ (\text{any } f \in L_\tau).$$

Since the convex-conjugate τ^* of τ given by $\tau^*(t) = (1+|t|)\log(1+|t|) - |t|$ ($t \in \mathbb{R}$) is also an even convex nonnegative function, one defines equivalently $\mathcal{L}_{\tau^*}$, $\mathcal{M}_{\tau^*}$ and $\|\cdot\|_{\tau^*}$.

A function f belongs to $\mathcal{L}_\tau$ (resp. $\mathcal{M}_\tau$) if and only if $P^\star \exp(a|f|) < \infty$ for *some* (resp. *all*) $a > 0$: thus, we say that the elements of $\mathcal{L}_\tau$ (resp. $\mathcal{M}_\tau$) admit *some* (resp. *all*) exponential moments. Particularly, $\mathcal{M}_\tau$ contains all bounded measurable functions.

Because $f \in \mathcal{L}_\tau$, $g \in \mathcal{L}_{\tau^*}$ yields $fg \in L^1(P^\star)$ and $P^\star|fg| \leq 2\,\|f\|_\tau \|g\|_{\tau^*}$, $\mathcal{L}_{\tau^*}$ is identified with a subspace of the topological dual $\mathcal{L}'_\tau$ of $(\mathcal{L}_\tau, \|\cdot\|_\tau)$. Thus, denoting $\mathcal{L}^*_\tau$ the algebraic dual of $\mathcal{L}_\tau$, the following inclusions will prevail in the sequel:

$$\mathcal{L}_{\tau^*} \subset \mathcal{L}'_\tau \subset \mathcal{L}^*_\tau. \tag{12}$$

Continuous and singular parts. By definition, a continuous linear form $Q \in \mathcal{L}'_\tau$ is $P^\star$*-singular* if there exists a sequence $\{A_p\}$ of measurable sets such that[2] $|Q|\mathbb{1}\{A_p^c\} = 0$ for all $p \geq 1$ while $\lim_{p\to\infty} P^\star(A_p) = 0$. We denote $\mathcal{L}^s_\tau$ the set of all $P^\star$-singular elements of $\mathcal{L}'_\tau$. Then, one can prove that $\mathcal{L}'_\tau$ is isomorphic to the direct sum $\mathcal{L}_{\tau^*} \oplus \mathcal{L}^s_\tau$.

Equivalently, any continuous linear form $Q \in \mathcal{L}'_\tau$ is uniquely decomposed into the sum $Q = Q^a + Q^s$, where Q^a (the *continuous part* of Q) and Q^s (its *singular part*) are both continuous, Q^s is $P^\star$-singular and

$$Q^a\, f = P^\star \frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}\, f \text{ (any } f \in \mathcal{L}_\tau).$$

Here, $\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}$ is the sole element of $\mathcal{L}_{\tau^*}$ such that the previous relation does hold. We finally state a result which aims at casting some light on the contributions of Q^a and Q^s to Q (its proof is based on the dominated convergence theorem):

LEMMA 4. *Let $Q \in \mathcal{L}'_\tau$ be decomposed into the sum $Q^a + Q^s$.*
Primo, $Q^s(\mathcal{M}_\tau) = \{0\}$. Secundo, *if $f \in \mathcal{L}_\tau$ and $\{f_p\}$ is a sequence of bounded functions that converges pointwise to f with* $\sup_p |f_p| = O(f)$*, then*

$$Q\, f_p \underset{p\to\infty}{\longrightarrow} Q^a\, f \quad \text{and} \quad Q\, \mathbb{1}\{|f| > p\} \underset{p\to\infty}{\longrightarrow} Q^s\, f.$$

Extended Sanov's theorem (Léonard & Najim, 2000). For any $f \in \mathcal{L}_\tau$, $Q \mapsto Q\,f$ defines an element of $\mathcal{L}^*_\tau$. We denote $\sigma(\mathcal{L}^*_\tau, \mathcal{L}_\tau)$ the coarsest topology on $\mathcal{L}^*_\tau$ that makes all those linear forms continuous (note that it is not metrizable), and $\varsigma(\mathcal{L}^*_\tau, \mathcal{L}_\tau)$ the smallest σ-field that makes all of them measurable. In view of a future proof, let us emphasize that the topology $\sigma(\mathcal{L}^*_\tau, \mathcal{L}_\tau)$ is generated by the collection of open sets

$$\mathcal{O}(f, x, \varepsilon) = \{Q \in \mathcal{L}^*_\tau : |Q\,f - x| < \varepsilon\} \text{ (any } f \in \mathcal{L}_\tau,\ x \in \mathbb{R},\ \varepsilon > 0).$$

Let us now define a crucial subspace of $\mathcal{L}^*_\tau$ for our purpose:

$$\mathcal{Q} = \{Q \in \mathcal{L}^*_\tau : Q \geq 0, Q\,1 = 1\}$$

[2] Since $\mathcal{L}_\tau$ is a Riesz space and all $P \in \mathcal{L}'_\tau$ are relatively bounded, then $\mathcal{L}'_\tau$ is also a Riesz space and forwardly, one can associate $|Q| = Q \vee O - Q \wedge 0 \in \mathcal{L}'_\tau$ with any $Q \in \mathcal{L}'_\tau$.

endowed with the topology $\mathcal{T}$ and the σ-field $\mathcal{S}$ induced by $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ and $\varsigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$.

We finally introduce the function $I(\cdot\,|P^\star)$ on $\mathcal{Q}$ such that

$$I(Q\,|P^\star) = Q^a \log \frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} + \sup\{Q^s f : f \in \mathcal{L}_\tau,\ P^\star \exp(f) < \infty\}$$

for any $Q \in \mathcal{Q} \cap \mathcal{L}'_\tau$, $I(Q\,|P^\star) = \infty$ otherwise. Here too, $I(\Pi\,|P^\star)$ denotes the infimum $\inf\{I(Q\,|P^\star) : Q \in \Pi\}$ (for $\Pi \subset \mathcal{Q}$). $I(\cdot\,|P^\star)$ is the rate function associated with the extended Sanov's theorem of Léonard & Najim (2000), which encompasses the following result:

THEOREM 4 (LÉONARD & NAJIM (2000)). *The function $I(\cdot\,|P^\star)$ is a convex, lower semicontinuous mapping from $\mathcal{Q}$ to $[0,\infty]$ whose level sets $\{Q \in \mathcal{Q} : I(Q\,|P^\star) \le \alpha\}$ are compact for all $\alpha > 0$.*
Moreover, for any $S \in \mathcal{S}$ ($\mathrm{cl}(S)$ is its closure wrt $\mathcal{T}$),

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\mathbb{P}_n \in S) \le -I(\mathrm{cl}(S)\,|P^\star).$$

Theorem 4 is the key of the derivation of upper bounds for the underestimation rates as stated in the next section.

3.3. *Upper bounds for the rate of underestimation.* Let us introduce some assumptions (**U** stands for Underestimation).

U1 The following three properties are satisfied:

(i) For any $P_\theta, P_t \in \Pi_{K^\star-1}$, $P_\theta << P_t$ and $H(P_\theta\,|P^\star) < \infty$.

(ii) $\{p_\theta\,\ell_\theta : \theta \in \Theta_{K^\star-1}\} \cup \{p_{\theta^\star}\,\ell_{\theta^\star}\} \subset L^1(\mu)$.

(iii) $\{\,\ell_\theta : \theta \in \Theta_{K^\star}\} \subset \mathcal{L}_\tau$.

U2 There exist $l, u \in \mathbb{R}^{\mathcal{Z}}$ such that $(u - l) \in \mathcal{L}_\tau$,

$$l \le \ell_\theta \le u \quad (\text{any } \theta \in \Theta_{K^\star})$$

and for all $\varepsilon > 0$ and $Q \in \mathcal{Q}$, there exists a compact set C of $\mathcal{Z}$ satisfying $Q\,(u-l)1\!\mathrm{l}\{C^c\} < \varepsilon$.

U3 For all $K \le K^\star$ and for any compact set C of $\mathcal{Z}$, the set $\{\,\ell_\theta 1\!\mathrm{l}\{C\} : \theta \in \Theta_K\}$ is precompact in the class $\mathrm{C}^0(C, \|\cdot\|_\infty)$ of all the continuous functions on C equipped with the uniform norm.

Let us define finally the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ (Λ stands for Local and Γ for Global):

$$\Lambda_{\alpha,K} = \{Q \in \mathcal{Q} : \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \ge -\alpha\} \quad (\text{any } \alpha \ge 0), \tag{13}$$

$$\Gamma_{\alpha,K} = \{Q \in \mathcal{Q} : \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta \ge -\alpha\} \quad (\text{any } \alpha \ge 0). \tag{14}$$

Observe that $\Gamma_{\alpha,K} \subset \Lambda_{\alpha,K}$ ($K < K^\star$), hence $-I(\Gamma_{\alpha,K}\,|P^\star) \le -I(\Lambda_{\alpha,K}\,|P^\star)$.

