INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2005-27

Nonasymptotic Bounds for Bayesian Order Identification with Application to Mixtures

A. CHAMBAZ[1]
J. ROUSSEAU[2]

[1] Laboratoire de Statistiques, Université Paris V, 45 rue des Saints-Pères, 75006 Paris. antoine.chambas@univ-paris5.fr
[2] CEREMADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75016 PARIS and CREST-INSEE, France. rousseau@ceremade.dauphine.fr

Nonasymptotic bounds for Bayesian order identification with application to mixtures

By

A. Chambaz [1] and J. Rousseau [2]

[1] Laboratoire de Statistique, Université Paris V,
45 rue des St-Pères, 75 006 Paris .
antoine.chambaz@univ-paris5.fr

[2] CEREMADE, Université Paris Dauphine,
Place du Maréchal deLattre de Tassigny, 75 016 Paris
and CREST - INSEE France
rousseau@ceremade.dauphine.fr

11/05

NONASYMPTOTIC BOUNDS FOR BAYESIAN ORDER IDENTIFICATION WITH APPLICATION TO MIXTURES

By Antoine Chambaz and Judith Rousseau

Université René Descartes and Université Dauphine

Abstract

The efficiency of two Bayesian order estimators is studied under weak assumptions. By using nonparametric techniques, we prove new nonasymptotic underestimation and overestimation bounds. The bounds compare favorably with optimal bounds yielded by the Stein lemma and also with other known asymptotic bounds. The results apply to mixture models. In this case, the underestimation probabilities are bounded by a constant times e^{-an} (some $a > 0$, all sample size $n \geq 1$). The overestimation probabilities are bounded by $1/\sqrt{n}$ (all n larger than a known integer), up to a $\log n$ factor.

Résumé

Dans cet article nous étudions deux estimateurs bayésien de l'ordre d'un modèle, le mode a posteriori et un estimateur basé sur un facteur de Bayes séquentiel. Nous étudions l'efficacité de ces estimateurs, en particulier nous obtenons, dans un cadre général des bornes non asymptotiques sur les probabilités de surestimation et de sousestimation de cet ordre. Nous appliquons ces résultats au cas des mélange où nous montrons que la probabilité de surestimer l'ordre décroit en $\sqrt{n}$ et celle de sousestimer décroit exponentiellement vite.

AMS 2000 subject classifications: Primary 62F05, 62F12, 62G05, 62G10

Keywords and phrases: Mixture, Model selection, Nonparametric Bayesian inference, Order estimation, Rate of convergence

1. Introduction. Model choice is an important and difficult topic and the literature on the subject is vast. In this paper, we consider a special case of model choice, namely the identification of the order of a model. Order identification deals with the estimation and test of a structural parameter which indexes the complexity of a model. In words, a most economical representation of a random phenomenon is sought. This problem is encountered in many situations and for instance: in mixture models (Titterington et al., 1985; McLachlan & Peel, 2000) with unknown number of components; in autoregressive models (Azencott & Dacunha-Castelle, 1986), when the process memory is not known; in cluster analysis (Hastie et al., 2001), when the number of clusters is unknown. One of the main difficulties lies in the fact that, although it is stated as a discrete problem, order identification is in essence continuous.

This paper is devoted to the study of two Bayesian estimators of the order of a model. Frequentist properties of efficiency are particularly investigated. We obtain new nonasymptotic efficiency bounds under mild assumptions. Those bounds provide a theoretical answer to the questions raised for instance in (Fraley & Raftery, 2002) (see their Section 4). Application of the main results to the notoriously difficult problem of order identification for mixture models illustrates their generality.

1.1. *Description of the problem.* We observe n i.i.d. random variables $Z_1, \ldots, Z_n$ with values in a measured sample space $(\mathcal{Z}, \mathcal{F}, \mu)$. These observations are defined on a common measurable space upon which all the random variables will be defined.

Let $\{(\Theta_k, d_k)\}_{k\geq 1}$ be an increasing family of nested parametric sets (d_k will abbreviate to d for simplicity). The dimension of Θ_k is denoted by $D(k)$. Let us introduce $\Theta_\infty = \cup_{k\geq 1}\Theta_k$. For every $\theta \in \Theta_\infty$, let f_θ be a probability density with respect to the measure μ. We denote by P_θ the probability measure whose density is f_θ. The expectation with respect to P_θ (resp. $P_\theta^{\otimes n} = P_\theta^n$) writes as E_θ (resp. E_θ^n).

The order of any distribution P_{θ_0} is the unique integer k such that $P_{\theta_0} \in \{P_\theta : \theta \in \Theta_k \setminus \Theta_{k-1}\}$ (with convention $\Theta_0 = \varnothing$). It is assumed that the distribution $P^\star$ of Z_1 belongs to one model $\{P_\theta : \theta \in \Theta_k\}$ for some $k \geq 1$.

The density of $P^\star$ is denoted by $f^\star = f_{\theta^\star}$ ($\theta^\star \in \Theta_{k^\star} \setminus \Theta_{k^\star-1}$). The order of $P^\star$ is denoted by $k^\star$, and is the quantity we want to estimate.

We are interested in frequentist properties of two Bayesian estimates of $k^\star$. In that perspective, the problem can be restated as an issue of composite hypotheses testing: we want to decide between the null hypothesis "$k^\star \leq k_0$" and its alternative "$k^\star > k_0$" (for some integer k_0), that is to test

$$\text{“}P^\star \in \{P_\theta : \theta \in \Theta_{k_0}\}\text{”} \quad \text{against} \quad \text{“}P^\star \notin \{P_\theta : \theta \in \Theta_{k_0}\}\text{”}.$$

This question is obviously crucial when the order is the quantity of interest. Furthermore, order identification may also be a prerequisite to consistent parameter estimation, when overestimation of the order causes loss of identifiability.

Efficiency issues. Let α_n and β_n be the type I and type II errors of a procedure that tests the hypotheses above. The efficiency of this procedure is measured in terms of rates of convergence of α_n and β_n to zero. Obviously, if $\widetilde{k}_n$ estimates $k^\star$, then the natural rule is to reject the null hypothesis if $\widetilde{k}_n > k_0$. Then

$$\alpha_n \leq P^\star\{\widetilde{k}_n > k^\star\} \quad \text{and} \quad \beta_n \leq P^\star\{\widetilde{k}_n < k^\star\}.$$

These upper bounds do not depend on k_0. Moreover, even from a Bayesian decision theoretical point of view, these quantities are of interest when considering 0-1 types of losses.

In most previous work, the behavior of the underestimation probability $P^\star\{\widetilde{k}_n < k^\star\}$ and overestimation probability $P^\star\{\widetilde{k}_n > k^\star\}$ are investigated as the sample size n grows to infinity. In this paper, we obtain new nonasymptotic upper bounds of these probabilities for two Bayesian order estimators.

Two Bayesian procedures. Let Π be a prior on Θ_∞ that writes as

$$d\Pi(\theta) = \pi(k)\, \pi_k(\theta)\, d\theta$$

for all $\theta \in \Theta_k$ (every $k \geq 1$). We denote by $\Pi(k|Z^n)$ the posterior probability of each $k \geq 1$.

In a Bayesian decision theoretic perspective, the Bayes estimator associated with the 0-1 loss function is the mode of the posterior distribution of the order k:

$$\widehat{k}_n^{\mathrm{G}} = \arg\max_{k\geq 1}\{\Pi(k|Z^n)\}.$$

This estimator is often used in practice. It is a global estimator in the sense that it takes into account the whole of the posterior distribution on k. Following the Ockam's razor principle and considering a more local and sequential approach, we propose another estimator:

$$\widehat{k}_n^{\mathrm{L}} = \inf\{k \geq 1 : \Pi(k|Z^n) \geq \Pi(k+1|Z^n)\} \leq \widehat{k}_n^{\mathrm{G}}.$$

If the posterior distribution on k is unimodal, then obviously both estimators are equal. The advantage of $\widehat{k}_n^{\mathrm{L}}$ over $\widehat{k}_n^{\mathrm{G}}$ is that $\widehat{k}_n^{\mathrm{L}}$ does not require the computation of the whole posterior distribution on k. It can also be slightly modified into a second sequential estimator, based on the *marginal likelihood* $\Pi(k|Z^n)/\pi(k)$:

$$\inf\left\{k \geq 1 : \frac{\Pi(k|Z^n)}{\pi(k)} \geq \frac{\Pi(k+1|Z^n)}{\pi(k+1)}\right\}. \tag{1}$$

In other words, this estimator equals the smallest integer k such that the *Bayes factor* comparing Θ_{k+1} to Θ_k is less than one. When considering a model comparison point of view, Bayes factors are often used to compare two models (Kass & Raftery, 1995). They are also to some extent Bayesian solutions to the 0-1 loss, in a two models testing problem. One of the advantages of Bayes factors over posterior probabilities is that the prior probability of each (sub) model does not need to be specified.

In the following, we shall focus on $\widehat{k}_n^{\mathrm{G}}$ and $\widehat{k}_n^{\mathrm{L}}$. The estimator defined by (1) differs very little from $\widehat{k}_n^{\mathrm{L}}$ and shares the same properties.

1.2. *Results in perspective.* The resort to nonparametric techniques in the spirit of (Barron et al., 1999; Ghosal et al., 2000) combined with a variant of the locally conic parameterization (first introduced in (Dacunha-Castelle & Gassiat, 1997b)) allows to prove that the underestimation probabilities are both a $O(e^{-an})$ (some $a > 0$), see Theorems 1 and 2. We also show that the overestimation probabilities are both a $O((\log n)^b/n^c)$ (some

$b \geq 0$, $c > 0$), see Theorems 3 and 4. The bounds we get actually hold for any sample size. All constants can be expressed explicitly. This particularly makes it possible to compare precisely our results with previous ones. Besides, it is interesting to analyze the contribution of each assumption to the final bounds. Finally, we show that our results apply to mixture models, for which order identification is notoriously difficult. Such models are notably characterized by their lack of identifiability when overestimating the order and the subsequent singularity of the Fisher information matrix, which prevents from using classical methods based on Taylor expansions. It is also known (Dacunha-Castelle & Gassiat, 1999) that the maximum likelihood statistic is not asymptotically Gaussian. However, we obtain for $\widehat{k}_n^{\mathrm{L}}$ that the underestimation probability is a $O(e^{-an})$ and that the overestimation probability is a $O(1/\sqrt{n})$, see Theorem 5. The case of location-scale mixture of Gaussian densities is particularly addressed in Corollary 1.

In most previous work, the choice of the framework is contingent on the need for tractable explicit calculus. Order identification for exponential models is studied in (Haughton, 1989), with a generalization to regular models (Keribin & Haughton, 2003). A method is developed ad hoc for mixture order identification in (Dacunha-Castelle & Gassiat, 1997a). In (Guyon & Yao, 1999), a common basic procedure is adapted then applied to various models characterized by the existence of an exhaustive finite dimensional statistic. (Boucheron & Gassiat, 2004) is devoted to order identification for autoregressive processes. All these papers deal with efficiency issues. In (Chambaz, 2003), general efficiency results are obtained thanks to the resort to powerful properties of empirical processes.

None of the papers cited above use Bayesian techniques. There is an extensive literature on Bayesian estimation of mixture models and in particular on the order selection in mixture models. However this literature is essentially devoted to determining coherent noninformative priors, see for instance (Moreno & Liseo, 2003) and to implementing procedures, see for instance (Mengersen & Robert, 1996). To the best of our knowledge, there is hardly any work on frequentist properties of Bayesian estimators such as $\widehat{k}_n^{\mathrm{G}}$ and $\widehat{k}_n^{\mathrm{L}}$ outside the regular case. In the case of mixture models, Ishwaran et al. (2001) suggest a Bayesian estimator of the mixing distribution when the

number of components is unknown and bounded. They deduce from this an estimator of the order based on a sort of penalized likelihood; they obtain rates of convergence on the mixing distribution but not on the order. Note that in (Finesso et al., 1996; Gassiat & Boucheron, 2003), asymptotic efficiency results are obtained for estimates that have a Bayesian flavor. They are indeed based on the so-called Krichevsky-Trofimov mixture, which is a product of Dirichlet priors in the information theory literature (Csiszár & Körner, 1981).

Not only our Bayesian approach to order identification yields nonasymptotic bounds in place of the usual asymptotic results. In addition, it provides an answer (at least theoretical) to the delicate question of the choice of the penalty which is central in all the papers cited above.

Indeed, posterior probabilities or Bayes factors naturally take into account the uncertainty on the parameter by integrating it out (Jefferys & Berger, 1992).

On the contrary, any bare information criterion chooses the largest model, since the latter necessarily fits the best the data at hand. A well balanced penalization term can compensate this drawback. In (Chambaz, 2003), two minimal penalization requirements yield two asymptotic overestimation behaviors (see Theorems 10 and 11). However, both requirements exclude penalization terms in $\log n$ and particularly the *Bayesian Information Criterion* (BIC) (Schwarz, 1978), that is the maximum likelihood penalized by a term $-\frac{1}{2}D(k)\log n$ (for Θ_k).

The BIC criterion is well behaved in regular models. Nonetheless, it is known that it can be inconsistent outside regular models, for instance in models where the number of parameters increases with n (Stone, 1979). The role of BIC criterion is preponderant due to its equivalence with marginal likelihood in regular models (Kass & Raftery, 1995). This equivalence does not always hold. For instance, Berger et al. (2003) have proved that in ANOVA models, marginal likelihood behaves like a penalized maximum likelihood, where the penalization term differs from the BIC one. Also, the equivalence between marginal likelihood and BIC criterion has not been established in mixture models, where the maximum likelihood statistic has a nonstandard asymptotic distribution. Although we have not sought to

determine an equivalent of the marginal likelihood, it appears (see Section 2.3), that generally speaking it behaves like a penalized likelihood estimator, where the penalization is equal to the logarithm of the prior probability of $1/n$-Kullback-Leibler neighborhoods of the true density. Perhaps more restrictively, it seems that the penalization can be expressed as $-\frac{1}{2}\widetilde{D}(k)\log n$, where $\widetilde{D}(k)$ represents an *effective dimension* of Θ_k relative to $\Theta_{k^\star}$.

1.3. *Organization of the paper.* In Section 2, we state our main results. Underestimation is addressed in Section 2.2 and overestimation in Section 2.3. We deal with Bayesian order identification for mixture models in Section 2.4. The main proofs are gathered in Section 3 (underestimation), Section 4 (overestimation) and Section 5 (mixtures). In Section A of the appendix, a useful family of tests is introduced. Section B is devoted to the subtle verification of an assumption, by using a variant of the locally conic parameterization. This section may be of general interest. Finally, entropy estimates are considered in Section C.

2. Nonasymptotic efficiency bounds.

2.1. *Notations and global assumptions.* Let us introduce some notations.

For every $a, b \in \mathbb{R}$, $\max(a,b)$ and $\min(a,b)$ are denoted by $a \vee b$ and $a \wedge b$, respectively. The integral $\int f d\lambda$ of a function f with respect to a measure λ is written as λf.

Let $L^1_+(\mu)$ be the subset of all nonnegative functions in $L^1(\mu)$. For every $f \in L^1_+(\mu)\setminus\{0\}$, the measure P_f is defined by its derivative $\mathrm{d}P_f/\mathrm{d}\mu = f$ with respect to μ. For every $f, f' \in L^1_+(\mu)$ such that $P_f \ll P_{f'}$ and $\mu f = \mu f' = 1$, $H(f,f')$ stands for the relative entropy (or Kullback-Leibler divergence) between P_f and $P_{f'}$ (Dupuis & Ellis, 1997):

$$H(f,f') = P_f\left(\log f - \log f'\right) \geq 0,$$

with equality if and only if $f = f'$ P_f-a.s. This definition is extended to functions $f, f' \in L^1_+(\mu) \setminus \{0\}$ for which $P_f(\log f - \log f')$ is defined.

Let finally $H(f^\star, f) = H(f)$ for every $f \in L^1_+(\mu) \setminus \{0\}$.

For every $f, f' \in L^1_+(\mu)$, let also

$$V(f,f') = P_f\left(\log f - \log f'\right)^2$$

(with convention $V(f,f')=\infty$ if the set $\{z : \log f(z)/f'(z)=\infty\}$ has positive P_f measure). It is worth noting that, for every $f, f' \in L_+^1(\mu)\setminus\{0\}$,

$$|H(f,f')| \le \left[V(f,f')\,\mu f\right]^{1/2}. \tag{2}$$

Besides, the quantity

$$V(f) = V(f^\star, f) \vee V(f, f^\star) \quad (\text{every } f \in L_+^1(\mu))$$

will play a central role in this study.

For every probability density $f \in L^1(\mu)$, the probability measure $P_{f^{\otimes n}}$ is denoted P_f^n for abbreviation. The expectation with respect to P_f (resp. P_f^n) is denoted by E_f (resp. E_f^n).

