A minimum description length approach to hidden Markov models with Poisson and Gaussian emissions. Application to order identification

A. Chambaz [a,*], A. Garivier [b], E. Gassiat [c]

[a] *MAP5, Université Paris Descartes, France*

[b] *CNRS & TELECOM ParisTech, France*

[c] *Laboratoire de Mathématiques, Université Paris-Sud, France*

Abstract

We address the issue of order identification for hidden Markov models 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. A simulation study and an application to postural analysis in humans are provided.

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

* Corresponding author.
Email addresses: `antoine.chambaz@univ-paris5.fr` (A. Chambaz), `garivier@telecom-paristech.fr` (A. Garivier), `elisabeth.gassiat@math.u-psud.fr` (E. Gassiat).

1 Introduction

Hidden Markov models (HMM) were formally introduced by Baum & Petrie in 1966. Since then, they have 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 into HMM, see (Ephraim & Merhav, 2002; Cappé et al., 2005). Mixture models with independent observations are a particular case of HMMs.

In most practical cases, the *order* of the model (*ie* the true number of hidden states) is unknown and has to be estimated. There is an extensive literature dedicated to the issue of order estimation. The particular case of order estimation for mixtures of continuous densities with independent identically distributed (abbreviated to i.i.d) observations is notoriously challenging (see (Chambaz, 2006) for a comprehensive bibliography). It has been addressed through various methods: ad hoc or minimum distance (Henna, 1985; Chen & Kalbfleisch, 1996; Dacunha-Castelle & Gassiat, 1997; James et al., 2001), maximum likelihood (Leroux, 1992b; Keribin, 2000; Gassiat, 2002; Chambaz, 2006) or Bayesian (Ishwaran et al., 2001; Chambaz & Rousseau, 2007). Actually, Bayesian literature on order selection in mixture models is essentially devoted to determining coherent non informative priors, see for instance (Moreno & Liseo, 2003) and to implementing procedures, see for instance (Mengersen & Robert, 1996). Order estimation in HMMs is much more difficult. It has been proved that, even if the null hypothesis is true, the maximum likelihood test statistic is unbounded (Gassiat & Kéribin, 2000) in the case of independent mixture only if parameters are unbounded, see (Azais et al., 2006) and references therein. This is why the choice of a penalty to obtain estimators using penalized maximum likelihood that do not over-estimate the order is a difficult problem. Earlier results on penalized maximum likelihood estimators (as in (Finesso, 1991)) and Bayesian procedures (as in (Liu & Narayan, 1994)) assume a prior upper bound on the order. In (McKay, 2002), the minimum distance estimator introduced by (Chen & Kalbfleisch, 1996) for mixtures is extended to HMMs. Regarding finite emission alphabet, Kieffer (1993) proves the consistency of the penalized maximum likelihood estimator with penalties increasing exponentially fast with the order with no prior upper bound. In the same context, Gassiat & Boucheron (2003) prove almost sure (abbreviated to "a.s.") consistency with penalties increasing as a power of the order. The question of the minimal penalty which is sufficient to obtain almost sure consistency with no prior upper bound remains open.

In this paper, we address the issue of order identification for HMM with Poisson and Gaussian emissions. In 1978, Rissanen introduced the Minimum Description Length (MDL) principle which connected model selection to coding theory via the following principle: "Choose the model that gives the shortest description of data." We prove here MDL-inspired mixture inequalities which lead to consistent penalized estimators requiring no prior bound on the order.

Let us recall basic ideas that sustain the MDL principle. Given any k-dimensional model (*ie* parametric family of densities indexed by Θ of dimension $k \geq 1$), let E_θ be the expectation with respect to a random variable X_1^n with distribution P_θ, whose 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

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

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*) caused by using the same q for compressing all g_θ ($\theta \in \Theta$).

If one assumes that the maximum likelihood estimator $\widehat{\theta}(X_1^n)$ achieves a $\sqrt{n}$-rate and that there exists a summable sequence $\{\delta_n\}$ of positive numbers which is such that, for every $\theta \in \Theta$,

$$P_\theta \left\{ \sqrt{n} \left\| \widehat{\theta}(X_1^n) - \theta \right\| \geq \log n \right\} \leq \delta_n,$$

then Theorem 1 in (Rissanen, 1986) guarantees that

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

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

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

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 called optimal if it achieves equality in (1). The following 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

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

- mixture coding, where q is a mixture of all densities g_θ $(\theta \in \Theta)$.

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

Now, let us consider the following problem: given a family of models $(\mathcal{M}_i)_{i\in I}$, which best represents some given data x_1^n? The MDL methodology suggests to choose model $\widehat{\mathcal{M}} = \mathcal{M}_{\widehat{i}}$ 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 two optimal coding distributions presented above selects a model:

- two-stage coding chooses

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

  where $\widehat{\theta}_i$ is the maximum likelihood estimator over model $\mathcal{M}_i$;
- mixture coding chooses

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

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

The challenging task is to prove that such estimators are consistent: if x_1^n is emitted 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 a.s. 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 (2006) and Garivier (2005).

Organization of the paper

In Section 2 we prove inequalities that compare maximum likelihood and a particular mixture coding distribution (see Theorems 1 and 2) for HMM mixture models and i.i.d models, with Poisson or Gaussian emissions. In Section 3, these inequalities are used to calibrate a penalty to obtain a.s consistent estimators using penalized likelihood or penalized mixture coding distributions. They require no prior bound on orders (see Theorems 5 and 6). The penalties are heavier than BIC penalties. The question whether BIC penalties lead to consistent estimation of the order remains open. In Section 4, we investigate this question through a simulation study. An application to postural analysis in humans is also presented. Proofs of two lemmas as well as a useful result demonstrated by Leroux (1992a) are contained in 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 non negative integers) is denoted by π_m.