THEOREM 5. *Assume* **U1**, **U2**, **U3**.

- *If* **C2** *is also satisfied, then*

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) \le -\inf_{K<K^\star} I(\Lambda_{0,K}\,|P^\star) < 0. \tag{15}$$

- *Without any further assumption,*

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} < K^\star) \le -\inf_{K<K^\star} I(\Gamma_{0,K}\,|P^\star) < 0. \tag{16}$$

The main contribution of this theorem to the already existing collection of results that cope with upper bounds for the rate of underestimation of an order relies in its mild assumptions. We particularly prove in Section 5 that the assumptions of the theorem above are satisfied in **MGM** for $\widehat{K}_n^{\mathrm{L}}$ and in *both* **MGM** and **VR** for $\widehat{K}_n^{\mathrm{G}}$: this is, as far as we know, a novel result in the framework of penalized maximum likelihood procedure.

For instance, the earlier upper bounds for the rate of underestimation proved by Haughton & Keribin (2001) in a framework of penalized maximum likelihood procedure require more restrictive assumptions than ours (their proof relies on a Taylor expansion): twice continuously differentiability of the log–likelihood and nonsingularity of the Fisher information matrix (which *de facto* excluded the mixtures of distributions as **MGM**), as well as an extra condition on the uniqueness of the projection (wrt the Kullback-Leibler divergence) of the distribution of the observations on the submodels (*i.e.* with our notations, of $P^\star$ on Π_K for $K < K^\star$).

We nevertheless must emphasize that Gassiat & Boucheron (2001) have obtained upper bounds for the underestimation rates that coincide with the optimal lower bound yielded by Stein's lemma when estimating the order of a Hidden Markov Model. On the contrary, the bounds exhibited in Theorem 5 are not optimal with regard to Proposition 4. Now, the latter can be made more accurate under more restrictive assumptions, *cf.* Theorem 6.

Let us reinforce the structure of the spaces Θ_K: now, the distance d on Θ_K derives from a norm $\|\cdot\|$ on the vector space Θ_K, so that the notion of differentiability wrt to $\theta \in \Theta_K$ is available. Furthermore, we introduce:

U4 Let $l, u \in \mathbb{R}^{\mathcal{Z}}$ as defined above in **U2**. There exists a compact set C_{lu} of $\mathcal{Z}$ such that $1\!\!1\{C_{lu}^c\}(u-l) \in \mathcal{M}_\tau$.

U5 For any $K \le K^\star$ and $z \in \mathcal{Z}$, the functions $\theta \mapsto \ell_\theta(z)$ are differentiable on the interior of Θ_K, with derivative $\dot{\ell}_\theta(z)$. Moreover, the coordinates of $\dot{\ell}_\theta$ are elements of $\mathcal{M}_\tau$ and there exists $F \in \mathcal{L}_\tau$ such that

$$|\ell_{\theta+h} - \ell_\theta - \dot{\ell}_\theta^{\,T} h| \le F \cdot o(h).$$

It suffices that h in the latter inequality has the particular form $h = \|h\|\, e_k$ where e_k is the k-th canonical base vector of Θ_K.

Let us denote $\mathcal{Q}\cap M_1(\mathcal{Z})$ the set of the linear forms $Q \in \mathcal{Q}$ such that $Q(A) = Q\,1\!\!1\{A\}$ (any $A \in \mathcal{A}$) defines a probability and finally introduce the subsets

$$\Lambda_{0,K}^a = \Lambda_{0,K} \cap M_1(\mathcal{Z}), \tag{17}$$

$$\Gamma_{0,K}^a = \Gamma_{0,K} \cap M_1(\mathcal{Z}). \tag{18}$$

THEOREM 6. *Assume* **U1**, **U2**, **U3**, **U4**, **U5**.

- *If in addition* **C2** *is fulfilled, then*

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) \le -\inf_{K<K^\star} H(\Lambda^a_{0,K}\,|P^\star) < 0.$$

- *Without any further assumption,*

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} < K^\star) \le -\inf_{K<K^\star} H(\Gamma^a_{0,K}\,|P^\star) < 0.$$

We prove in Section 5 that the assumptions of the theorem above are satisfied in **MGM** for $\widehat{K}_n^{\mathrm{L}}$ and in both **MGM** and **VR** for $\widehat{K}_n^{\mathrm{G}}$. Assumption **U3** is verified *via* Ascoli's theorem. The lemma below allows to verify **U2** and **U4** (its proof is postponed to Appendix A.2):

LEMMA 5. *Let $\psi \in \mathbb{R}_+^{*\mathbb{R}}$ be an increasing positive function such that the ratio function $x \mapsto \psi(x)/x$ tends to infinity as $|x|$ also does. Assume that $\psi(u-l) \in \mathcal{L}_\tau$ and that the level sets of $(u-l)$ $\{z \in \mathcal{Z} : (u-l)(z) \le M\}$ are compact for all $M > 0$.*

Then, for all $\varepsilon > 0$ and $Q \in \mathcal{L}_\tau^*$, there exists a compact subset C of $\mathcal{Z}$ such that $Q\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$.

In case $\mathcal{Z} = \mathbb{R}^q$ and $(u-l)$ is continuous, the level sets of $(u-l)$ are compact. Furthermore, $(u-l) \in M_\tau$ and therefore, **U4** *holds true.*

The following lemma allows comparisons with Theorem 5:

LEMMA 6. *Under* **U1** (i), (ii), *for all $K < K^\star$, $\Pi_K \subset \Lambda^a_{0,K} \subset \Lambda_{0,K}$ and $\Pi_K \subset \Gamma^a_{0,K} \subset \Gamma_{0,K}$, hence*

$$-\inf_{K<K^\star} H(\Pi_K\,|P^\star) \le -\inf_{K<K^\star} H(\Lambda^a_{0,K}\,|P^\star) \le -\inf_{K<K^\star} I(\Lambda_{0,K}\,|P^\star),$$
$$-\inf_{K<K^\star} H(\Pi_K\,|P^\star) \le -\inf_{K<K^\star} H(\Gamma^a_{0,K}\,|P^\star) \le -\inf_{K<K^\star} I(\Gamma_{0,K}\,|P^\star).$$

In conclusion, we get under yet more restrictive assumptions that the upper bound turns out to be optimal for $\widehat{K}_n^{\mathrm{G}}$ with a view to Proposition 4 in exponential models. This is another novel result. It applies particularly to **VR**, as shown in Section 5.

PROPOSITION 5. *Under the assumptions of Theorem 6 and* for exponential models, *the best underestimation rate with a view to Proposition 4 is achieved by $\widehat{K}_n^{\mathrm{G}}$.*

In this case, the proof (postponed to Appendix A.3.) involves *H-projections* as defined and studied by Csiszár (1975): the H-projection $\overline{Q}$ of the probability measure Q on a convex set of probability measures $\mathcal{C}$ is defined (under conditions) by

$$\overline{Q} = \arg\min\Big\{H(P\,|Q) : P \in \mathcal{C}\Big\}.$$

It enjoys some remarkable properties, including a useful Pythagorean inequality (*cf.* Lemma 17). Nonetheless, the proof also involves much less tractable projections with reversed order of P and Q in the definition above. We can conclude in exponential models in the spirit of (Čencov, 1982), where the author derives another Pythagorean equality for reversed projections.

3.4. *Proofs.*

Proof of Theorem 5. We first state five useful lemmas.

LEMMA 7. *Consider $Q \in \mathcal{Q}$ such that $I(Q\,|P^\star) < \infty$. Then $Q^s \geq 0$ and $Q^a(A) = Q^a\, \mathbb{1}\{A\}$ (any $A \in \mathcal{A}$) defines a probability measure dominated by $P^\star$ with density $\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}$. Consequently, $I(Q^a\,|P^\star) = H(Q^a\,|P^\star)$.*

In greater generality, if $Q \in \mathcal{Q} \cap M_1(\mathcal{Z})$, then $I(Q\,|P^\star) = H(Q\,|P^\star)$.

The proof relies essentially on Lemma 4 and is omitted.

