ESAIM: PS January 2009, Vol. 13, p. 38–50
DOI: 10.1051/ps:2007048

ESAIM: Probability and Statistics
www.esaim-ps.org

NUMBER OF HIDDEN STATES AND MEMORY: A JOINT ORDER ESTIMATION PROBLEM FOR MARKOV CHAINS WITH MARKOV REGIME

Antoine Chambaz[1] and Catherine Matias[2]

Abstract. This paper deals with order identification for Markov chains with Markov regime (MCMR) in the context of finite alphabets. We define the joint order of a MCMR process in terms of the number k of states of the hidden Markov chain and the memory m of the conditional Markov chain. We study the properties of penalized maximum likelihood estimators for the unknown order (k, m) of an observed MCMR process, relying on information theoretic arguments. The novelty of our work relies in the joint estimation of two structural parameters. Furthermore, the different models in competition are not nested. In an asymptotic framework, we prove that a penalized maximum likelihood estimator is strongly consistent without prior bounds on k and m. We complement our theoretical work with a simulation study of its behaviour. We also study numerically the behaviour of the BIC criterion. A theoretical proof of its consistency seems to us presently out of reach for MCMR, as such a result does not yet exist in the simpler case where $m = 0$ (that is for hidden Markov models).

Mathematics Subject Classification. 62B10, 62B15, 62M07.

Received October 17, 2006. Revised July 6, 2007.

1. Introduction

Markov chains with Markov regime

Let $\mathcal{X} = \{1, \ldots, k\}$ and $\mathcal{Y} = \{1, \ldots, r\}$ be two finite sets and m be some integer. Here, $\mathbb{N}^\star$ denotes the set of positive integers and for any $i \le j$, we use x_i^j to denote the sequence $x_i, x_{i+1}, \ldots, x_j$. We consider a process $\{X_j, Y_j\}_{j\ge1}$ on $(\mathcal{X} \times \mathcal{Y})^{\mathbb{N}^\star}$ with distribution as follows. Process $\{X_j\}_{j\ge1}$ is a Markov chain with memory one on $\mathcal{X}$ with transition matrix $A = (a(i,j))_{1\le i,j\le k}$. Besides, conditionally on $\{X_j\}_{j\ge1}$, process $\{Y_j\}_{j\ge1}$ is a Markov chain with memory m [abbreviated to MC(m)], and the conditional distribution of Y_s conditional on $(\{X_j\}_{j\ge1}, \{Y_j\}_{j<s})$ is given by $b(Y_s|Y_{s-m}^{s-1}, X_s)$, for any $s > m$. The process has some initial distribution μ on $\mathcal{X} \times \mathcal{Y}^m$.

Keywords and phrases. Markov regime, order estimation, hidden states, conditional memory, hidden Markov model.

[1] Laboratoire MAP5, UMR CNRS 8145, Université René Descartes, 45 rue des Saints-Pères, 75270 Paris Cedex 06, France; Antoine.Chambaz@univ-paris5.fr

[2] Laboratoire Statistique et Génome, UMR CNRS 8071, Tour Évry 2, 523 pl. des Terrasses de l'Agora, 91000 Évry, France; matias@genopole.cnrs.fr

Article published by EDP Sciences

© EDP Sciences, SMAI 2009

The set $\Pi^{k,m}$ denotes the set of all such probability measures $\mathbb{P}$ on $(\mathcal{X}\times\mathcal{Y})^{\mathbb{N}^\star}$ formally described by, for all $n\in\mathbb{N}^\star$ and $(x_1^n,y_1^n)\in(\mathcal{X}\times\mathcal{Y})^n$,

$$\mathbb{P}(x_1^n,y_1^n)=\mu(x_1,y_1^m)\left\{\prod_{i=1}^{n-1}a(x_i,x_{i+1})\right\}\left\{\prod_{i=m+1}^{n}b(y_i|y_{i-m}^{i-1};x_i)\right\}. \tag{1}$$