For simplicity of notations, for every $f \in L_+^1(\mu)\setminus\{0\}$ and $\theta,\theta' \in \Theta^\infty$, the following shortcuts will be used throughout the paper: $H(f,f_\theta)=H(f,\theta)$, $H(f_\theta,f)=H(\theta,f)$, $H(f_\theta,f_{\theta'})=H(\theta,\theta')$ and $H(f_\theta)=H(\theta)$. Similarly, for every $f \in L_+^1(\mu)$ and $\theta,\theta' \in \Theta^\infty$: $V(f,f_\theta)=V(f,\theta)$, $V(f_\theta,f)=V(\theta,f)$, $V(f_\theta,f_{\theta'})=V(\theta,\theta')$ and $V(f_\theta)=V(\theta)$.

For all $\theta \in \Theta_\infty$, we introduce the following quantities: $\ell_\theta = \log f_\theta$, $\ell^\star = \log f^\star$, $\ell_n(\theta)=\sum_{i=1}^n \ell_\theta(Z_i)$, $\ell_n^\star = \sum_{i=1}^n \ell^\star(Z_i)$ and, for every $k \ge 1$,

$$\mathbb{B}_n(k) = \pi(k)\int_{\Theta_k} e^{\ell_n(\theta)-\ell_n^\star} d\pi_k(\theta).$$

Obviously, if $k<k'$ are two integers, then $\widehat{k}_n^{\mathrm{L}}=k$ yields $\mathbb{B}_n(k)\ge\mathbb{B}_n(k+1)$ and $\widehat{k}_n^{\mathrm{G}}=k$ implies that $\mathbb{B}_n(k)\ge\mathbb{B}_n(k')$.

Basic assumptions. The dimension of Θ_k is denoted by $D(k)$. The two following assumptions will be needed throughout this paper.

A1 Compactness. For every $k \ge 1$, the parameter set (Θ_k, d) is a compact metric set.

A2 Parameterization. The parameterization $\theta \mapsto \ell_\theta(z)$ from Θ_k to $\mathbb{R}$ is continuous for every $z \in \mathcal{Z}$ and $k \ge 1$.

The continuous parameterization assumption **A2** is standard in Statistics (see for instance (van der Vaart, 1998)). Another standard assumption is the boundedness of the parameter sets. Assumption **A1** is slightly stronger.

2.2. *Underestimation efficiency.* Let us define, for every $k \geq 1$, $\alpha, \delta > 0$ and $\theta \in \Theta^\infty$,

$$\begin{aligned} H_k^\star &= \inf\{H(\theta) : \theta \in \Theta_k\}, \\ S_k(\delta) &= \{\theta \in \Theta_k : H(\theta) \leq H_k^\star + \delta/2\}, \\ q(\theta, \alpha) &= P^\star(\ell^\star - \ell_\theta)^2 e^{\alpha(\ell^\star - \ell_\theta)} + V(\theta^\star, \theta) \in [0, \infty]. \end{aligned}$$

The definition of the order $k^\star$ yields that $H_k^\star = 0$ for all $k \geq k^\star$.

Three main assumptions are required when dealing with underestimation efficiency.

U-prior. $\pi_k\{S_k(\delta)\} > 0$ for all $\delta > 0$ and $k = 1, \ldots, k^\star$.

U-moment. There exist $\delta_0 > 0$ and $\alpha, M > 0$ such that, for all $\delta \in (0, \delta_0]$,

$$\sup_{1 \leq k \leq k^\star} \sup_{\theta \in S_k(\delta)} q(\theta, \alpha) \leq M.$$

U-local brackets. For $k = 1, \ldots, k^\star - 1$, for all $\theta \in \Theta_k$, let us define for every $\eta > 0$:

$$\begin{aligned} l_{\theta,\eta} &= \inf\{f_{\theta'} : \theta' \in \Theta_k, d(\theta, \theta') < \eta\} \quad \text{and} \\ u_{\theta,\eta} &= \sup\{f_{\theta'} : \theta' \in \Theta_k, d(\theta, \theta') < \eta\}. \end{aligned}$$

For every $k = 1, \ldots, k^\star - 1$ and $\theta \in \Theta_k$, there exists $\eta_\theta > 0$ such that

$$V(u_{\theta,\eta_\theta}, \theta^\star) + V(\theta^\star, l_{\theta,\eta_\theta}) + V(\theta^\star, u_{\theta,\eta_\theta}) + V(u_{\theta,\eta_\theta}, \theta) < \infty.$$

THEOREM 1. *Let us assume that assumptions* **U-prior**, **U-moment** *and* **U-local brackets** *are satisfied.*

If in addition $H_k^\star > H_{k+1}^\star$ for $k = 1, \ldots, k^\star - 1$, then there exist $c_1, c_2 > 0$ such that, for every $n \geq 1$,

$$P^{\star n}\left\{\widehat{k}_n^{\mathrm{L}} < k^\star\right\} \leq c_1 e^{-nc_2}. \tag{3}$$

Similarly,

THEOREM 2. *If assumptions* **U-prior**, **U-moment** *and* **U-local brackets** *are valid, then there exist $c_1', c_2' > 0$ such that, for every $n \geq 1$,*

$$P^{\star n}\left\{\widehat{k}_n^{\mathrm{G}} < k^\star\right\} \leq c_1' e^{-nc_2'}. \tag{4}$$

Exploring the assumptions. Assumption **U-prior** is classical in the Bayesian literature. It requires that prior probabilities π_k do put some mass around the Kullback-Leibler projections of $\theta^\star$ upon Θ_k for all $k = 1, \dots, k^\star$.

Assumption **U-moment** is a condition of existence of *some* (rather than *any*) exponential moment for log ratios of densities $(\ell^\star - \ell_\theta)$ for θ ranging over some neighborhoods of the same projections of $\theta^\star$. Although some kind of uniformity is required of these exponential moments, we want to emphasize that assumption **U-moment** is mild. As explained in (Chambaz, 2003), the underestimation phenomenon is tightly bound to large deviations of the log-likelihood process (see Section 4.2 therein). Such large deviations results are quite easy to obtain under the classical Cramér condition of existence of *any* exponential moments. They are much more delicate when existence of *some* exponential moments is guaranteed, as it is here or in (Chambaz, 2003).

As for assumption **U-local brackets**, it is less restrictive than the existence of $l, u \in \mathbb{R}^{\mathcal{Z}}$ such that $(u - l) \in L^2(u)$ and $l \leq l_\theta \leq u$ (μ-almost everywhere for all $\theta \in \Theta_{k^\star}$), which is also a standard assumption.

Comment. According to inequalities (3) and (4), both underestimation probabilities decay exponentially with respect to the sample size n. This is the best achievable rate. Indeed, a variant of the Stein lemma (see Theorem 2.1 in (Bahadur et al., 1980)) guarantees that, if $\widetilde{k}_n$ is an order estimator which ultimately overestimates it with a probability bounded away from one, then the underestimation rate is at most exponential (an optimal exponent is provided), see Lemma 3 in (Chambaz, 2003)).

Exact values of constants c_1, c_1', c_2, c_2' can be found in the proofs of Theorems 1 and 2. We think that they shed some light on the underestimation phenomenon. Their expressions involve quantities among which: a common finite exponential moment for log ratios of densities $(\ell^\star - \ell_\theta)$ (constant α of assumption **U-moment**); prior masses of neighborhoods of the projections of $\theta^\star$ upon Θ_k for $k < k^\star$ (constant c in the proof of Theorem 1); covering numbers (constant N_ε in the same proof); differences $H_k^\star - H_{k+1}^\star > 0$ for $\widehat{k}_n^{\mathrm{L}}$ (or $H_k^\star - H_{k^\star}^\star = H_k^\star > 0$ for $\widehat{k}_n^{\mathrm{G}}$); infima of ratios $\inf_{\theta \in \Theta_k} [H(\theta) - H_{k+1}^\star]/V(\theta)$ for $\widehat{k}_n^{\mathrm{L}}$ (or $\inf_{\theta \in \Theta_k} H(\theta)/V(\theta)$ for $\widehat{k}_n^{\mathrm{G}}$).

It is natural to compare c_2 and c_2' (known as *underestimation error expo-*

nents in the information theory literature) to the constant $\inf_{\theta\in\Theta_{k^\star-1}} H(\theta,\theta^\star)$ which appears in Stein's lemma. The constants do not match.

We want to emphasize that this does not mean that $\widehat{k}_n^{\mathrm{L}}$ and $\widehat{k}_n^{\mathrm{G}}$ are not optimal. See (Chambaz, 2003) for a discussion about optimality.

2.3. *Overestimation efficiency.* Given $\delta>0$ and two functions $l\le u$, the bracket $[l,u]$ is the set of all functions f with $l\le f\le u$. The bracket $[l,u]$ is a δ-bracket if $l,u\in L_+^1(\mu)$ and

$$\mu(u-l)\le\delta,\quad P^\star(\log u-\log l)^2\le\delta^2,$$
$$P_{u-l}(\log u-\log f^\star)^2\le\delta\log^2\delta\quad\text{and}\quad P_l(\log u-\log l)^2\le\delta\log^2\delta.$$

For $\mathcal{C}$ a class of functions (or $\mathcal{C}$ a set which parameterizes a class of functions), the δ-entropy with bracketing of $\mathcal{C}$ is the logarithm $\mathcal{E}(\mathcal{C},\delta)$ of the minimum number of δ-brackets needed to cover $\mathcal{C}$ (or the class of functions that $\mathcal{C}$ parameterizes). A set of cardinality $\exp(\mathcal{E}(\mathcal{C},\delta))$ of δ-brackets which cover $\mathcal{C}$ is called a δ-bracketing net and written as $\mathcal{H}(\mathcal{C},\delta)$.

Let us state the five main assumptions which are required for controlling the overestimation efficiency. Let $K\ge k^\star+1$ be an integer.

O-comparison. There exist $h_1\in(0,1]$ and $C_1\ge 1/32$ such that, for all $k=k^\star+1,\dots,K$, for every $\theta\in\Theta_k$, $H(\theta)\le h_1$ yields

$$V(\theta)\le C_1 H(\theta)\log^2 H(\theta).$$

O-prior. There exist $C_2>0$ and dimensional indices $\widetilde{D}(k)>D(k^\star)$ for all $k=k^\star+1,\dots,K$ such that, for every sequence $\{\delta_n\}$ that decreases to zero, for all $n\ge 1$,

$$\pi_k\Big\{\theta\in\Theta_k : H(\theta)\le\delta_n\Big\}\le C_2\,{\delta_n}^{\widetilde{D}(k)/2}.$$

O-approximation. There exists $C_3>0$ such that, for each $k=k^\star+1,\dots,K$, there exists a sequence of approximating sets $\mathcal{F}_n^k\subset\Theta_k$ such that, for all $n\ge 1$,

$$\pi_k\left\{(\mathcal{F}_n^k)^c\right\}\le C_3\,n^{-\widetilde{D}(k)/2}.$$

O-local brackets. For all $k = k^\star + 1, \dots, K$, for all $\theta \in \Theta_k$, there exists $\eta_\theta > 0$ such that

$$V(u_{\theta,\eta_\theta}, \theta^\star) + V(\theta^\star, l_{\theta,\eta_\theta}) + V(\theta^\star, u_{\theta,\eta_\theta}) + V(u_{\theta,\eta_\theta}, \theta) < \infty.$$

O-Laplace. There exist $\beta_1, \beta_2 \geq 0$ and $L > 0$ such that, for all $n \geq 1$,

$$P^{\star n}\left\{\mathbb{B}_n(k^\star) < \left(\beta_1 (\log n)^{\beta_2} n^{D(k^\star)/2}\right)^{-1}\right\} \leq L \frac{(\log n)^{3\widetilde{D}(k^\star+1)/2+\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}.$$

THEOREM 3. *Let us suppose that assumptions* **O-comparison**, **O-prior**, **O-approximation** *and* **O-local brackets** *are satisfied for* $K = k^\star + 1$. *Let us also assume that* **O-Laplace** *is valid.*

Let n_0 *be the smallest integer* n *such that*

$$\delta_0 = 4 \max_{m \geq n} \left\{ m^{-1} \log \left[\beta_1 (\log m)^{\beta_2} m^{D(k^\star)/2}\right]\right\} \leq \frac{h_1 \wedge e^{-2}}{2}.$$

Let $\delta_n = \delta_1 n^{-1} \log^3 n$ *for each* $n \geq 2$, *with*

$$\delta_1 \geq 512(C_1 + 2)[\widetilde{D}(k^\star + 1) - D(k^\star)] \vee 128 C_1 \widetilde{D}(k^\star + 1) \vee \log^{-3} n_0. \tag{5}$$

If in addition, for all integers $n \geq n_0$ *such that* $\delta_n < \delta_0$ *and for every* $j \leq \lfloor \delta_0/\delta_n \rfloor$,

$$\mathcal{E}\left(\mathcal{F}_n^{k^\star+1} \cap \left[S_{k^\star+1}\Big(2(j+1)\delta_n\Big) \setminus S_{k^\star+1}\Big(2j\delta_n\Big)\right], \frac{j\delta_n}{4}\right) \leq \frac{nj\delta_n}{256(C_1+2)\log^2(j\delta_n)}, \tag{6}$$

then there exists $c_3 > 0$ *(which depends on* $C_2, C_3, \pi(k^\star + 1), \widetilde{D}(k^\star + 1)$ *and* δ_1*) such that, for all* $n \geq n_0$,

$$P^{\star n}\left\{\widehat{k}_n^{\mathrm{L}} > k^\star\right\} \leq c_3 \frac{(\log n)^{3\widetilde{D}(k^\star+1)/2+\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}. \tag{7}$$

Here, the largest integer which is strictly smaller than $u \in \mathbb{R}$ is denoted by $\lfloor u \rfloor$. Similarly,

THEOREM 4. *Let $k_{\max}$ be a prior bound on $k^\star$. Let us suppose that assumptions* **O-comparison**, **O-prior**, **O-approximation** *and* **O-local brackets** *are satisfied for $K = k_{\max}$. Let us also assume that* **O-Laplace** *is valid.*

Let n_0 be the smallest integer n such that

$$\delta_0 = 4 \max_{m \geq n} \left\{ m^{-1} \log \left[\beta_1 (\log m)^{\beta_2} m^{D(k^\star)/2} \right] \right\} \leq \frac{h_1 \wedge e^{-2}}{2}.$$

Let us set for every $k = k^\star + 1, \ldots, k_{\max}$ and for all $n \geq 2$, $\delta_{k,n} = \delta_{k,1} n^{-1} \log^3 n$, with

$$\delta_{k,1} \geq 512(C_1 + 2)[\widetilde{D}(k) - D(k^\star)] \vee 128 C_1 \widetilde{D}(k) \vee \log^{-3} n_0.$$

If in addition, for every $k = k^\star + 1, \ldots, k_{\max}$, for all integers $n \geq n_0$ such that $\delta_{k,n} < \delta_0$ and for every $j \leq \lfloor \delta_0 / \delta_{k,n} \rfloor$,

$$\mathcal{E}\left(\mathcal{F}_n^k \cap S_k\Big(2(j+1)\delta_{k,n}\Big) \setminus \mathcal{F}_n^k \cap S_k\Big(2j\delta_{k,n}\Big), \frac{j\delta_{k,n}}{4} \right) \leq \frac{nj\delta_{k,n}}{256(C_1+2)\log^2(j\delta_{k,n})}, \tag{8}$$

then there exists $c_3' > 0$ such that, for all $n \geq n_0$,

$$P^{\star n}\left\{ \widehat{k}_n^{\mathrm{L}} > k^\star \right\} \leq c_3' \frac{(\log n)^{3 \max_k \widetilde{D}(k)/2 + \beta_2}}{n^{\min_k [\widetilde{D}(k) - D(k^\star)]/2}}. \tag{9}$$

In the formula above, index k in the maximum and minimum ranges between $k^\star + 1$ and $k_{\max}$.

Exploring the assumptions. Proofs of Theorems 3 and 4 rely on tests of $P^\star$ versus complements $\{P_\theta : \theta \in \Theta_k,\ H(\theta) \geq \varepsilon\}$ of Kullback-Leibler balls around $P^\star$ for $k > k^\star$. By doing this, we follow the paradigm recently presented by Ghosal et al. (2000) in the spirit of earlier works by Schwartz (1965) and Le Cam (1973). Assumption **O-comparison** and the entropy conditions stated in inequalities (6) and (8) are at the core of the construction of those tests. Such a comparison of $V(\theta)$ and $H(\theta)$ is proved in (Wong & Shen, 1995) under mild conditions. The entropy is known to quantify the complexity of a model. Thus, the related conditions warrant

that (a critical region of) Θ_k ($k = k^\star+1$ for $\widehat{k}_n^{\mathrm{L}}$ and each $k = k^\star+1, \ldots, k_{\max}$ for $\widehat{k}_n^{\mathrm{G}}$) is not too large.

Besides, the reader should not pay too much attention to the various multiples of 2 that appear in inequalities (5), (6) and (8). They are by-products of the repeated use of inequality $(a+b)^2 \leq 2(a^2+b^2)$ (all $a, b \in \mathbb{R}$) in the proofs, and could be replaced by smaller numbers at extra calculations cost.