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

We denote by $\mathbb{R}$ the set of real numbers and by $\mathbb{R}_+$ that of non-negative real numbers. 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 possible transition probabilities $\mathbf{p} = (p_{jj'} : j, j' \leq k) \in \mathbb{R}_+^{k^2}$ $(\sum_{j'=1}^k p_{jj'} = 1$ for all $j \leq k)$. Let $\mathcal{C} \subset \mathbb{R}$ be a bounded set. Then the parameter set is

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

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 $\{1, \ldots, k\}$, 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 $X_1^n = (X_1, \ldots, X_n)$ 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,

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

The related mixture statistic is defined by

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

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**.

Denote the positive part of a real number t by $(t)_+$. Let $X_{(n)}$ and $|X|_{(n)}$ be the maxima of $X_1, \ldots, X_n$ and $|X_1|, \ldots, |X_n|$, respectively. Let us also introduce, for all $k, n \geq 1$,

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

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

GE

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

PE

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

Particular case of i.i.d mixture models

The i.i.d mixture model is a particular case of the HMM model. Here, $\{Z_n\}_{n \geq 0}$ is a sequence of i.i.d random variables.

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

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

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$(1/2, \ldots, 1/2)$ distributed,
- $m_1, \ldots, m_k$ are independent, identically distributed with density ϕ_{0,τ^2} in example **GE** and with density Gamma$(\tau, 1/2)$ in example **PE**.

Equality (3) with ν'_k in place of ν_k and Θ'_k in place of Θ_k defines a mixture statistic $q_k(X_1^n)$ in this framework. The second main result is another comparison between the maximum log-likelihood and the mixture statistics in examples **GE** and **PE**.

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

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

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

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 (4), (5), (6), (7), the upper bounds are written as a sum of $\frac{1}{2}\dim(\Theta_k)\log n$, a bounded term and a random term which involves the maximum of $|X_1|, \ldots, |X_n|$. 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** (for HMM and i.i.d mixture models). Indeed, the probability that $|X|_{(n)}$ or $X_{(n)}$ exceeds some level u_n may be written as the expectation of the same probability conditionally on the hidden variables. As soon as this conditional probability has an upper bound that does not depend on the hidden variables, the same upper bound holds for the unconditional probability.

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

First, let us introduce some notations.

For all $\theta \in \Theta_k$ or $\theta \in \Theta'_k$ (any $k \geq 1$), as appropriate, and for all $x_1^n \in \mathcal{X}^n$, $z_0^n = (z_0, \ldots, z_n) \in \{1, \ldots, k\}^{n+1}$, 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 a substitution of $g_\theta(x_1^n|z_1^n)$ for $g_\theta(X_1^n)$.

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 shorthand 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 $j \leq k$ such that $n_j > 0$, let us set

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

By convention, we set $\bar{x}_j = 0$ whenever $n_j = 0$.

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

Straightforwardly, using twice the 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 non negative $\alpha_1, \ldots, \alpha_k$ and positive $\beta_1, \ldots, \beta_k$) yields

$$\begin{aligned}\sup_{\theta\in\Theta_k}\log\frac{g_\theta(x_1^n)}{q_k(x_1^n)} &= \log k + \sup_{\theta\in\Theta_k}\log\frac{\sum_{z_0\le k} g_\theta(x_1^n|z_0)p^o_{z_0}}{\sum_{z_0\le k} q_k(x_1^n|z_0)} \\ &\le \log k + \sup_{\theta\in\Theta_k}\max_{z_0\le k}\log\frac{g_\theta(x_1^n|z_0)p^o_{z_0}}{q_k(x_1^n|z_0)} \\ &\le \log k + \sup_{\theta\in\Theta_k}\max_{z_0\le k}\log\frac{g_\theta(x_1^n|z_0)}{q_k(x_1^n|z_0)} \\ &\le \log k + \sup_{\theta\in\Theta_k}\max_{z_0\le k}\log\frac{\sum_{z_1^n\in\{1,\ldots,k\}^n} g_\theta(x_1^n|z_0^n)P_\theta\{z_1^n|z_0\}}{\sum_{z_1^n\in\{1,\ldots,k\}^n} q_k(x_1^n|z_0^n)q_k(z_1^n|z_0)} \\ &\le \log k + \sup_{\theta\in\Theta_k}\max_{z_0^n\in\{1,\ldots,k\}^{n+1}}\log\frac{g_\theta(x_1^n|z_1^n)}{q_k(x_1^n|z_1^n)}\cdot\frac{P_\theta\{z_1^n|z_0\}}{q_k(z_1^n|z_0)}. \qquad (8)\end{aligned}$$

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

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

where the second inequality is derived from the following Robbins-Stirling approximation formula, valid for all $z>0$,

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

This concludes the study of the second ratio in the right-hand term of (8). 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 $j\le k$, so that the following bound holds for every $x_1^n\in\mathcal{X}^n$ and $z_1^n\in\{1,\ldots,k\}^n$:

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

Besides, simple calculations yield

$$q_k(x_1^n|z_1^n) = \prod_{j=1}^k \frac{1}{(\sigma\sqrt{2\pi})^{n_j}} \int \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{m^2}{2\tau^2} - \frac{1}{2\sigma^2}\sum_{i\in I_j}(x_i - m)^2\right) dm$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{j=1}^k \frac{1}{\sqrt{1+\frac{n_j\tau^2}{\sigma^2}}} \exp\left(-\frac{\sum_{i\in I_j} x_i^2}{2\sigma^2} + \frac{n_j^2}{2\sigma^2(n_j + \frac{\sigma^2}{\tau^2})}(\bar{x}_j)^2\right). \quad (11)$$