Let us denote by $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y}^m)$ the set of probability measures on $\mathcal{X}\times\mathcal{Y}^m$. The set $\Pi^{k,m}$ is naturally parametrized by $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y}^m)\times\Theta^{k,m}$, where

$$\Theta^{k,m}=\Bigg\{\theta=(A,B)\ :A=(a(i,j))_{1\le i,j\le k},\ a(i,j)\ge 0,\ \sum_{j=1}^{k}a(i,j)=1 \text{ and}$$
$$B=(b(y|y_1^m;x))_{y\in\mathcal{Y},y_1^m\in\mathcal{Y}^m,x\in\mathcal{X}};\ b(y|y_1^m;x)\ge 0,\ \sum_{y=1}^{r}b(y|y_1^m;x)=1\Bigg\}. \tag{2}$$

Thus, $\Pi^{k,m}=\left\{\mathbb{P}=\mathbb{P}_{\mu,\theta}:(\mu,\theta)\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y}^m)\times\Theta^{k,m}\right\}$. Moreover, for stationary processes with stationary measure π_θ on $\mathcal{X}\times\mathcal{Y}^m$, we use the notation $\mathbb{P}_\theta=\mathbb{P}_{\pi_\theta,\theta}$ to remind that the initial probability is fixed.

The observations consist in $\{Y_j\}_{1\le j\le n}$ which is called a Markov chain with Markov regime (abbreviated to MCMR). Note that $\{Y_j\}_{j\ge1}$ is not a Markov process. We assume that its distribution is the marginal onto $\mathcal{Y}^n$ of some $\mathbb{P}_{\theta_0}$ (θ_0 is the true and unknown parameter value), which is stationary, ergodic and belongs to Π^{k_0,m_0} for some unknown $(k_0,m_0)\in\mathbb{N}^\star\times\mathbb{N}$. In other words, it is assumed that there exists a hidden stationary process $\{X_j\}_{j\ge1}$ such that the complete process $\{(X_j,Y_j)\}_{j\ge1}$ has distribution $\mathbb{P}_{\theta_0}\in\Pi^{k_0,m_0}$. When there is no ambiguity, $\mathbb{P}_{\theta_0}$ will abbreviate to $\mathbb{P}_0$. In this setup, the cardinality r of the observed alphabet is known.

While HMMs can model the heterogeneity of a sequence by distinguishing different segments with different i.i.d. distributions (*i.e.* $m=0$), MCMRs enable furthermore a Markovian modelling of each segment ($m\ge1$). HMMs and MCMRs are widely used in practical applications among which genomics, econometrics and speech recognition. We refer to [4,8] for recent and comprehensive overviews on this topic. Note that more flexibility could be added to these models by authorising different memory lengths for the different regimes but the choice of these lengths is a problem which is as delicate as the one we address here.

When the couple (k_0,m_0) associated with the distribution $\mathbb{P}_0$ of a MCMR is *a priori* known, inference on the parameters has been investigated to a great extent (most recent results can be found in [10]). However, in many applications where MCMR are used as a modeling device, there is no clear indication about a good choice for (k_0,m_0). So, inference about (k_0,m_0) is a crucial issue, for even consistency may fail to hold in a wrong model. In this paper, we propose a sound definition of the *order* of a MCMR which we substitute to (k_0,m_0) as main quantity of interest. We explain why below.

Defining the order of a MCMR

Model selection for MCMRs already appears in [3]. The authors propose a reversible jump MCMC procedure to select the memory m as well as the number of regimes k. However, no simulations were given to establish the correctness of the procedure (the method was rather directly applied to real biological data) and it is still an open question to know whether such a procedure is consistent or not.

Model selection for HMMs is a more widely studied subject (see for instance [9,11,16,17,21,22]). The order of a HMM simply is the minimal number of hidden states (here $m=0$). Our approach to model selection for MCMRs draws its inspiration from [11].

One of the interesting problems raised by HMM modeling is the question of identifiability: when do two different Markov chains generate the same stochastic process? This question first raised by [1] can be solved for HMM using linear algebra (see [9,13]). To our knowledge, such a complete solution does not exist in the context of MCMR models. As an immediate consequence, the definition of the order of a MCMR has to be clarified.