Assumption **O-prior** is concerned with the decay to zero of the prior mass of shrinking Kullback-Leibler neighborhoods of $\theta^\star$. Verifying that this assumption holds in the mixture setting is a demanding task (more on this in Section 2.4). Note that dimensional indices $\widetilde{D}(k)$ ($k > k^\star$) are introduced, which might be different from the usual dimensions $D(k)$. They should be understood as *effective* dimensions of Θ_k relative to $\Theta_{k^\star}$. In models of mixtures of g_γ densities ($\gamma \in \Gamma \subset \mathbb{R}^d$) for instance, $\widetilde{D}(k^\star+1) = D(k^\star)+1$ while $D(k^\star+1) = D(k^\star)+(d+1)$.

Assumption **O-approximation** is standard in Bayesian Statistics. It can be useful to get rid of difficulties which are not relevant to the main problem, see Section 5.2 for the mixture example.

Assumption **O-local brackets** is similar to assumption **U-local brackets**.

Finally, assumption **O-Laplace** is milder than the existence of a Laplace expansion of the marginal likelihood(which holds in regular models). As explained in (Chambaz, 2003), the overestimation phenomenon is tightly bound to moderate deviations of the log-likelihood process (see Section 4.3 therein). The bound stated in assumption **O-Laplace** plays here the role that moderate deviations play in (Chambaz, 2003).

Comment. The upper bounds we get in the proofs are actually tighter than the one stated in the theorems. Each time, we actually chose the largest of several terms to make the formulas more readable. Besides, the possibility in Theorem 3 to tune the value of δ_1 makes it easier to apply the theorem to the mixture model example. Naturally, the best value is given by the right-hand side of inequality (5) and the larger δ_1, the larger c_3 and the less accurate the overestimation bound.

Furthermore, assuming the existence of a bound $k_{\max}$ on $k^\star$ when studying $\widehat{k}_n^{\mathrm{G}}$ is convenient but not mandatory. It is possible to adapt the scheme of proof with a prior bound on $k^\star$ that increases as a function of the sample size. We omit the details of the adaptation.

According to inequalities (7) and (9), both overestimation probabilities decay as a negative power of the sample size n (up to a power of a $\log n$ factor). This is in stark contrast with the exponential rate exhibited for underestimation probabilities in Theorems 1 and 2. It is however another consequence of the Stein lemma that, if $\widetilde{k}_n$ is an order estimator which ultimately underestimates it with a probability bounded away from one, then the overestimation rate is necessarily slower than exponential, see Lemma 3 in (Chambaz, 2003). We want to emphasize that the overestimation rates obtained in Theorems 3 and 4 depend on intrinsic quantities (such as dimensions $D(k)$ and $\widetilde{D}(k)$, power β_2 from assumption **O-Laplace**). On the contrary, the rates obtained in Theorems 10 and 11 of (Chambaz, 2003) depend directly on the choice of a penalty term.

2.4. *Application to mixture models.* This section is devoted to the particular case of mixture models. Because order identification in mixture models is notoriously difficult, we think that this section illustrates the generality of our results. We prove that Theorems 1 and 3 apply here with

$$\widetilde{D}(k^\star+1) = D(k^\star) + 1,$$

yielding an overestimation rate of order $O((\log n)^c/\sqrt{n})$ for some positive c.

Let Γ be a compact subset of $\mathbb{R}^d$. Let us denote by $|\cdot|_1$ and $|\cdot|_2$ the ℓ^1 and ℓ^2 norms on $\mathbb{R}^d$.

For all $\gamma \in \Gamma$, let g_γ be a density. In this section, mixtures of g_γ's are studied. Formally, $\Theta_1 = \Gamma$ and for every $k \geq 1$,

$$\Theta_k = \left\{ \theta = (\boldsymbol{p}, \boldsymbol{\gamma}) : \boldsymbol{p} = (p_1, \ldots, p_{k-1}) \in \mathbb{R}_+^{k-1},\ \sum_{j=1}^{k-1} p_j \leq 1, \boldsymbol{\gamma} \in \Gamma^k \right\}.$$

We point out that, obviously, $D(k) = k(d+1) - 1$ for each $k \geq 1$. Besides, assumptions **A1** (compactness of the parameter sets) and **A2** (continuous parameterization) are satisfied.

Now, let us state six assumptions. The first, second and third order differentiation operators are denoted by ∇, D^2 and D^3, and $|\cdot|$ stands for any norm on the space of second and third order derivatives (upon which all norms are equivalent).

M-prior Prior probabilities π_k (all $k \geq 1$) write as

$$d\pi_k(\theta) = \pi_k^{\boldsymbol{p}}(\boldsymbol{p})\, \pi_k^{\boldsymbol{\gamma}}(\boldsymbol{\gamma})\, d\boldsymbol{p} d\boldsymbol{\gamma}$$

for all $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_k$. For every $k \geq 1$, π_k is continuously differentiable over Θ_k. Here, $\pi_k^{\boldsymbol{p}}$ is probability density on the simplex. Moreover, if

$$\Delta = \left\{ \mathbf{g} = (g_1, \ldots, g_k) \in \Gamma^k : \min_{j<j'} |g_j - g_{j'}|_2 = 0 \right\},$$

then π_k is bounded away from zero on the complementary of an open neighborhood of Δ. Furthermore, *when* Γ *is one-dimensional* ($d = 1$), for all $\boldsymbol{\gamma}$ in that open neighborhood of Δ,

$$\pi_k^{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \propto \prod_{j<j'} |\gamma_j - \gamma_{j'}|_2.$$

M-local brackets For all $\gamma \in \Gamma$, let us define for all $\eta > 0$:

$$\underline{g}_{\gamma,\eta} = \inf\{g_{\gamma'} : |\gamma - \gamma'|_1 \leq \eta\} \quad \text{and} \quad \overline{g}_{\gamma,\eta} = \sup\{g_{\gamma'} : |\gamma - \gamma'|_1 \leq \eta\}.$$

There exist $\eta_1, M > 0$ such that, for every $\gamma_1, \gamma_2 \in \Gamma$, there exists $\eta_2 > 0$ such that

$$\begin{aligned} P_{\overline{g}_{\gamma_1,\eta_1} - \underline{g}_{\gamma_1,\eta_1}} (1 + \log^2 g_{\gamma_2}) &\leq M\eta_1, \\ P_{g_{\gamma_2}} (\log \overline{g}_{\gamma_1,\eta_1} - \log \underline{g}_{\gamma_1,\eta_1})^2 &\leq M\eta_1^2, \\ P_{\overline{g}_{\gamma_1,\eta_1} - \underline{g}_{\gamma_1,\eta_1}} \log^2 \overline{g}_{\gamma_1,\eta_1} &\leq M\eta_1 \log^2 \eta_1 \quad \text{and} \end{aligned}$$

$$P_{\overline{g}_{\gamma_1,\eta_1}} \log^2 \overline{g}_{\gamma_1,\eta_1} + P_{\overline{g}_{\gamma_1,\eta_1}} \log^2 g_{\gamma_2} + P_{g_{\gamma_2}} \log^2 \overline{g}_{\gamma_1,\eta_1} + P_{g_{\gamma_2}} \log^2 \underline{g}_{\gamma_1,\eta_1} \leq M.$$

M-moment For every $\gamma_1, \gamma_2 \in \Gamma$, there exists $\alpha > 0$ such that

$$\sup_{\gamma \in \Gamma} P_{\gamma_1} \left(\frac{g_{\gamma_2}}{g_\gamma} \right)^\alpha < \infty.$$

M-regularity The parameterization $\gamma \mapsto g_\gamma(z)$ is twice continuously differentiable for μ-almost every $z \in \mathcal{Z}$. Moreover, the set $\{|\nabla g_\gamma|_1, |D^2 g_\gamma| : \gamma \in \Gamma\}$ is bounded.

The parameterization $\gamma \mapsto \log g_\gamma(z)$ is three times continuously differentiable for μ-almost every $z \in \mathcal{Z}$. Besides, for all $\gamma_1, \gamma_2 \in \Gamma$, there exists $\eta > 0$ for which

$$P_{\gamma_1} |D^2 \log g_{\gamma_2}|^2 < \infty,$$
$$P_{\gamma_1} \sup_{|\gamma - \gamma_2|_1 \leq \eta} |D^3 \log g_\gamma|^2 < \infty.$$

M-Fisher For every $\gamma \in \Gamma$, the Fisher information matrix $I(\gamma)$ is positive definite.

M-linear independence Let $\mathcal{I} = \{(r, s) : 1 \leq r \leq s \leq d\}$. There exist a nonempty subset $\mathcal{A}$ of $\mathcal{I}$ and two constants $\eta_0, a > 0$ such that, for every $k \geq 2$, for every k-tuple $(\gamma_1, \ldots, \gamma_k)$ of pairwise distinct elements of Γ,

(a) the functions

$$g_{\gamma_j}, (\nabla g_{\gamma_j})_l \quad (\text{for } j = 1, \ldots, k \text{ and } l = 1, \ldots, d)$$

are linearly independent;

(b) for every $j = 1, \ldots, k$, the functions

$$g_{\gamma_j}, (\nabla g_{\gamma_j})_l, (D^2 g_{\gamma_j})_{rs} \quad (\text{for } l = 1, \ldots, d \text{ and all } (r, s) \in \mathcal{A})$$

are linearly independent;

(c) for every $j = 1, \ldots, k$ and each $(r, s) \in \mathcal{I} \setminus \mathcal{A}$, there exist real numbers $\lambda_{rs}^{0j}, \ldots, \lambda_{rs}^{dj}$ such that

$$(D^2 g_{\gamma_j})_{rs} = \lambda_{rs}^{0j} g_{\gamma_j} + \sum_{l=1}^{d} \lambda_{rs}^{lj} (\nabla g_{\gamma_j})_l;$$

(d) for all $\eta \leq \eta_0$ and all $u, v \in \mathbb{R}^d$, for every $j = 1, \ldots, k$, if

$$\sum_{(r,s)\in\mathcal{A}} (|u_r u_s| + |v_r v_s|) + \left| \sum_{(r,s)\notin\mathcal{A}} \lambda_{rs}^{0j} (u_r u_s + v_r v_s) \right| \leq \eta,$$

then $|u|_2^2 + |v|_2^2 \leq a\eta$.

These assumptions suffice to guarantee that the conclusions of the theorem below are valid.

THEOREM 5. *Let us assume that assumptions* **M-prior**, **M-local brackets**, **M-moment**, **M-regularity**, **M-Fisher** *and* **M-linear independence** *hold. Then for all* $n \geq n_0$,

$$P^\star \left\{ \widehat{k}_n^{\mathrm{L}} < k^\star \right\} \leq c_1 e^{-nc_2}, \tag{10}$$

$$P^\star \left\{ \widehat{k}_n^{\mathrm{L}} > k^\star \right\} \leq c_3 \frac{(\log n)^{[3(d+1)k^\star/2]}}{\sqrt{n}}. \tag{11}$$

The positive constants c_1, c_2 *are defined in Theorem 1. Constant* c_3 *corresponds to a convenient choice of* δ_1 *(both are defined in Theorem 3).*

As an illustration, the previous theorem applies to the case of location-scale mixtures of Gaussian distributions.

COROLLARY 1. *Set* $A, B > 0$ *and* $\Gamma = \left\{ (\mu, \sigma^2) \in [-A, A] \times [\frac{1}{B}, B] \right\}$. *For every* $\gamma = (\mu, \sigma^2) \in \Gamma$, *let us denote by* g_γ *the Gaussian density with mean* μ *and variance* σ^2.

Inequalities (10) and (11) with $d = 2$ *hold for all* $n \geq n_0$ *(as defined in Theorem 3).*

Comments. We emphasize that all assumptions involve the mixed densities g_γ ($\gamma \in \Gamma$) rather than the resulting mixture densities f_θ ($\theta \in \Theta_\infty$). In assumption **M-prior**, the required form of prior distributions π_k is specified for all $k \geq 1$. Assumption **M-local brackets** states conditions for **U-local brackets** and **O-local brackets** to hold. Assumption **M-moment** is also a condition that guarantees the validity of assumption **U-moment**. Assumptions **M-regularity** and **M-Fisher** are common.

On the contrary, assumption **M-linear independence** is more original. It draws its inspiration from a similar assumption that appears in (Dacunha-Castelle & Gassiat, 1997b; Dacunha-Castelle & Gassiat, 1999), where the locally conic parameterization is introduced. Indeed, proving that assumption **O-prior** holds in the mixture setting relies on a variant of this method (see Proposition 1 in Section 5.2 and its proof in Section B). It is quite involved and we think our method may be of general interest.

Assumption **M-linear independence** is less stringent than its parent in (Dacunha-Castelle & Gassiat, 1997b; Dacunha-Castelle & Gassiat, 1999). This is due to the introduction of the subset of indices $\mathcal{A}$, which implicitly equals $\mathcal{I}$ in (Dacunha-Castelle & Gassiat, 1997b; Dacunha-Castelle & Gassiat, 1999). Property (d) guarantees that *enough* indices $(r,s) \in \mathcal{I}$ (namely, those from $\mathcal{A}$) are involved in property (b), when property (c) deals with the remaining indices.

When $\mathcal{A}$ is set to $\mathcal{I}$, then assumption **M-linear independence** is not verified in the setting of Corollary 1. Luckily, it is for $\mathcal{A} = \mathcal{I} \setminus \{(1,1)\}$. Indeed, Properties (a) and (b) are obviously satisfied. Besides, for any $\gamma \in \Gamma$, $(D^2 g_\gamma)_{11}$ is a linear combination of g_γ and $(\nabla g_\gamma)_2$, with coefficients bounded away from zero independently of γ. Therefore, Properties (c) and (d) also hold.

3. Underestimation proofs. Theorem 1 relies on the following lower bound for $\mathbb{B}_n(k)$.

LEMMA 1. *Let us assume that assumptions* **U-prior** *and* **U-moment** *of Theorem 1 are satisfied. Let us set $k \leq k^\star$ and some positive $\delta \leq \alpha M \wedge \delta_0$.*

With probability at least $1 - 2\exp\{-n\delta^2/8M\}$,

$$\mathbb{B}_n(k) \geq \frac{\pi(k)\pi_k\{S_k(\delta)\}}{2}\, e^{-n[H_k^\star + \delta]}.$$

The proof of this lemma is based on a comparison of $\ell_n(\theta) - \ell_n^\star$ for $\theta \in S_k(\delta)$ to its expectation $\mathrm{E}^{\star n}[\ell_n(\theta) - \ell_n^\star] = -nH(\theta) \geq -nH_k^\star - n\delta/2$.

PROOF. Let us set $1 \leq k \leq k^\star$, $0 < \delta \leq \alpha M \wedge \delta_0$ and define

$$B = \{(\theta, Z^n) \in \Theta_k \times \mathcal{Z}^n : \ell_n(\theta) - \ell_n^\star \geq -n[H_k^\star + \delta]\}.$$

Then

$$\begin{aligned}\mathbb{B}_n(k) &\geq \pi(k)\int_{S_k(\delta)} e^{\ell_n(\theta)-\ell_n^\star} d\pi_k(\theta)\\ &\geq \pi(k)\int_{S_k(\delta)\cap B} e^{-n[H_k^\star+\delta]} d\pi_k(\theta)\\ &\geq \pi(k)e^{-n[H_k^\star+\delta]}\Big(\pi_k\{S_k(\delta)\}-\pi_k\{S_k(\delta)\cap B^c\}\Big).\end{aligned}$$

The goal is therefore to provide a convenient lower bound for the factor between parentheses above.

The Markov inequality and Fubini theorem yield

$$P^{\star n}\left\{\pi_k\{S_k(\delta)\cap B^c\}\geq\frac{\pi_k\{S_k(\delta)\}}{2}\right\}\leq\int_{S_k(\delta)}\frac{2P^{\star n}\{B^c\}}{\pi_k\{S_k(\delta)\}}d\pi_k(\theta). \tag{12}$$

Set $s\in[0,\alpha]$ and $\theta\in S_k(\delta)$ and let $\varphi_\theta(t)=P^\star e^{t(\ell^\star-\ell_\theta)}$ (every $t\in\mathbb{R}$). By virtue of **U-moment**, function φ_θ is $\mathcal{C}^\infty$ over $[0,\alpha]$ and φ_θ'' is bounded by $q(\theta,\alpha)\leq M$ on that interval. Moreover, a Taylor expansion implies that

$$\varphi_\theta(s)=1+sH(\theta)+s^2\int_0^1(1-t)\varphi_\theta''(st)dt\leq 1+sH(\theta)+\frac{1}{2}\,s^2M.$$

Applying the Chernoff bounding method and inequality $\log t\leq t-1$ $(t>0)$ implies that

$$\begin{aligned}\mathrm{E}^\star\mathbb{1}\{\theta\in S_k(\delta)\cap B^c\} &= P^{\star n}\{\ell_n^\star-\ell_n(\theta)>n[H_k^\star+\delta]\}\\ &\leq \exp\{-ns[H_k^\star+\delta]+n\log\varphi_\theta(s)\}\\ &\leq \exp\Big\{-ns[H_k^\star+\delta-H(\theta)]+ns^2M/2\Big\}.\end{aligned}$$

Let us finally choose $s=[H_k^\star+\delta-H(\theta)]/M$, then emphasize that:

- since $\theta\in S_k(\delta)$, $s\geq\delta/2\geq 0$;
- since $H_k^\star\leq H(\theta)$ and $\delta\leq\alpha M$, $s\leq\alpha$;
- since $\theta\in S_k(\delta)$, this value of s yields a bound which is less than $\exp\{-n\delta^2/8M\}$.