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

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

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

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

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

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

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

PE The same argument as in example **GE** implies that, for each $j \le k$, for every $x_1^n \in \mathcal{X}^n$ and $z_1^n \in \{1,\dots,k\}^n$:

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

if $P_n = 1/\prod_{i=1}^n (x_i)!$. In particular, the factor associated with some $j \le k$ for which $\bar{x}_j = 0$ equals one. Furthermore, the following can easily be derived:

$$q_k(x_1^n|z_1^n) = P_n \prod_{j=1}^k \sqrt{\frac{\tau}{2\pi}} \int m^{n_j\bar{x}_j - 1/2} \exp\Big(-(n_j+\tau)m\Big) dm$$

$$= P_n \prod_{j=1}^k \sqrt{\frac{\tau}{2\pi}} \frac{\Gamma(n_j\bar{x}_j + 1/2)}{(n_j+\tau)^{n_j\bar{x}_j+1/2}}. \quad (15)$$

Here, the factor associated with some $j \le k$ 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 $j \le k$, 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, the following holds:

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

(the second inequality follows by convexity). Combining (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.

□

Note that (12) cannot be improved, since equality is attained when the n_j are equal.

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

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

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

In addition,

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

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

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

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 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. □

3 Application to order identification

Let k_0 be the sole integer such that the distribution P_0 of process $\{X_n\}_{n\geq 1}$ 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 to describe the distribution P_0. Our goal in this section is to estimate k_0.

Let us denote by $\mathrm{pen}(n,k)$ a positively valued increasing function of $n, k \geq 1$ such that, for each $k \geq 1$, $\mathrm{pen}(n,k) = o(n)$. We define hereby the estimators:

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

Convenient choices of the penalty term involve the following quantities: 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 $\widehat{k}_n^{\mathrm{ML}}$) *Set $\alpha > 2$, and for each $n \geq 3, k \geq 1$,*

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

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

$$S_{kn} = D_{kn} + 5\sigma^2 k(k+1) \log n,$$

PE

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

Under the assumptions described above, $\widehat{k}_n^{\mathrm{ML}} = k_0$ eventually P_0-a.s.

Similarly,

Theorem 6 (consistency of $\widehat{k}_n^{\mathrm{MIX}}$) *Set $\alpha > 2$, and for each $n \geq 3, k \geq 1$,*

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

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

$$S_{kn} = 5\sigma^2 k(k+1)\log n,$$

PE

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

Under the assumptions described above, $\widehat{k}_n^{\text{MIX}} = k_0$ eventually P_0-a.s.

Theorems 5 and 6 thus guarantee that $\widehat{k}_n^{\text{ML}}$ and $\widehat{k}_n^{\text{MIX}}$ are consistent estimators of k_0. We emphasize that *no prior bound on k_0 is required.*

The penalty function satisfies $\text{pen}(n,k) = O(\log n)$ for every $k \geq 1$ in both examples. 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.

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, 2007) 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 $\widehat{k}_n^{\text{ML}}$ is consistent while $\widehat{k}_n^{\text{MIX}}$ 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 $\widehat{k}_n^{\text{ML}} \geq k_0$ eventually P_0-a.s 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}-a.s 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 more difficult part is to obtain that $\widehat{k}_n^{\text{ML}} \leq k_0$ eventually P_0-a.s.

Let $P_0 = P_{\theta_0}$ for $\theta_0 \in \Theta_{k_0} \setminus \Theta_{k_0-1}$. Let us consider a positively valued sequence $\{t_n\}_{n\geq 3}$ to be chosen conveniently later on. Let $k > k_0$ and $n \geq 3$. Obviously, if $\widehat{k}_n^{\text{ML}} = k$, 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 (4), (5), (6) or (7) (with $\tau = 1/2$ in example **GE** and $\tau = 2$ in example **PE**), $\widehat{k}_n^{\text{ML}} = k$ yields

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

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 $U_n = |X|^2_{(n)}$, $b_{kn} = d_{kn}$ 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. Let us choose $t_n = 5\sigma^2 \log n$ in example **GE** and $t_n = \log n / \sqrt{\log \log n}$ in example **PE**, so that as soon as $U_n \leq t_n$, then

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

Obviously, we have

$$P_0 \left\{ \widehat{k}_n^{\text{ML}} > k_0 \right\} \leq P_0 \left\{ \widehat{k}_n^{\text{ML}} > k_0, U_n \leq t_n \right\} + P_0\{U_n \geq t_n\}. \tag{19}$$

Because q_k defines a probability measure, we have

$$\begin{aligned} &P_0 \left\{ \widehat{k}_n^{\text{ML}} = k, U_n \leq t_n \right\} \\ &\quad \leq \int_{x_1^n \in \mathcal{X}^n} \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \mathbb{1} \left\{ \log \frac{g_{\theta_0}(x_1^n)}{q_k(x_1^n)} \leq \Delta_{nk}, U_n \leq t_n \right\} q_k(x_1^n) d\mu(x_1^n) \\ &\qquad \qquad \leq \exp \left\{ -\frac{\alpha}{2}(k - k_0) \log n \right\}, \end{aligned}$$

hence

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

As a consequence of Lemmas 3 and 4, $P_0\{\widehat{k}_n^{\text{ML}} > k_0\}$ is $O(n^{-\alpha/2} + n^{-3/2})$ in example **GE** and $O(n^{-\alpha/2} + n^{-2})$ in example **PE**: we apply the Borel-Cantelli lemma to complete the proof. □

The proof of Theorem 6 uses the following

Lemma 7 *There exists a sequence $\{\varepsilon_n\}_{n \geq 1}$ of random variables that converges* *to 0 P_0-a.s such that, for any $n \geq 1$, if $\widehat{k}_n^{\text{MIX}} < k_0$ then*

$$\frac{1}{n}\left[\sup_{\theta\in\Theta_k} \log g_\theta(X_1^n) - \log g_{\theta_0}(X_1^n)\right] \geq \varepsilon_n. \tag{20}$$

Proof of Lemma 7. Set $k < k_0$. It is sufficient to show the existence of $\{\varepsilon_n\}_{n\geq 1}$ that converges to 0 P_0-a.s such that, for any $n \geq 1$, $\widehat{k}_n^{\text{MIX}} = k$ implies that (20) holds.

