

A MDL approach to HMM with Poisson and Gaussian emissions. Application to order identification

A. Chambaz^{a,*}, A. Garivier^b, E. Gassiat^b

^a*MAP5, Université René Descartes, Paris, France*

^b*Laboratoire de Mathématiques, Université Paris-Sud, Orsay, France*

Abstract

We address the issue of order identification for HMM with Poisson and Gaussian emissions. We prove information-theoretic BIC-like mixture inequalities in the spirit of (Finesso, 1991; Liu & Narayan, 1994; Gassiat & Boucheron, 2003). These inequalities lead to consistent penalized estimators that need no prior bound on the order nor on the parameters of the mixture components.

Key words: BIC, HMM, infinite alphabet, MDL, model selection, order estimation

1 Introduction

Formally introduced as *probabilistic functions of Markov Chains* in 1966 by Baum & Petrie, hidden Markov models (HMM) have known since then a growing interest as they proved useful in various applications, from speech recognition (Levinson et al., 1983) to blind deconvolution of unknown communication channels (Kaleh & Vallet, 1994), biostatistics (Koski, 2001) or meteorology (Hughes & Guttorp, 1994). For a mathematical survey on HMM, see (Ephraim & Merhav, 2002; Cappé et al., 2005).

In most practical cases, the *order* of the model (*ie* the true number of hidden states) is unknown and has to be estimated. For that purpose, two approaches

* Corresponding author.

Email addresses: Antoine.Chambaz@univ-paris5.fr (A. Chambaz), aurelien.garivier@math.u-psud.fr (A. Garivier), elisabeth.gassiat@math.u-psud.fr (E. Gassiat).

have been proposed: penalized maximum likelihood estimators as in (Finesso, 1991; Kieffer, 1993) and Bayesian procedures as in (Liu & Narayan, 1994). As Gassiat & Boucheron (2003) did for finite emission alphabet HMM, we follow the methodologies of Finesso and Liu & Narayan for some instances of infinite (continuous and discrete) emission alphabets. Following the work of these authors, we address the issue of order identification for HMM with Poisson and Gaussian emissions: we prove MDL-inspired mixture inequalities which lead to consistent penalized estimators requiring no prior bound on the order nor on the parameters of the mixture components.

In 1978, Rissanen introduced the Minimum Description Length (MDL) principle with motto:

“Choose the model that gives the shortest description of data.”

More precisely, given any k -dimensional model (*ie* parametric family of densities indexed by Θ of dimension $k \geq 1$):

$$\mathcal{M} = \{g_\theta : \theta \in \Theta\},$$

let E_θ be the expectation with respect to a random variable X_1^n with distribution P_θ , which density is g_θ (with respect to Lebesgue measure). For any density q such that $q(x_1^n) = 0$ implies $g_\theta(x_1^n) = 0$, the Kullback-Leibler divergence between g_θ and q is

$$K_n(g_\theta, q) = E_\theta \log \frac{g_\theta(X_1^n)}{q(X_1^n)} = E_\theta [-\log q(X_1^n) - (-\log g_\theta(X_1^n))].$$

In Information theory, $-\log q(X_1^n)$ is interpreted as the code length for X_1^n when using coding distribution q , so $E_\theta[-\log g_\theta(X_1^n)]$ is the *ideal code length* for X_1^n . In this perspective, $K_n(g_\theta, q)$ is the average additional cost (or *redundancy*) for compressing some source in \mathcal{M} without knowing which one.

When assuming that the maximum likelihood estimator $\hat{\theta}(x_1^n)$ achieves a \sqrt{n} -rate and that the distribution of $\hat{\theta}(X_1^n)$ has uniformly summable tail probabilities: there exists a summable sequence $\{\delta_n\}$ of positive numbers such that, for every $\theta \in \Theta$,

$$P_\theta \left\{ \sqrt{n} \|\hat{\theta}(X^n) - \theta\| \geq \log n \right\} \leq \delta_n,$$

Rissanen proved (1986) that

$$\liminf_{n \rightarrow \infty} \frac{K_n(g_\theta, q)}{\frac{k}{2} \log n} \geq 1 \tag{1}$$

for all $\theta \in \Theta$ except on a set with Lebesgue measure 0 (that depends on q and k). Here, k is the dimension of the parameter space Θ . This result has a minimax counterpart for i.i.d sequences (Clarke & Barron, 1990): under mild

assumptions,

$$K_n^* = \min_q \sup_{\theta \in \Theta} K_n(g_\theta, q) \geq \frac{k}{2} \log \frac{n}{2\pi e} + O(1). \quad (2)$$

Both (1) and (2) put forward a leading term $\frac{k}{2} \log n$ that has taken a great importance in Information theory and Statistics. The coding density q is said optimal if it achieves equality in inequation (1). Three optimal coding distributions are often encountered in Information theory (we refer to (Barron et al., 1998; Hansen & Yu, 2001) for surveys):

- two-stage coding, that yields description length

$$-\log q(x_1^n) = -\log g_{\hat{\theta}(x_1^n)}(x_1^n) + \frac{k}{2} \log n;$$

- mixture coding, where q is a mixture of all densities g_θ ($\theta \in \Theta$);
- predictive coding, where q is based on an iterative prediction scheme of x_j given x_1^{j-1} ($j = 1, \dots, n$), namely $q_{\text{PDL}}(x_1^n) = \prod_{j=1}^n g_{\hat{\theta}(x_1^{j-1})}(x_j)$.

We want to highlight that the quantity $-\log g_{\hat{\theta}(x_1^n)}(x_1^n) + \frac{k}{2} \log n$, also called Bayesian Information Criterion (BIC), has been considerably studied since its first introduction in (Schwarz, 1978) for purpose of model dimension estimation.

Now, let us consider the following problem: given a family of models $(\mathcal{M}_i)_{i \in I}$, which one best represents some given data x_1^n ? The MDL methodology suggests to choose model $\widehat{\mathcal{M}} = \mathcal{M}_{\hat{i}}$ that yields the shortest description length of x_1^n .

Let k_i be the dimension of model \mathcal{M}_i for every $i \in I$. Each of the three optimal coding distributions presented above selects a model:

- two-stage coding chooses

$$\widehat{\mathcal{M}}_{\text{BIC}} = \arg \min_{\mathcal{M}_i (i \in I)} \left\{ -\log g_{\hat{\theta}_i(x_1^n)}(x_1^n) + \frac{k_i}{2} \log n \right\},$$

where $\hat{\theta}_i$ is the maximum likelihood estimator over model \mathcal{M}_i ;

- mixture coding chooses

$$\widehat{\mathcal{M}}_{\text{MIX}} = \arg \min_{\mathcal{M}_i (i \in I)} \{ -\log q_i(x_1^n) \},$$

where q_i is a particular mixture to be specified later – we will actually introduce a penalized version of this estimation procedure;

- predictive coding chooses

$$\widehat{\mathcal{M}}_{\text{PDL}} = \arg \min_{\mathcal{M}_i (i \in I)} \{-\log q_{\text{PDL},i}(x_1^n)\},$$

where $q_{\text{PDL},i}$ is the predictive distribution relative to \mathcal{M}_i .

The challenging task is to prove that such estimators are consistent: if x_1^n is output by a source of density g_{θ_0} such that $g_{\theta_0} \in \mathcal{M}_{i_0}$ and $g_{\theta_0} \in \mathcal{M}_i$ implies $\mathcal{M}_{i_0} \subset \mathcal{M}_i$, then $\widehat{\mathcal{M}} = \mathcal{M}_{i_0}$ eventually almost surely. This has been successfully accomplished for Markov Chains by Csiszár & Shields (2000), and for Context Tree Models (or Variable Length Markov Chains) by Csiszár & Talata (2005) and Garivier (2005).