Thus, the last statement above and (12) imply the lemma. □

The proof of Theorem 1 is now at hand. It will particularly rely on nets of upper bounds for the f_θ ($\theta\in\Theta_k$, $k=1,\ldots,k^\star-1$) whose construction is detailed below. Similar nets have been first introduced in a context of nonparametric Bayesian estimation in (Barron et al., 1999).

PROOF OF THEOREM 1. Since $P^{\star n}\{\widehat{k}_n^{\mathrm{L}} < k^\star\} = \sum_{k=1}^{k^\star} P^{\star n}\{\widehat{k}_n^{\mathrm{L}} = k\}$, it is sufficient to study $P^{\star n}\{\widehat{k}_n^{\mathrm{L}} = k\}$ for k arbitrarily chosen between 1 and $k^\star - 1$.

Let $\delta < \alpha M \wedge \delta_0$ be a small positive number and $c = \frac{1}{2}\pi(k)\pi_k\{S_k(\delta)\} \in (0, 1]$. Let us define for convenience $\varepsilon = 2\delta/[H_k^\star - H_{k+1}^\star] > 0$. Lemma 1 yields

$$\begin{aligned} P^{\star n}\left\{\widehat{k}_n^{\mathrm{L}} = k\right\} &\leq P^{\star n}\left\{\mathbb{B}_n(k) \geq \mathbb{B}_n(k+1)\right\} \\ &\leq 2e^{-n\frac{\delta^2}{8M}} + P^{\star n}\left\{\mathbb{B}_n(k) \geq ce^{-n[H_{k+1}^\star + \delta]}\right\}. \end{aligned} \tag{13}$$

Let us construct a net of upper bounds for the f_θ ($\theta \in \Theta_k$) in order to control the rightmost term of (13). Let $\theta, \theta' \in \Theta_k$.

Because $V(\theta^\star, u_{\theta,\eta_\theta}) + V(\theta^\star, l_{\theta,\eta_\theta})$ is finite, the dominated convergence theorem ensures that, first $H(\theta')$ tends to $H(\theta)$ when θ' goes to θ, second $H(u_{\theta,\eta})$ tends to $H(\theta)$ when η goes to zero. Therefore, for some $\eta'_\theta < \eta_\theta$, for all $\eta < \eta'_\theta$, $d(\theta, \theta') < \eta$ yields

$$H(u_{\theta,\eta}) \leq H(\theta') \leq H(u_{\theta,\eta}) + \delta. \tag{14}$$

Similarly, because $V(\theta^\star, u_{\theta,\eta_\theta}) + V(\theta^\star, l_{\theta,\eta_\theta})$ is finite, first $V(\theta^\star, \theta')$ tends to $V(\theta^\star, \theta) > 0$ when θ' goes to θ and second $V(\theta^\star, u_{\theta,\eta})$ tends to $V(\theta^\star, \theta)$ when η goes to zero. Therefore, for some $\eta''_\theta < \eta'_\theta$, for all $\eta < \eta''_\theta$, $d(\theta, \theta') < \eta$ yields

$$V(\theta^\star, u_{\theta,\eta}) \leq (1 + \varepsilon)V(\theta^\star, \theta'). \tag{15}$$

Finally, by using the same arguments as before, there exists $\eta^o_\theta < \eta''_\theta$, such that for all $\eta < \eta^o_\theta$, $d(\theta, \theta') < \eta$ yields

$$V(u_{\theta,\eta}, \theta^\star) \leq (1 + \varepsilon)V(\theta', \theta^\star). \tag{16}$$

In summary, for every $\theta \in \Theta_k$, there exists $\eta^o_\theta > 0$ such that (14), (15), (16) hold for $\eta = \eta^o_\theta$ as soon as $d(\theta, \theta') < \eta^o_\theta$. Let us define $\mathcal{B}(\theta, \eta^o_\theta) = \{\theta' \in \Theta_k : d(\theta, \theta') < \eta^o_\theta\}$ for all $\theta \in \Theta_k$. The collection of open sets $\{\mathcal{B}(\theta, \eta^o_\theta)\}_{\theta \in \Theta_k}$ covers Θ_k, which is a compact set (by virtue of **A1**). So, there exist $\theta_1, \ldots, \theta_{N_\varepsilon} \in \Theta_k$

such that $\Theta_k = \cup_{j=1}^{N_\varepsilon} \mathcal{B}(\theta_j, \eta^o_{\theta_j})$. For $j = 1, \ldots, N_\varepsilon$, let $u_j = u_{\theta_j, \eta^o_{\theta_j}}$,

$$\widetilde{T}_{kj} = \Big\{\theta \in \Theta_k : \ell_\theta \leq \log u_j,\, H(\theta) \leq H(u_j) + \delta, \\ V(\theta^\star, u_j) \leq (1+\varepsilon) V(\theta^\star, \theta), \\ V(u_j, \theta^\star) \leq (1+\varepsilon) V(\theta, \theta^\star)\Big\}.$$

then $T_{k1} = \widetilde{T}_{k1}$ and $T_{kj} = \widetilde{T}_{kj} \cap (\cup_{j'<j} \widetilde{T}_{kj'})^c$ $(j = 2, \ldots, N_\varepsilon)$. The family $\{T_{k1}, \ldots, T_{kN_\varepsilon}\}$ is a partition of Θ_k.

Accordingly, with $\ell_{n,u_j} = \sum_{i=1}^n \log u_j(Z_i)$ $(j = 1, \ldots, N_\varepsilon)$, the rightmost term of (13) satisfies:

$$\begin{aligned}
&P^{\star n}\left\{\mathbb{B}_n(k) \geq c e^{-n[H^\star_{k+1}+\delta]}\right\} \\
&\quad = P^{\star n}\left\{\sum_{j=1}^{N_\varepsilon} \int_{T_{kj}} e^{\ell_n(\theta) - \ell^\star_n} d\pi_k(\theta) \geq c\, e^{-n[H^\star_{k+1}+\delta]}\right\} \\
&\quad \leq \sum_{j=1}^{N_\varepsilon} P^{\star n}\left\{e^{\ell_{n,u_j} - \ell^\star_n} \int_{T_{kj}} e^{\ell_n(\theta) - \ell_{n,u_j}} d\pi_k(\theta) \geq c\, e^{-n[H^\star_{k+1}+\delta]} \pi_{k+1}\{T_{kj}\}\right\} \\
&\quad \leq \sum_{j=1}^{N_\varepsilon} P^{\star n}\Big\{\ell_{n,u_j} - \ell^\star_n \geq -n[H^\star_{k+1} + \delta] + \log c\Big\} \\
&\quad \leq \sum_{j=1}^{N_\varepsilon} P^{\star n}\Big\{\ell_{n,u_j} - \ell^\star_n + nH(u_j) \geq n\rho_j + \log c\Big\}
\end{aligned}$$

for $\rho_j = [H(u_j) - H^\star_{k+1} - \delta]$. Let us point out that $\rho_j \geq (1-\varepsilon)[H(\theta_j) - H^\star_{k+1}] > 0$ for $j = 1, \ldots, N_\varepsilon$ by construction. Applying (30) of Proposition 3 (whose assumptions are satisfied) finally implies that

$$P^{\star n}\left\{\mathbb{B}_n(k) \geq c e^{-n[H^\star_{k+1}+\delta]}\right\} \leq \\ \frac{N_\varepsilon}{c} \exp\left\{-n \frac{(1-\varepsilon)^2}{4(1+\varepsilon)} [H^\star_k - H^\star_{k+1}] \min\left(\inf_{\theta \in \Theta_k} \frac{H(\theta) - H^\star_{k+1}}{V(\theta)}, \frac{2(1+\varepsilon)}{1-\varepsilon}\right)\right\}.$$

This bound and (13) conclude the proof, since N_ε does not depend on n. □

Remark 1. *The proof of Theorem 2 is a straightforward adaptation of the preceding one. It relies on the bound*

$$P^{\star n}\left\{\widehat{k}^{\mathrm{G}}_n < k^\star\right\} = \sum_{k=1}^{k^\star} P^{\star n}\left\{\widehat{k}^{\mathrm{G}}_n = k\right\} \leq \sum_{k=1}^{k^\star} P^{\star n}\left\{\mathbb{B}_n(k) \geq \mathbb{B}_n(k^\star)\right\}.$$

Let us emphasize that the quantity $H^\star_{k+1}$ which arise when dealing with $P^{\star n}\{\mathbb{B}_n(k) \geq \mathbb{B}_n(k^\star)\}$ is replaced by $H^\star_{k^\star} = 0$ for all $k = 1, \ldots, k^\star - 1$.

It is worth pointing out that, as far as underestimation is concerned, the posterior mode $\widehat{k}^{\mathrm{G}}_n$ does not require that the models be nested, whereas $\widehat{k}^{\mathrm{L}}_n$ needs that property.

4. Overestimation proofs.

Proof of Theorem 3. Set

$$\delta_0 = 4 \max_{n \geq n_0} \{n^{-1} \log \left[\beta_1 (\log n)^{\beta_2} n^{D(k^\star+1)/2}\right]\},$$

where n_0 is chosen so that $2\delta_0$ be smaller than the constant h_1 of assumption **O-comparison** and e^{-2} (note that function $u \mapsto u \log^2 u$ increases on interval $(0, e^{-2})$). Obviously, for every $n \geq n_0$,

$$-\log \left(\beta_1 (\log n)^{\beta_2} n^{D(k^\star)/2}\right) \geq -n \frac{\delta_0}{4}.$$

By definition of $\widehat{k}^{\mathrm{L}}_n$,

$$\begin{aligned} P^{\star n}\left\{\widehat{k}^{\mathrm{L}}_n > k^\star\right\} &\leq P^{\star n}\{\mathbb{B}_n(k^\star) < \mathbb{B}_n(k^\star + 1)\} \leq \\ &P^{\star n}\left\{\mathbb{B}_n(k^\star) \leq \left(\beta_1 (\log n)^{\beta_2} n^{D(k^\star)/2}\right)^{-1}\right\} + \\ &\qquad P^{\star n}\left\{\mathbb{B}_n(k^\star + 1) \geq \left(\beta_1 (\log n)^{\beta_2} n^{D(k^\star)/2}\right)^{-1}\right\}. \end{aligned} \tag{17}$$

Thus, the left-hand side term is bounded as described in **O-Laplace**. Let us focus on the right-hand side term.

To this end, the parameter set $\Theta_{k^\star+1}$ is decomposed into the union of the following three disjoint sets: let us set δ_1 satisfying inequality (5) of Theorem 3 and $\delta_n = \delta_1 n^{-1} \log^3 n$,

$$\begin{aligned} S_{k^\star+1}(2\delta_0)^c &= \{\theta \in \Theta_{k^\star+1} : H(\theta) > \delta_0\}, \\ S_n = S_{k^\star+1}(2\delta_0) \cap S_{k^\star+1}(2\delta_n)^c &= \{\theta \in \Theta_{k^\star+1} : \delta_n < H(\theta) \leq \delta_0\}, \\ S_{k^\star+1}(2\delta_n) &= \{\theta \in \Theta_{k^\star+1} : H(\theta) \leq \delta_n\}. \end{aligned}$$

Note that S_n can be empty. According to this decomposition, the quantity of interest is bounded by the sum of three terms (of which the second is zero when S_n is empty): if we define $w_n = 3\pi(k^\star+1)\beta_1(\log n)^{\beta_2} n^{D(k^\star)/2}$, then

$$
\begin{aligned}
& P^{\star n}\left\{\mathbb{B}_n(k^\star+1) \geq \left(\beta_1(\log n)^{\beta_2} n^{D(k^\star)/2}\right)^{-1}\right\} \leq \\
& \quad P^{\star n}\left\{\int_{S_{k^\star+1}(2\delta_0)^c} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/w_n\right\} + \\
& \quad P^{\star n}\left\{\int_{S_n} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/w_n\right\} + \\
& \quad P^{\star n}\left\{\int_{S_{k^\star+1}(2\delta_n)} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/w_n\right\}. \qquad (18)
\end{aligned}
$$

The Markov inequality, Fubini theorem and **O-prior** yield the following bound for the third term of (18) denoted by $p_{n,3}$ (a similar argument appeared in the proof of Lemma 1):

$$
\begin{aligned}
p_{n,3} &\leq w_n \, \pi_{k^\star+1}\{S_{k^\star+1}(2\delta_n)\} \\
&\leq C_2 w_n \, {\delta_n}^{\widetilde{D}(k^\star+1)/2} \\
&\leq 3\beta_1 C_2 \pi(k^\star+1) {\delta_1}^{\widetilde{D}(k^\star+1)/2} \, \frac{(\log n)^{3\widetilde{D}(k^\star+1)/2+\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}. \qquad (19)
\end{aligned}
$$

The first term of (18), denoted by $p_{n,1}$, is alike the quantity $P^{\star n}\{\mathbb{B}_n(k) \geq ce^{-n[H^\star_{k+1}+\delta]}\}$ that has already been bounded in the proof of Theorem 1. Indeed, the infima for $\theta \in S_{k^\star+1}(2\delta_0)^c$ of $H(\theta)$, $V(\theta^\star,\theta)$ and $V(\theta,\theta^\star)$ are positive (obviously, if $\inf V(\theta^\star,\theta)$ or $\inf V(\theta,\theta^\star)$ were not, then $\theta^\star$ would belong to $S_{k^\star+1}(2\delta_0)^c$). It is readily seen that the scheme of proof of Theorem 1 also applies here. Thus, there exist a finite number N_{δ_0} of functions $u_1,\ldots,u_{N_{\delta_0}}$ and $c_1,\ldots,c_{N_{\delta_0}} \in (0,1]$ such that $\inf_j H(u_j) \geq \delta_0/2$ and (with notations $\ell_{n,u_j} = \sum_{i=1}^n \log u_j(Z_i)$)

$$
\begin{aligned}
p_{n,1} &\leq \sum_{j=1}^{N_{\delta_0}} P^{\star n}\Big\{\ell_{n,u_j} - \ell_n^\star + nH(u_j) \geq nH(u_j) - \\
&\qquad\qquad \log\left(\beta_1(\log n)^{\beta_2} n^{-D(k^\star)/2}\right) + \log c_j\Big\} \\
&\leq \sum_{j=1}^{N_{\delta_0}} P^{\star n}\left\{\ell_{n,u_j} - \ell_n^\star + nH(u_j) \geq n\frac{\delta_0}{4} + \log c_j\right\}.
\end{aligned}
$$

By virtue of (30) in Proposition 3 (whose assumptions are satisfied) and some simple calculations, there finally exist $c_4, c_5 > 0$ which do not depend on n and guarantee that

$$p_{n,1} \leq c_4 e^{-nc_5}. \tag{20}$$

Bounding the second term of (18), which is denoted by $p_{n,2}$, is the most demanding step of this proof. Of course, this study is required only when $\delta_n < \delta_0$.

Let $\Delta_n = \lfloor \delta_0/\delta_n \rfloor$ be the largest integer strictly smaller than δ_0/δ_n. For all $j = 1, \ldots, \Delta_n$, let us introduce $S_{n,j} = \{\theta \in \mathcal{F}_n \cap S_n : j\delta_n < H(\theta) \leq (j+1)\delta_n\}$. Let us choose $[l_i, u_i] \in \mathcal{H}(S_{n,j}, j\delta_n/4)$, define $\overline{u}_i = u_i/\mu u_i$ and introduce the local tests

$$\phi_{i,j} = \mathbb{1}\left\{\ell_{n,\overline{u}_i} - \ell_n^\star + nH(\overline{u}_i) \geq n\frac{j\delta_n}{2}\right\}.$$

In the perspective of Proposition 3, $\phi_{i,j} = \phi_{n,f,\rho,c}$ for $f = \overline{u}_i$, $\rho = j\delta_n/2$ and $c = 1$.