Because $\text{pen}(n, k) = o(n)$ and $\text{pen}(n, k_0) = o(n)$, $\widehat{k}_n^{\text{MIX}} = k$ 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 (20) where

$$\varepsilon_n = \frac{1}{n}\log\frac{\sup_{\theta\in\Theta_k} g_\theta(X_1^n)}{q_k(X_1^n)} - \frac{1}{n}\log\frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)} + o(1).$$

Now, by virtue of (4), (5), (6), (7) and Lemmas 3, 4, P_0-a.s,

$$\frac{1}{n}\log\frac{\sup_{\theta\in\Theta_k} g_\theta(X_1^n)}{q_k(X_1^n)} \underset{n\to\infty}{\longrightarrow} 0.$$

The same inequalities and lemmas also guarantee that, P_0-a.s,

$$\frac{1}{n}\left(\log\frac{g_{\theta_0}(X_1^n)}{q_{k_0}(X_1^n)}\right)_+ \underset{n\to\infty}{\longrightarrow} 0.$$

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-a.s,

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

This completes the proof. $\square$

Proof of Theorem 6. A straightforward combination of Lemma 7 with the strong law of large numbers (in the i.i.d framework) or Lemma 8 from the

Appendix (in the HMM framework) yields that $\widehat{k}_n^{\text{MIX}} \geq k_0$ eventually P_0-a.s.

From now on, we use the same notations as those used in the preceding proof except when notified. Let $k > k_0$. If $\widehat{k}_n^{\text{MIX}} = k$, then

$$-\log q_k(X_1^n) + \text{pen}(n,k) \leq -\log q_{k_0}(X_1^n) + \text{pen}(n,k_0).$$

By using (4), (5), (6), (7), the latter inequality 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, as soon as $U_n \leq t_n$, one has (18). Consequently,

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

The result follows by virtue of the Borel-Cantelli lemma, the previous bound and Lemmas 3, 4: $\widehat{k}_n^{\text{MIX}} \leq k_0$ eventually P_0-a.s. $\square$

4 Simulations and experimentation

In this section, we focus on the penalized maximum likelihood estimator $\widehat{k}_n^{\text{ML}}$. In Section 4.1 we investigate the importance of the choice of the penalty term. We first illustrate that the penalty given in Theorems 5 and 6 is heavy enough to obtain a.s consistency with no prior upper bound. Then we try to understand whether a smaller penalty could be chosen to retain a.s consistency in the same context. Section 4.2 is dedicated to the presentation of an application to postural analysis in humans within framework **GE**. In order to compute the maximum likelihood estimates, we use standard EM algorithm (Baum et al., 1970; Cappé et al., 2005). The algorithm is run with several random starting points, and iterations are stopped whenever the parameter estimates hardly differ from one iteration to the other.

4.1 A simulation study of the penalty calibration

We first propose to illustrate the a.s convergence of $\widehat{k}_n^{\mathrm{ML}}$ in a toy-model of HMM with Poisson emissions. We simulate 5 samples of distribution P_θ for $\theta = (\mathbf{p}, \mathbf{m}) \in \Theta_6$, where $m_j = 3j$ (each $j \leq 6$), and $p_{6,1} = 1$, $p_{j,j+1} = 1 - p_{j,1} = 0.9$ (each $j \leq 5$). As estimator $\widehat{k}_n^{\mathrm{ML}}$ requires no upper bound on the order, the question arises to determine at which values of k the penalized maximum likelihood should be evaluated. Figure 1 illustrates the behavior of criterion $-\sup_{\theta\in\Theta_k} \log g_\theta(x_1^n) + \mathrm{pen}(n, k)$ with a sample size $n = 1,000$ versus the number k of hidden states. The criterion looks very regular: it first decreases rapidly, then stabilizes, and finally increases slowly but systematically. Thus, identifying the maximizer $\widehat{k}_n^{\mathrm{ML}}$ is an easy task. The values of $\widehat{k}_n^{\mathrm{ML}}$ are displayed in Figure 2. We emphasize that only under-estimation and never over-estimation occur with our choice of penalty. This may indicate that our penalty is as small as possible.

We also study the examples considered in Section 5 (pp 582–585) of (McKay, 2002). As expected, estimator $\widehat{k}_n^{\mathrm{ML}}$ has a good behavior for sample sizes which are large enough. Figure 3 represents the evolution of the penalized maximum likelihood criteria $-\sup_{\theta\in\Theta_k} \log g_\theta(x_1^n) + \mathrm{pen}(n, k)$ for $k \leq 4$ as the sample size n grows for a realization x_1^n of the so-called "well separated, unbalanced"