LEMMA 8. *Under* **U1** (i)*, the density $\frac{\mathrm{d}P_\theta}{\mathrm{d}P^\star}$ of $P_\theta \in \Pi_{K^\star-1}$ belongs to $\mathcal{M}_{\tau^\star}$. Therefore, $\Pi_{K^\star-1} \subset \mathcal{L}'_\tau \cap \mathcal{Q}$ in virtue of (12) and $I(\cdot\,|P^\star)$ coincides with $H(\cdot\,|P^\star)$ on $\Pi_{K^\star-1}$ in virtue of Lemma 7.*

LEMMA 9. *For $\alpha \leq \alpha'$, $\Lambda_{\alpha,K} \subset \Lambda_{\alpha',K}$ and $\Gamma_{\alpha,K} \subset \Gamma_{\alpha',K}$. Besides, under* **U1** (iii)*, the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ are $\mathcal{S}$-measurable.*

LEMMA 10. *Assume* **U1** (ii) *and set $K < K^\star$.*
- *If* **C2** *is also fulfilled, then $P^\star \notin \Lambda_{0,K}$.*
- *Without any additional assumption, $P^\star \notin \Gamma_{0,K}$.*

PROOF. If we had $P^\star \in \Lambda_{0,K}$, we would also have

$$0 \leq \sup_{\theta\in\Theta_K} P^\star \ell_\theta - \sup_{\theta\in\Theta_{K+1}} P^\star \ell_\theta = -H(P^\star\,|\Pi_K) + H(P^\star\,|\Pi_{K+1}) \leq 0,$$

in contradiction with **C2**. In the same spirit, whenever $P^\star \in \Gamma_{0,K}$,

$$0 \leq \sup_{\theta\in\Theta_K} P^\star \ell_\theta - \sup_{\theta\in\Theta_{K^\star}} P^\star \ell_\theta \leq -H(P^\star\,|\Pi_K).$$

But the right hand term above is negative because $H(P^\star\,|\cdot)$ achieves its infimum on compact sets (for the weak topology on $M_1(\mathcal{Z})$), *e.g.* on Π_K, while $P^\star \notin \Pi_K$.

LEMMA 11. *Assume* **U1** (iii)*,* **U2** *and* **U3**. *Then the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ are closed wrt the topology $\mathcal{T}$ for every $\alpha > 0$, $K < K^\star$.*

This lemma (whose proof is postponed to Appendix A.4) is of central importance in our scheme of proof. Further comments follow the proof of Theorem 5.

PROOF OF THEOREM 5. Let us focus on $\widehat{K}_n^{\mathrm{L}}$: the proof for $\widehat{K}_n^{\mathrm{G}}$ is almost identical, so we omit it. Since $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) = \sum_{K<K^\star} \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K)$, Lemma 1.2.15 of

(1998) ensures that

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) = \sup_{K<K^\star}\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K).$$

Thus, it suffices to choose $K < K^\star$ and show that

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K) \le -I(\Lambda_{0,K}\,|P^\star).$$

Now, for any $\alpha > 0$,

$$\limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K) \le \limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\mathbb{P}_n \in \Lambda_{\alpha,K}) \le -I(\Lambda_{\alpha,K}\,|P^\star)$$

in virtue of Theorem 4 and Lemma 11.

Furthermore, Lemma 9 ensures that $\{\Lambda_{\alpha,K}\}$ nonincreases as $\alpha\downarrow 0$, so that $\{I(\Lambda_{\alpha,K}\,|P^\star)\}$ nondecreases as $\alpha\downarrow 0$, and it is bounded above by $I(\Lambda^a_{0,K}\,|P^\star) = H(\Lambda^a_{0,K}\,|P^\star)$ thanks to Lemma 7. Observe that Lemmas 6 and 8 finally ensure that the former upper bound is bounded itself by $H(\Pi_{K^\star-1}\,|P^\star)$, which is finite. Therefore, $\{I(\Lambda_{\alpha,K}\,|P^\star)\}$ converges as $\alpha\downarrow 0$ to some limit denoted L, $L \le I(\Lambda_{0,K}\,|P^\star)$.

Let $\{\alpha_p\}$ be a decreasing sequence of positive numbers such that the sequence $\{I(\Lambda_{\alpha_p,K}\,|P^\star)\}$ increases to L. Since $I(\cdot\,|P^\star)$ is a good rate function and $\Lambda_{\alpha_p,K}$ is closed (see Lemma 11 and recall that a good lower semicontinuous function achieves its infimum on closed sets), there exists $Q_p\in\Lambda_{\alpha_p,K}$ such that $I(Q_p\,|P^\star) = I(\Lambda_{\alpha_p,K}\,|P^\star)$.

Let $\mathcal{K}$ be the compact subset of $\mathcal{Q}$ consisting of the Q's with $I(Q\,|P^\star) \le H(\Pi_{K^\star-1}\,|P^\star)$. For any $q\ge 1$, the set $\mathrm{cl}(\{Q_p : p\ge q\})$ is a closed subset of $\Lambda_{\alpha_q,K}\cap\mathcal{K}$ which is compact as a closed subset of a compact set: thus $\{\mathrm{cl}(\{Q_p : p\ge q\})\}_q$ is a nonincreasing sequence of non void compact sets. The Borel-Lebesgue property therefore ensures that their intersection is non void either: let $\overline{Q}$ be in the intersection.

Since $\overline{Q}\in\Lambda_{\alpha_p,K}$ for any $p\ge 1$, $\overline{Q}\in\Lambda_{0,K}$. Moreover, $I(Q_p\,|P^\star)\le L$ (any $p\ge 1$) yields that $\{Q_p : p\ge q\}$ is included in the closed set of the Q's of $\mathcal{Q}$ with $I(Q\,|P^\star)\le L$, and so does its closure. Forwardly, $\overline{Q}$ also satisfies $I(\overline{Q}\,|P^\star)\le L$. Then $L\le I(\Lambda_{0,K}\,|P^\star)\le I(\overline{Q}\,|P^\star)$ concludes the main part of this proof.

It still remains to show that $I(\overline{Q}\,|P^\star) > 0$. Suppose on the contrary that equality holds true. Then

$$\sup\{\overline{Q}^s f : f\in\mathcal{L}_\tau,\, P^\star\exp(f)<\infty\} = 0. \tag{19}$$

Choose $f\in\mathcal{L}_\tau$: there exists $a>0$ such that $P^\star\exp(a|f|)<\infty$. Then $\overline{Q}^s a|f|\ge 0$ (recall that $\overline{Q}^s\ge 0$ thanks to Lemma 7) and equality does hold thanks to (19). Consequently, $\overline{Q}^s|f| = \overline{Q}^s f\vee 0 - \overline{Q}^s f\wedge 0 = 0$ and $\overline{Q}^s\ge 0$ implies that $\overline{Q}^s f\vee 0 = \overline{Q}^s f\wedge 0 = 0$, hence $\overline{Q}^s f = 0$ as their sum.

Thus, $\overline{Q} = \overline{Q}^a$. Following Lemma 7, we can think of $\overline{Q}$ as a probability measure satisfying $H(\overline{Q}\,|P^\star) = I(\overline{Q}\,|P^\star) = 0$, *i.e.* $\overline{Q} = P^\star$. This contradicts Lemma 10 and concludes the proof for the estimator $\widehat{K}_n^{\mathrm{L}}$.

We can now emphasize why the sophisticated setting of the extended Sanov Theorem 4 is not superfluous.

- *Primo*, a classical Sanov's theorem on $M_1(\mathcal{Z})$ is insufficient for our purpose. Indeed, the proof requires that some linear forms $Q \mapsto Q\,f$ on $\mathcal{Q}$ for some functions $f \in \mathcal{L}_\tau$ (which may belong to $\mathcal{L}_\tau \setminus \mathcal{M}_\tau$) are continuous. Now, Schied (1998) has shown that the extension of a Sanov's theorem on $M_1(\mathcal{Z})$ to a topology on $M_1(\mathcal{Z})$ that renders the above mapping continuous is possible if and only if $f \in \mathcal{M}_\tau$ (this is the classical Cramér's condition).
- *Secundo*, provided that we need the continuity of $Q \mapsto Q\,f$ for some $f \in \mathcal{L}_\tau \setminus \mathcal{M}_\tau$, the $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$-topology on $\mathcal{L}_\tau^*$ is natural.
- *Tertio*, another consequence of (Schied, 1998) is that the simpler rate function $I'(Q\,|P^\star) = I(Q^a\,|P^\star)$ for $Q \in \mathcal{Q} \cap \mathcal{L}'_\tau$, $I'(Q\,|P^\star) = \infty$ otherwise, does not have compact level sets (I' is not good). Now, our scheme of proof requires that the rate function is good, so Theorem 4 really is a convenient tool.

Proof of Theorem 6. The two lemmas below are readily proved.

Lemma 12. *Assume* **U1** (iii), **U2**, **U3**, **U4**. *If $Q \in \mathcal{Q} \cap L'_\tau$, then the function $\theta \mapsto Q\,\ell_\theta$ mapping $\Theta_{K^\star}$ to $\mathbb{R}$ is continuous over $\Theta_{K^\star}$.*

Lemma 13. *Assume* **U1** (iii) *and* **U5**. *Set $Q \in \mathcal{Q}$ and $K \leq K^\star$. Then the function $\theta \mapsto Q\,\ell_\theta$ mapping Θ_K to $\mathbb{R}$ is differentiable on the interior of Θ_K, with derivative $\theta \mapsto Q\,\dot{\ell}_\theta$.*

For Lemma 12, since $Q \in \mathcal{L}'_\tau$, we prove that $\|\ell_{\theta_q} - \ell_{\theta_0}\|_\tau \to 0$ when $\theta_q \to \theta_0$ ($\theta_q, \theta_0 \in \Theta_{K^\star}$). For Lemma 13, we simply invoke the positivity of any $Q \in \mathcal{Q}$.