Organization of the paper

We start in Section 2 by stating and proving inequalities that compare BIC criterion and a particular mixture coding distribution (see Theorems 1 and 2). Four related models are involved, namely HMM mixture models and i.i.d models, with Poisson or Gaussian emissions. These inequalities are used in Section 3 for order identification purposes in the four models cited above. We notably obtain the consistency of two order estimators based on two-stage and mixture coding without assuming the existence of any prior bound on orders (see Theorems 5 and 6). We give a hint in Section 4 why order estimation based on raw (that is *not* penalized) predictive coding procedure cannot likely be proven consistent along the same lines than the two others (see Theorem 7). The proof of two lemmas and a useful result due to Leroux (1992a) are postponed to Appendix A and Appendix B.

2 Mixture inequalities

Mixture inequalities for HMM mixture model

Let σ^2 be a positive number. The Gaussian density with mean m and variance σ^2 (with respect to the Lebesgue measure on the real line) is denoted by ϕ_{m,σ^2} . The Poisson density with mean m (with respect to the counting measure on the set of nonnegative integers) is denoted by π_m .

Let $\{X_n\}_{n \geq 1}$ be a sequence of random variables with values in the measured space $(\mathcal{X}, \mathcal{A}, \mu)$ and defined on a measurable set upon which all random variables will be defined. Let us denote by $\{Z_n\}_{n \geq 0}$ a sequence of hidden random variables such that, conditionally on $Z_1^n = (Z_1, \dots, Z_n)$, X_1, \dots, X_n are independent and the distribution of each X_i only depends on Z_i (all $i \leq n$).

For every $k \geq 1$, let $(p_j^o : j \leq k) \in \mathbb{R}_+^k$ be an initial distribution and let \mathcal{S}_k be the set of all $\mathbf{p} = (p_{jj'} : j, j' \leq k) \in \mathbb{R}_+^{k^2}$ such that, for all $j \leq k$, $\sum_{j'=1}^k p_{jj'} = 1$, then the parameter set

$$\Theta_k = \left\{ \theta = (\mathbf{p}, \mathbf{m}) : \mathbf{p} \in \mathcal{S}_k, \mathbf{m} = (m_1, \dots, m_k) \in \mathbb{R}^k \right\}.$$

Under parameter $\theta = (\mathbf{p}, \mathbf{m}) \in \Theta_k$ (some $k \geq 1$), $\{Z_n\}_{n \geq 0}$ is a Markov chain with values in $\{1, \dots, k\}$, initial distribution $P_\theta\{Z_0 = j'\} = p_{j'}^o$ and transition probabilities $P_\theta\{Z_{i+1} = j' | Z_i = j\} = p_{jj'}$ (all $j, j' \leq k$). Therefore, $\{X_n\}_{n \geq 1}$ is a HMM under parameter θ .

We shall consider two examples of emission distributions:

Gaussian emission (GE) For every $n \geq 1$, X_n has density $\phi_{m_{Z_n}, \sigma^2}$ conditionally on Z_n .

Poisson emission (PE) For every $n \geq 1$, X_n has density $\pi_{m_{Z_n}}$ conditionally on Z_n .

For all parameter $\theta \in \Theta_k$ (any $k \geq 1$), let g_θ be the density of $X_1^n = (X_1, \dots, X_n)$ under θ . For every $k \geq 1$, let ν_k be a prior probability on Θ_k such that, for some chosen $\tau > 0$, under ν_k :

- \mathbf{p} and \mathbf{m} are independent,
- $p_{j'}^o = 1/k$ for all $j' \leq k$ are determinist,
- the vectors $(p_{jj'} : j' \leq k)$ ($j \leq k$) are independently Dirichlet(1/2, ..., 1/2) distributed,
- m_1, \dots, m_k are independent, identically distributed with density $\phi_{0, \tau}$ in example **GE** and with density Gamma($\tau, 1/2$) in example **PE**.

The related mixture statistics is defined by

$$q_k(X_1^n) = \int_{\Theta_k} g_\theta(X_1^n) d\nu_k(\theta). \quad (3)$$

It is worth noting that q_k is a positive function of $x_1^n \in \mathcal{X}^n$ in examples **GE** and **PE**.

The main results of this section are comparisons between the maximum log-likelihood and the mixture statistics in examples **GE** and **PE**.

Let $X_{(n)}$ and $|X|_{(n)}$ be the maxima of X_1, \dots, X_n and $|X_1|, \dots, |X_n|$, respectively. Let us also introduce, for all $k, n \geq 1$,

$$\begin{aligned}
c_{kn} &= \log k - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k^2(k-1)}{4n} + \frac{k}{12n}, \\
d_{kn} &= \frac{k}{2} \log \left(\frac{\tau^2}{k\sigma^2} + \frac{1}{n} \right), \\
e_{kn} &= \frac{k}{2} \left(1 + \tau - \log(k\tau) \right).
\end{aligned}$$

Theorem 1 (HMM mixture models) *Under the assumptions described above, for every integers $k, n \geq 1$,*

GE

$$0 \leq \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leq \frac{k^2}{2} \log n + \frac{k}{2\tau^2} |X|_{(n)}^2 + c_{kn} + d_{kn}. \quad (4)$$

PE

$$0 \leq \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leq \frac{k^2}{2} \log n + k\tau X_{(n)} + c_{kn} + e_{kn}. \quad (5)$$

Particular case of i.i.d mixture models

The i.i.d mixture model is a particular case of the HMM model. Here, the sequence $\{Z_n\}_{n \geq 0}$ is made of mutually independent random variables. In other words, for all $j, j' \leq k$, $p_{j'}^o = p_{jj'}$.

For every $k \geq 1$, let us introduce the set \mathcal{S}'_k of all $\mathbf{p} = (p_j^o : j \leq k) \in \mathbb{R}_+^k$ such that $\sum_{j=1}^k p_j^o = 1$, then the parameter set

$$\Theta'_k = \left\{ \theta = (\mathbf{p}, \mathbf{m}) : \mathbf{p} \in \mathcal{S}'_k, \mathbf{m} = (m_1, \dots, m_k) \in \mathbb{R}^k \right\}.$$

Again, g_θ is the density of X_1^n under parameter $\theta \in \Theta'_k$. For every $k \geq 1$, a new mixing probability ν_k on Θ'_k is chosen such that, under ν'_k :

- \mathbf{p} and \mathbf{m} are independent,
- \mathbf{p} is Dirichlet(1/2, ..., 1/2) distributed,
- m_1, \dots, m_k are independent, identically distributed with density $\phi_{0,\tau}$ in example **GE** and with density Gamma($\tau, 1/2$) in example **PE**.

Equality (3) defines a mixture statistics $q_k(X_1^n)$ in this framework. The main result of this section is another comparison between the maximum log-likelihood and the mixture statistics in examples **GE** and **PE**.

Let us introduce, for all $n, k \geq 1$,

$$c'_{kn} = -\log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k(k-1)}{4n} + \frac{1}{12n}.$$

Theorem 2 (i.i.d mixture models) *Under the assumptions described above, for every integers $k, n \geq 1$,*

GE

$$0 \leq \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leq \frac{2k-1}{2} \log n + \frac{k}{2\tau^2} |X|_{(n)}^2 + c'_{kn} + d_{kn}. \quad (6)$$

PE

$$0 \leq \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) - \log q_k(X_1^n) \leq \frac{2k-1}{2} \log n + k\tau X_{(n)} + c'_{kn} + e_{kn}. \quad (7)$$

Comment

In inequations (4), (5), (6) and (7), the upper bounds write as a sum of $\frac{1}{2} \dim(\Theta_k) \log n$, a bounded term and a random term which involves the maximum of $|X_1|, \dots, |X_n|$. The following lemmas guarantee that these random terms are bounded in probability at rate $\log n$ in example **GE** and slower than $\log n$ in example **PE**.