In the convenient case where each model $\mathcal{M}_\alpha$ is characterized by $\alpha \in \mathbb{N}$, the order of the distribution of the observations is the smallest α such that this distribution belongs to $\mathcal{M}_\alpha$. This definition is motivated by the will to guarantee that the statistician is looking for the *most economical* representation of the process (the number of parameters required for its description is minimized). In contrast, the definition of the order may be more involved when the above notion of minimality does not have a natural meaning anymore. Two examples follow.

First, order identification for autoregressive moving average ARMA(p, q) models is a well-known example where the structural parameter is bivariate (see for example [12,20]). Nevertheless, this problem is very different from the one studied here because there exists a minimal representation (p_0, q_0) thus defined as the true one. Indeed, the spectral density of an ARMA process admits a unique representation of the form $\lambda \mapsto |Q/P(\mathrm{e}^{-\mathrm{i}\lambda})|^2/2\pi$ where P and Q are polynomial functions with no common factors, $P(z) \neq 0$, for all $|z| \leq 1$ and $Q(z) \neq 0$, for all $|z| < 1$. Then the true order of the ARMA process is defined as the couple (p_0, q_0) of degrees of the polynomials P and Q respectively.

Second, when dealing with model selection for context trees, the order to be selected is a tree. However, there exists a natural ordering (given by the inclusion) which is not a total ordering. Csiszàr and Talata [5] establish the consistency of both penalized (with Bayesian Information Criterion, alias BIC, penalization) maximum likelihood and minimum description length procedures.

A particularity of MCMR modeling is that the sets $\Pi^{k,m}$ are not globally nested, even though $\{\Pi^{k,m}\}_{k\geq 1}$ and $\{\Pi^{k,m}\}_{m\geq 0}$ are nested. In general, for a given probability $\mathbb{P} \in \cup_{(k,m)\in\mathbb{N}^\star\times\mathbb{N}}\Pi^{k,m}$, there is no unique $(k_0, m_0) \in \mathbb{N}^\star \times \mathbb{N}$ such that $\mathbb{P} \in \Pi^{(k_0,m_0)}$ and $\mathbb{P}$ does not belong to any of its subsets (that is, for any $(k, m) \in \mathbb{N}^\star \times \mathbb{N}$ such that $(k < k_0, m = m_0)$, or $(k = k_0, m < m_0)$, or $(k < k_0, m < m_0)$, one has $\mathbb{P} \notin \Pi^{k,m}$).

So, we decide to rely on the point of view of minimizing the number of parameters in order to determine which of the possibly multiple representations is to be selected. Let us denote by $N(k, m)$ the number of parameters required to describe an element of $\Theta^{k,m}$

$$N(k, m) = \dim(\Theta^{k,m}) = k(k - 1) + kr^m(r - 1). \tag{3}$$

This induces an ordering of the set $\mathbb{N}^\star \times \mathbb{N}$. For all $(k_1, m_1), (k_2, m_2) \in \mathbb{N}^\star \times \mathbb{N}$,

$$(k_1, m_1)\prec(k_2, m_2) \quad \text{if and only if} \quad \{N(k_1, m_1) < N(k_2, m_2)\} \text{ or } \{N(k_1, m_1) = N(k_2, m_2) \text{ and } k_1 < k_2\}.$$

Note that we made an arbitrary choice between k and m to get a total order. Obviously, all the results remain valid when using m instead of k. In the following, $a\succ b$ means $b\prec a$ and $a\preccurlyeq b$ means ($a\prec b$ or $a = b$).

We are now able to define the true order (k_0, m_0) of a probability $\mathbb{P}$ belonging to $\cup_{k\geq 1,m\geq 0}\Pi^{k,m}$ as

$$(k_0, m_0) = \min\left\{(k, m) \in (\mathbb{N}^\star \times \mathbb{N}, \prec) : \mathbb{P} \in \Pi^{k,m}\right\}.$$

Content of the paper

Next, we tackle the issue of estimating the true order of a MCMR by penalized maximum likelihood procedure. In Section 2, we introduce our penalized maximum likelihood estimator and two others code-based estimators. The two latter are not computable in practice, but their behaviour is strongly connected to that of our estimator. Its strong consistency (as well as that of the two other estimators) is established in two steps in Section 3: Section 3.1 is dedicated to overestimation and Section 3.2 to underestimation. We present in Section 4 the results of a simulation study.