Proof of Theorem 6. Let us start where we stopped in the proof of Theorem 5. Since $\overline{Q} \in L'_\tau$, so does $\overline{Q}^s$. Apply Lemma 13: $\theta \mapsto \overline{Q}^s\,\ell_\theta$ is differentiable on the interiors of Θ_K and Θ_{K+1}, with derivative $\overline{Q}^s\,\dot{\ell}_\theta$, which is zero since $\dot{\ell}_\theta \in \mathcal{M}_\tau$ in virtue of **U5**. So, $\overline{Q}^s\,\ell_\theta$ is constant over Θ_K and Θ_{K+1} with possibly distinct values on each of them. Now, Lemma 12 yields that $\overline{Q}^s\,\ell_\theta = \overline{Q}^s\,\ell_t$ for any $\theta \in \Theta_K$, $t \in \Theta_{K+1}$.

Finally, $\overline{Q}^a \in \Lambda^a_{0,K}$, Lemma 7 and the obvious inequality $I(Q\,|P^\star) \geq I(Q^a\,|P^\star)$ (any $Q \in \mathcal{Q} \cap L'_\tau$) give

$$H(\Lambda^a_{0,K}\,|P^\star) \leq H(\overline{Q}^a\,|P^\star) = I(\overline{Q}^a\,|P^\star) \leq I(\overline{Q}\,|P^\star) = I(\Lambda_{0,K}\,|P^\star) \leq H(\Lambda^a_{0,K}\,|P^\star)$$

and this completes the proof for the estimator $\widehat{K}_n^{\mathrm{L}}$. Once again, the proof when studying $\widehat{K}_n^{\mathrm{G}}$ is very similar, so we again omit it.

4. Overestimation.

4.1. *Lower bound for the rate of overestimation* via *Stein' lemma.*

Proposition 6. *Let $\widetilde{K}_n$ be any estimator of the order of the common distribution of $Z_1, \ldots, Z_n$. One assumes that, for all $K_0 \geq 2$ and $P_0 \in \Pi_{K_0} \setminus \Pi_{K_0-1}$, $\widetilde{K}_n$*

does not asymptotically almost surely underestimate K_0, i.e.

$$\limsup_{n\to\infty} \mathrm{P}_{P_0}(\widetilde{K}_n < K_0) < 1,$$

whenever the Z_i*'s are* P_0*-distributed.*

Let us denote again $P^\star$ *the actual distribution of the* Z_i*'s. We suppose that for all* $\theta \in \Theta_{K^\star+1}$, $\ell_\theta, \ell_{\theta^\star} \in L^1(P_\theta)$. *Then*

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n > K^\star) = \limsup_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n > K^\star) = 0.$$

The rate of overestimation is slower than exponential in n, according to Proposition 6 whose proof parallels the proof of Proposition 4. We get upper bounds for this rate in the next Section: the scheme of proof is alike the previous in the case of underestimation, with the substitution of Wu's MDP (1994) to the Sanov's LDP Theorem 4.

4.2. *Upper bounds for the rate of overestimation.* Let us introduce some assumptions (**O** stands for Overestimation).

Oa Let $K > K^\star$ be an integer to be chosen later. The class $\mathcal{G}_K^{\mathrm{a}}$ defined in **C5a** is $P^\star$-Donsker. Besides, there exist $l, u \in \mathbb{R}^{\mathcal{Z}}$ such that $G = (u - l) \in \mathcal{L}_\tau$, with

$$l \le \ell_\theta \le u \quad (\text{any } \theta \in \Theta_{K^\star+1}).$$

Hence, $\mathcal{G}_K^{\mathrm{a}}$ admits the envelope function $G = (u - l) \in \mathcal{L}_\tau$.

POa The penalty function is of the form $\mathrm{pen}(n, K) = v_n D(K)$ for $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ also increasing such that $n^{1/2} v_n^{-1} = o(1)$, $v_n = o(n)$ and finally,

$$v_{nk} \le A k^{1-\delta} v_n$$

for some $A \ge 1$, $0 < \delta < 1$, any $n, k \ge 1$.

Assumption **POa** is satisfied *e.g.* for $v_n = n^{1-\delta}$ $(0 < \delta < 1/2)$. Now, one can state the following result:

THEOREM 7. *Assume* **POa**.

- *If* **Oa** *holds true for* $K = K^\star + 1$, *then*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) < 0;$$

- *If we know a prior bound* $K_{\max}$ *on* $K^\star$ *and if* **Oa** *holds true for* $K = K_{\max}$, *then*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) < 0.$$

The main contribution of this theorem relies again in its mild assumptions. We particularly prove in Section 5 that it applies to *both* **MGM** and **VR**.

In (Dacunha-Castelle & Gassiat, 1997), the under- and overestimation rates are handled together, hence an exponential rate in $n^{-1}v_n^2$. We emphasize that this rate is precisely the lowest of the underestimation and overestimation rates exhibited by us in Theorems 6 and 7. The latter paper does not require any prior bound on $K^\star$, contrarily to (Haughton & Keribin, 2001).

The proof of (Haughton & Keribin, 2001) relies on a Taylor expansion and then on MDP and LDP (respectively for the leading and the remainder terms), hence some assumptions of three times continuously differentiability of the log-likelihood, existence of some exponential moments for the gradient and for the supremum of the Hessian. Mixture models are excluded.

Once again, Huber's trick enhances the previous result wrt the conditions on the penalty function, up to strengthened assumptions on the likelihoods.

Ob Let $K > K^\star$ be an integer to be chosen later. The class $\mathcal{G}_K^{\mathrm{b}}$ defined in **C5b** is $P^\star$-Donsker. Besides, $\mathcal{G}_K^{\mathrm{b}}$ admits an envelope function $G \in \mathcal{L}_\tau$.

POb The penalty function is of the form $\mathrm{pen}(n, K) = v_n D(K)$ for $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ increasing to infinity such that both $(\log n)\, v_n{}^{-1} = o(1)$ and $v_n = o(n)$ with moreover

$$v_{nk} \leq A k^{1-\delta} v_n$$

for some $A \geq 1$, $0 < \delta < 1$, any $n, k \geq 1$.

THEOREM 8. *Assume* **POb**.

- *If* **Ob** *holds true for* $K = K^\star + 1$, *then*

$$\limsup_{n\to\infty} v_n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) < 0.$$

- *If we know a prior bound* $K_{\max}$ *on* $K^\star$ *and if* **Ob** *holds true for* $K = K_{\max}$, *then*

$$\limsup_{n\to\infty} v_n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) < 0.$$

However, we must emphasize that the existence of an envelope $G \in \mathcal{L}_\tau$ for $\mathcal{G}_K^{\mathrm{b}}$ is constraining.

4.3. *Proof of Theorems 7, 8.* The proof is quite alike the proofs of consistency (*cf.* Section 2.2). Let us denote $D^\star = D(K^\star+1) - D(K^\star) > 0$. The event $[\widehat{K}_n^{\mathrm{L}} > K^\star]$ is included in

$$\left[\sup_{\theta\in\Theta_{K^\star+1}} \mathbb{P}_n\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n\,\ell_\theta \geq n^{-1} v_n\, D^\star\right]$$

which in turn is included in (we set $K = K^\star + 1$)

$$\left[\sup_{g\in\mathcal{G}_K^{\mathrm{a}}} |(\mathbb{P}_n - P^\star) g| \geq n^{-1} v_n\, D^\star\right] \text{ or } \left[\left(\sup_{g\in\mathcal{G}_K^{\mathrm{b}}} |(\mathbb{P}_n - P^\star) g|\right)^2 \geq n^{-1} v_n\, D^\star\right].$$

The latter are obtained by application of either (4) or (5) of Proposition 1.

Set $\mathcal{G} = \mathcal{G}_K^{\mathrm{a}}$ or $\mathcal{G} = \mathcal{G}_K^{\mathrm{b}}$. Now, any $Q \in M(\mathcal{Z})$ (the space of all signed measures on $\mathcal{Z}$) such that $G \in L^1(Q)$ defines an element Q^∞ in $\ell^\infty(\mathcal{G})$ (the space of all uniformly bounded real functions on $\mathcal{G}$ equipped with the topology and the σ-field induced by the uniform norm $\|\cdot\|_\infty$), along $Q^\infty g = Q\,g$ (any $g \in \mathcal{G}$). Particularly, $(n\,v_n^{-1})(\mathbb{P}_n - P^\star)$ is associated with $(n\,v_n^{-1})(\mathbb{P}_n - P^\star)^\infty$.

Let us define the nonnegative map $J^\infty(\cdot\,|P^\star)$ on the space $\ell^\infty(\mathcal{G})$ by

$$J^\infty(B\,|P^\star) = \inf\left\{ P^\star \left(\frac{\mathrm{d}Q}{\mathrm{d}P^\star}\right)^2 /2 : Q \in M(\mathcal{Z}), Q << P^\star, Q\,1 = 0, Q^\infty = B \right\}$$

(with the usual convention $\inf \varnothing = \infty$).