Lemma 3 *Let $\{Y_n\}_{n \geq 1}$ be a sequence of independent Gaussian random variables with variance σ^2 . The mean of Y_n is denoted by m_n . If $\sup_{n \geq 1} |m_n|$ is finite, then for n large enough,*

$$P \left\{ |Y|_{(n)}^2 \geq 5\sigma^2 \log n \right\} \leq \frac{1}{n^{3/2}}.$$

Lemma 4 *Let $\{Y_n\}_{n \geq 1}$ be a sequence of independent Poisson random variables. The mean of Y_n is denoted by m_n . If $\sup_{n \geq 1} m_n$ is finite, then for n large enough,*

$$P \left\{ Y_{(n)} \geq \frac{\log n}{\sqrt{\log \log n}} \right\} \leq \frac{1}{n^2}.$$

The proofs of Lemmas 3 and 4 are postponed to Section A of the Appendix.

Proof of Theorems 1 and 2

In the first place, let us introduce some notations.

For all $\theta \in \Theta_k$ (any $k \geq 1$) and for all $x_1^n \in \mathcal{X}^n$, $z_0^n = (z_0, \dots, z_n) \in \{1, \dots, k\}^{n+1}$, we denote by $g_\theta(x_1^n | z_1^n)$ the density of X_1^n at x_1^n conditionally

on $Z_1^n = z_1^n$. The mixture density $q_k(x_1^n|z_1^n)$ at x_1^n conditionally on $Z_1^n = z_1^n$ is defined as in (3), with substitution of $g_\theta(x_1^n|z_1^n)$ to $g_\theta(X_1^n)$.

Similarly, we denote by $g_\theta(x_1^n|z_0)$ the density of X_1^n at x_1^n conditionally on $Z_0 = z_0$, and $q_k(\cdot|z_0)$ the corresponding conditional mixture density. Besides, if $P_\theta\{z_1^n|z_0\}$ is a shortcut for $P_\theta\{Z_1^n = z_1^n|Z_0 = z_0\}$, then the mixture density at z_1^n $q_k(z_1^n|z_0)$ is defined as in (3), with replacement of $g_\theta(X_1^n)$ by $P_\theta\{z_1^n|z_0\}$. Finally, for every $j \leq k$, let us set

$$n_j = \sum_{i=1}^n \mathbb{1}\{z_i = j\}, \quad I_j = \{i \leq n : z_i = j\} \quad \text{and} \quad \bar{x}_j = n_j^{-1} \sum_{i \in I_j} x_i.$$

Proof of Theorem 1 Let us set $x_1^n \in \mathcal{X}^n$. The left-hand inequalities of (4) and (5) are obvious.

Quite straightforwardly, using twice inequality $\sum_{j \leq k} \alpha_j / \sum_{j \leq k} \beta_j \leq \max_{j \leq k} \alpha_j / \beta_j$ (valid for all nonnegative $\alpha_1, \dots, \alpha_k$ and positive β_1, \dots, β_k) yields

$$\begin{aligned} \sup_{\theta \in \Theta_k} \log \frac{g_\theta(x_1^n)}{q_k(x_1^n)} &= \log k + \sup_{\theta \in \Theta_k} \log \frac{\sum_{z_0 \leq k} g_\theta(x_1^n|z_0) p_{z_0}^o}{\sum_{z_0 \leq k} q_k(x_1^n|z_0)} \\ &\leq \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leq k} \log \frac{g_\theta(x_1^n|z_0) p_{z_0}^o}{q_k(x_1^n|z_0)} \\ &\leq \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leq k} \log \frac{g_\theta(x_1^n|z_0)}{q_k(x_1^n|z_0)} \\ &\leq \log k + \sup_{\theta \in \Theta_k} \max_{z_0 \leq k} \log \frac{\sum_{z_1^n \in \{1, \dots, k\}^n} g_\theta(x_1^n|z_0^n) P_\theta\{z_1^n|z_0\}}{\sum_{z_1^n \in \{1, \dots, k\}^n} q_k(x_1^n|z_0^n) q_k(z_1^n|z_0)} \\ &\leq \log k + \sup_{\theta \in \Theta_k} \max_{z_0^n \in \{1, \dots, k\}^{n+1}} \log \frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)} \cdot \frac{P_\theta\{z_1^n|z_0\}}{q_k(z_1^n|z_0)}. \quad (8) \end{aligned}$$

Now, as shown in ((Davisson et al., 1981)) (see equations (52)-(61) therein),

$$\begin{aligned} \sup_{\theta \in \Theta_k} \max_{z_0^n \in \{1, \dots, k\}^{n+1}} \log \frac{P_\theta\{z_1^n|z_0\}}{q_k(z_1^n|z_0)} &\leq k \log \frac{\Gamma(n+k/2)\Gamma(1/2)}{\Gamma(k/2)\Gamma(n+1/2)} \\ &\leq k \left(\frac{k-1}{2} \log n - \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \frac{k(k-1)}{4n} + \frac{1}{12n} \right), \quad (9) \end{aligned}$$

where the second inequality is derived from the following Robbins-Stirling approximation formula, valid for all $z \in \mathbb{R}_+^*$,

$$\sqrt{2\pi} e^{-z} z^{z-1/2} \leq \Gamma(z) \leq \sqrt{2\pi} e^{-z+1/12z} z^{z-1/2}.$$

So, the second ratio in the right-hand term of inequality (8) is controlled. The last step of the proof is dedicated to bounding the first ratio. The same scheme

of proof applies to both examples **GE** and **PE**. It is nevertheless simpler to address each of them at a time.

GE Conditionally on $Z_1^n = z_1^n$ the maximum likelihood estimator of m_j is \bar{x}_j for every $j \leq k$, so that the following bound holds for every $x_1^n \in \mathcal{X}^n$ and $z_1^n \in \{1, \dots, k\}^n$:

$$g_\theta(x_1^n | z_1^n) \leq \prod_{j=1}^k \prod_{i \in I_j} \phi_{\bar{x}_j, \sigma^2}(x_i) = \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^k \exp\left(-\frac{\sum_{i \in I_j} x_i^2}{2\sigma^2} - \frac{n_j(\bar{x}_j)^2}{2\sigma^2}\right). \quad (10)$$

Besides, simple calculations yield

$$\begin{aligned} q_k(x_1^n | z_1^n) &= \prod_{j=1}^k \frac{1}{(\sigma\sqrt{2\pi})^{n_j}} \int \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{m^2}{2\tau^2} - \frac{1}{2\sigma^2} \sum_{i \in I_j} (x_i - m)^2\right) dm \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^k \frac{1}{\sqrt{1 + \frac{n_j\tau^2}{\sigma^2}}} \exp\left(-\frac{\sum_{i \in I_j} x_i^2}{2\sigma^2} + \frac{n_j^2}{2\sigma^2(n_j + \frac{\sigma^2}{\tau^2})}(\bar{x}_j)^2\right). \end{aligned} \quad (11)$$

We now get, as a by-product of inequalities (10) and (11),

$$\frac{g_\theta(x_1^n | z_1^n)}{q_k(x_1^n | z_1^n)} \leq \prod_{j=1}^k \sqrt{1 + \frac{n_j\tau^2}{\sigma^2}} \exp\left(\sum_{j=1}^k \frac{n_j}{2\sigma^2(1 + n_j\tau^2/\sigma^2)}(\bar{x}_j)^2\right).$$

By convexity, the first factor in the right-hand side expression above satisfies

$$\prod_{j=1}^k \sqrt{1 + \frac{n_j\tau^2}{\sigma^2}} \leq \left(1 + \frac{n\tau^2}{k\sigma^2}\right)^{k/2}, \quad (12)$$

while the ratios $n_j/(1 + n_j\tau^2/\sigma^2)$ are upper bounded by σ^2/τ^2 for all $j \leq k$. Therefore,

$$\sup_{\theta \in \Theta_k} \max_{z_0^n \in \{1, \dots, k\}^{n+1}} \log \frac{g_\theta(x_1^n | z_1^n)}{q_k(x_1^n | z_1^n)} \leq \frac{k}{2} \log\left(1 + \frac{n\tau^2}{k\sigma^2}\right) + \frac{k}{2\tau^2} |x|_{(n)}^2. \quad (13)$$

Combining inequalities (8), (9) and (13) yields the result.