2. Estimation procedure

The general form of our estimators writes as

$$\widehat{(k,m)}_n = \underset{(k,m)\in(\mathbb{N}^\star\times\mathbb{N},\prec)}{\operatorname{argmin}} \Big(-\log \mathbb{Q}_{k,m}(Y_1^n) + \operatorname{pen}(n,k,m) \Big), \tag{4}$$

where $\mathbb{Q}_{k,m}$ is a (coding) measure on $\mathcal{Y}^n$ and $\operatorname{pen}(n,k,m)$ is a penalty term. Three different coding measures are considered: $\mathrm{KT}_{k,m}, \mathrm{NML}_{k,m}$ and $\mathrm{ML}_{k,m}$ defined below.

Let us consider the distribution density $\nu_{k,m}$ on $\Theta^{k,m}$, given for all $\theta \in \Theta^{k,m}$ by

$$\nu_{k,m}(\theta) = \prod_{i=1}^{k} \frac{\Gamma(k/2)\Gamma(r/2)}{\Gamma(1/2)^k\Gamma(1/2)^r} \left(\prod_{j=1}^{k} \frac{1}{a(i,j)^{1/2}} \right) \left(\prod_{t_1^m\in\mathcal{Y}^m} \prod_{t=1}^{r} \frac{1}{b(t|t_1^m;i)^{1/2}} \right),$$

where $\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\mathrm{d}x$.

The Krichevsky-Trofimov mixture is the probability measure $\mathrm{KT}_{k,m}$ on $(\mathcal{X}\times\mathcal{Y})^{\mathbb{N}^\star}$ whose marginals have density

$$(x_1^n, y_1^n) \mapsto \int_{\theta\in\Theta^{k,m}} \mathbb{P}_{\bar{\mu}^X\otimes\bar{\mu}^{Y,m},\theta}(x_1^n,y_1^n)\nu_{k,m}(\theta)\mathrm{d}\theta, \tag{5}$$

where $\bar{\mu}^X$ and $\bar{\mu}^{Y,m}$ are the uniform distributions on $\mathcal{X}$ and $\mathcal{Y}^m$, respectively. Note that we use for simplicity of notation the same symbol for the probability measure and its marginals on $\mathcal{Y}^n$. The maximum likelihood ($\mathrm{ML}_{k,m}$) and the normalized maximum likelihood ($\mathrm{NML}_{k,m}$) coding measures are defined in a natural way:

$$\mathrm{ML}_{k,m}(y_1^n) = \sup_{\theta\in\Theta^{k,m}} \mathbb{P}_\theta(y_1^n),$$

and if we set $\mathcal{C}(n,k,m) = \sum_{y_1^n\in\mathcal{Y}^n} \sup_{\theta\in\Theta^{k,m}} \mathbb{P}_\theta(y_1^n)$, then

$$\mathrm{NML}_{k,m}(y_1^n) = \sup_{\theta\in\Theta^{k,m}} \frac{\mathbb{P}_\theta(y_1^n)}{\mathcal{C}(n,k,m)} = \frac{\mathrm{ML}_{k,m}(y_1^n)}{\mathcal{C}(n,k,m)}\cdot$$

We will use later that $\mathrm{KT}_{k,m}$ and $\mathrm{NML}_{k,m}$ (but not $\mathrm{ML}_{k,m}$) are probability measures.

The so-called penalized maximum likelihood estimator of the order that we focus on corresponds to the coding measure $\mathrm{ML}_{k,m}$ and to a particular choice of penalty. It is computable, contrarily to the estimators based on $\mathrm{KT}_{k,m}$ and $\mathrm{NML}_{k,m}$ (which are not computable for large sample sizes even in the HMM framework). Nevertheless, studying the two latter is important here because coding measures $\mathrm{KT}_{k,m}$ and $\mathrm{NML}_{k,m}$ are strongly related to $\mathrm{ML}_{k,m}$ (see Lem. 3.4). Note finally that Liu and Narayan dedicated an article [16] to the asymptotic study of the order estimator based on $\mathrm{KT}_{k,m}$ in the HMM framework.