- Set $\theta \in S_{n,j}$ such that $f_\theta \in [l_i, u_i]$, $g = f_\theta$ and $\rho' = \log \mu u_i$. Then $\mu g = 1$, $V(g) = V(\theta) > 0$ and

  $$H(\overline{u}_i) - (\rho + \rho') = P^\star(\ell^\star - \log \overline{u}_i) - \log \mu u_i - \rho = P^\star(\ell^\star - \log u_i) - \rho =$$
  $$H(\theta) + P^\star(\ell_\theta - \log u_i) - \rho \geq H(\theta) - P^\star(\log u_i - \log l_i) - \rho \geq \frac{j\delta_n}{4} > 0.$$

  Thus, according to (31) of Proposition 3,

  $$\mathrm{E}_\theta^n(1 - \phi_{i,j}) \leq \exp\left\{-\frac{n[H(\overline{u}_i) - (\rho + \rho')]}{2}\left(\frac{H(\overline{u}_i) - (\rho + \rho')}{V(\theta)} \wedge 1\right)\right\}.$$

  Besides, since

  $$H(\theta) \leq (j+1)\delta_n \leq \delta_0 + \delta_n \leq 2\delta_0 \leq h_1 \vee e^{-2},$$

  then $\log^2 \delta_n \geq \log^2(j\delta_n) \geq \log^2((j+1)\delta_n)$ and assumption **O-comparison** ensures that

  $$\begin{aligned} V(\theta) &\leq C_1 H(\theta) \log^2 H(\theta) \\ &\leq C_1 (j+1)\delta_n \log^2((j+1)\delta_n) \\ &\leq C_1 (j+1)\delta_n \log^2(j\delta_n). \end{aligned}$$

Consequently, using $8C_1 \log^2(j\delta_n) \geq 1$ (this justifies the condition $C_1 \geq 1/32$ in **O-comparison**) and $j/(j+1) \geq 1/2$ implies that the following bound holds:

$$\mathrm{E}_\theta^n(1-\phi_{i,j}) \leq \exp\left\{-\frac{nj\delta_n}{64C_1\log^2(j\delta_n)}\right\}. \tag{21}$$

- According to (30) of Proposition 3, (the assumptions of the proposition are satisfied),

$$\mathrm{E}^{\star n}\phi_{i,j} \leq \exp\left\{-\frac{nj\delta_n}{4}\left(\frac{j\delta_n}{4V(\overline{u}_i)} \wedge 1\right)\right\}.$$

The point is now to bound $V(\overline{u}_i) = V(\theta^\star, \overline{u}_i) \vee V(\overline{u}_i, \theta^\star)$. Let again $\theta \in S_{n,j}$ be such that $f_\theta \in [l_i, u_i]$. First, using repeatedly $(a+b)^2 \leq 2(a^2+b^2)$ $(a, b \in \mathbb{R})$, the definition of a δ-bracket and assumption **O-comparison** yields

$$\begin{aligned}
V(\theta^\star, \overline{u}_i) &= P^\star(\log f^\star - \log u_i + \log \mu u_i)^2 \\
&\leq 2P^\star(\log f^\star - \log u_i)^2 + 2\log^2 \mu u_i \\
&\leq 4P^\star(\log f^\star - \log f_\theta)^2 + 4P^\star(\log f_\theta - \log u_i)^2 + \\
&\qquad 2(\mu(u_i - l_i))^2 \\
&\leq 4V(\theta) + 4P^\star(\log u_i - \log l_i)^2 + 2(\mu(u_i - l_i))^2 \\
&\leq 2(2C_1+3)\delta_n \log^2((j+1)\delta_n). \qquad (22)
\end{aligned}$$

Second, similar arguments imply that

$$\begin{aligned}
V(\overline{u}_i, \theta^\star) &= P_{\overline{u}_i}(\log u_i - \log f^\star - \log \mu u_i)^2 \\
&\leq 2V(u_i, \theta^\star) + 2\log^2 \mu u_i \\
&\leq 2P_{l_i}(\log u_i - \log f^\star)^2 + 2P_{u_i - l_i}(\log u_i - \log f^\star)^2 + \\
&\qquad 2(\mu(u_i - l_i))^2
\end{aligned}$$

and the first term in the line above is bounded by

$$\begin{aligned}
4P_{l_i}(\log u_i - \log f_\theta)^2 + 4P_{l_i}(\log f_\theta - \log f^\star)^2 \leq \\
4P_{l_i}(\log u_i - \log l_i)^2 + 4P_\theta(\log f_\theta - \log f^\star)^2 \leq \\
4V(\theta) + 4P_{l_i}(\log u_i - \log l_i)^2.
\end{aligned}$$

Thus, combining the two previous bounds finally yields

$$\begin{aligned} V(\overline{u}_i, \theta^\star) &\leq 4(C_1+2)(j+1)\delta_n \log^2((j+1)\delta_n) \\ &\leq 4(C_1+2)(j+1)\delta_n \log^2(j\delta_n). \end{aligned} \tag{23}$$

A bound for $V(\overline{u}_i)$ is derived from (22) and (23), which yields in turn

$$\mathrm{E}^{\star n}\phi_{i,j} \leq \exp\left\{-\frac{nj\delta_n}{128(C_1+2)\log^2(j\delta_n)}\right\}. \tag{24}$$

Now, let us define the global test

$$\phi_n = \max\{\phi_{i,j} : i \leq \exp\{\mathcal{E}(S_{n,j}, j\delta_n/4)\},\ j \leq \Delta_n\}.$$

- Of course, for every $j \leq \Delta_n$ and $\theta \in S_{n,j}$, $\mathrm{E}_\theta^n(1-\phi_n) \leq \mathrm{E}_\theta^n(1-\phi_{1,j})$, hence Inequality (21) implies that

$$\mathrm{E}_\theta^n(1-\phi_n) \leq \exp\left\{-\frac{nj\delta_n}{64C_1\log^2(j\delta_n)}\right\}. \tag{25}$$

- Furthermore, bounding ϕ_n by the sum of all $\phi_{i,j}$, invoking (24) and assumption (6) in Theorem 3 yield

$$\begin{aligned} \mathrm{E}^{\star n}\phi_n &\leq \sum_{j=1}^{\Delta_n} \exp\left\{\mathcal{E}(S_{n,j}, j\delta_n/4) - \frac{nj\delta_n}{128(C_1+2)\log^2(j\delta_n)}\right\} \\ &\leq \sum_{j=1}^{\Delta_n} \exp\left\{-\frac{nj\delta_n}{256(C_1+2)\log^2(j\delta_n)}\right\} \\ &\leq \sum_{j=1}^{\Delta_n} \exp\left\{-\frac{nj\delta_n}{256(C_1+2)\log^2\delta_n}\right\}. \end{aligned} \tag{26}$$

Let us set $\rho_n = n\delta_n/256(C_1+2)\log^2\delta_n$. Inequality (26) writes as

$$\mathrm{E}^{\star n}\phi_n \leq \sum_{j=1}^{\Delta_n} \exp\{-j\rho_n\} \leq \frac{\exp\{-\rho_n\}}{1-\exp\{-\rho_n\}}.$$

Now, the choice of $\delta_1 \geq 512(C_1+2)[\widetilde{D}(k^\star+1) - D(k^\star)] \vee \log^{-3}(n_0)$ yields $\log^2\delta_n \leq 4\log^2 n$, hence $\rho_n \geq \frac{1}{2}[\widetilde{D}(k^\star+1) - D(k^\star)]\log n$. The final bound for $\mathrm{E}^{\star n}\phi_n$ is thus

$$\mathrm{E}^{\star n}\phi_n \leq \frac{1}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}-1}. \tag{27}$$

Bounding term $p_{n,2}$ is now at hand. By virtue of a simple decomposition,

$$\begin{aligned} p_{n,2} &= \mathrm{E}^{\star n}\Big(\phi_n + (1-\phi_n)\Big)\mathbb{1}\left\{\int_{S_n} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/w_n\right\} \\ &\leq \mathrm{E}^{\star n}\phi_n + P^{\star n}\left\{\int_{S_n\cap\mathcal{F}_n^c} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/2w_n\right\} + \\ &\qquad \mathrm{E}^{\star n}(1-\phi_n)\mathbb{1}\left\{\int_{S_n\cap\mathcal{F}_n} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/2w_n\right\}. \end{aligned}$$

The first term of the right-hand side expression above is bounded according to (27). Moreover, applying the Markov inequality and Fubini theorem to the second term ensures that

$$\begin{aligned} P^{\star n}\left\{\int_{S_n\cap\mathcal{F}_n^c} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/2w_n\right\} &\leq 2w_n\pi_{k^\star+1}\{\mathcal{F}_n^c\} \\ &\leq 6\beta_1 C_3\,\frac{(\log n)^{\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}. \end{aligned} \tag{28}$$

As for the third term, by invoking again the Markov inequality and Fubini theorem, then inequality (25), it satisfies

$$\begin{aligned} &\mathrm{E}^{\star n}(1-\phi_n)\mathbb{1}\left\{\int_{S_n\cap\mathcal{F}_n} e^{\ell_n(\theta)-\ell_n^\star} d\pi_{k^\star+1}(\theta) \geq 1/2w_n\right\} \\ &\qquad \leq 2w_n\int_{S_n\cap\mathcal{F}_n} \mathrm{E}_\theta^n(1-\phi_n)d\pi_{k^\star+1}(\theta) \\ &\qquad \leq 2w_n\sum_{j=1}^{\Delta_n}\int_{S_{n,j}} \mathrm{E}_\theta^n(1-\phi_n)d\pi_{k^\star+1}(\theta) \\ &\qquad \leq 2w_n\sum_{j=1}^{\Delta_n}\exp\left\{-\frac{nj\delta_n}{64C_1\log^2(j\delta_n)}\right\}\pi_{k^\star+1}\{S_{n,j}\} \\ &\qquad \leq 2w_n\exp\left\{-\frac{n\delta_n}{64C_1\log^2\delta_n}\right\} \\ &\qquad \leq 2w_n\exp\left\{-\frac{\delta_1}{256C_1}\log n\right\} \\ &\qquad \leq 6\,\beta_1\pi(k^\star+1)\frac{(\log n)^{\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}. \end{aligned} \tag{29}$$

Combining inequalities (27), (28) and (29) yields

$$p_{n,2} \leq \frac{1}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}-1} + 6\beta_1\Big(\pi(k^\star+1)+C_3\Big)\frac{(\log n)^{\beta_2}}{n^{[\widetilde{D}(k^\star+1)-D(k^\star)]/2}}.$$

Inequalities (19), (20) and the one above conclude the proof. □

REMARK 2. *The proof of Theorem 4 is very similar to the proof of Theorem 3. Indeed, by virtue of the union bound,*

$$\begin{aligned} P^{\star n}\left\{\widehat{k}_n^{\mathrm{G}} > k^\star\right\} &\leq \sum_{k=k^\star+1}^{k_{\max}} P^{\star n}\{\mathbb{B}_n(k) \geq \mathbb{B}_n(k^\star)\} \\ &\leq P^{\star n}\left\{\mathbb{B}_n(k^\star) \leq \left(\beta_1(\log n)^{\beta_2} n^{D(k^\star)/2}\right)\right\} + \\ &\qquad \sum_{k=k^\star+1}^{k_{\max}} P^{\star n}\left\{\mathbb{B}_n(k) \geq \left(\beta_1(\log n)^{\beta_2} n^{D(k^\star)/2}\right)\right\}, \end{aligned}$$

so the parallel with the previous proof is obvious.

5. Mixtures proofs.

5.1. *Mixtures proofs: underestimation.* This section is dedicated to the proof of Inequality (10) in Theorem 5, which asserts that the underestimation error exponent is positive. It is a consequence of Theorem 1. Let us verify that its assumptions are satisfied.

In the sequel, we shall use the notation $f^\star = f_{\theta^\star}$, $\theta^\star = (\boldsymbol{p}^\star, \boldsymbol{\gamma}^\star)$, $\boldsymbol{p}^\star = (p_1^\star, \ldots, p_{k^\star-1}^\star)$ and $p_{k^\star}^\star = 1 - \sum_{j=1}^{k^\star-1} p_j^\star$. In greater generality, if $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_k$, then $1 - \sum_{j=1}^{k-1} p_j$ will be denoted by p_k.

In the first place, let us point out a convenient property satisfied by general mixtures (see Lemma 3 in (Leroux, 1992)):

LEMMA 2. *Let F be a mixing distribution on Γ. Let P_0 have density $p_0 = \int g_\gamma dF(\gamma)$. For any $k \geq 1$, $P_0 \notin \{P_\theta : \theta \in \Theta_k\}$ yields $H_k^\star > H_{k+1}^\star$.*

Lemma 2 particularly implies that the additional condition in Theorem 1 is satisfied in the mixture models. Some useful results due to assumption **M-local brackets** now follow.

LEMMA 3. *In the setting of the mixture model, assumption* **U-local brackets** *is valid. Besides, function $\theta \mapsto H(\theta)$ is continuous on the interior* $\mathrm{int}(\Theta_k)$ *of Θ_k for all $k \geq 1$. Finally, $\theta \mapsto V(\theta^\star, \theta)$ is continuous on Θ_k for all $k \geq 1$.*

PROOF. The first part of Lemma 3 is a straightforward consequence of assumption **M-local brackets**.

Let us set $k \geq 1$ and $\theta^o \in \mathrm{int}(\Theta_k)$. Let us fix some positive $\eta_1 \leq \min_{j \leq k} p_j^o/2$. Let $\theta = (\boldsymbol{p}, \boldsymbol{\gamma})$ satisfy $|p_j - p_j^o| \leq \eta_1$ and $|\gamma_j - \gamma_j^o|_1 \leq \eta_1$ for all $j \leq k$. Since $P^\star \ell^\star$ is finite, $|H(\theta) - H(\theta^o)| = |P^\star(\ell_{\theta^o} - \ell_\theta)|$. Now, simple calculation yield

$$\begin{aligned}
|\ell_{\theta^o} - \ell_\theta| &\leq \log \left(\frac{\sum_{j=1}^k p_j^o g_{\gamma_j^o}}{\sum_{j=1}^k (p_j^o - \eta_1) \underline{g}_{\gamma_{j,\eta_1}^o}} \vee \frac{\sum_{j=1}^k (p_j^o + \eta_1) \overline{g}_{\gamma_{j,\eta_1}^o}}{\sum_{j=1}^k p_j^o g_{\gamma_j^o}} \right) \\
&\leq \max_{j \leq k} \log \left(\frac{p_j^o g_{\gamma_j^o}}{(p_j^o - \eta_1) \underline{g}_{\gamma_{j,\eta_1}^o}} \vee \frac{(p_j^o + \eta_1) \overline{g}_{\gamma_{j,\eta_1}^o}}{p_j^o g_{\gamma_j^o}} \right) \\
&\leq \log 2 + \max_{j \leq k} \log(g_{\gamma_j^o} / \underline{g}_{\gamma_j^o,\eta_1}) + \max_{j \leq k} \log(\overline{g}_{\gamma_j^o,\eta_1} / g_{\gamma_j^o}).
\end{aligned}$$

So, the second part of Lemma 3 is a consequence of the dominated convergence theorem.

Finally, let $\theta^o = (\boldsymbol{p}^o, \boldsymbol{\gamma}^o) \in \Theta_k$ and $\eta_2 > 0$. Let $\theta = (\boldsymbol{p}, \boldsymbol{\gamma})$ satisfy $|p_j - p_j^o| \leq \eta_2$ and $|\gamma_j - \gamma_j^o|_1 \leq \eta_2$ for every $j \leq k$. A crude bound for ℓ_θ^2 writes as

$$\begin{aligned}
\ell_\theta^2 &\leq \log^2 \left(k \max_{j \leq k} \overline{g}_{\gamma_j^o,\eta_2} \right) + \log^2 \left(\frac{1}{k} \min_{j \leq k} \underline{g}_{\gamma_j^o,\eta_2} \right) \\
&\leq \sum_{j=1}^k \log^2 \left(k \overline{g}_{\gamma_j^o,\eta_2} \right) + \sum_{j=1}^k \log^2 \left(\frac{1}{k} \min_{j \leq k} \underline{g}_{\gamma_j^o,\eta_2} \right)
\end{aligned}$$

By invoking assumption **M-local brackets**, η_2 can be chosen small enough so that each term in the right-hand side of the previous display (which do not depend on θ) belongs to $L^1(P^\star)$. Because $\ell^{\star 2} \in L^1(P^\star)$, the dominated convergence theorem implies that $V(\theta^\star, \theta) = P^\star(\ell^\star - \ell_\theta)^2$ tends to $V(\theta^\star, \theta^o)$ when θ tends to θ^o. This completes the proof. □

The two lemmas below complete the verification.

LEMMA 4. *Assumption* **U-prior** *is valid in the setting of the mixture model.*

PROOF. Let us set $k \leq k^\star$ and $\delta, \eta > 0$.

By virtue of Lemma 2 and the continuity of $\theta \mapsto H(\theta)$ on $\mathrm{int}(\Theta_k)$ (see Lemma 3), there exists $\theta^o \in \mathrm{int}(\Theta_k)$ such that $H(\theta^o) = H_k^\star$.

Let $\theta = (\boldsymbol{p}, \boldsymbol{\gamma})$ satisfy $|p_j - p_j^o| \leq \eta$ and $|\gamma_j - \gamma_j^o|_1 \leq \eta$ for all $j \leq k$. Since $\theta \mapsto H(\theta)$ is continuous, η can be chosen so that $H(\theta) - H(\theta^o) \leq \delta/2$. This concludes the proof, because the set of all such θ's has a positive π_k-probability. □

LEMMA 5. *In the setting of the mixture model, there exists $\alpha, M > 0$ such that, for all $\theta \in \Theta_{k^\star}$, $P^\star e^{\alpha(\ell^\star - \ell_\theta)} \leq M$. Consequently, assumption* **U-moment** *is valid.*

PROOF. By virtue of Lemma 3, the continuous function $\theta \mapsto V(\theta^\star, \theta)$ is bounded on the compact set $\Theta_{k^\star}$. So, assumption **U-moment** holds if we prove for instance the existence of $\alpha > 0$ such that function $\theta \mapsto P^\star e^{\alpha(\ell^\star - \ell_\theta)}$ is bounded on $\Theta_{k^\star}$.