PE As justified above, for each $j \leq k$, for every $x_1^n \in \mathcal{X}^n$ and $z_1^n \in \{1, \dots, k\}^n$:

$$g_\theta(x_1^n | z_1^n) \leq \prod_{j=1}^k \prod_{i \in I_j} \pi_{\bar{x}_j}(x_i) = P_n \prod_{j=1}^k \exp\left(-n_j \bar{x}_j (1 - \log \bar{x}_j)\right) \quad (14)$$

if $P_n = 1/\prod_{i=1}^n (x_i)!$. In particular, the factor associated with some $j \leq k$ for which $\bar{x}_j = 0$ equals one. Furthermore, it is readily seen that

$$\begin{aligned}
q_k(x_1^n | z_1^n) &= P_n \prod_{j=1}^k \sqrt{\frac{\tau}{2\pi}} \int m^{n_j \bar{x}_j - 1/2} \exp\left(- (n_j + \tau)m\right) dm \\
&= P_n \prod_{j=1}^k \sqrt{\frac{\tau}{2\pi}} \frac{\Gamma(n_j \bar{x}_j + 1/2)}{(n_j + \tau)^{n_j \bar{x}_j + 1/2}}.
\end{aligned} \tag{15}$$

Here, the factor associated with some $j \leq k$ for which $\bar{x}_j = 0$ equals $\sqrt{\tau/(n_j + \tau)}$.

At this stage, the ratio $g_\theta(x_1^n | z_1^n)/q_k(x_1^n | z_1^n)$ is naturally decomposed into the product of k ratios: for each $j \leq k$, the right-hand side factor of (14) divided by the right-hand side factor of (15) is upper bounded by

$$\sqrt{\frac{e}{\tau}} \times \exp\left(\frac{1}{2} \log n_j + \left(n_j \bar{x}_j + \frac{1}{2}\right) \log\left(1 + \frac{\tau}{n_j}\right)\right)$$

whether $\bar{x}_j = 0$ or not. This simple calculation relies again on the lower bound for $\Gamma(n_j \bar{x}_j + 1/2)$ yielded by the Robbins-Stirling approximation formula.

Consequently, it holds that

$$\begin{aligned}
\log \frac{g_\theta(x_1^n | z_1^n)}{q_k(x_1^n | z_1^n)} &\leq \frac{k}{2}(1 - \log \tau) + \sum_{j=1}^k \left[\frac{1}{2} \log n_j + \tau \left(x_{(n)} + \frac{1}{2}\right) \right] \\
&\leq \frac{k}{2} \log \frac{n}{k} + k\tau x_{(n)} + \frac{k}{2} (1 + \tau - \log \tau)
\end{aligned} \tag{16}$$

(the second inequality follows by convexity). Combining inequalities (8), (9) and (16) (we emphasize that the right-hand term in (16) does not depend on z_0^n nor on θ) gives the result.

□

Remark 1 *We want to point out that inequality (12) can not be improved, since it is optimal when all n_j 's are roughly equal.*

The scheme of proof for Theorem 2 is similar to the one of Theorem 1.

Proof of Theorem 2 Let $x_1^n \in \mathcal{X}^n$. Straightforwardly, for every $\theta \in \Theta_k$,

$$g_\theta(x_1^n) = \sum_{z_1^n \in \{1, \dots, k\}^n} g_\theta(x_1^n | z_1^n) \prod_{j=1}^k (p_j^\theta)^{n_j} \leq \sum_{z_1^n \in \{1, \dots, k\}^n} g_\theta(x_1^n | z_1^n) \prod_{j=1}^k \left(\frac{n_j}{n}\right)^{n_j}.$$

Besides, it is readily seen that

$$\begin{aligned}
q_k(x_1^n) &= \sum_{z_1^n \in \{1, \dots, k\}^n} q_k(x_1^n | z_1^n) \int_{S_k'} \prod_{j=1}^k (p_j^o)^{n_j} d\nu'_k(\mathbf{p}) \\
&= \sum_{z_1^n \in \{1, \dots, k\}^n} \frac{\Gamma(k/2)}{\Gamma(n+k/2)} q_k(x_1^n | z_1^n) \prod_{j=1}^k \frac{\Gamma(n_j + 1/2)}{\Gamma(1/2)}.
\end{aligned}$$

Consequently, using the same argument that yielded inequality (8) implies that

$$\log \frac{g_\theta(x_1^n)}{q_k(x_1^n)} \leq \sup_{z_1^n \in \{1, \dots, k\}^n} \left(\log \frac{\Gamma(n+k/2)\Gamma(1/2)^k}{\Gamma(k/2)} + \log \prod_{j=1}^k \frac{\binom{n_j}{n}^{n_j}}{\Gamma(n_j + 1/2)} + \log \frac{g_\theta(x_1^n | z_1^n)}{q_k(x_1^n | z_1^n)} \right).$$

Handling the second term in the right-hand side of the display above has already been done in the proof of Theorem 1. As for the first term, it holds that it is bounded by

$$\log \frac{\Gamma(n+k/2)\Gamma(1/2)}{\Gamma(k/2)\Gamma(n+1/2)} \leq \frac{k-1}{2} \log n + c'_{kn}$$

(by virtue of (Davisson et al., 1981), equations (52-61) again and the Robbins-Stirling approximation formula). This completes the proof. \square

3 Application to order identification

Let k_0 be the sole integer such that the common distribution P_0 of all X_n ($n \geq 1$) satisfies

$$P_0 \in \{P_\theta : \theta \in \Theta_{k_0}\} \setminus \{P_\theta : \theta \in \Theta_{k_0-1}\}$$

(with convention $\Theta_0 = \emptyset$). By definition, k_0 is the order of P_0 . In examples **GE** and **PE**, k_0 is the minimal number of Gaussian or Poisson densities needed for describing the distribution P_0 . Our goal in this section is to estimate k_0 .

There is a large amount of literature dedicated to the issue of order estimation. The particular case of i.i.d order estimation for mixtures of continuous densities is notoriously challenging (refer to (Chambaz, 2003) for a comprehensive bibliography). It has been addressed through various methods that can be classified into three categories: ad hoc (Henna, 1985; Dacunha-Castelle & Gassiat, 1997; James et al., 2001), maximum likelihood (Leroux, 1992b; Keribin, 2000; Gassiat, 2002; Chambaz, 2003) or Bayesian (Ishwaran et al., 2001; Chambaz & Rousseau, 2005). Actually, Bayesian literature on order selection in mixture

models is essentially devoted to determining coherent noninformative priors, see for instance (Moreno & Liseo, 2003) and to implementing procedures, see for instance (Mengersen & Robert, 1996). The particular case of HMM order estimation was addressed for instance in (Finesso, 1991; Kieffer, 1993; Gassiat, 2002) from a maximum likelihood point of view and in (Liu & Narayan, 1994) from a Bayesian one. Both approaches are combined in (Gassiat & Boucheron, 2003), and we refer to the comprehensive bibliography therein for further references.

Here, we study two estimators: one is based on maximum likelihood and the other is Bayesian flavored.

Let us denote by $\text{pen}(n, k)$ a so-called penalization term, which is a positive valued increasing function of $n, k \geq 1$ such that, for each $k \geq 1$, $\text{pen}(n, k) = o(n)$. This device is required for defining our estimators:

$$\begin{aligned}\widehat{k}_{\text{ML}} &= \arg \min_{k \geq 1} \left\{ - \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) + \text{pen}(n, k) \right\} \quad \text{and} \\ \widehat{k}_{\text{MIX}} &= \arg \min_{k \geq 1} \{ - \log q_k(X_1^n) + \text{pen}(n, k) \}.\end{aligned}$$

Convenient choices of the penalty term involve the following quantities. Let us set $\alpha > 2$, $\{\varphi_n\}_{n \geq 1}$ a sequence of positive numbers that increases slowly to infinity with $\varphi_n = o(n)$. For every $n, k \geq 1$, we introduce the cumulative sums: $C_{kn} = \sum_{\ell=1}^k c_{\ell n}$, $C'_{kn} = \sum_{\ell=1}^k c'_{\ell n}$, $D_{kn} = \sum_{\ell=1}^k d_{\ell n}$ and $E_{kn} = \sum_{\ell=1}^k e_{\ell n}$. All of them are bounded functions of n .