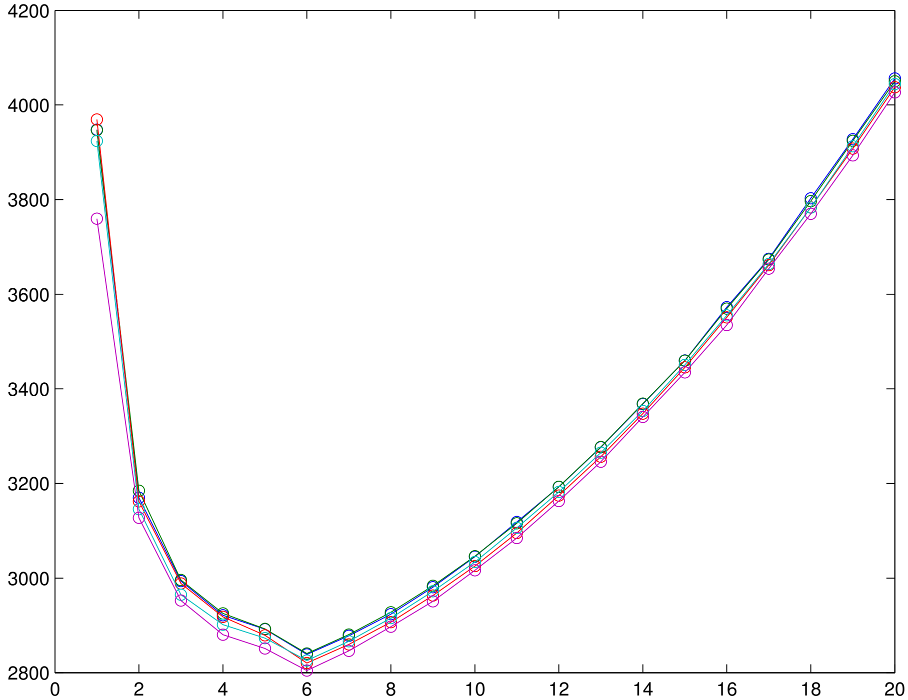

Fig. 1. For each of 5 samples of length $n = 1,000$, penalized maximum likelihood criteria $-\sup_{\theta\in\Theta_k} \log g_\theta(x_1^n) + \mathrm{pen}(n,k)$ for k varying from 1 to 20.

model of order 2 taken from Section 5 in (McKay, 2002).

For small samples, smaller models are systematically chosen, and this agrees with our presumption that our penalty is too heavy. Note that the BIC criterion suffers from the same defect, as can be seen in Figure 2 of (McKay, 2002). In that perspective, one may search for some minimal penalty leading to a consistent estimator. We address this issue by computing the differences $[\sup_{\theta\in\Theta_2} \log g_\theta(x_1^n) - \sup_{\theta\in\Theta_k} \log g_\theta(x_1^n)]$ for $k = 1, 3, 4$, see Figure 4. For $k = 1$, the difference grows linearly so that any sub-linear penalty prevents from under-estimation (see also the beginning of the proof of Theorem 5). For $k = 3, 4$, the differences seem almost constant in expectation. A convenient penalty should dominate (eventually almost surely) their extreme values. For instance, it is proved in (Chambaz, 2006) that a $\log\log n$ penalty guarantees

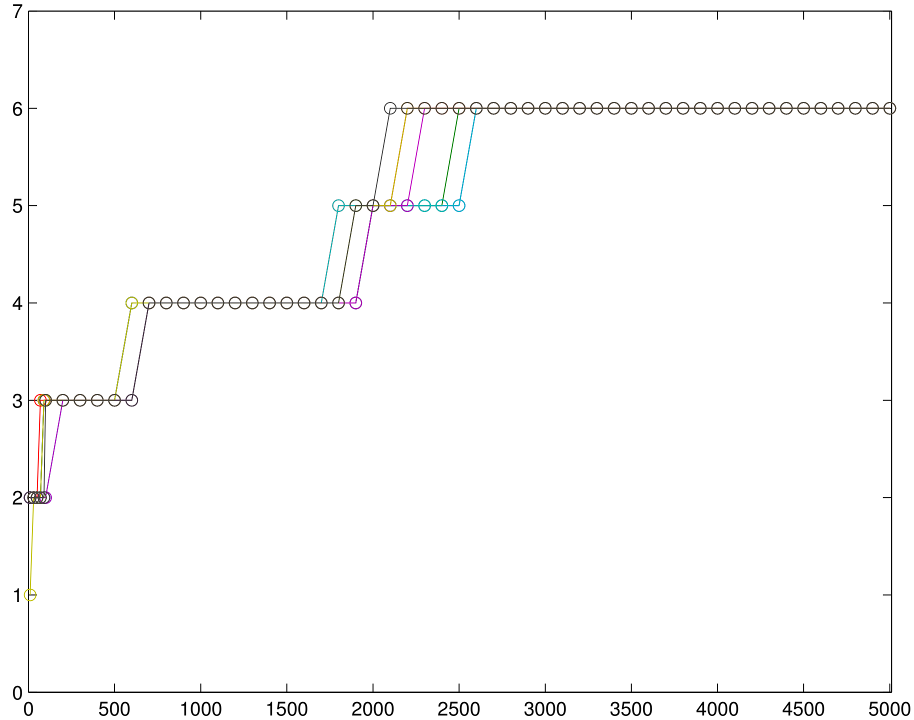

Fig. 2. Almost sure convergence of $\widehat{k}_n^{\mathrm{ML}}$. As the sample size grows (x-axis), the values of $\widehat{k}_n^{\mathrm{ML}}$ (y-axis) increase to the true order $k_0 = 6$.

consistency when an upper bound on the order is known. Without such a bound, it remains open whether a $\log n$ penalty is optimal or not.

4.2 Application to postural analysis in humans

Maintaining posture efficiently is achieved by dynamically resorting to the best available sensory information. The latter is divided in three categories: vestibular, proprioceptive, and visual information. Every individual has developed his/her own preferences according to his/her sensorimotor experience.