3. Consistency issue

This section is dedicated to the statement and proof of the main consistency result.

Theorem 3.1. *Let $\mathbb{P}_0$ be stationary, ergodic and belong to $\cup_{k\geq1,m\geq0}\Pi^{k,m}$ with unknown true order (k_0,m_0). Let $\{Y_j\}_{1\leq j\leq n}$ be a stationary process drawn from the marginal of $\mathbb{P}_0$ on $\mathcal{Y}^n$.*

Let us denote by φ an increasing function which maps $(\mathbb{N}^\star \times \mathbb{N}, \prec)$ to $\mathbb{N}$. Let us choose $\alpha > 1$ and introduce, for all $n \in \mathbb{N}^\star$, $k \geq 1$ and $m \geq 0$,

$$\tau(n,k,m) = \max\Bigg(0, \log k + m \log r - k \log \frac{\Gamma(k/2)}{\Gamma(1/2)} - kr^m \log \frac{\Gamma(r/2)}{\Gamma(1/2)} + \frac{k^2(k-1)}{4n} + \frac{kr^{m+1}(r-1)}{4n} + \frac{5k}{24n}(1+r^m)\Bigg). \quad (6)$$

Let $\widehat{(k,m)}_n$ be defined by (4), with $\mathbb{Q}_{k,m} = \mathrm{ML}_{k,m}$ and

$$\mathrm{pen}(n,k,m) = \sum_{(k',m')\preccurlyeq(k,m)} \left(\frac{1}{2}N(k',m')\log n + \tau(n,k',m')\right) + \alpha\varphi(k,m)\log n. \quad (7)$$

Then, $\mathbb{P}_0$-almost surely, $\widehat{(k,m)}_n = (k_0,m_0)$ eventually.

Put in other words, $\widehat{(k,m)}_n$ does not overestimate, nor underestimate the true order (k_0,m_0) eventually, $\mathbb{P}_0$-almost surely. The proof is naturally divided accordingly: overestimation is considered in Section 3.1 and underestimation in Section 3.2. Note that a simple way to choose φ is to set $\varphi(k,m) = \mathrm{card}\{(k',m') \in \mathbb{N}^\star \times \mathbb{N} : (k',m')\preccurlyeq(k,m)\}$.

Remark 3.2. The theorem is valid more generally for $\mathbb{Q}_{k,m} = \mathrm{KT}_{k,m}$ or $\mathrm{NML}_{k,m}$ with the penalty

$$\mathrm{pen}(n,k,m) = \sum_{(k',m')\preccurlyeq(k,m)} \left(\frac{1}{2}N(k',m')\log n\right) + \alpha\varphi(k,m)\log n.$$

Note also that the precise form of the penalty is used in the non-overestimation step (see the proof of Prop. 3.3).

Any reader familiar with the BIC criterion will immediately interpret our penalty in terms of cumulated sum of BIC penalty terms (*i.e.* of the form $\frac{1}{2}N(k,m)\log n$). We do not prove here the consistency of the BIC procedure. We think this would be a very difficult task in our setup, and such a result does not even exist in the simpler HMM case. One explanation of this lack is that no explicit expression exists for the maximum likelihood estimate, turning explicit computations unfeasible. Thus our penalty is heavier than the BIC one but it is inspired by the penalty studied in [11] for order estimation in the HMM framework. However, if we cannot propose a theoretical study of the BIC estimator, we provide an original numerical study of the consistency of both our estimator and the BIC one in Section 4.

3.1. No overestimation

In this section, we prove that, $\mathbb{P}_0$-almost surely, $\widehat{(k,m)}_n$ does not overestimate the true order (k_0,m_0) eventually. Besides, a rate of decrease to zero of the overestimation probability is also obtained.

Proposition 3.3. *Under the assumptions and notations of Theorem 3.1, $\mathbb{P}_0$-almost surely, $\widehat{(k,m)}_n \preccurlyeq(k_0,m_0)$ eventually. Moreover,*

$$\mathbb{P}_0\left\{\widehat{(k,m)}_n \succ (k_0,m_0)\right\} = O(n^{-\alpha}),$$

where $\alpha > 1$ is chosen in Theorem 3.1.