Theorem 5 (consistency of \widehat{k}_{ML}) *Under the assumptions described above, $\widehat{k}_{\text{ML}} = k_0$ eventually almost surely as soon as for every $n \geq 3, k \geq 1$,*

$$\text{pen}(n, k) = \sum_{\ell=1}^k \frac{D(\ell) + \alpha}{2} \log n + R_{kn} + S_{kn},$$

where $D(k) = \dim(\Theta_k) = k^2$ and $R_{kn} = C_{kn}$ for HMM mixtures models, $D(k) = \dim(\Theta'_k) = (2k - 1)$ and $R_{kn} = C'_{kn}$ for i.i.d mixtures models and

GE

$$S_{kn} = D_{kn} + k(k + 1)\varphi_n \log n,$$

PE

$$S_{kn} = E_{kn} + k(k + 1) \frac{\log n}{\sqrt{\log \log n}}.$$

Similarly,

Theorem 6 (consistency of \widehat{k}_{MIX}) Under the assumptions described above, $\widehat{k}_{\text{MIX}} = k_0$ eventually almost surely as soon as for every $n \geq 3, k \geq 1$,

$$\text{pen}(n, k) = \sum_{\ell=1}^{k-1} \frac{D(\ell) + \alpha}{2} \log n + S_{kn},$$

where $D(k) = \dim(\Theta_k) = k^2$ for HMM mixtures models, $D(k) = \dim(\Theta'_k) = (2k - 1)$ for i.i.d mixtures models and

GE

$$S_{kn} = k(k+1)\varphi_n \log n,$$

PE

$$S_{kn} = k(k+1) \frac{\log n}{\sqrt{\log \log n}}.$$

Theorems 5 and 6 thus guarantee that \widehat{k}_{ML} and \widehat{k}_{MIX} are consistent estimators of k_0 . We emphasize that *no prior bound on k_0 is required*. This is particularly interesting for \widehat{k}_{ML} , since such a bound was generally needed when studying its overestimation properties in the mixture of continuous densities order estimation setting, as in (Keribin, 2000; Gassiat, 2002; Chambaz, 2003). This feature illustrates the fact that our contribution to order estimation is mainly related to the overestimation phenomenon. In the HMM framework, few results are known for infinite alphabet emissions.

The conditions required on the penalty function in Theorems 5 and 6 should be discussed too. It is worth noting that, in example **GE**, the penalty is designed without knowing σ^2 . This is why $\text{pen}(n, k)$ is larger than $\log n$ when n grows to infinity, for every $k \geq 1$, which is not satisfactory in reference to the BIC criterion. Assuming known an upper-bound for σ^2 would allow to choose a penalty function that satisfies $\text{pen}(n, k) = O(\log n)$ for every $k \geq 1$. In example **PE**, the behavior of the random term in inequalities (5) and (7) allows to choose penalties satisfying $\text{pen}(n, k) = O(\log n)$ for every $k \geq 1$.

It is also important to compare the dependency of $\text{pen}(n, k)$ with respect to k with that of the BIC criterion. We do not get a single term $\frac{1}{2}D(k)$ on the $\log n$ scale, but rather a cumulative sum of terms $\frac{1}{2}[D(\ell) + \alpha]$ for ℓ ranging from 1 to k .

The few lines of comment above are devoted to \widehat{k}_{ML} . It is well understood that Bayesian estimators naturally take into account the uncertainty on the parameter by integrating it out (Jefferys & Berger, 1992), thus providing an example of auto-penalization. This is illustrated by the equivalence between marginal likelihood and BIC criterion that holds, for instance, in regular models:

$$-\log q_k(X_1^n) = -\log \sup_{\theta \in \Theta_k} g_\theta(X_1^n) + \frac{1}{2}D(k) \log n + O_P(1),$$

as n goes to infinity, valid for every $k \geq 1$. It is proven in (Chambaz & Rousseau, 2005) that efficient order estimation can be achieved by comparing marginal likelihoods (implicitly, without additional penalization) even in non-regular models (and for instance for mixtures of continuous densities). However, Csiszár & Shields (2000) provide an example where \widehat{k}_{ML} is consistent while \widehat{k}_{MIX} is not when its penalty term is set to zero. Here, we (over-) penalize $q_k(X_1^n)$ so that the proofs of Theorems 5 and 6 mainly rely on the mixture inequalities stated in Theorems 1 and 2.

Proof of Theorem 5 In the i.i.d framework, showing that $\widehat{k}_{\text{ML}} \geq k_0$ eventually almost surely is a rather simple consequence of the strong law of large numbers and $\min_{k < k_0} \inf_{\theta \in \Theta'_k} K(g_{\theta_0}, g_\theta) > 0$ for any $\theta_0 \in \Theta'_{k_0} \setminus \Theta'_{k_0-1}$ (see (Leroux, 1992b) for a proof of the latter), where

$$K(g_{\theta_0}, g_\theta) = \int_{x_1 \in \mathcal{X}} g_{\theta_0}(x_1) \log \frac{g_{\theta_0}(x_1)}{g_\theta(x_1)} d\mu(x_1)$$

is the P_{θ_0} -almost sure limit of $n^{-1}[\log g_{\theta_0}(X_1^n) - \log g_\theta(X_1^n)]$.

In the HMM framework, it is a consequence of Lemma 8 (see Appendix B), which contains a Shannon-Breiman-McMillan theorem for HMM that holds in examples **GE** and **PE** (see Theorem 2 in (Leroux, 1992a)) and a useful by-product of the proof of Theorem 3 in the same paper.

The difficult part is to get that $\widehat{k}_{\text{ML}} \leq k_0$ eventually almost surely.

Let $P_0 = P_{\theta_0}$ for $\theta_0 \in \Theta_{k_0} \setminus \Theta_{k_0-1}$. Let us consider a positive valued sequence $\{t_n\}_{n \geq 3}$ to be chosen conveniently later on. Let $k > k_0$. Obviously, if $\widehat{k}_{\text{ML}} = k$, then $\log g_{\theta_0}(X_1^n) \leq \sup_{\theta \in T_k} \log g_\theta(X_1^n) + \text{pen}(n, k_0) - \text{pen}(n, k)$. Here, T_k equals Θ_k for HMM mixture models and equals Θ'_k for i.i.d mixture models. Consequently, using inequalities (4), (5), (6) or (7) (with $\tau = 1/2$ in example **GE** and $\tau = 2$ in example **PE**), $\widehat{k}_{\text{ML}} = k$ yields

$$\log g_{\theta_0}(X_1^n) \leq \log q_k(X_1^n) + \Delta_{nk} \quad (17)$$

with

$$\Delta_{nk} = \text{pen}(n, k_0) - \text{pen}(n, k) + \frac{D(k)}{2} \log n + a_{kn} + b_{kn} + 2kU_n,$$

where $U_n = |X|_{(n)}^2$, $b_{kn} = d_{kn}$ in example **GE** and $U_n = X_{(n)}$, $b_{kn} = e_{kn}$ in example **PE**, while $a_{kn} = c_{kn}$ for HMM mixture models and $a_{kn} = c'_{kn}$ for i.i.d mixture models.