In the sequel, we focus on the alternative **a**: the other one is similar. Thus, let us define $\Lambda^\infty = \left\{ B \in \ell^\infty(\mathcal{G}_K^{\mathrm{a}}) : \|B\|_{\mathcal{G}_K^{\mathrm{a}}} \geq D^\star \right\}$. Since the assumptions of Theorem 5 of (Wu, 1994) are fulfilled, one has

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) \leq$$
$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}\left((n\,v_n^{-1})\,(\mathbb{P}_n - P^\star)^\infty \in \Lambda^\infty \right) \leq -\inf\{J^\infty(B\,|P^\star) : B \in \Lambda^\infty\}.$$

So, we have to prove that the right hand term is negative.

We first emphasize that Λ^∞ is closed for the uniform topology on $\ell^\infty(\mathcal{G}_K^{\mathrm{a}})$ equipped with $\|\cdot\|_{\mathcal{G}_K^{\mathrm{a}}}$ Let us now suppose on the contrary that the infimum is zero. Then, there exists a sequence $\{B_p\}$ of elements of $\ell^\infty(\mathcal{G}_K^{\mathrm{a}})$ such that $B_p \in \Lambda^\infty$ and $J^\infty(B_p\,|P^\star) \leq p^{-1}$. Consequently, there exists a sequence $\{Q_p\}$ of elements of $M(\mathcal{Z})$ such that both $P^\star(\frac{\mathrm{d}Q_p}{\mathrm{d}P^\star})^2/2 \leq J^\infty(B_p\,|P^\star) + p^{-1} \leq 2\,p^{-1}$ and $Q_p^\infty = B_p$ for any $p \geq 1$. In particular, $Q_p << P^\star$ thus, for any $g \in \mathcal{G}$, one has

$$(B_p^\infty\,g)^2 = \left(P^\star \frac{\mathrm{d}Q_p}{\mathrm{d}P^\star}\,g\right)^2 \leq (P^\star g^2)\,P^\star\left(\frac{\mathrm{d}Q_p}{\mathrm{d}P^\star}\right)^2 \leq 2\left(\sup_{g\in\mathcal{G}_K^{\mathrm{a}}} P^\star g^2\right) J(Q_p\,|P^\star)$$

thanks to Cauchy-Schwarz's inequality. Now, **Oa** ensures that $\mathcal{G}_K^{\mathrm{a}}$ is (totally) bounded in $L^2(P^\star)$, hence $\|B_p^\infty\|_{\mathcal{G}_K^{\mathrm{a}}} = o(1)$ and forwardly, $0 \in \Lambda^\infty$ as a limit of a sequence of elements of the closed set Λ^∞. Since this is false, the infimum must be positive and the proof for $\widehat{K}_n^{\mathrm{L}}$ is complete.

• As for $\widehat{K}_n^{\mathrm{G}}$, we can write

$$\mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) = \sum_{K=K^\star+1}^{K_{\max}} \mathrm{P}(\widehat{K}_n^{\mathrm{G}} = K),$$

so that Lemma 1.2.15 of (Dembo & Zeitouni, 1998) yields the equality below (we denote $D_K^\star = D(K) - D(K^\star)$)

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) = \sup_{K^\star<K\leq K_{\max}} \limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} = K) \leq$$
$$\sup_{K^\star<K\leq K_{\max}} \limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}\left(\sup_{\theta\in\Theta_K} \mathbb{P}_n\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n\,\ell_\theta \geq n^{-1} v_n\,D_K^\star \right),$$

while the inequality stems from (4) of Proposition 1 (we would apply (5) for the alternative **b**). The remainder of the proof goes along the same lines than above, so we omit it.

5. Benchmark examples. We have already introduced our four benchmark examples **MGM**, **RKO**, **AC** and **VR** in Section 1.4. Let us now investigate them, omitting the simplest or more technical proofs.

5.1. *Mixtures of Gaussians.*
Further description of the models. In **MGM** (resp. **RKO**), Π_1 is the class of all real Gaussian densities γ_u with mean $u \in \mathcal{U} = \Theta_1$ compact set of $\mathbb{R}$ and variance σ^2 (resp. the class of all real Gaussian densities with mean m and variance σ^2, $u = (m, \sigma^2) \in \mathcal{U} = \Theta_1$ compact set of $\mathbb{R} \times \mathbb{R}_+^*$). We also define for all $K \geq 2$

$$\Pi_K = \left\{ p_\theta \,\mathrm{d}\mu = \sum_{k=1}^{K-1} \pi_k \gamma_{u_k} + \left(1 - \sum_{k=1}^{K-1} \pi_k\right) \gamma_{u_K} \,\mathrm{d}\mu : \theta = (\boldsymbol{\pi}, \boldsymbol{u}) \in \Theta_K \right\},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})$ is a vector of nonnegative numbers satisfying $\sum_k \pi_k \leq 1$ and $\boldsymbol{u}$ is a vector of parameters with coordinates in $\mathcal{U}$. We emphasize that **A1** is fulfilled in virtue of Lévy's continuity theorem.
Exploring the various assumptions.

Proposition 7. *Let F be a mixing distribution on $\mathcal{U}$ (of support possibly infinite) and $P^\star$ have density $p^\star(z) = \int_{\mathcal{U}} \gamma_u(z) \,\mathrm{d}F(u)$ (any $z \in \mathcal{Z}$). Then* **C1** *and* **C2** *hold true in* **MGM** *and* **RKO**.

Proof. First, observe that in both cases, $\ell^\star \in L^1(P^\star)$ and **C1** holds true. Concerning **C2**, suppose on the contrary that

$$H(P^\star | \Pi_K) \leq H(P^\star | \Pi_{K+1}), \tag{20}$$

i.e. that we have an equality. Lower semicontinuity of $H(P^\star|\cdot)$ and compactness of Π_K ensure the existence of $P_0 \in \Pi_K$ such that $H(P^\star|\Pi_K) = H(P^\star|P_0)$. Then, (20) yields that for any $Q \in \Pi_{K+1}$,

$$H(P^\star|Q) \geq H(P^\star|P_0). \tag{21}$$

Now let γ denote any γ_u and let us define $Q_h = (1-h)P_0 + h\,\gamma\mu \in \Pi_{K+1}$ (any $h \in [0,1]$). Then $H(P^\star|Q_h) \leq (1-h)H(P^\star|P_0) + h\,H(P^\star|\gamma\mu)$ by convexity, hence $H(P^\star|Q_h)$ is finite. Furthermore, $\ell^\star \in L^1(P^\star)$, so (21) yields

$$h^{-1} P^\star \log \frac{(1-h)p_0 + h\,\gamma}{p_0} \leq 0 \text{ (any } h \in [0,1]),$$

where p_0 is the density of P_0 wrt μ. Applying Fatou's lemma, we derive from the previous inequality that $P^\star(\gamma - p_0)/p_0 \leq 0$. Now, Fubini's theorem ensures that $\int_{\mathcal{U}} P^\star(\gamma_u - p_0)/p_0 \,\mathrm{d}F(u) = P^\star(p^\star - p_0)/p_0 \leq 0$. Finally, $\log x \leq x - 1$ (any $x > 0$) implies

$$0 \leq H(P^\star|P_0) = P^\star \log(p^\star/p_0) \leq P^\star(p^\star/p_0) - 1 \leq 0,$$

and forwardly $P^\star = P_0 \in \Pi_K$. This completes the proof.

Furthermore, it is readily seen that **C3** and **C4** hold true in **MGM** and **RKO**. Besides in **MGM**, the natural envelope $(u-l)$ obtained for $u = \sup \ell_\theta$ and $l = \inf \ell_\theta$ verifies $(u-l)^{1+\delta} \in \mathcal{M}_\tau$ for some $\delta \in [0, 1[$. Consequently, **U4** is fulfilled, and so does **U2** (thanks to Lemma 5) in **MGM**.

Now, back to the consistency assumptions, one can prove that **C5a** holds true for all $K > K^\star$ in **MGM**. Indeed, the class $\mathcal{G}_K^{\mathrm{a}}$ has finite bracketing entropy integral in virtue of Example 19.7 of (van der Vaart, 1998) – *cf.* the referred book for definitions. We must insist on the fact that we can conclude because the function $(\sup \gamma_u)(\inf \gamma_u)^{-1}$ belongs to $L^2(P^\star)$. This is not true anymore in **RKO**. Finally, one can prove tediously by hand that the class $\mathcal{G}_K^{\mathrm{b}}$ has ε-bracketing numbers bounded by a polynomial in ε^{-1} in **MGM**, hence finite bracketing entropy integral and **C5b**.

Concerning the underestimation assumptions in **MGM**, **U1** is straightforward, **U2** and **U4** are already proved, **U3** derives from Ascoli's theorem. A crude yet careful application of Taylor's integral remainder theorem yields **U5**.

In conclusion, **Oa** in **MGM** is a simple by-product of the properties above. On the contrary, we can not prove that **Ob** holds true in **MGM**.