Let us set $\alpha > 0$ and $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_{k^\star}$. By convexity, the following bound holds

$$\ell^\star - \ell_\theta \leq \ell^\star - \sum_{j=1}^{k^\star} p_j \log g_{\gamma_j} = \sum_{j=1}^{k^\star} p_j (\ell^\star - \log g_{\gamma_j}),$$

hence it is sufficient to show that α can be chosen so that function $\gamma \mapsto P^\star e^{\alpha(\ell^\star - \log g_\gamma)}$ is bounded on Γ. Now, let us observe that

$$\begin{aligned} P^\star e^{\alpha(\ell^\star - \log g_\gamma)} &\leq \sum_{j=1}^{k^\star} P_{g_{\gamma_j^\star}} \left(\sum_{j'=1}^{k^\star} \frac{g_{\gamma_{j'}^\star}}{g_\gamma} \right)^\alpha \\ &\leq (k^\star)^\alpha \sum_{j,j' \leq k^\star} P_{g_{\gamma_j^\star}} \left(\frac{g_{\gamma_{j'}^\star}}{g_\gamma} \right)^\alpha . \end{aligned}$$

Therefore, invoking assumption **M-moment** yields the existence of α for which the right-hand side of the inequality above is finite. Thus, the proof is complete. □

Inequality (10) of Theorem 5 is true by virtue of Theorem 1 and Lemmas 2, 3, 4 and 5.

5.2. *Mixtures proofs: overestimation.* This section is dedicated to the proof of Inequality (11) in Theorem 5, according to which the overestimation probability declines as $1/\sqrt{n}$, up to a $\log n$ factor. It is a consequence of Theorem 1. Let us verify that its assumptions are satisfied.

LEMMA 6. *Assumptions* **O-local brackets** *and* **O-comparison** *are verified in the setting of mixture models.*

The second assumption follows from Lemma 5 and Theorem 5 in (Wong & Shen, 1995). We show that assumption **O-Laplace** is valid by invoking a Laplace expansion.

LEMMA 7. *Assumption* **O-Laplace** *is valid for some* $\beta_1 > 0$ *and* $\beta_2 = 0$ *in the setting of mixture models.*

PROOF. Let us write $\widehat{\theta}$ for a maximizer of ℓ_n over $\Theta_{k^\star}$ and β_1 for a positive number. For every $u \in \mathbb{R}^{D(k^\star)}$ and $\mathbf{g} = (g_{ijk})_{i,j,k \leq D(k^\star)}$, $u^{(3)}\mathbf{g} = \sum_{i,j,k \leq D(k^\star)} u_i u_j u_k g_{ijk}$.

Obviously, for any $\delta > 0$,

$$\begin{aligned} \mathbb{B}_n(k^\star) &\geq \int_{\Theta_{k^\star}} e^{\ell_n(\theta) - \ell_n(\widehat{\theta})} d\pi_{k^\star}(\theta) \\ &\geq \int_{|\theta - \widehat{\theta}|_1 \leq \delta} e^{\ell_n(\theta) - \ell_n(\widehat{\theta})} d\pi_{k^\star}(\theta). \end{aligned}$$

Using the usual techniques of Laplace expansions (Guihenneuc & Rousseau, 2004) and invoking assumptions **M-prior**, **M-regularity** and **M-Fisher** yield

$$\begin{aligned} \mathbb{B}_n(k^\star) &\geq n^{-D(k^\star)/2} \pi_{k^\star}(\widehat{\theta}) \int_{|u|_1 \leq \delta\sqrt{n}} e^{-\frac{u^T J u}{2}} \left(1 - C \frac{|u|_1^3 + |u|_1}{\sqrt{n}}\right) du \\ &\geq C\, n^{-D(k^\star)/2} \end{aligned}$$

for some $C > 0$. Therefore, choosing $\beta_1 = 1/C$ implies the result. □

Two results are still needed. They are stated in the two propositions below. On the one hand, in relation with assumption **O-prior**, the following holds:

PROPOSITION 1. *There exists $C_2 > 0$ such that, in the setting of mixture models, for every sequence $\{\delta_n\}$ that decreases to zero, for all $n \geq 1$,*

$$\pi_{k^\star+1}\Big\{\theta \in \Theta_{k^\star+1} : H(\theta) \leq \delta_n\Big\} \leq C_2\, {\delta_n}^{[D(k^\star)+1]/2}.$$

On the other hand, in relation with assumption **O-approximation** and the related condition on δ-entropy, we shall show that

PROPOSITION 2. *If $\mathcal{F}_n^{k^\star+1} = \{(\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_{k^\star+1} : \min_{j \leq k^\star+1} p_j \geq e^{-n}\}$ approximates the set $\Theta_{k^\star+1}$, then assumption* **O-approximation** *is fulfilled for $K = k^\star + 1$. Furthermore, the entropy condition given by inequality (6) in Theorem 3 holds as soon as δ_1 is chosen large enough.*

Proofs of Propositions 1 and 2 are rather technical. Thus, we postpone them to Section B and Section C, respectively.

Finally, inequality (11) of Theorem 5 is a consequence of Theorem 1, Lemmas 6, 6, 7 and Propositions 1 and 2.

Appendix

APPENDIX A: CONSTRUCTION OF TESTS

PROPOSITION 3. *Let (ρ, c) belong to $\mathbb{R}_+^* \times (0, 1]$ and $f \in L_+^1(\mu) \setminus \{0\}$. Let us assume that $V(f)$ is positive and finite. Thus, $H(f)$ exists and is finite too.*

Let $\ell_{n,f} = \sum_{i=1}^n \log f(Z_i)$ and

$$\phi_{n,f,\rho,c} = \mathbb{1}\{\ell_{n,f} - \ell_n^\star + nH(f) \geq n\rho + \log c\}.$$

The following bound holds true:

$$\mathrm{E}^{\star n}\phi_{n,f,\rho,c} \leq \frac{1}{c} \exp\left\{-\frac{n\rho}{2}\left(\frac{\rho}{2V(f)} \wedge 1\right)\right\}. \tag{30}$$

Let $\rho' \in \mathbb{R}_+$ and $g \in L_+^1(\mu)$ be such that $\mu g = 1$, $g \leq e^{\rho'} f$ and $V(g)$ is finite. If in addition $(\rho + \rho') < H(f)$, then the following bound holds true:

$$\mathrm{E}_g^n(1-\phi_{n,f,\rho,c}) \leq \exp\left\{-\frac{n[H(f) - (\rho+\rho')]}{2}\left(\frac{H(f) - (\rho+\rho')}{V(g)} \wedge 1\right)\right\}. \tag{31}$$

PROOF. $H(f)$ is finite because of (2). Let us denote $\log f$ by ℓ_f and $\log g$ by ℓ_g. For every $s \in [0,1]$,

$$\begin{aligned} c\,\mathrm{E}^{\star n}\phi_{n,f,\rho,c} &= c\,P^{\star n}\left\{\ell_{n,f} - \ell_n^\star \geq n\rho - nH(f) + \log c\right\} \\ &\leq e^{-ns(\rho - H(f))}\left(P^\star e^{s(\ell_f - \ell^\star)}\right)^n. \end{aligned}$$

Now, invoking the Taylor formula with integral remainder applied to the $\mathcal{C}^\infty$ function $s \mapsto P^\star e^{s(\ell_f - \ell^\star)}$ implies that, for every $s \in [0,1]$,

$$\begin{aligned} P^\star e^{s(\ell_f - \ell^\star)} &= 1 - sH(f) + \\ &\quad s^2 \int_0^1 (1-t) P^\star\left[(\ell^\star - \ell_f)^2 (\mathbb{1}\{\ell_f \geq \ell^\star\} + \mathbb{1}\{\ell_f < \ell^\star\}) e^{st(\ell_f - \ell^\star)}\right] dt \\ &\leq 1 - sH(f) + s^2 V(f). \end{aligned}$$

Consequently, because $\log t \leq t - 1$ (every $t > 0$), the following bound holds true:

$$c\,\mathrm{E}^{\star n}\phi_{n,f,\rho,c} \leq \exp\left[-ns\rho + ns^2 V(f)\right].$$

The choice $s = 1 \wedge \frac{\rho}{2V(f)}$ yields (30).

Similarly, for all $s \in [0,1]$,

$$\begin{aligned} \mathrm{E}_g^n(1 - \phi_{n,f,\rho,c}) &\leq P^{\star n}\{\ell_n^\star - \ell_{n,f} > n[H(f) - \rho]\} \\ &= P^{\star n}\{\ell_n^\star - \ell_{n,g} > n[H(f) - (\rho + \rho')]\} \\ &\leq e^{-ns[H(f) - (\rho+\rho')]}\left(P_g e^{s(\ell^\star - \ell_g)}\right)^n. \end{aligned}$$

Again, the Taylor expansion of the function $s \mapsto P_g e^{s(\ell^\star - \ell_g)}$ and the Hölder inequality imply that, for every $s \in [0,1]$,

$$\begin{aligned} P_g e^{s(\ell^\star - \ell_g)} &\leq 1 - sH(g,\theta^\star) + s^2 \int_0^1 (1-t) \int (\ell^\star - \ell_g)^2 (f^\star)^{st} g^{1-st} d\mu dt \\ &\leq 1 + s^2 \int_0^1 (1-t)\left[P^\star(\ell^\star - \ell_g)^2\right]^{st}\left[P_g(\ell^\star - \ell_g)^2\right]^{(1-st)} dt \\ &\leq 1 + \frac{s^2}{2} V(g). \end{aligned}$$

Therefore, because $\log t \leq t - 1$ (every $t > 0$),

$$\mathrm{E}_g^n(1 - \phi_{n,f,\rho,c}) \leq \exp\left\{-ns\left[H(f) - (\rho + \rho')\right] + n\frac{s^2}{2} V(g)\right\},$$

and the choice $s = 1 \wedge \frac{H(f) - (\rho+\rho')}{V(g)}$ yields (31). □

APPENDIX B: PROOF OF PROPOSITION 1

The key-tool in this section is a variant of the locally conic parameterization first introduced in (Dacunha-Castelle & Gassiat, 1997b). The main difference is the substitution of the L^1 norm to the L^2 norm.

In the sequel, c, C will be positive constants whose values can change from one line to another.

Let $\{\delta_n\}$ be a decreasing sequence of positive numbers which tend to zero. Let us denote by $\|\cdot\|$ the $L^1(\mu)$ norm. Our first move is to point out that, for all $\theta \in \Theta_{k^\star+1}$,

$$\sqrt{H(\theta)} \geq \frac{\|f^\star - f_\theta\|}{2}.$$

So, assumption **M-prior** ensures that Proposition 1 holds if

$$\pi_{k^\star+1}\Big\{\theta \in \Theta_{k^\star+1} : \|f^\star - f_\theta\| \leq \sqrt{\delta_n}\Big\} \leq C_2 \sqrt{\delta_n}^{D(k^\star)+1} \tag{32}$$

for some $C_2 > 0$ which does not depend on $\{\delta_n\}$.

The hard work is now to translate the condition $\|f^\star - f_\theta\| \leq \sqrt{\delta_n}$ (each $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \mathrm{int}(\Theta_{k^\star+1})$) in terms of parameters $\boldsymbol{p}$ and $\boldsymbol{\gamma}$.

Introducing the L^1 locally conic parameterization. For each $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \mathrm{int}(\Theta_{k^\star+1})$, let us define iteratively the permutation σ_θ upon $\{1, \ldots, k^\star+1\}$ as follows:

- let $(j_1, \sigma_\theta(j_1))$ be such that

  $$|\gamma_{\sigma_\theta(j_1)} - \gamma^\star_{j_1}|_1 = \min_{j \leq k^\star} \min_{j' \leq k^\star+1} |\gamma^\star_j - \gamma_{j'}|_1,$$

  subject to j_1 and $\sigma_\theta(j_1)$ are minimal;
- let us assume that $(j_1, \sigma_\theta(j_1)), \ldots, (j_{l-1}, \sigma_\theta(j_{l-1}))$ with $l < k^\star$ have been defined; then $(j_l, \sigma_\theta(j_l))$ is chosen such that

  $$|\gamma_{\sigma_\theta(j_l)} - \gamma^\star_{j_l}|_1 = \min_{j} \min_{j'} |\gamma^\star_j - \gamma_{j'}|_1,$$

  subject to j_l and $\sigma_\theta(j_l)$ are minimal, where index $j \leq k^\star$ does not belong to $\{j_1, \ldots, j_{l-1}\}$ and index $j' \leq k^\star + 1$ does not belong to $\{\sigma_\theta(j_1), \ldots, \sigma_\theta(j_{l-1})\}$;

- once $(j_1, \sigma_\theta(j_1)), \ldots, (j_{k^\star}, \sigma_\theta(j_{k^\star}))$ are defined, $\sigma_\theta(j)$ is known for each $j \leq k^\star$. The value of $\sigma_\theta(k^\star+1) \in \{j : j \leq k^\star+1\} \setminus \{\sigma_\theta(j) : j \leq k^\star\}$ is imposed.

At this stage, let us present a simple yet useful decomposition of the quantity of interest:

LEMMA 8. *For every $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_{k^\star+1}$ and each permutation ς onto $\{1, \ldots, k^\star+1\}$, let us denote by $\theta^\varsigma = (\boldsymbol{p}^\varsigma, \boldsymbol{\gamma}^\varsigma) \in \Theta_{k^\star+1}$ the parameter with coordinates $p_j^\varsigma = p_{\varsigma(j)}$, $\gamma_j^\varsigma = \gamma_{\varsigma(j)}$ (all $j \leq k^\star+1$) and set $\pi_{k^\star+1}^\varsigma(\theta) = \pi_{k^\star+1}(\theta^\varsigma)$.*
Since for all θ and ς, $\|f^\star - f_\theta\| = \|f^\star - f_{\theta^\varsigma}\|$,

$$\pi_{k^\star+1}\Big\{\theta \in \Theta_{k^\star+1} : \|f^\star - f_\theta\| \leq \sqrt{\delta_n}\Big\} = \sum_\varsigma \pi_{k^\star+1}^\varsigma\Big\{\theta \in \Theta_{k^\star+1} : \sigma_\theta = \mathrm{id},\ \|f^\star - f_\theta\| \leq \sqrt{\delta_n}\Big\}, \tag{33}$$

where index ς in the sum above ranges through the set of all permutations onto $\{1, \ldots, k^\star+1\}$.

We show below that the term in the sum above associated with $\varsigma = \mathrm{id}$ is bounded by a constant times $\sqrt{\delta_n}^{D(k^\star)+1}$. Besides, the proof involves only properties that all $\pi_{k^\star+1}^\varsigma$ share. Studying the latter term is therefore sufficient to conclude that Proposition 1 holds.

PROOF. We abbreviate $\mathbb{1}\{\|f^\star - f_\theta\| \leq \sqrt{\delta_n}\}$ to $h(\theta)$. Note that, for all ς, $h(\theta^\varsigma) = h(\theta)$. The left-hand side term in Equation (33) equals (index ς in the sums below ranges over the permutations onto $\{1, \ldots, k^\star+1\}$)

$$\begin{aligned}\int_{\Theta_{k^\star+1}} h(\theta) d\pi_{k^\star+1}(\theta) &= \sum_\varsigma \int_{\Theta_{k^\star+1}} h(\theta) \mathbb{1}\{\sigma_\theta = \varsigma\} d\pi_{k^\star+1}(\theta) \\ &= \sum_\varsigma \int_{\Theta_{k^\star+1}} h(\theta) \mathbb{1}\{\sigma_\theta = \mathrm{id}\} d\pi_{k^\star+1}^\varsigma(\theta),\end{aligned}$$

hence the result. □

Let us set $\Theta^\star = \{\theta \in \Theta_{k^\star+1} : \sigma_\theta = \mathrm{id}\}$. For all $\theta \in \Theta^\star$, let $\gamma_\theta = \gamma_{k^\star+1}$, $p_\theta = p_{k^\star+1}$ and $R_\theta = (\rho_1, \ldots, \rho_{k^\star-1}, r_1, \ldots, r_{k^\star})$, where

$$\rho_j = \frac{p_j - p_j^\star}{p_\theta} \quad \text{and} \quad r_j = \frac{\gamma_j - \gamma_j^\star}{p_\theta} \quad (\text{all } j \leq k^\star)$$

(we emphasize that $\sum_{j\leq k^\star}\rho_j=-1$). Now, let us define

$$N(\gamma_\theta,R_\theta)=\left\|g_{\gamma_\theta}+\sum_{j=1}^{k^\star}p_j^\star r_j^T\nabla g_{\gamma_j^\star}+\sum_{j=1}^{k^\star}\rho_j g_{\gamma_j^\star}\right\|,$$

then $t_\theta=p_\theta N(\gamma_\theta,R_\theta)$.