Furthermore, because q_k defines a probability measure, the event defined by

inequality (17) has P_0 -probability

$$\int_{x_1^n \in \mathcal{X}^n} \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \mathbb{1} \left\{ \log \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \leq \Delta_{nk} \right\} q_k(x_1^n) d\mu(x_1^n) \leq \exp(\Delta_{nk}).$$

Now, if we choose $t_n = \varphi_n \log n$ in example **GE** and $t_n = \log n / \sqrt{\log \log n}$ in example **PE**, then $U_n \leq t_n$ implies

$$\Delta_{nk} \leq \frac{\alpha}{2} (k_0 - k) \log n. \quad (18)$$

Therefore

$$P_0 \left\{ \hat{k}_{\text{ML}} > k_0 \text{ and } U_n \leq t_n \right\} \leq \sum_{k > k_0} \exp \left\{ -\frac{\alpha}{2} (k - k_0) \log n \right\} = O(n^{-\alpha/2}).$$

In conclusion, by virtue of the Borel-Cantelli lemma, the previous bound and Lemmas 3 and 4 guarantee that $\hat{k}_{\text{ML}} \leq k_0$ eventually almost surely. This completes the proof. \square

Remark 2 *The specific form chosen for the penalty term in each setting is meant for ensuring inequality (18).*

Proof of Theorem 6 In that case, proving that $\hat{k}_{\text{MIX}} \geq k_0$ eventually almost surely is a little more involved (but still well known). The key-point is to show the existence of a random variables sequence $\{\varepsilon_n\}_{n \geq 1}$ that converges to 0 P_0 -almost surely and, for each $k < k_0$, $\hat{k}_{\text{MIX}} = k$ yields

$$\frac{1}{n} \left[\sup_{\theta \in \Theta_k} \log g_{\theta}(X_1^n) - \log g_{\theta_0}(X_1^n) \right] \geq \varepsilon_n \text{ i.o.}$$

(i.o. stands for “infinitely often”). Then, the conclusion follows again from the strong law of large numbers again in the i.i.d framework and from Lemma 8 in the HMM framework.

Let us set $k < k_0$. Because $\text{pen}(n, k) = o(n)$ and $\text{pen}(n, k_0) = o(n)$, $\hat{k}_{\text{MIX}} = k$ yields

$$0 \geq \frac{1}{n} \log \frac{q_{k_0}(X_1^n)}{q_k(X_1^n)} + o(1).$$

By adding the same quantity to both sides, we get

$$\begin{aligned} \frac{1}{n} \left[\sup_{\theta \in \Theta_k} \log g_{\theta}(X_1^n) - \log g_{\theta_0}(X_1^n) \right] &\geq \frac{1}{n} \log \frac{\sup_{\theta \in \Theta_k} g_{\theta}(X_1^n)}{q_k(X_1^n)} \\ &\quad - \frac{1}{n} \log \frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} + o(1). \end{aligned}$$

Now, by virtue of inequalities (4), (5), (6) and (7) and Lemmas 3 and 4, P_0 -almost surely

$$\frac{1}{n} \log \frac{\sup_{\theta \in \Theta_k} g_\theta(X_1^n)}{q_k(X_1^n)} \xrightarrow[n \rightarrow \infty]{} 0.$$

The same inequalities and lemmas also guarantee that, P_0 -almost surely,

$$\frac{1}{n} \left(\log \frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} \right)_+ \xrightarrow[n \rightarrow \infty]{} 0$$

(the positive part of $t \in \mathbb{R}$ is denoted by $(t)_+$). The final step is a variant of the so-called Barron's lemma taken from ((Finesso, 1991), Theorem 4.4.1): another application of the Borel-Cantelli lemma implies that, P_0 -almost surely,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} \geq \liminf_{n \rightarrow \infty} \frac{-2 \log n}{n} = 0.$$

Therefore, the key-point holds, hence the conclusion for underestimation.

From now on, we use the same notations than in the preceding proof except when notified. Let $k > k_0$. If $\hat{k}_{\text{MIX}} = k$, then $-\log q_k(X_1^n) + \text{pen}(n, k) \leq -\log q_{k_0}(X_1^n) + \text{pen}(n, k_0)$ and using inequalities (4), (5), (6) and (7) implies that

$$\log g_{\theta_0}(X_1^n) \leq \log q_k(X_1^n) + \Delta_{nk}$$

with

$$\Delta_{nk} = \text{pen}(n, k_0) - \text{pen}(n, k) + \frac{D(k_0)}{2} \log n + a_{k_0 n} + 2k_0 U_n,$$

where $\{a_{k_0 n}\}_{n \geq 1}$ is a bounded sequence. The definition of the penalty guarantees that $U_n \leq t_n$ implies that inequality (18) still holds in this setting. Consequently,

$$P_0 \left\{ \hat{k}_{\text{MIX}} > k_0 \text{ and } U_n \leq t_n \right\} \leq \sum_{k > k_0} \exp \left\{ -\frac{\alpha}{2} (k - k_0) \log n \right\} = O(n^{-\alpha/2}).$$

The result follows by virtue of the Borel-Cantelli lemma, the previous bound and Lemmas 3 and 4: $\hat{k}_{\text{ML}} \leq k_0$ eventually almost surely. This completes the proof. \square

4 Predictive MDL

In this section, we are interested in predictive MDL applied to mixture order identification in the i.i.d framework. This is an alternative to methods based on maximum likelihood (a variant of two-stage MDL) or Bayesian estimation (a variant of mixture MDL), see (Hansen & Yu, 2001), considered in Section 3.

For each $k \geq 1$, let $\hat{\theta}_k(x_1^n) \in \Theta_k$ be the maximum likelihood estimator of θ over Θ_k based on the observation of $x_1^n \in \mathcal{X}^n$. The predictive MDL statistics is

$$r_k(X_1^n) = \prod_{i=1}^n g_{\hat{\theta}_k(X_1^{i-1})}(X_i),$$

where $\hat{\theta}_k(x_1^0)$ equals some fixed parameter in the interior of Θ_k .

This statistics yields another order estimator, which is the smallest maximizer of $r_k(X_1^n)$ over integers $k \geq 1$. We merely aim at giving a hint why the scheme of proof used in Section 3 does not apply for this estimator. In words, if the true distribution P_{θ_0} is of order k_0 , then for all $k > k_0$, the null measure set outside of which Rissanen's lower bound over Θ_k is verified will include θ_0 .

Let us denote by \mathcal{F} a set of densities with respect to a measure μ over \mathbb{R}^d . For any two densities $g_1, g_2 \in \mathcal{F}$, $H^2(g_1, g_2) = \int (\sqrt{g_1} - \sqrt{g_2})^2 d\mu$ is the square Hellinger distance between g_1 and g_2 . We recall that $K(g_1, g_2) = \int g_1 \log(g_1/g_2) d\mu$ is the Kullback-Leibler divergence between g_1 and g_2 .

Let X_1, \dots, X_n be i.i.d random variables whose distribution P has density $f \in \mathcal{F}$ with respect to μ . We introduce, for every $\varepsilon > 0$,

$$\mathcal{F}_\varepsilon = \{g \in \mathcal{F} : H^2(g, f) \leq \varepsilon^2\}$$

and

$$\mathcal{D}_\varepsilon = \left\{ \frac{\sqrt{g/f} - 1}{H(g, f)} : g \in \mathcal{F}_\varepsilon \right\}.$$

We recall that the bracket-entropy $H_{[]} (u, \mathcal{S}, L^2(P))$ of the set $\mathcal{S} \subset L^2(P)$ is the logarithm of the minimal number of brackets of length u needed to cover \mathcal{S} (see (van der Vaart, 1998)). Let us state three assumptions (see (van der Vaart, 1998) for definitions of Glivenko-Cantelli and Donsker classes).

Envelope There exists $C \in L^2(P)$ such that, for all $g_1, g_2 \in \mathcal{F}_\varepsilon$, $|\log g_1 - \log g_2| \leq CH(g_1, g_2)$ P -almost surely.

Glivenko The set $\{\log g : g \in \mathcal{F}\}$ is P -Glivenko-Cantelli.