5.2. *Abrupt changes and various regressions.*

Further description of the model in **AC**. Let $\mathcal{T}$ be the set of all finite Caccioppoli partitions of $\mathcal{X}$ whose perimeters are uniformly bounded by a constant fixed throughout the sequel. We refer the reader to (Leonardi & Tamanini, 2002) for definitions and properties of the Caccioppoli partitions. $\mathcal{T}_K$ is the subset of $\mathcal{T}$ of all the partitions of cardinality K, and $\mathcal{T}_{\leq K} = \cup_{k \leq K} \mathcal{T}_k$. Thanks to (Leonardi & Tamanini, 2002), we know that there exists a metric d_P on $\mathcal{T}$ that makes $(\mathcal{T}, d_P)$ a compact metric set and $\mathcal{T}_{\leq K}$ a closed subset of $\mathcal{T}$ equipped with the topology induced by d_P.

Let us recall that $\mathcal{M} \subset \mathbb{R}$ is compact. We define, for any $K \geq 1$, any $\tau \in \mathcal{T}_K$, any $\boldsymbol{m} = (m_1, \ldots, m_K) \in \mathcal{M}^K$: $\theta = (\tau, \boldsymbol{m})$, $f_\theta(x) = \sum_{k=1}^K m_k 1\!\!1\{x \in \tau_k\}$ and $p_\theta(z) = \gamma(y\,; f_\theta(x))\, p(x)$ (for any $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). Here, $\gamma(\cdot\,; m)$ is a Gaussian density wrt μ with mean m and known variance σ^2.

Then, we set for any $K \geq 1$,

$$\begin{aligned}
\Pi_1 &= \left\{p_\theta \,\mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_1\right\} &&\text{with} && \Theta_1 = \{\mathcal{X}\} \times \mathcal{M} \\
\Pi_{K+1} &= \Pi_K \cup \left\{p_\theta \,\mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_{K+1}\right\} &&\text{with} && \Theta_{K+1} = \Theta_K \cup \left(\mathcal{T}_{K+1} \times \mathcal{M}^{K+1}\right).
\end{aligned}$$

We emphasize that **A1** is fulfilled in virtue of Lévy's continuity theorem.

Further description of the model in **VR**. Let us recall that $\mathcal{U}$ is a compact subset of $\mathbb{R}$ that contains 0. We define $\Theta_K = \mathcal{U}^K$ and for any $\theta \in \Theta_K$ and $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$: $f_\theta(x) = \sum_{k=1}^K \theta_k t_k(x)$ and $p_\theta(z) = \gamma(y\,; f_\theta(x))$ (with the same notation for $\gamma(\cdot; m)$ than above). Once again, all $\Pi_K = \{p_\theta \,\mathrm{d}\mu : \theta \in \Theta_K\}$ are compact and **A1** holds true thanks to Lévy's continuity theorem.

Exploring the various assumptions. We denote $\|\cdot\|_2$ the $L^2(P)$-norm. The simple lemma below particularly yields that **C1** holds true in **AC** and **VR**:

LEMMA 14. *In* **AC** *and* **VR***:* $\forall\, P_\theta, P_t \in \Pi_\infty,\ 2\,\sigma^2\, H(P_\theta\,|P_t) = \|f_\theta - f_t\|_2^2.$

PROPOSITION 8. *The* **VR** *example satisfies* **C1**. *In the* **AC** *example, both* **C1** *and* **C2** *hold true.*

PROOF. Suppose on the contrary that

$$H(P^\star\,|\Pi_K) \le H(P^\star\,|\Pi_{K+1}), \tag{22}$$

i.e. that we have an equality. Lower semicontinuity of $H(P^\star\,|\cdot)$ and compactness of Π_K ensure the existence of $P_0 \in \Pi_K$ such that $H(P^\star\,|P_0) = H(P^\star\,|\Pi_K)$. Denote $f_0 = \sum_{k=1}^K m_k \mathbb{1}\{x \in \tau_k\}$ (any $x \in \mathcal{X}$) the mean function of P_0, where $\mathbf{m} \in \mathcal{M}^K$ and $\tau \in \mathcal{T}_{\le K}$. Now, Lemma 14 and the simple equality

$$2\sigma^2 H(P^\star\,|P_0) = \|f^\star - f_0\|^2 = \sum_{k=1}^K P(f^\star - m_k)^2 \mathbb{1}\{\tau_k\},$$

yield that $m_k = Pf^\star \mathbb{1}\{\tau_k\}/P(\tau_k)$ for all $1 \le k \le K$.

We prove below that $f^\star = f_0$, hence $P^\star \in \Pi_K$.

Indeed, (22) ensures that, for any $1 \le k_0 \le K$, for any subset S of τ_{k_0} with positive P-measure,

$$Pf^{\star 2} - \sum_{k=1}^K \frac{(Pf^\star \mathbb{1}\{\tau_k\})^2}{P(\tau_k)} \le Pf^{\star 2} - \sum_{1 \le k \ne k_0 \le K} \frac{(Pf^\star \mathbb{1}\{\tau_k\})^2}{P(\tau_k)} \\ - \left(\frac{(Pf^\star \mathbb{1}\{S\})^2}{P(S)} + \frac{(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\})^2}{P(\tau_{k_0} \setminus S)}\right)$$

or equivalently,

$$\frac{(Pf^\star \mathbb{1}\{S\})^2}{P(S)} + \frac{(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\})^2}{P(\tau_{k_0} \setminus S)} \le \frac{(Pf^\star \mathbb{1}\{\tau_{k_0}\})^2}{P(\tau_{k_0})}.$$

Thus, we get, expanding first the right hand term and then factorizing

$$\frac{P(\tau_{k_0} \setminus S)}{P(S)} \left(Pf^\star \mathbb{1}\{S\}\right)^2 + \frac{P(S)}{P(\tau_{k_0} \setminus S)} \left(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}\right)^2 \le \\ 2\left(Pf^\star \mathbb{1}\{S\}\right)\left(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}\right). \tag{23}$$

Now, the basic inequality $2\,ab \le (au)^2 + (bu^{-1})^2$ (any $a, b \in \mathbb{R}$ and positive u) together with (23) ensure (take $u^2 = P(\tau_{k_0} \setminus S)/P(S)$) that the equality holds in (23). Consequently, for any subset S of τ_{k_0} with positive P-measure:

$$\frac{Pf^\star \mathbb{1}\{S\}}{P(S)} = \frac{Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}}{P(\tau_{k_0} \setminus S)} = \frac{Pf^\star \mathbb{1}\{\tau_{k_0}\} - Pf^\star \mathbb{1}\{S\}}{P(\tau_{k_0 \setminus S})},$$

hence, for any subset S of τ_{k_0}:

$$Pf^\star 1\!\!1\{S\} = \frac{P(S)}{P(\tau_{k_0})} Pf^\star 1\!\!1\{\tau_{k_0}\}.$$

Let us choose $S = S_+ = \{x \in \tau_{k_0} : f^\star(x) > Pf^\star 1\!\!1\{\tau_{k_0}\}P(\tau_{k_0})^{-1}\}$: we obtain $P(S_+) = 0$. Let us choose now $S = S_- = \{x \in \tau_{k_0} : f^\star(x) < Pf^\star 1\!\!1\{\tau_{k_0}\}P(\tau_{k_0})^{-1}\}$: we obtain $P(S_-) = 0$, hence finally $P(S_0) = P(\tau_{k_0})$ where $S_0 = \{x \in \tau_{k_0} : f^\star(x) = Pf^\star 1\!\!1\{\tau_{k_0}\}P(\tau_{k_0})^{-1}\}$ (*i.e.* $f^\star$ P-almost surely constant on τ_{k_0}). We conclude because k_0 is arbitrary.

Furthermore, it is readily seen that **C3** and **C4** hold true in **AC** and **VR**. Besides, the natural envelope $(u - l)(z) = (2\sigma^2)^{-1}[(\sup f_\theta(x))^2 + (\inf f_\theta(x))^2] + 2|y|\,[\sup f_\theta(x) - \inf f_\theta(x)]$ verifies $(u-l)^{1+\delta} \in \mathcal{M}_\tau$ for some $\delta \in [0, 1[$. Consequently, **U4** is fulfilled, and so does **U2** (thanks to Lemma 5) in **AC** and **VR**.

Now, back to the consistency assumptions, one can prove that **C5a** holds true for all $K > K^\star$ in **VR** mimicking the proof in **MGM**.

Concerning the underestimation assumptions in **VR**, **U1** is straightforward, **U2** and **U4** are already proved, **U3** derives from Ascoli's theorem. A crude yet careful application of Taylor's integral remainder theorem yields **U5**.

In conclusion, **Oa** in **VR** is a simple by-product of the properties above. On the contrary, we can not prove that **Ob** holds true in **VR**.

We finally emphasize that **VR** is an example of exponential model, so Proposition 5 applies: $\widehat{K}_n^{\mathrm{G}}$ achieves the best underestimation rate with a view to Proposition 4.