PROPOSITION 4. *For all $\theta\in\Theta^\star$, let $\Psi(\theta)=(t_\theta,\gamma_\theta,R_\theta)$. Function Ψ is a bijection between $\Theta^\star$ and $\Psi(\Theta^\star)$. Furthermore, $T=\sup_{\theta\in\Theta^\star}t_\theta$ is finite, so that the projection of $\Psi(\Theta^\star)$ along its first coordinate is included in $[0,T]$. Finally, for all $\varepsilon>0$, there exists $\eta>0$ such that, for every $\theta\in\Theta^\star$, $\|f^\star-f_\theta\|\leq\eta$ yields $t_\theta\leq\varepsilon$.*

PROOF. It is readily seen that Ψ is a bijection. We point out that $N(\gamma,R)$ is necessarily positive for all $(t,\gamma,R)\in\Psi(\Theta^\star)$, by virtue of assumption **M-linear independence**. As for the finiteness of T, let us note that, for any $\theta\in\Theta^\star$,

$$\begin{aligned}t_\theta &= \left\|p_\theta g_{\gamma_\theta}+\sum_{j=1}^{k^\star}p_j^\star(\gamma_j-\gamma_j^\star)^T\nabla g_{\gamma_j^\star}+\sum_{j=1}^{k^\star}(p_j-p_j^\star)g_{\gamma_j^\star}\right\| \\ &\leq 2+\sum_{j=1}^{k^\star}p_j^\star\|(\gamma_j-\gamma_j^\star)^T\nabla g_{\gamma_j^\star}\|.\end{aligned}\tag{34}$$

The right-hand side term above is finite because Γ is bounded and $\|(\nabla g_{\gamma_j^\star})_l\|$ ($j\leq k^\star$, $l\leq d$) are finite thanks to assumption **M-regularity**. Besides, the term does not depend on θ, so T is finite indeed.

In order to show that the last point is valid, let us consider a sequence $\{\theta^n\}$ with values $\theta^n=(\boldsymbol{p}^n,\boldsymbol{\gamma}^n)\in\Theta^\star$ such that $\|f^\star-f_{\theta^n}\|$ goes to zero as n tends to infinity. Let $\{\theta^{\varphi(n)}\}$ be a convergent subsequence to some θ^o. Then $f_{\theta^{\varphi(n)}}$ converges almost-surely to f_{θ^o} (see assumption **A2**) so that $f_{\theta^o}=fs$.

Now, since t_{θ^n} belongs to the compact set $[0,T]$, it suffices to prove that any convergent subsequence $\{t_{\theta^{\varphi(n)}}\}$ tends to zero. Because Γ is compact, there exists a subsequence of $\{\gamma_{\theta^{\varphi(n)}}\}$ (still denoted by $\{\gamma_{\theta^{\varphi(n)}}\}$ for simplicity of notations) which converges to γ^o

(a) Let us assume that $\gamma^o \notin \{\gamma_j^\star : j \leq k^\star\}$. Then necessarily, $p_{\theta^{\varphi(n)}}$, $p_j^{\varphi(n)}$ and $\gamma_j^{\varphi(n)}$ respectively tend to zero, $p_j^\star$ and $\gamma_j^\star$ for all $j \leq k^\star$. In this case, $t_{\theta^{\varphi(n)}}$ tends to zero.

(b) Let us assume on the contrary that $\gamma^o = \gamma_{k^\star}^\star$, say. Then necessarily, $(p_{\theta^{\varphi(n)}} + p_{k^\star}^{\varphi(n)})$, $\gamma_{k^\star}^{\varphi(n)}$, $p_j^{\varphi(n)}$ and $\gamma_j^{\varphi(n)}$ respectively tend to $p_{k^\star}^\star$, $\gamma_{k^\star}^\star$, $p_j^\star$ and $\gamma_j^\star$ for all $j < k^\star$. In this case too, $t_{\theta^{\varphi(n)}}$ tends to zero.

This completes the proof. □

Using the L^1 locally conic parameterization.

PROPOSITION 5. *For every $\tau > 0$ there exists $C > 0$ such that, for all $n \geq 1$,*

$$\pi_{k^\star+1}\left\{\theta \in \Theta^\star : \min_{j \leq k^\star} |\gamma_\theta - \gamma_j^\star|_1 > \tau,\ \|f^\star - f_\theta\| \leq \sqrt{\delta_n}\right\} \leq C\ \sqrt{\delta_n}^{k^\star(d+1)}.$$

We shall need the following lemma of equivalence of norms while proving Proposition 5:

LEMMA 9. *Let Λ be a compact subset of $\mathbb{R}^d$ and, for all $\lambda \in \Lambda$, $g_\lambda \in L^1(\mu)$. Let us suppose that $\|g_\lambda - g_{\lambda'}\|$ tends to zero when λ' tends to λ. Let $g_1, \ldots, g_k \in L^1(\mu)$ be k functions such that, for every $\lambda \in \Lambda$, $g_\lambda, g_1, \ldots, g_k$ are linearly independent.*

There exists $C > 0$ such that, for all $a = (a_0, \ldots, a_k) \in \mathbb{R}^{k+1}$ and $\lambda \in \Lambda$,

$$\left\| a_0 g_\lambda + \sum_{j=1}^k a_j g_j \right\| \geq C\ \sum_{j=0}^k |a_j|.$$

PROOF. Set $\lambda \in \Lambda$ and let $|a|_\lambda = \|a_0 g_\lambda + \sum_{j=1}^k a_j g_j\|$ for all $a \in \mathbb{R}^{k+1}$. This defines a norm $|\cdot|_\lambda$ on $\mathbb{R}^{k+1}$. Because $\mathbb{R}^{k+1}$ is finite dimensional, the $|\cdot|_1$ and $|\cdot|_\lambda$ norms are equivalent: particularly, if S_1 is the $|\cdot|_1$-sphere of $\mathbb{R}^{k+1}$ with radius one, then for all $a \in S_1$,

$$|a|_\lambda \geq C_\lambda\ |a|_1,$$

where $C_\lambda = \inf\{|a|_\lambda : a \in S_1\} > 0$. Furthermore, function $(\lambda, a) \mapsto |a|_\lambda$ is continuous and positive over the compact set $\Lambda \times S_1$, so that $\inf\{C_\lambda : \lambda \in \Lambda\}$ is positive, hence the result. □

PROOF OF PROPOSITION 5. Let $\tau > 0$, let $(t, \gamma, R) \in \Psi(\Theta^\star)$ and $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) = \Psi^{-1}(t, \gamma, R)$ satisfy $|\gamma_\theta - \gamma_j^\star|_1 > \tau$ for all $j \leq k^\star$ and $\|f^\star - f_\theta\| \leq \sqrt{\delta_n}$. A Taylor-Lagrange expansion (in t) of $(f^\star - f_\theta)$ yields the existence of $t^o \in (0, t)$ such that

$$|f^\star - f_\theta| \geq \frac{t}{N} \left| g_\gamma + \sum_{j=1}^{k^\star} p_j^\star r_j^T \nabla g_{\gamma_j^\star} + \sum_{j=1}^{k^\star} \rho_j g_{\gamma_j^\star} \right| - \frac{t^2}{N^2} \left| \sum_{j=1}^{k^\star} \rho_j r_j^T \nabla g_{\gamma_j^o} + \frac{1}{2} \sum_{j=1}^{k^\star} p_j^o r_j^T D^2 g_{\gamma_j^o} r_j \right|,$$

where $\gamma_j^o = \gamma_j^\star + t^o r_j / N$ and $p_j^o = p_j^\star + t^o \rho_j / N$ (all $j \leq k^\star$). Therefore, by virtue of assumption **M-regularity**, there exists a constant $C > 0$ such that

$$\|f^\star - f_\theta\| \geq t \left(1 - C \, \frac{t}{N^2} \left[\sum_{j=1}^{k^\star} \left(|\rho_j| |r_j|_1 + |r_j|_2^2 \right) \right] \right). \tag{35}$$

Furthermore, assumption **M-linear independence** and Lemma 9 imply that, for some constant $C > 0$ (that depends on τ),

$$N \geq C \, \left(1 + \sum_{j=1}^{k^\star} \left(|\rho_j| + p_j^\star |r_j|_1 \right) \right), \tag{36}$$

so the following lower bounds for $\|f^\star - f_\theta\|$ are deduced from inequality (35):

$$\begin{aligned} \|f^\star - f_\theta\| &\geq t \left(1 - C \, \frac{t}{N} \, \frac{\sum_{j=1}^{k^\star} \left(|\rho_j| |r_j|_1 + |r_j|_2^2 \right)}{1 + \sum_{j=1}^{k^\star} \left(|\rho_j| + p_j^\star |r_j|_1 \right)} \right) \\ &\geq t \left(1 - C \, \frac{\sum_{j=1}^{k^\star} \left(|p_j - p_j^\star| |\gamma_j - \gamma_j^\star|_1 + |\gamma_j - \gamma_j^\star|_2^2 \right)}{p_\theta + \sum_{j=1}^{k^\star} \left(|p_j - p_j^\star| + p_j^\star |\gamma_j - \gamma_j^\star|_1 \right)} \right) \\ &\geq t \left(1 - C \, \frac{\sum_{j=1}^{k^\star} \left(|p_j - p_j^\star| |\gamma_j - \gamma_j^\star|_1 + |\gamma_j - \gamma_j^\star|_2^2 \right)}{\sum_{j=1}^{k^\star} \left(|p_j - p_j^\star| + p_j^\star |\gamma_j - \gamma_j^\star|_1 \right)} \right). \end{aligned} \tag{37}$$

Now, it is readily seen by mimicking the last part of the proof of Proposition 4 that for all $\varepsilon > 0$, there exists $\eta > 0$ such that, for every $\theta \in \Theta^\star$, $\|f^\star - f_\theta\| \leq \eta$ yields $|p_j - p_j^\star| \leq \varepsilon$, $|\gamma_j - \gamma_j^\star|_1 \leq \varepsilon$ and $|\gamma_j - \gamma_j^\star|_2 \leq \varepsilon$. Consequently, for n

large enough, the quantity between parentheses in inequality (37) is lower bounded by $1/2$, hence the validity for n large enough of

$$t \leq 2 \, \|f^\star - f_\theta\| \leq 2 \, \sqrt{\delta_n}. \tag{38}$$

Finally, combining equality $t = p_\theta N$ with inequalities (36) and (38) yields the existence of a constant $C > 0$ such that, for n large enough,

$$p_\theta + \sum_{j=1}^{k^\star} \Big(|p_j - p_j^\star| + p_j^\star |\gamma_j - \gamma_j^\star|_1 \Big) \leq C \, \sqrt{\delta_n}.$$

Therefore, for large values of n,

$$\pi_{k^\star+1} \left\{ \theta \in \Theta^\star : \min_{j \leq k^\star} |\gamma_\theta - \gamma_j^\star|_1 > \tau, \ \|f^\star - f_\theta\| \leq \sqrt{\delta_n} \right\} \leq$$
$$\pi_{k^\star+1} \left\{ \theta \in \Theta^\star : \sum_{j=1}^{k^\star} \Big(|p_j - p_j^\star| + p_j^\star |\gamma_j - \gamma_j^\star|_1 \Big) \leq C \, \sqrt{\delta_n} \right\}.$$

Finally, by virtue of assumption **M-prior**, the latter is bounded by a constant (independent of n) times $\sqrt{\delta_n}^{k^\star(d+1)}$. The conclusion of Proposition 5 follows. □

PROPOSITION 6. *There exist $\tau, C > 0$ such that, for all $n \geq 1$,*

$$\pi_{k^\star+1} \left\{ \theta \in \Theta^\star : \min_{j \leq k^\star} |\gamma_\theta - \gamma_j^\star|_1 \leq \tau, \ \|f^\star - f_\theta\| \leq \sqrt{\delta_n} \right\} \leq C \, \sqrt{\delta_n}^{k^\star(d+1)}.$$

PROOF. Let $\tau > 0$, let $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta^\star$ satisfy $\|f^\star - f_\theta\| \leq \sqrt{\delta_n}$. Let us also assume that $|\gamma_\theta - \gamma_j^\star|_1 \leq \tau$ for some $j \leq k^\star$, say $j = 1$. By construction of $\Theta^\star$, $|\gamma_1 - \gamma_1^\star|_1 \leq |\gamma_\theta - \gamma_1^\star|_1 \leq \tau$. The first inequality also implies that τ can be chosen small enough so that γ_θ must be different from $\gamma_j^\star$ for every $j = 2, \ldots, k^\star$. We consider without loss of generality that $\gamma_\theta \notin \{\gamma_j^\star : j \leq k^\star\}$.

Moreover, using the result stated in the proof of Proposition 4, it holds that $|\gamma_j - \gamma_j^\star|_1$ and $|p_j - p_j^\star|$ go to zero as n goes to infinity for every $j = 2, \ldots, k^\star$. This implies that $|p_1 + p_\theta - p_1^\star|$ goes to zero as n goes to infinity. Therefore, by virtue of assumption **M-linear independence** and Lemma 9,

there exist constants $c, C > 0$ such that, for n large enough,

$$
\begin{aligned}
(39)\quad \|f^\star - f_\theta\| \geq C \Bigg(& \sum_{j=2}^{k^\star} |p_j - p_j^\star| + \sum_{j=2}^{k^\star} p_j^\star |\gamma_j - \gamma_j^\star|_1 + \Bigg| (p_1 + p_\theta - p_1^\star) + \\
& \sum_{(r,s)\notin\mathcal{A}} \lambda_{rs}^0 \Big[p_\theta (\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s + p_1 (\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s \Big] \Bigg| + \\
& \sum_{(r,s)\in\mathcal{A}} \Big[p_\theta |(\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s| + p_1 |(\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s| \Big] + \\
& \sum_{l=1}^{d} \Bigg| p_1 (\gamma_1 - \gamma_1^\star)_l + p_\theta (\gamma_\theta - \gamma_1^\star)_l + \\
& \sum_{(r,s)\notin\mathcal{A}} \lambda_{rs}^l \Big[p_\theta (\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s + p_1 (\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s \Big] \Bigg| \Bigg) \\
& - c \left(p_\theta |\gamma_\theta - \gamma_1^\star|_1^3 + p_1 |\gamma_1 - \gamma_1^\star|_1^3 + \sum_{j=2}^{k^\star} |\gamma_j - \gamma_j^\star|_2^2 \right).
\end{aligned}
$$

Let us denote by A_1 and A_2 the first and second expressions between parentheses above. Since $|\gamma_j - \gamma_j^\star|_1$ goes to zero for $j = 2, \ldots, k^\star$, when n is large enough, $\sum_{j=2}^{k^\star} |\gamma_j - \gamma_j^\star|_2^2$ can be neglected in A_2. If $CA_1 \leq 2cA_2$, then $\sum_{j=2}^{k^\star} |p_j - p_j^\star| \leq 2cA_2$, so that $|p_1 + p_\theta - p_1^\star| \leq 2cA_2$, which yields in turn

$$
\begin{aligned}
& \Bigg| \sum_{(r,s)\notin\mathcal{A}} \lambda_{rs}^0 \Big[p_\theta (\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s + p_1 (\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s \Big] \Bigg| + \\
& \sum_{(r,s)\in\mathcal{A}} \Big[p_\theta |(\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s| + p_1 |(\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s| \Big] \leq 3cA_2.
\end{aligned}
$$

Consequently, assumption **M-linear independence** guarantees the existence of a constant $C' > 0$ such that

$$
p_\theta |\gamma_1 - \gamma_1^\star|_2^2 + p_1 |\gamma_1 - \gamma_1^\star|_2^2 \leq C' \left(p_\theta |\gamma_1 - \gamma_1^\star|_1^3 + p_1 |\gamma_1 - \gamma_1^\star|_1^3 \right).
$$

By choosing $\tau > 0$ small enough, we can ensure that the latter is impossible. Therefore, $CA_1 > 2cA_2$ and using inequality (39) and assumption **M-linear**

independence gives, for another constant $C > 0$,

$$\|f^\star - f_\theta\| \geq C \left(\sum_{j=2}^{k^\star} |p_j - p_j^\star| + \sum_{j=2}^{k^\star} p_j^\star |\gamma_j - \gamma_j^\star|_1 + |p_1 + p_\theta - p_1^\star| + p_\theta |\gamma_\theta - \gamma_1^\star|_2^2 + p_1 |\gamma_1 - \gamma_1^\star|_2^2 + \sum_{l=1}^{d} \left| p_1 (\gamma_1 - \gamma_1^\star)_l + p_\theta (\gamma_\theta - \gamma_1^\star)_l + \sum_{(r,s) \notin \mathcal{A}} \lambda_{rs}^l \Big[p_\theta (\gamma_\theta - \gamma_1^\star)_r (\gamma_\theta - \gamma_1^\star)_s + p_1 (\gamma_1 - \gamma_1^\star)_r (\gamma_1 - \gamma_1^\star)_s \Big] \right| \right),$$

hence finally (up to a change of C),

$$|p_1 + p_\theta - p_1^\star| + \sum_{j=2}^{k^\star} |p_j - p_j^\star| + p_1 |\gamma_1 - \gamma_1^\star|_2^2 + p_\theta |\gamma_\theta - \gamma_1^\star|_2^2 + |p_1 (\gamma_1 - \gamma_1^\star) + p_\theta (\gamma_\theta - \gamma_1^\star)|_1 + \sum_{j=2}^{k^\star} p_j^\star |\gamma_j - \gamma_j^\star|_1 \leq C \sqrt{\delta_n}. \tag{40}$$