Donsker There exist $\delta, \varepsilon > 0$ such that

$$\int_0^\delta \sqrt{H_{[]} (u, \mathcal{D}_\varepsilon, L^2(P))} du < \infty$$

and, if $\{g_p\}_{p \geq 1}$ is a sequence of elements of \mathcal{F}_ε such that $H(f, g_p) = o(1)$, then $K(f, g_p) = 2H^2(f, g_p)(1 + o(1))$.

Let \hat{f} be the maximum likelihood estimator of f on \mathcal{F} .

We emphasize that, by virtue of Theorems 5.7 and 5.52 in (van der Vaart, 1998), \widehat{f} is consistent in Hellinger distance when assumption **Glivenko** holds, that is

$$H^2(\widehat{f}, f) = o_P(1),$$

and, under assumption **Envelope**,

$$nH^2(\widehat{f}, f) = O_P(1).$$

Under assumption **Donsker**, we can define the set \mathcal{D}_0 of all limit points of sequences $\{d_\varepsilon\}$, $d_\varepsilon \in \mathcal{D}_\varepsilon$, $\varepsilon \rightarrow 0$, as a compact subset of the unit sphere in $L^2(P)$. Moreover, \mathcal{D}_ε and \mathcal{D}_0 are P -Donsker classes. Notice that for all $g \in \mathcal{F}_\varepsilon$,

$$2 \int \frac{1 - \sqrt{g/f}}{H(g, f)} f d\mu = H(g, f) \leq \varepsilon,$$

so that for all $d \in \mathcal{D}_0$, $E(d(X_1)) = 0$. Let us introduce the centered empirical process \mathbb{G}_n on \mathcal{D}_0 defined, for every $d \in \mathcal{D}_0$, by

$$\mathbb{G}_n(d) = \frac{1}{\sqrt{n}} \sum_{i=1}^n d(X_i).$$

This process has unit variance and covariance $\langle d_1, d_2 \rangle_{L^2(P)}$. We denote by W the centered Gaussian process on \mathcal{D}_0 with the same covariance.

Theorem 7 (expansion of $nK(f, \widehat{f})$) *Under assumptions **Envelope**, **Glivenko** and **Donsker**, the following expansion holds:*

$$nK(f, \widehat{f}) = \frac{1}{2} \sup_{d \in \mathcal{D}_0} \left(\mathbb{G}_n(d) \mathbb{1}\{\mathbb{G}_n(d) \geq 0\} \right)^2 + o_P(1).$$

Therefore, $nK(f, \widehat{f})$ converges in distribution to

$$\frac{1}{2} \sup_{d \in \mathcal{D}_0} \left(W(d) \mathbb{1}\{W(d) \geq 0\} \right)^2.$$

Theorem 7 applies to mixture of densities from a smooth enough parametric family $\{\gamma_{\alpha_i, \beta} : \alpha \in A \subset \mathbb{R}^m\}$ with possibly unknown common nuisance parameter $\beta \in B \subset \mathbb{R}^q$. In that case, any k -mixture density g_θ writes as

$$g_\theta = \sum_{j=1}^k p_j \gamma_{\alpha_j, \beta},$$

where $\mathbf{p} = (p_j : j \leq k) \in \mathcal{S}'_k$ (introduced in Section 2), $\mathbf{m} = (\alpha_1, \dots, \alpha_k, \beta) \in A^k \times B$ and $\theta = (\mathbf{p}, \mathbf{m}) \in \Theta'_k = \mathcal{S}'_k \times A^k \times B$. Here, the number of parameters is $D_k = km + q + k - 1$.

In (Azais et al., 2004), general assumptions are given on parametric family $\{\gamma_{\alpha,\beta} : \alpha \in A, \beta \in B\}$ so that assumptions **Envelope**, **Glivenko** and **Donsker** be satisfied. In particular, they hold for Binomial, Poisson and multidimensional Gaussian mixtures.

Let $\mathcal{F} = \{g_\theta : \theta \in \Theta'_k\}$ for some integer $k \geq 2$.

If $f \in \mathcal{F} \setminus \{g_\theta : \theta \in \Theta'_{k-1}\}$ (all p_j are positive and $\alpha_1, \dots, \alpha_k$ are mutually distinct), then \mathcal{D}_0 is a linear space, $\sup_{d \in \mathcal{D}_0} (W(d) \mathbb{1}\{W(d) \geq 0\})^2$ has a chi-square distribution with D_k degrees of freedom and

$$E \left[\sup_{d \in \mathcal{D}_0} (W(d) \mathbb{1}\{W(d) \geq 0\})^2 \right] = D_k,$$

as in identifiable parametric situations.

Now if, on the contrary, $f \in \{g_\theta : \theta \in \Theta'_{k_0}\}$ for some $k_0 < k$, then

$$E \left[\sup_{d \in \mathcal{D}_0} (W(d) \mathbb{1}\{W(d) \geq 0\})^2 \right] < D_k$$

(see for instance (Delmas, 2001)).

Proof of Theorem 7 Let us define the log-likelihood

$$\ell_n(g) = \sum_{i=1}^n \log g(X_i)$$

and, for every $g \in \mathcal{F}$,

$$d_g = \frac{\sqrt{g/f} - 1}{H(g, f)}.$$

Using the same tricks as in Section 3 of (Gassiat, 2002) yields, for some $\varepsilon_n = o(1)$,

$$\begin{aligned} \ell_n(\hat{f}) - \ell_n(f) &= \sup_{g \in \mathcal{F}_{\varepsilon_n}} \left(2H(f, g) \sum_{i=1}^n d_g(X_i) - H^2(f, g) \sum_{i=1}^n d_g^2(X_i) \right) (1 + o_P(1)) \\ &= \sup_{g \in \mathcal{F}_{\varepsilon_n}} \left(2H(f, g) \sum_{i=1}^n d_g(X_i) - 2nH^2(f, g) \right) (1 + o_P(1)). \end{aligned}$$

Because $nH^2(f, \hat{f}) = O_P(1)$ and (as in Section 3 of (Gassiat, 2002)), for all $d \in \mathcal{D}_0$, there is a submodel $g_{c,d}$ with normalized score d and $c = H(f, g_{c,d})$, it holds that

$$nH^2(f, \hat{f}) = \frac{1}{4} \sup_{d \in \mathcal{D}_0} \left(\mathbb{G}_n(d) \mathbb{1}_{\mathbb{G}_n(d) \geq 0} \right)^2 (1 + o_P(1)),$$

so that Theorem 7 follows from assumption **Donsker**. \square

A Proofs of Lemmas 3 and 4

Proof of Lemma 3 Let $m = \sup_{n \geq 1} |m_n|$ and $t_n = \sqrt{5\sigma^2 \log n}$ (all $n \geq 1$). Let n be large enough, so that $t_n \geq m$. For every $i \leq n$,

$$\begin{aligned} P\{|Y_i| \leq t_n\} &= P\{|m_i + Y_i - m_i| \leq t_n\} \\ &\geq P\{|Y_i - m_i| \leq t_n - |m_i|\} \\ &\geq P\{|Y_i - m_i| \leq t_n - m\} \\ &= \int_{-t_n+m}^{t_n-m} \phi_{0,\sigma^2}(y) dy \\ &= \left(1 - \sigma \frac{\phi_{0,\sigma^2}(t_n)}{t_n}\right) (1 + o(1)). \end{aligned}$$