A. Appendix.

A.1. *Proof of Proposition 1.* Inequality (4) is readily proved, since one has

$$\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n \ell_\theta \le \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) =$$

$$\sup_{\theta\in\Theta_{K_2}} \Big\{(\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) + P^\star(\ell_\theta - \ell^\star)\Big\} \le \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star).$$

Inequality (5) is more (à la Huber) tricky. Define for all $\theta \in \Theta_{K_2}$ such that $H(\theta) > 0$ (*i.e.* $P^\star \ne P_\theta$) the scaled likelihood ratio

$$g_\theta = \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}}$$

and $g_\theta = 0$ otherwise. Now, for any $\theta \in \Theta_{K_2}$, $H(\theta)$ nonnegative yields

$$\mathbb{P}_n(\ell_\theta - \ell^\star) + H(\theta) = (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) \le H(\theta)^{1/2} \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)g_\theta. \quad (24)$$

Let us set some $\theta_0 \in \Theta_{K_2}$ such that both $\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \le \mathbb{P}_n(\ell_{\theta_0} - \ell^\star) + \varepsilon$ and $\mathbb{P}_n(\ell_{\theta_0} - \ell^\star) \ge 0$. Then, (24) gives for θ_0

$$\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \le H(\theta_0)^{1/2} \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)g_\theta + \varepsilon.$$

Furthermore, $\mathbb{P}_n(\ell_{\theta_0} - \ell^\star) \geq 0$ combined with (24) implies in turn

$$H(\theta_0) \leq H(\theta_0)^{1/2} \sup_{\theta \in \Theta_{K_2}} (\mathbb{P}_n - P^\star) g_\theta,$$

hence

$$\sup_{\theta \in \Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_{K_1}} \mathbb{P}_n \ell_\theta \leq \sup_{\theta \in \Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq \left(\sup_{\theta \in \Theta_{K_2}} (\mathbb{P}_n - P^\star) g_\theta \right)^2 + \varepsilon$$

which completes the proof, since $\varepsilon > 0$ is arbitrary.

A.2. *Proof of Lemma 5.* Set $Q \in \mathcal{L}_\tau^*$. If $(u - l)$ is bounded, then $(u - l) \in \mathcal{M}_\tau$ hence $Q^s(u - l) = 0$ and Lemma 4 yields the first result. If $(u - l)$ is not bounded, then we simply write

$$Q\,(u - l)\mathbb{1}\{u - l > M\} \leq \frac{Q\,\psi(u - l)}{\psi(M)/M}$$

and conclude. The remainder of the proof is straightforward.

A.3. *Proof of Proposition 5.* Choose $K < K^\star$ and consider $\overline{Q}^a \in M_1(\mathcal{Z}) \cap \mathcal{L}'_\tau$ satisfying $H(\overline{Q}^a \,|P^\star) = H(\Gamma^a_{0,K}) < \infty$ (*cf.* the proof of Theorem 6 in Section 3.4). We aim at proving that $\overline{Q}^a \in \Pi_K$, which is a sufficient condition for the achievement of the best rate of underestimation by $\widehat{K}_n^{\mathrm{G}}$, since then $H(\overline{Q}^a \,|P^\star) = H(\Pi_K \,|P^\star)$ and K is arbitrarily smaller than $K^\star$.

The following technical lemmas are easily proved:

Lemma 15. Primo, $\overline{Q}^a = \bar{q}\,\mathrm{d}\mu$ *and* $\overline{Q}^a \log \bar{q} < \infty$. Secundo, *there exists* $\overline{P} = \bar{p}\,\mathrm{d}\mu \in \Pi_K$ *such that* $H(\overline{Q}^a \,|\overline{P}) = H(\overline{Q}^a \,|\Pi_K) = H(\overline{Q}^a \,|\Pi_{K^\star}) < \infty$.

Let us introduce now the sets

$$\mathcal{C}_1 = \Big\{ Q \in M_1(\mathcal{Z}) : Q << \mu,\ Q = q\,\mu \text{ and } Q\,\log q < \infty \Big\},$$

$$\mathcal{C}_2 = \Big\{ Q \in M_1(\mathcal{Z}) : Q\,\log \bar{p} = \sup_{\theta \in \Theta_{K^\star}} Q\,\ell_\theta \Big\} \quad \text{and}$$

$$\mathcal{C}_3 = \Big\{ Q \in M_1(\mathcal{Z}) : H(Q\,|\overline{P}) < \infty \Big\}.$$

Lemma 16. *The set* $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3 \cap \mathcal{L}'_\tau$ *is a convex set. It contains* $\overline{Q}^a$ *and it is included in* $\Gamma_{0,K}$. *Besides, for all* $Q \in \mathcal{C}$, $H(Q\,|\overline{P}) = H(Q\,|\Pi_{K^\star}) < \infty$.

The combination of Lemmas 15 and 16 with the Pythagorean characterization of H-projections (*cf.* Theorem 2.2 in (Csiszár, 1975)) straightforwardly yields:

Lemma 17. $\overline{Q}^a$ *is the H-projection of $P^\star$ on the convex set* $\mathcal{C}$, i.e. $H(\overline{Q}^a \,|P^\star) = H(\mathcal{C}\,|P^\star)$. *Consequently, for all* $Q \in \mathcal{C}$, $H(Q\,|P^\star) \geq H(Q\,|\overline{Q}^a) + H(\overline{Q}^a \,|P^\star)$ *and this property uniquely characterizes* $\overline{Q}^a$.

We do not have used the exponential nature of our models yet: here, the distributions $P_\theta = p_\theta \,\mathrm{d}\mu$ satisfy $p_\theta(z) = h(z)\exp(\theta^T t(z) - \phi(\theta))$ (any $z \in \mathcal{Z}$), where: $t = (t_1, \dots, t_{K^\star})$ is a known function on $\mathcal{Z} \subset \mathbb{R}^q$ equipped with the Lebesgue measure μ on Borel sets; $h \in \mathbb{R}^{\mathcal{Z}}$ is measurable; $\theta \in \Theta_K$, convex subset of the convex and supposed open natural parameter space $\Theta = \{\theta \in \mathbb{R}^{K^\star} : \mu(h\exp(\theta^T t)) < \infty\}$; $\phi(\theta) = \log\mu(h\exp(\theta^T t))$ (any $\theta \in \Theta$). We emphasize that ϕ is convex, differentiable on Θ, with $\dot\phi(\theta) = P_\theta\, t$.

It is remarkable that $\overline{P}$ also satisfies a Pythagorean property, *cf.* Lemma 18, whose proof draws its inspiration from (Čencov, 1982).

Lemma 18. $\forall\, Q \in \mathcal{C}$, $\forall\, P_\theta \in \Pi_{K^\star}$, $H(Q\,|P_\theta) \geq H(Q\,|\overline{P}) + H(\overline{P}\,|\Pi_\theta)$.

Proof. First, since $Q\log q$ and $H(Q\,\overline{P})$ are finite, $H(Q\,|\overline{P}) = Q\log q - Q\log\bar p$. Furthermore, $Q \in \mathcal{L}_\tau'$ and $\ell_\theta \in \mathcal{L}_\tau$ yield $H(Q\,|P_\theta) = Q\log q - Q\,\ell_\theta$ (any $\theta \in \Theta_{K^\star}$). Now, $H(Q\,|\overline{P}) = H(Q\,|\Pi_{K^\star}) < \infty$ implies that

$$0 \leq Q\log\frac{\bar p}{p_\theta} < \infty \text{ (any } \theta \in \Theta_{K^\star}). \tag{25}$$

Let Θ_c be the subset of Θ of all the parameters θ such that $H(Q\,|P_\theta) < \infty$: it is open (thanks to the decomposition of $H(Q\,|P_\theta)$ above and Lemma 12) and convex (this is readily seen). It suffices to show that the inequality stated in the lemma we are proving holds true on Θ_c (it is obvious on $\Theta \setminus \Theta_c$). Observe that $Q\log q < \infty$ and $H(Q\,|\overline{P}) < \infty$ ensures that the latter inequality is equivalent to $(Q - \overline{P})\log\bar p/p_\theta \geq 0$ (any $\theta \in \Theta_c$). Since the model is exponential, it turns out that this is in turn equivalent to (with notation $\overline{P} = P_{\bar\theta}$)

$$(\bar\theta - \theta)^T(Q - \overline{P})\,t \geq 0 \quad \text{(any } \theta \in \Theta_c). \tag{26}$$

Let us finally define f on Θ_c by $f(\theta) = Q\log\bar p/p_\theta = (\bar\theta - \theta)^T\,Q\,t + \phi(\theta) - \phi(\bar\theta)$. Then, the convexity of ϕ and (25) yield that f is a proper convex function on Θ_c. Furthermore, f is differentiable at $\bar\theta$, with gradient $\dot f(\bar\theta) = (\overline{P} - Q)\,t$. Since f attains its minimum in $\bar\theta$ in virtue of (25), Theorem 27.4 of (Rockafellar, 1970) applies, hence

$$(\theta - \bar\theta)^T\,\left(-\dot f(\bar\theta)\right) = (\theta - \bar\theta)^T\,(Q - \overline{P})\,t \leq 0 \text{ (any } \theta \in \Theta_c).$$

This is exactly (26), so the proof is complete.