Sometimes, a sole kind of information –usually, visual– is processed in all situations. This occurs in healthy individuals, but it is more common in elderly people, in people having suffered from a stroke, in people afflicted by Parkinson

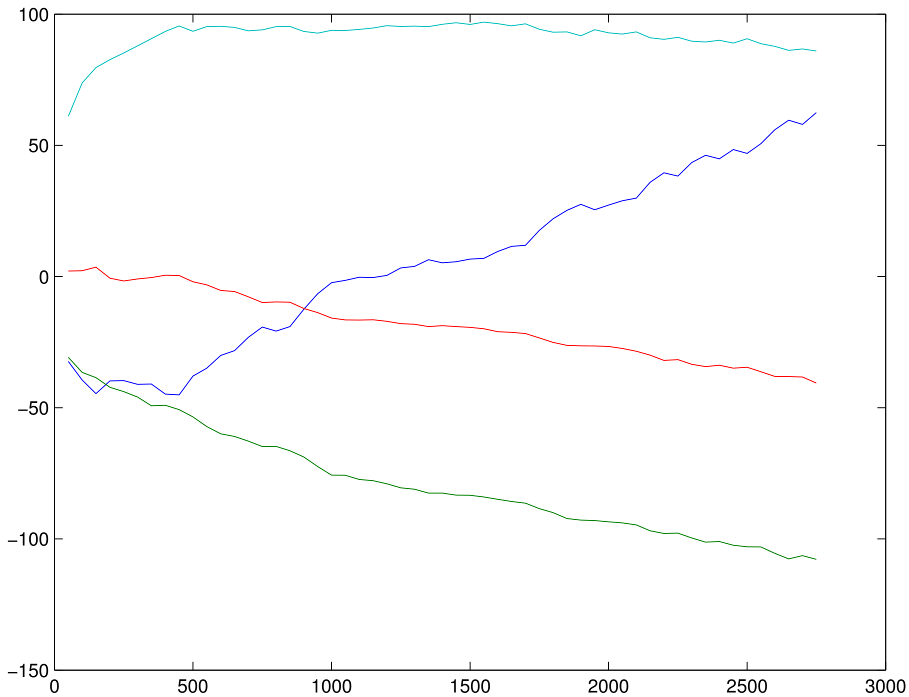

Fig. 3. Values of $-\sup_{\theta\in\Theta_k}\log g_\theta(x_1^n)+\mathrm{pen}(n,k)$ $(k\leq 4)$ as n grows. From top to bottom, for large values of n: $k=4$, $k=1$, $k=3$, $k=2$.

Disease for instance. Although processing a sole kind of information may be efficient for maintaining posture in one's usual environment, it is likely not to be adapted to new or unexpected situations, and may result in a fall. Therefore, it is of primordial importance to learn how to detect such a sensory typology, so as to propose an adapted reeducation program.

Postural analysis in humans at stable equilibrium has already been addressed using fractional Brownian motion (see (Bardet & Bertrand, 2007) and references therein), or diffusion processes (Rozenholc et al., 2007). We illustrate now how the study of this difficult issue can be addressed within the theoretical framework of HMM with Gaussian emission. Data are collected during a 70-second experiment. Every $\Delta = 0.025$ second, the position where a control subject exerts maximal pressure on a force platform is recorded. We denote by T_n the distance between the latter at time $n\Delta$ and a reference position.