Hence, by virtue of the independence of Y_1, \dots, Y_n ,

$$\begin{aligned} P\{|Y|_{(n)}^2 \geq t_n^2\} &= 1 - \prod_{i=1}^n P\{|Y_i| \leq t_n\} \\ &\leq 1 - \left(1 - \sigma \frac{\phi_{0,\sigma^2}(t_n)}{t_n} (1 + o(1))\right)^n \\ &= 1 - \exp\left\{-\frac{n \exp\left(-\frac{t_n^2}{2\sigma^2}\right)}{t_n \sqrt{2\pi}} (1 + o(1))\right\} \\ &= -\frac{n \exp\left(-\frac{5\sigma^2 \log n}{2\sigma^2}\right)}{\sqrt{5\sigma^2 \log n} \sqrt{2\pi}} (1 + o(1)) \\ &\leq n^{-3/2}, \end{aligned}$$

as soon as n is large enough. \square

Proof of Lemma 4 Let $m = \sup_{n \geq 1} m_n$ and $t_n = \log n / \sqrt{\log \log n}$ (all $n \geq 3$). Let Y be a Poisson random variable with mean m . The logarithmic moment generating function Ψ of $(Y - m)$ satisfies $\Psi(\lambda) = \log Ee^{\lambda(Y-m)} = m(e^\lambda - \lambda - 1)$ (all $\lambda \geq 0$). Its Legendre transform Ψ^* is given for all $t \geq 0$ by

$$\Psi^*(t) = \sup_{\lambda \geq 0} \{\lambda t - \Psi(\lambda)\} = (t + m) \log \frac{t + m}{m} - t.$$

Now, it is obvious that $P\{Y_i \geq t\} \leq P\{Y \geq t\}$ (for each $i \leq n$ and $t > m$).

Therefore, by using the Chernoff bounding method,

$$P\{Y_{(n)} \geq t_n\} \leq nP\{Y \geq t_n\} = nP\{Y - m \geq t_n - m\} \leq n \exp\{-\Psi^*(t_n - m)\}. \quad (\text{A.1})$$

Besides,

$$\Psi^*(t_n - m) = t_n \log \frac{t_n}{m} - t_n - m = (\log n) \sqrt{\log \log n (1 + o(1))} \geq 3 \log n$$

as soon as n is large enough. We conclude by plugging this lower bound into inequality (A.1). \square

B A useful lemma for HMM mixture models

Lemma 8 (Leroux) *For HMM mixture models, both in examples **GE** and **PE**, for every $k \geq 1$ and $\theta_0, \theta \in \Theta_k$, there exists a constant $K_\infty(g_{\theta_0}, g_\theta) < \infty$ such that, P_{θ_0} -almost surely, $n^{-1}[\log g_{\theta_0}(X_1^n) - \log g_\theta(X_1^n)]$ tends to $K_\infty(g_{\theta_0}, g_\theta)$ as n goes to infinity. Besides, for any $\theta_0 \in \Theta_{k_0} \setminus \Theta_{k_0-1}$,*

$$\min_{k < k_0} \inf_{\theta \in \Theta_k} K_\infty(g_{\theta_0}, g_\theta) > 0.$$

Sketch of proof of Lemma 8 The Shannon-Breiman-McMillan part of the lemma is a straightforward consequence of Theorem 2 in (Leroux, 1992a). The second part of the lemma is a by-product of the proof of Theorem 3 of the same paper. Indeed, Leroux proved that, for each $\theta \in \Theta_{k_0}$ such that $g_\theta \neq g_{\theta_0}$, there exists an open neighborhood \mathcal{O}_θ of θ (for the euclidean topology of the one-point compactification of Θ_{k_0}) and $\varepsilon > 0$ such that $\inf_{\theta' \in \mathcal{O}_\theta} K_\infty(g_{\theta_0}, g_{\theta'}) > \varepsilon$. Because Θ_{k_0-1} is precompact, it is covered by the finite union of $\mathcal{O}_{\theta_1}, \dots, \mathcal{O}_{\theta_I}$ (each of them associated with $\varepsilon_i > 0$) and therefore

$$\inf_{\theta \in \Theta_{k_0-1}} K_\infty(g_{\theta_0}, g_\theta) \geq \min_{i \leq I} \inf_{\theta \in \mathcal{O}_{\theta_i}} K_\infty(g_{\theta_0}, g_\theta) \geq \min_{i \leq I} \varepsilon_i > 0.$$

\square

References

- AZAIS, J.-M., GASSIAT, E. and MERCADIER, C. (2004). Asymptotic distribution and power of the likelihood ratio test for mixtures: bounded and unbounded case. Accepted for publication in *Bernoulli*.
- AZENCOTT, R. and DACUNHA-CASTELLE, D. (1986). *Series of irregular observations*. Springer-Verlag, New-York.

- BARRON, A. R., RISSANEN, J. and YU, B. (1998). The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. Inf. Theory*, **44** 2743–2760.
- BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37** 1554–1563.
- CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag.
- CHAMBAZ, A. (2003). Testing the order of a model. Accepted for publication in *Ann. Statist.*
- CHAMBAZ, A. and ROUSSEAU, J. (2005). Nonasymptotic bounds for Bayesian order identification with application to mixtures. Submitted.
- CLARKE, B. S. and BARRON, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, **36** 453–471.
- CSISZÁR, I. and SHIELDS, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.*, **6** 1601–1619.
- CSISZÁR, I. and TALATA, Z. (2005). Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL. *IEEE Trans. Inf. Theory*, accepted.
- DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). The estimation of the order of a mixture model. *Bernoulli*, **3**(3) 279–299.
- DAVISSON, L. D., MCELIECE, R. J., PURSLEY, M. B. and WALLACE, M. S. (1981). Efficient universal noiseless source codes. *IEEE Trans. Inf. Theory*, **27** 269–279.
- DELMAS, C. (2001). *Distribution du maximum d'un champ alatoire et applications statistiques*. PhD thesis, Université Paul Sabatier.
- EPHRAIM, Y. and MERHAV, N. (2002). Hidden Markov Processes. *IEEE Trans. Inform. Theory* **48** 1518–1569.
- FINESSO, L. (1991). *Consistent estimation of the order for Markov and hidden Markov chains*. PhD Thesis, University of Maryland.
- GARIVIER, A. (2005). Consistency of the unlimited BIC Context Tree Estimator. Submitted.
- GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**(6):897–906.
- GASSIAT, E. and BOUCHERON, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory*, **49**(4) 964–980.
- HANSEN, M. H. and YU, B. (2001). Model Selection and the Principle of Minimum Description Length. *JASA* **96**(454) 746–774
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The elements of statistical learning*. Springer-Verlag, New-York.
- HENNA, J. (1985). On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.*, **37**(2) 235–240.
- HUGHES, J. P. and GUTTORP, P. (1994). A class of stochastic models for

- relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research* **30** 1535–1546.
- 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.
- JAMES, L. F., PRIEBE, C. E. and MARCHETTE, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.*, **29**(5) 1281–1296.
- JEFFERYS, W. and BERGER, J. (1992). Ockam’s razor and Bayesian analysis. *American Scientist*, **80** 64–72.
- KALEH, G. K. and VALLET, R. (1994). Joint parameter estimation and symbol detection for linear or nonlinear unknown channels. *IEEE Trans. Commun.* **42** 2406–2413.
- KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, **62**(1) 49–66.
- KIEFFER, J. C. (1993). Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory*, **39** 893–902.
- KOSKI, T. (2001). *Hidden Markov Models For Bioinformatics*. Kluwer Academic Publishers Group.
- LEROUX, B. G. (1992a). Maximum-likelihood estimation for Hidden Markov models. *Stochastic Processes Their Applic.* **40** 127–143.
- LEROUX, B. G. (1992b). Consistent estimation of a mixing distribution. *Ann. Statist.*, **20**(3) 1350–1360.
- LEVINSON, S. E., RABINER, L. R. and SONDHI, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal* **62** 1035–1074.
- LIU, C. C. and NARAYAN, P. (1994). Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Trans. Inf. Theory*, **40**(4) 1167–1180.
- MCLACHLAN, G. and PEEL, D. (2000). *Finite mixture models*. Wiley-Interscience, New York.
- 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.
- 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.
- RISSANEN, J. (1978). Modelling by shortest data description. *Automatica*, **14** 465–471.
- RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.*, **14**(3) 1080–1100.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**(2) 461–464.
- 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.