Proof of Proposition 5. Let Q be an element of $\mathcal{C}$. Then Lemma 18 applied to $P^\star \in \Pi_{K^\star}$ yields

$$H(Q\,|P^\star) \geq H(Q\,|\overline{P}) + H(\overline{P}\,|P^\star).$$

Since the latter is true for all $Q \in \mathcal{C}$, $\overline{P}$ is the H-projection of $P^\star$ on the convex set $\mathcal{C}$ (thanks the characterization of Lemma 17), *i.e.* $\overline{Q}^a = \overline{P}$. Consequently, $\overline{Q}^a \in \Pi_K$, which is the final conclusion.

A.4. *Proof of Lemma 11.* Set $\alpha > 0$, $K < K^\star$ and $\varepsilon > 0$. We shall actually prove that $\Lambda^c_{\alpha,K}$ is an open set. The same proof applies to $\Gamma_{\alpha,K}$, up to minor changes in indices.

Choose $Q_0 \in \Lambda^c_{\alpha,K}$ and $\alpha' > \alpha$ such that $\sup_{\theta\in\Theta_K} Q_0\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q_0\,\ell_\theta < -\alpha'$: can we construct an open neighbourhood of Q_0 in $\Lambda^c_{\alpha,K}$?

First, observe that for any subset A of $\mathcal{Y}$, T of Θ_{K+1} and $Q \in \mathcal{L}_\tau$, $Q \geq 0$,

$$\sup_{\theta\in T} Q\,\ell_\theta \mathbb{1}\{A\} \leq \sup_{\theta\in T} Q\,\ell_\theta - Q\,l\mathbb{1}\{A^c\},$$
$$\sup_{\theta\in T} Q\,\ell_\theta \leq \sup_{\theta\in T} Q\,\ell_\theta \mathbb{1}\{A\} + Q\,u\mathbb{1}\{A^c\}$$

and forwardly

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{A\} - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{A\} \leq \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta + Q\,(u-l)\mathbb{1}\{A^c\}, \tag{27}$$

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \leq \sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{A^c\} - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{A^c\} + Q\,(u-l)\mathbb{1}\{A^c\}. \tag{28}$$

Consequently, if C is a compact set of $\mathcal{Y}$ such that $Q_0\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$ in virtue of **U2**, (27) with $Q = Q_0$ and $A = C$ gives

$$\sup_{\theta\in\Theta_K} Q_0\,\ell_\theta \mathbb{1}\{C\} - \sup_{\theta\in\Theta_{K+1}} Q_0\,\ell_\theta \mathbb{1}\{C\} \leq \alpha' + \varepsilon. \tag{29}$$

Secondly, since $\{\,\ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_K\}$ ($\{\,\ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_{K+1}\}$, respectively) is precompact in $\mathrm{C}^0(C, \|\cdot\|_\infty)$, the Borel-Lebesgue property ensures that there exist $\theta_1, \ldots, \theta_r \in \Theta_K$ ($\theta_{r+1}, \ldots, \theta_{r+s} \in \Theta_{K+1}$, respectively) such that the union over $1 \leq i \leq r$ ($r+1 \leq i \leq r+s$, respectively) of the balls $B_{\|\cdot\|_\infty}(\ell_{\theta_i}\mathbb{1}\{C\}, \varepsilon)$ covers it. Observe now that $f\mathbb{1}\{A\} \in \mathcal{L}_\tau$ as soon as $f \in \mathcal{L}_\tau$ and introduce the open neighbourhood V of Q_0 defined by

$$V = \Big\{Q \in \mathcal{Q} : \big|Q\,(u-l)\mathbb{1}\{C^c\} - Q_0\,(u-l)\mathbb{1}\{C^c\}\big| < \varepsilon\Big\} \cap \bigcap_{i=1}^{r+s} \Big\{Q \in \mathcal{Q} : \big|Q\,\ell_{\theta_i}\mathbb{1}\{C\} - Q_0\,\ell_{\theta_i}\mathbb{1}\{C\}\big| < \varepsilon\Big\}.$$

Then for all $Q \in V$, since $Q1 = 1$ and $Q \geq 0$, one can verify easily that

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{C\} \leq \sup_{\theta\in\Theta_K} Q_0\,\ell_\theta \mathbb{1}\{C\} + 2\varepsilon,$$
$$\sup_{\theta\in\Theta_{K+1}} Q_0\,\ell_\theta \mathbb{1}\{C\} \leq \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{C\} + 3\varepsilon.$$

Combining those two previous inequalities with (28) and (29) implies that $V \subset \Lambda^c_{\alpha,K}$ since $\varepsilon > 0$ is arbitrarily small.

Acknowledgments. This work is part of my PhD thesis, and I wish to express my gratitude to my advisor Elisabeth Gassiat. I would also like to thank Stéphane Boucheron, Christian Léonard and Jamal Najim; Raphaël Cerf and Pascal Massart for helpful and luminous advises.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723. System identification and time-series analysis.

Antoniadis, A. and Berruyer, J. (1986). On estimating the number of components in a finite mixture of power series distributions. *Comput. Statist. Data Anal.*, 4(4):229–241.

Bahadur, R. R. (1967). An optimal property of the likelihood ratio statistic. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 13–26. Univ. California Press, Berkeley, Calif.

Bahadur, R. R. (1971). *Some limit theorems in statistics.* Society for Industrial and Applied Mathematics, Philadelphia, Pa. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 4.

Bahadur, R. R. and Zabell, S. L. and Gupta, J. C. (1980). Large deviations, tests, and estimates. In *Asymptotic theory of statistical tests and estimation (Proc. Adv. Internat. Sympos., Univ. North Carolina, Chapel Hill, N.C., 1979)*, pages 33–64. Academic Press, New York.

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

Čencov, N. N. (1982). *Statistical decision rules and optimal inference.* American Mathematical Society, Providence, R.I. Translation from the Russian edited by Lev J. Leifman.

Chambaz, A. (2002). Detecting abrupt changes in random fields. *ESAIM P&S*, 6:189–209.

Chernoff, H. (1956). Large sample theory: parametric case. *Ann. Math. Statist.*, 27:1–22.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158.

Csiszár, I. and Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.*, 6:1601–1619.

Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli*, 3(3):279–299.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.*, 27(4):1178–1209.

Dembo, A. and Zeitouni, O. (1998). *Large deviations techniques and applications.* Springer-Verlag, New York.

Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete*, 62(4):509–552.

Dupuis, P. and Ellis, R. S. (1997). A weak convergence approach to the theory of large deviations. Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. To appear in Ann. Inst. H. Poincaré Probab. Statist.

Gassiat, E. and Boucheron, S. (2001). Optimal error exponents in Hidden Markov Models order estimation. Preprint, submitted.

Guyon, X. and Yao, J. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal.*, 70(2):221–249.

Haughton, D. and Keribin, C. (2001). Asymptotic probabilities of overestimating and underestimating the order of a model in general regular families. Submitted.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā Ser. A*, 51(1):45–58.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16(1):342–355.

Henna, J. (1985). On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.*, 37(2):235–240.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif.

James, L. F. and Priebe, C. E. and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.*, 29(5):1281–1296.

Kallenberg, W. C. M. and Kourouklis, S. (1992). Hodges-Lehmann optimality of tests. *Statist. Probab. Lett.*, 14(1):31–38.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62(1):49–66.

Kourouklis, S. (1991). Bahadur efficiency of likelihood ratio and related tests in nonregular models. *Austral. J. Statist.*, 33(3):291–298.

Léonard, C. and Najim, J. (2000). An extension of Sanov's theorem. Application to the Gibbs conditioning principle. Preprint.

Leonardi, G. P. and Tamanini, I. (2002). Metric spaces of partitions. Preprint.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.*, 20(3):1350–1360.

Lindsay, B. G. (1989). Moment matrices: applications in mixtures. *Ann. Statist.*, 17(2):722–740.

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

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9 (2):245–303.

Merhav, N. (1989). The estimation of the model order in exponential families. *IEEE Trans. Inform. Theory*, 35 (5): 1109–1114.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory*, 1:295–314.

Rao, M. M. and Ren, Z. D. (1991). *Theory of Orlicz spaces.* New York: Marcel Dekker Inc.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica 14*, 465–471.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton University Press.

Schied, A. (1998). Cramer's condition and Sanov's theorem. *Statist. Probab. Lett. 39*, 55–60.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Senoussi, R. (1990). Statistique asymptotique presque sûre de modèles statistiques convexes. *Ann. Inst. H. Poincaré Probab. Statist.*, 26(1):19–44.

Titterington, D. M. and Smith, A. F. M. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* John Wiley & Sons Ltd., Chichester.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

Wu, L. (1994). Large deviations, moderate deviations and LIL for empirical processes. *Ann. Probab.*, 22(1):17–27.

Antoine Chambaz
Université Paris-Sud
Bât. 425
91405 Orsay cedex
France.