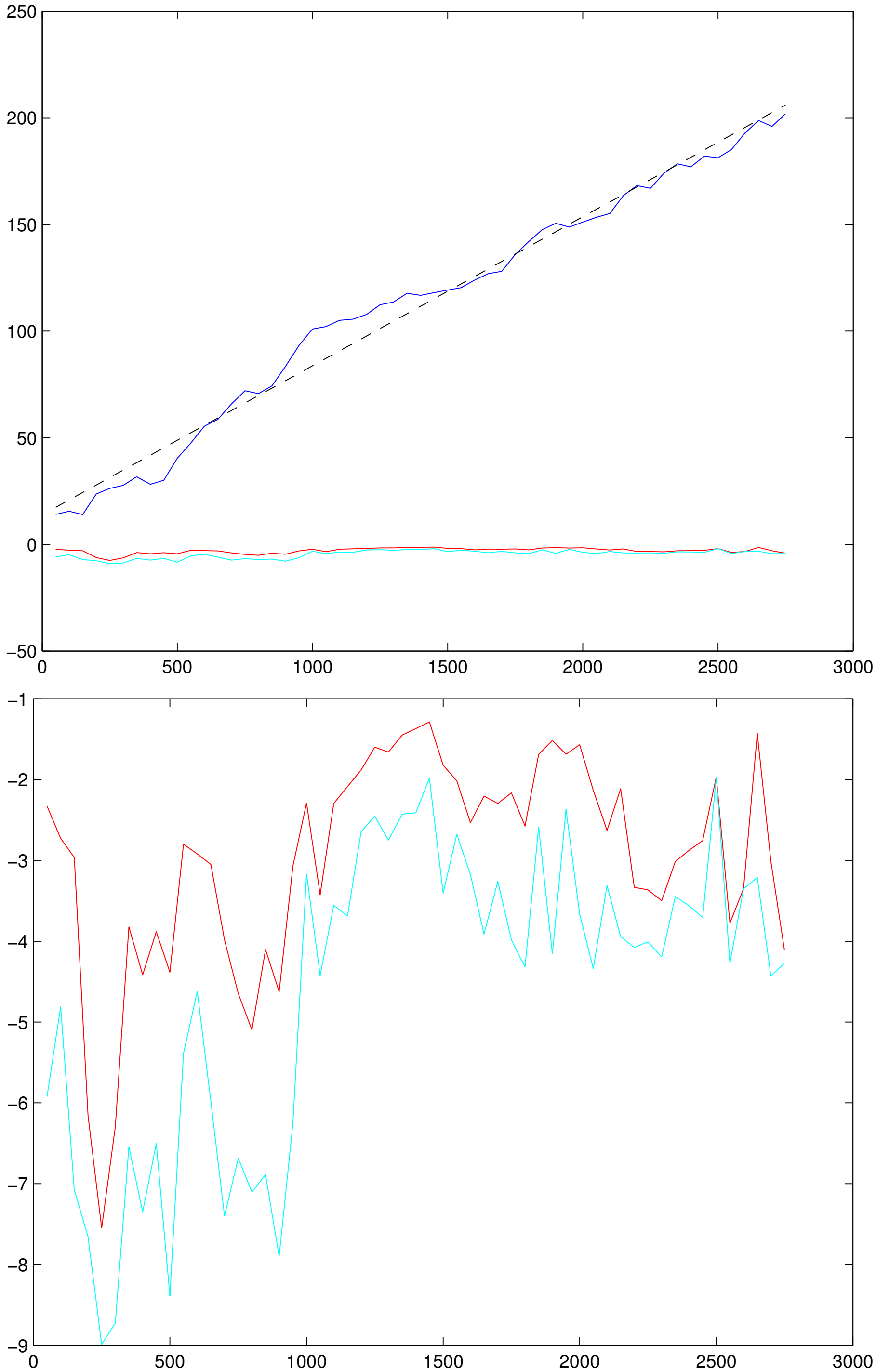

Fig. 4. Representation of differences $[\sup_{\theta\in\Theta_2}\log g_\theta(x_1^n) - \sup_{\theta\in\Theta_k}\log g_\theta(x_1^n)]$ (for $k = 1, 3, 4$) as n grows. Top: all curves (from top to bottom, for large n: $k = 1$, $k = 3$, $k = 4$). Bottom: curves for $k = 3, 4$ (from top to bottom for large n: $k = 3$, $k = 4$; note the change in scale along y-axis).

The experimental protocol we choose to present here is decomposed into three phases: a phase of 35 seconds during which the subject's balance is perturbed (by vibratory stimulation of the left tendon, known to force to tilt forward) is preceded by 15 seconds and followed by 20 seconds of recording without stimulation.

According to the medical background and a preliminary analysis, the process $(X_n)_{n\geq 1}$ of interest derives from the differenced process $(\nabla T_n)_{n\geq 1} = (T_{n+1} - T_n)_{n\geq 1}$, which is arguably stationary: for all $n \geq 1$,

$$X_n = \log\{(\nabla T_n)^2\}$$

(in any continuous model, $\nabla T_n = 0$ has probability 0). We hereafter assume that $(X_n)_{n\geq 1}$ is a HMM with Gaussian emission. Heuristically, we focus on the evolution of the volatility of process $(T_n)_{n\geq 1}$.

The estimated order $\widehat{k}_n^{\mathrm{ML}}$ equals 3. The result coincides with that of the BIC criterion. In order to compute $\widehat{k}_n^{\mathrm{ML}}$, we estimated σ on an independent experiment (same subject, eyes open, no perturbation). We assume that the variance of the volatility process remains the same all over the three-phase experiment. We are also interested in the inference concerning the unobservable sequence of hidden states. We compute the *a posteriori* most likely sequence of states by the Viterbi algorithm. In words, we find the sequence z_1^n which maximizes (with respect to $\xi_1^n \in \{1, 2, 3\}^n$) the joint conditional probability $P_{\widehat{\theta}}\{Z_1^n = \xi_1^n | X_1^n = x_1^n\}$, $\widehat{\theta} \in \Theta_3$ denoting the value of θ output by the EM algorithm on that model. Figure 5 represents the data and z_1^n.

Sequence z_1^n carries (non distributional) information about the model, and helps interpreting the event "$\widehat{k}_n^{\mathrm{ML}} = 3$". The three hidden states HMM proves very satisfactory from a medical point of view. Figure 5 suggests the following interpretation: a reference behavior in standard conditions of standing up (time intervals $[0; 15]$ and $[\sim 65; 70]$) is a combination of two regimes (indexed by 1 and 2); a learning behavior to adapt to new conditions when standing up corresponds to the third regime (indexed by 3). The first, second, and third regimes are respectively associated with medium ($m_1 = -3.90$), small ($m_2 = -6.13$), and large ($m_3 = -1.52$) volatility for process $(T_n)_{n\geq 1}$. The empirical proportions $\hat{\pi}_i(\xi)$ of each regime $\xi \in \{1, 2, 3\}$ on each phase $i \in \{1, 2, 3\}$ are as follows: $\hat{\pi}_1(1) = 0.69$, $\hat{\pi}_1(2) = 0.31$, $\hat{\pi}_1(3) = 0$; $\hat{\pi}_2(1) = 0.64$, $\hat{\pi}_2(2) = 0.04$, $\hat{\pi}_2(3) = 0.32$; $\hat{\pi}_3(1) = 0.50$, $\hat{\pi}_3(2) = 0.25$, $\hat{\pi}_3(3) = 0.25$.

The whole description (characterization of the three regimes and their succession through the duration of the experiment) coincides with the expectations of the medical team.