Therefore, for the chosen τ and large values of n,

$$\pi_{k^\star+1} \Big\{ \theta \in \Theta^\star : \min_{j \leq k^\star} |\gamma_\theta - \gamma_j^\star|_1 \leq \tau,\ \|f^\star - f_\theta\| \leq \sqrt{\delta_n} \Big\} \leq \pi_{k^\star+1} \Big\{ \theta \in \Theta^\star : \text{inequality (40) holds} \Big\}.$$

The conclusion is now at hand. The conditions on p_j and γ_j ($j = 2, \ldots, k^\star$) and a symmetry argument imply that the right-hand side term above is bounded by a constant times $\sqrt{\delta_n}^{[(d+1)(k^\star-1)]}$ times w_n, where

$$w_n = \int \mathbb{1}\{p_\theta \geq p_1\} \mathbb{1}\{|p_1 + p_\theta - p_1^\star| + p_1 |\gamma_1 - \gamma_1^\star|_2^2 + p_\theta |\gamma_\theta - \gamma_1^\star|_2^2 + |p_1 (\gamma_1 - \gamma_1^\star) + p_\theta (\gamma_\theta - \gamma_1^\star)|_1 \leq C \sqrt{\delta_n}\}\, d\pi_{k^\star+1}^{\boldsymbol{\gamma}}(\boldsymbol{\gamma})\, d\pi_{k^\star+1}^{\boldsymbol{p}}(\boldsymbol{p}).$$

Let us also point out that under the conditions of the formula above, $|\gamma_\theta - \gamma_1|_2^2 \leq 4C\sqrt{\delta_n}/p_1$ and $p_\theta \geq p_1^\star/4$ as soon as n is large enough to ensure $C\sqrt{\delta_n} \leq p_1^\star/2$. Now, according to assumption **M-prior**, let $\varepsilon > 0$ be such that $\min_{j<j'} |\gamma_j - \gamma_{j'}|_2^2 \leq \varepsilon$ implies $\pi_{k^\star+1}^{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \propto \prod_{j<j'} |\gamma_j - \gamma_{j'}|_2$ when $d = 1$. We split the integral by considering the cases $4C\sqrt{\delta_n}/p_1 > \varepsilon$ and $4C\sqrt{\delta_n}/p_1 \leq \varepsilon$. Let $h = 1$ when $d = 1$ and $h = 0$ otherwise. There exist

constants $c, c' > 0$ such that w_n is bounded (up to a constant factor) by the sum of $w_{n,1}$ and $w_{n,2}$, where

$$\begin{aligned} w_{n,1} &\leq \int \mathbb{1}\left\{|\gamma_\theta - \gamma_1^\star|_1 \leq c\sqrt{\delta_n},\ |p_\theta - p_1^\star| \leq c\sqrt{\delta_n}\right\}\ d\pi_{k^\star+1}^{\boldsymbol{p}}(\boldsymbol{p})\ d\pi_{k^\star+1}^{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \\ &\leq c'\sqrt{\delta_n}^{(d+1)}, \end{aligned}$$

and

$$\begin{aligned} w_{n,2} &\leq \sqrt{\delta_n}^{h/2} \int \mathbb{1}\left\{\left|\gamma_\theta - \frac{p_\theta + p_1}{p_\theta}\gamma_1^\star + \frac{p_1}{p_\theta}\gamma_1\right|_1 \leq c\sqrt{\delta_n}\right\} \\ &\qquad\qquad \mathbb{1}\left\{|\gamma_1 - \gamma_1^\star|_2^2 \leq c\sqrt{\delta_n}/p_1\right\}\ p_1^{-h/2}\ d\pi_{k^\star+1}^{\boldsymbol{p}}(\boldsymbol{p})\ d\boldsymbol{\gamma} \\ &\leq c'\sqrt{\delta_n}^{[d+(d+h)/2]} \int_{4C\sqrt{\delta_n}/\varepsilon}^{1/2} \frac{dp_1}{{p_1}^{(d+h)/2}} \\ &\leq c'\sqrt{\delta_n}^{[d+(d+h)/2]} \left(\frac{\varepsilon}{4C\sqrt{\delta_n}}\right)^{(d+h-2)/2}. \end{aligned}$$

This completes the proof of Proposition 6, because in summary,

$$\pi_{k^\star+1}\left\{\theta \in \Theta^\star : \min_{j \leq k^\star} |\gamma_\theta - \gamma_j^\star|_1 \leq \tau,\ \|f^\star - f_\theta\| \leq \sqrt{\delta_n}\right\} = O\left(\sqrt{\delta_n}\right)^{(d+1)k^\star}.$$

□

APPENDIX C: PROOF OF PROPOSITION 2

It is readily seen that assumption **O-approximation** holds for the chosen approximating set. Let us focus now on the entropy condition.

Constructing δ-brackets. Let δ_1 satisfy inequality (5) of Theorem 3. A convenient value will be chosen later on. Let us set $j' \leq \lfloor \delta_0/\delta_n \rfloor$ and $\varepsilon = j'\delta_n/4$. Let $\tau \geq 1$ be a constant.

Let $\theta = (\boldsymbol{p}, \boldsymbol{\gamma}) \in \Theta_{k^\star+1}$ be arbitrarily chosen. Let $\eta \in (0, \eta_1)$ be small enough so that, for every $j = 1, \ldots, k^\star + 1$, $u_j = \overline{g}_{\gamma_j,\eta}$ and $v_j = \underline{g}_{\gamma_j,\eta}$ (as defined in assumption **M-local brackets**) satisfy, for all $\gamma \in \Gamma$,

$$P_{u_j - v_j}(1 + \log^2 g_\gamma) \leq \varepsilon/\tau, \tag{41}$$

$$P_{g_\gamma}(\log u_j - \log v_j)^2 \leq (\varepsilon/\tau)^2, \tag{42}$$

$$P_{u_j - v_j} \log^2 u_j \leq (\varepsilon/\tau)\log^2(\varepsilon/\tau). \tag{43}$$

Let us introduce

$$v_\theta = (1-\varepsilon/\tau)\sum_{j=1}^{k^\star+1} p_j v_j \qquad \text{and} \qquad u_\theta = (1+\varepsilon/\tau)\sum_{j=1}^{k^\star+1} p_j u_j.$$

LEMMA 10. *There exists $\tau \geq 1$ (which depends only on $k^\star$ and constant M introduced in assumption* **M-local brackets***) such that the bracket $[v_\theta, u_\theta]$ be an ε-bracket as defined in Section 2.3.*

PROOF. Let us verify that the first condition that defines an ε-bracket is satisfied. Invoking (41) yields

$$\mu(u_\theta - v_\theta) = (1+\varepsilon/\tau)\sum_{j=1}^{k^\star+1} p_j \mu(u_j - v_j) + (2\varepsilon/\tau)\sum_{j=1}^{k^\star+1} p_j \mu v_j \leq 4\varepsilon/\tau. \tag{44}$$

As for the second condition, let us point out that (index j ranges below between 1 and $k^\star+1$)

$$\log u_\theta - \log v_\theta = \log\frac{1+\varepsilon/\tau}{1-\varepsilon/\tau} + \log\frac{\sum_j p_j u_j}{\sum_j p_j v_j} \leq \log\frac{1+\varepsilon/\tau}{1-\varepsilon/\tau} + \max_j \log\frac{u_j}{v_j},$$

hence

$$(\log u_\theta - \log v_\theta)^2 \leq 2\left(\log^2\frac{1+\varepsilon/\tau}{1-\varepsilon/\tau} + \sum_j \log^2\frac{u_j}{v_j}\right). \tag{45}$$

Using $(1+\varepsilon/\tau)/(1-\varepsilon/\tau) \leq 1+4\varepsilon/\tau$, $\log(1+t) \leq t$ (all $t>0$) and inequality (42) then ensures that

$$P^\star(\log u_\theta - \log v_\theta)^2 \leq 2(k^\star+17)(\varepsilon/\tau)^2. \tag{46}$$

Let us consider now the third condition. Straightforwardly,

$$P_{u_\theta - v_\theta}\log^2(u_\theta/f^\star) \leq (1+\varepsilon/\tau)\sum_{j=1}^{k^\star+1} p_j P_{u_j - v_j}\log^2(u_\theta/f^\star) + 2\varepsilon/\tau\sum_{j=1}^{k^\star+1} p_j P_{g_{\gamma_j}}\log^2(u/f^\star). \tag{47}$$

It is readily seen that $\log^2(u_\theta/f^\star) \leq 2^{k^\star+1}\sum_{j=1}^{k^\star}\log^2 g_{\gamma_j^\star} + 4\log^2(1+\varepsilon/\tau) + 2^{k^\star+3}\sum_{j=1}^{k^\star+1}\log^2 u_j$. Besides, assumption **M-local brackets** ensures that,

for all $\gamma \in \Gamma$ and $j = 1, \ldots, k^\star+1$, $P_{g_{\gamma_j}} \log^2 g_{\gamma_j^\star}$ and $P_{g_{\gamma_j}} \log^2 u_j$ are bounded by M. Now, because $\log^2(1+\varepsilon/\tau) \leq \varepsilon/\tau$ and inequalities (41), (43) hold, inequality (47) yields

$$P_{u_\theta - v_\theta} \log^2(u_\theta/f^\star) \leq \Big(2^{k^\star+2} k^\star (M+1) + 4(1+\log^2 2) + 2^{k^\star+4}(k^\star+1)(M+1)\Big)(\varepsilon/\tau)\log^2(\varepsilon/\tau) \tag{48}$$

The last step of this proof is dedicated to the verification of the fourth condition. Since the following holds:

$$P_{v_\theta}(\log u_\theta - \log v_\theta)^2 \leq \sum_{j=1}^{k^\star+1} p_j P_{g_{\gamma_j}}(\log u_\theta - \log v_\theta)^2,$$

inequalities (42) and (45) imply

$$P_{v_\theta}(\log u_\theta - \log v_\theta)^2 \leq 2(k^\star+17)(\varepsilon/\tau)^2. \tag{49}$$

In conclusion, inequalities (44), (46), (48) and (49) show that τ can be chosen large enough (independently of θ) so that $[v_\theta, u_\theta]$ be an ε-bracket. □

Controlling the δ-entropy. The rule $x_1(1-\varepsilon/\tau) = e^{-n}$ and $x_{j+1}(1-\varepsilon/\tau) = x_j(1+\varepsilon/\tau)$ is used for defining a net for the interval $(e^{-n}, 1)$. Such a net has at most $1+n/\log(1+\varepsilon/\tau)/(1-\varepsilon/\tau) \leq 1+2n\tau/\varepsilon$ support points. Using repeatedly this construction on each dimension of the $(k^\star+1)$-dimensional simplex yields a net for $\{\boldsymbol{p} \in \mathbb{R}_+^{k^\star} : \min_{j \leq k^\star} p_j \geq e^{-n}, 1 - \sum_{j \leq k^\star} p_j \geq e^{-n}\}$ with at most a $O((n/\varepsilon)^{(k^\star+1)})$ support points. Besides, we can choose a net for $\Gamma^{k^\star+1}$ with at most a $O(\varepsilon^{-d(k^\star+1)})$ support points such that each $\boldsymbol{\gamma} \in \Gamma^{k^\star+1}$ is within $|\cdot|_1$-distance ε of some element of the net. Consequently, the minimum number of ε-brackets needed to cover $\mathcal{F}_n^{k^\star+1}$ is a $O(n^{(k^\star+1)}/\varepsilon^{(d+1)(k^\star+1)})$, so that there exist positive constants a, b, c for which the δ-entropy with bracketing of $\mathcal{F}_n^{k^\star+1}$ (as defined in Section 2.3) is bounded as follows:

$$\mathcal{E}\left(\mathcal{F}_n^{k^\star+1}, \frac{j'\delta_n}{4}\right) \leq a \log n - b \log(j'\delta_n) + c. \tag{50}$$

Now, let us note that the following simple inequalities hold

$$\frac{nj'\delta_n}{\log^2(j'\delta_n)} \geq \frac{n\delta_n}{(\log \delta_n)\log(j'\delta_n)} \geq \frac{n\delta_n}{\log^2 \delta_n}$$

and consider each term of equation (50) in turn:

- it is readily proven that $a \log n \leq n\delta_n / \log^2 \delta_n$ is equivalent to

$$\delta_1 \geq \frac{1}{(\log^3 n) n^{(\delta_1/a)^{1/2}-1}}. \tag{51}$$

  When $\delta_1 \geq a$, the largest value of the right-hand side of inequality (51) is achieved at n_0. Hence, δ_1 can be chosen large enough (independently of j' and n) so that inequality (51) holds for all $n \geq n_0$ and $j' \leq \lfloor \delta_0/\delta_n \rfloor$.
- now,

$$-b \log(j'\delta_n) \leq \frac{n\delta_n}{(\log \delta_n) \log(j'\delta_n)} \quad \text{iff} \quad -b \log \delta_n \leq \delta_1 \log^3 n.$$

  Since $\log^2 \delta_n \leq 4 \log^2 n$, both inequalities above are valid as soon as

$$\delta_1 \geq \frac{2b}{\log^2 n}. \tag{52}$$

  The largest value of the right-hand side of inequality (52) is achieved at n_0. Therefore, δ_1 can be chosen large enough (independently of j' and n) so that inequality (52) holds for all $n \geq n_0$ and $j' \leq \lfloor \delta_0/\delta_n \rfloor$.
- finally, using again $\log^2 \delta_n \leq 4 \log^2 n$ yields that $c \leq n\delta_n / \log^2 \delta_n$ is implied by

$$\delta_1 \geq \frac{4c}{\log n}. \tag{53}$$

  Obviously, δ_1 can be chosen large enough (independently of j' and n) so that inequality (53) holds for all $n \geq n_0$ and $j' \leq \lfloor \delta_0/\delta_n \rfloor$.

In summary, δ_1 can be chosen according to inequality (5) in order to guarantee that $\mathcal{E}(\mathcal{F}_n^{k^\star+1}, j'\delta_n/4)$ be bounded by the right-hand side of inequality (6) (with j' substituted to j). This completes the proof of Proposition 2, because $\mathcal{E}(\mathcal{F}_n^{k^\star+1}, j'\delta_n/4)$ is larger than the left-hand side of inequality (6) (with the same substitution).

REFERENCES

Azencott, R. and Dacunha-Castelle, D. (1986). *Series of irregular observations.* Springer-Verlag, New-York.

Bahadur, R. R., 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., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, **27**(2) 536–561.

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference*, **112**(1-2) 241–258.

Boucheron, S. and Gassiat, E. (2004). Error exponents for AR order testing. *Preprint Orsay 2004-35*.

Chambaz, A. (2003). Testing the order of a model. Accepted for publication in *Ann. Statist.*

Csiszár, I. and Körner, J. (1981). *Information theory: coding theorems for discrete memoryless systems*. Academic Press, New York.

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

Dacunha-Castelle, D. and Gassiat, E. (1997b). Testing in locally conic models, and application to mixture models. *ESAIM Probab. Statist.*, **1** 285–317 (electronic).

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.

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

Finesso, L., Liu, C-C. and Narayan, P. (1996). The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, **42**(5) 1488–1497.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, 97 611-631.

Gassiat, E. and Boucheron, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory*, **49**(4) 964–980.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, **28**(2) 500–531.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning.* Springer-Verlag, New-York.

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.

Ishwaran, H., James, L. F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.*, **96**(456)

1316–1332.

Jefferys, W. and Berger, J. (1992). Ockam's razor and Bayesian analysis. *American Scientist*, **80** 64-72.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90 773-795.

Keribin, C. and Haughton, D. (2003). Asymptotic probabilities of overestimating and underestimating the order of a model in general regular families. *Communications in Statistics*, **32**(7) 1373–1404.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, **1** 38-53.

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

Moreno, E. and Liseo, B. (2003). A default Bayesian test for the number of components in a mixture. *J. Statist. Plann. Inference*, **111**(1-2) 129–142.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley-Interscience, New York.

Mengersen, K. and Robert, C. (1996). Testing for mixtures: a Bayesian entropy approach. In *Bayesian Statistics 5*, ed. J. O. Berger, J. M. Bernardo and A. P. Dawid.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **4** 10-26.

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

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, **29**(3) 687–714.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **41** 276-278.

Titterington, D. M., 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.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, **23**(2) 339–362.

Antoine Chambaz
MAP5 CNRS UMR 8145
Université René Descartes
45 rue des Saints-Pères
75270 Paris cedex 06, France.
e-mail: chambaz@univ-paris5.fr

Judith Rousseau
Cérémade, UMR CNRS 7534
Université Dauphine
Place de Lattre de Tassigny
75775 Paris cedex 16, France.
e-mail: rousseau@ensae.fr