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 $Y_1, \ldots, Y_n$,

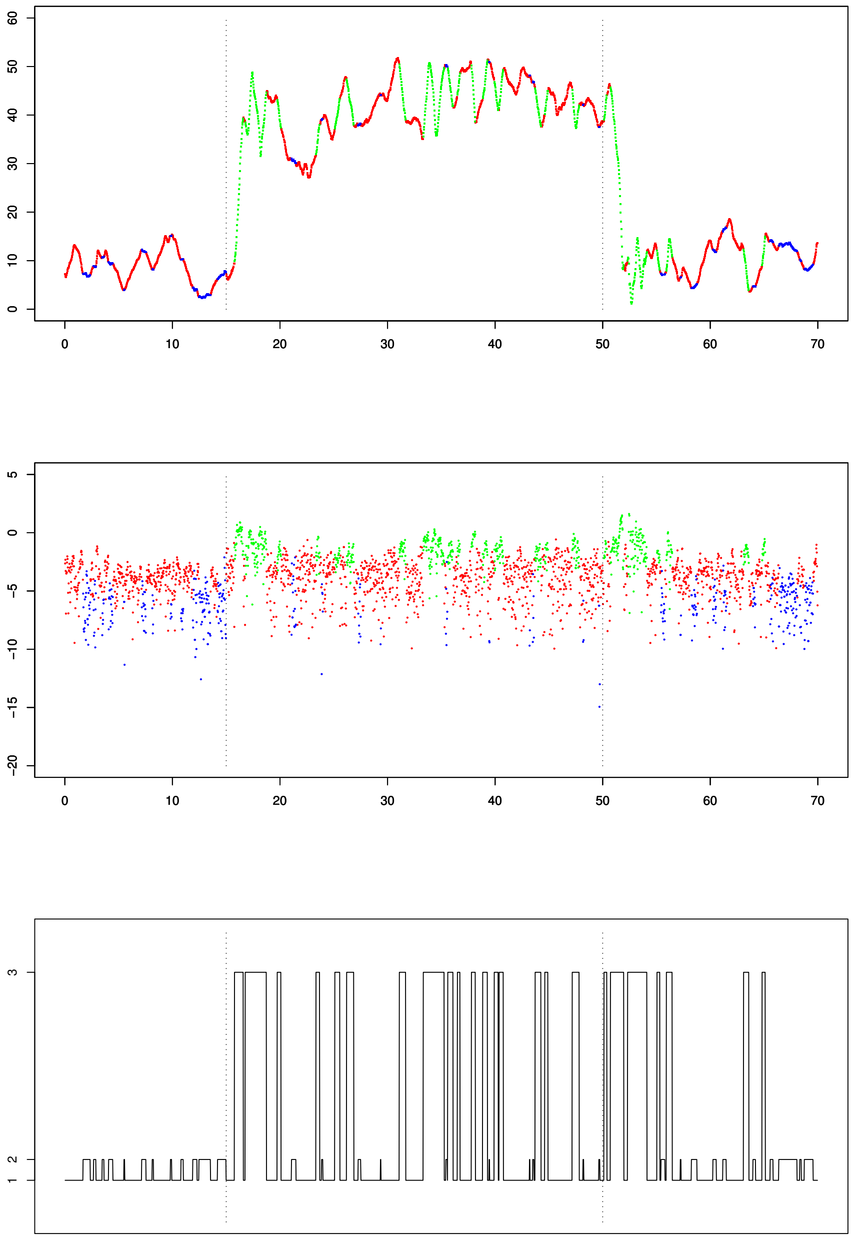

Fig. 5. Realizations t_1^n (top) and x_1^n (middle), and *a posteriori* most likely sequence of hidden states z_1^n (bottom). The vertical dotted lines indicate the limits of the vibratory stimulation phase. Top: points $(n\Delta, t_n)$. Middle: points $(n\Delta, x_n)$. Bottom: points $(n\Delta, z_n)$; each of the three postulated hidden states is associated with a particular level of volatility for ∇T_n. Note that the scale on the y-axis is not linear.

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

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,

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

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 (A.1). $\square$

B A useful lemma for HMM mixture models

Lemma 8 (Leroux) *For* HMM *mixture models with bounded parameter sets, 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}*-a.s,* $n^{-1}[\log g_{\theta_0}(X_1^n) - \log g_\theta(X_1^n)]$ *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 $\mathcal{O}_\theta$ 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 $\mathcal{O}_{\theta_1}, \ldots, \mathcal{O}_{\theta_I}$ (each of them associated with $\varepsilon_i > 0$) and therefore

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

□

Acknowledgment

We want to thank Isabelle Bonan (LNRS, UMR CNRS 7060, Université Paris Descartes, and APHP Lariboisière–Fernand-Widal) for providing us with the postural analysis problem and dataset. We also want to thank the associate editor and referees for their suggestions.

References

Azais, J.-M., Gassiat, E. and Mercadier, C. (2006). Asymptotic distribution and power of the likelihood ratio test for mixtures: bounded and unbounded case. *Bernoulli*, **12**(5) 775–799.

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

Bardet, J-M., Bertrand, P. (2007). Identification of the multiscale fractional brownian motion with biomechanical applications. *J. Time Ser. Anal.*, **28** 1–52.

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.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41** 164–171.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag.

Chambaz, A. (2003). Testing the order of a model. *Ann. Statist.*, **34** (3) 1166–1203.

Chambaz, A. and Rousseau, J. (2007). Bounds for Bayesian order identification with application to mixtures. To appear in *Ann. Statist.*

Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canad. J. Statist.*, **24** (2), 167–175.

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. (2006). Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL. *IEEE Trans. Inf. Theory*, **3** 123–145.

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. (2006). Consistency of the unlimited BIC Context Tree Estimator. *IEEE Trans. Inform. Theory*, **52**(10) 4630–4635.

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.

Gassiat, E. and Kéribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM P&S, 2000.*

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 Ressources 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. *Canad. J. Statist.* , **30**(4) 573–589.

MCKAY, R. (2002). Estimating the order of a hidden Markov model. , **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.

ROZENHOLC, Y., CHAMBAZ, A., BONAN, I. (2007). Penalized nonparametric mean square estimation for diffusion processes with application to postural analysis in Human. *Proceedings of ISI 2007 held in Lisboa.*

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.