The proof of Proposition 3.3 heavily relies on the following

Lemma 3.4. *Let us fix $(k,m) \in \mathbb{N}^\star \times \mathbb{N}$ and denote by $\mathbb{Q}_{k,m}$ the coding probability $\mathrm{KT}_{k,m}$ or $\mathrm{NML}_{k,m}$. Let us recall that τ is defined by (6). Then the following bounds hold:*

$$0 \leq \max_{y_1^n \in \mathcal{Y}^n} \left\{\log \frac{\mathrm{ML}_{k,m}(y_1^n)}{\mathbb{Q}_{k,m}(y_1^n)}\right\} \leq \frac{1}{2}N(k,m)\log n + \tau(n,k,m).$$

Lemma 3.4 is a combination of results which essentially go back to [23] and [6]. The proof is similar to the proof of [16], Lemma 3.4 and thus omitted.

Applying Lemma 3.4 allows to control the distribution of $\widehat{(k,m)}_n$ under $\mathbb{P}_0$ with respect to the dimensions of the involved models. More precisely, we have

Proposition 3.5. *Under the assumptions of Theorem 3.1, for fixed* $(k,m) \in \mathbb{N}^\star \times \mathbb{N}$,

$$\begin{aligned}\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m)\right\} &\leq \exp\{-\mathrm{pen}(n,k,m) + \mathrm{pen}(n,k_0,m_0)\} \\ &\times \left(\exp\left\{\frac{1}{2}N(k_0,m_0)\log n + \tau(n,k_0,m_0)\right\} \mathbb{1}\{\mathbb{Q}_{k,m} = \mathrm{NML}_{k,m} \text{ or } \mathrm{KT}_{k,m}\}\right. \\ &\left. + \exp\left\{\frac{1}{2}N(k,m)\log n + \tau(n,k,m)\right\} \mathbb{1}\{\mathbb{Q}_{k,m} = \mathrm{ML}_{k,m}\}\right).\end{aligned}$$

Proof of Proposition 3.5. Let $\mathbb{Q}_{k,m}$ be the probability measure $\mathrm{NML}_{k,m}$ or $\mathrm{KT}_{k,m}$. Using Definition (4) of $\widehat{(k,m)}_n$ and Lemma 3.4 implies that

$$\begin{aligned}\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m)\right\} &\leq \mathbb{P}_0\left\{\log\frac{\mathbb{Q}_{k,m}}{\mathbb{Q}_{k_0,m_0}}(Y_1^n) \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0)\right\} \\ &\leq \mathbb{P}_0\left\{\log\frac{\mathbb{Q}_{k,m}}{\mathrm{ML}_{k_0,m_0}}(Y_1^n) \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0) - \frac{1}{2}N(k_0,m_0)\log n - \tau(n,k_0,m_0)\right\}.\end{aligned}$$

Because $\mathbb{P}_0 \in \Pi^{k_0,m_0}$, we may use that $-\log \mathrm{ML}_{k_0,m_0}(Y_1^n) \leq -\log \mathbb{P}_0(Y_1^n)$, $\mathbb{P}_0$-almost surely, hence we have,

$$\begin{aligned}&\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m)\right\} \leq \\ &\qquad \mathbb{P}_0\left\{\log\frac{\mathbb{Q}_{k,m}}{\mathbb{P}_0}(Y_1^n) \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0) - \frac{1}{2}N(k_0,m_0)\log n - \tau(n,k_0,m_0)\right\} \\ &= \sum_{y_1^n \in \mathcal{Y}^n} \mathbb{P}_0(y_1^n)\mathbb{1}\left\{\log\frac{\mathbb{Q}_{k,m}(y_1^n)}{\mathbb{P}_0(y_1^n)} \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0) - \frac{1}{2}N(k_0,m_0)\log n - \tau(n,k_0,m_0)\right\} \\ &\qquad \leq \exp\left\{\tfrac{1}{2}N(k_0,m_0)\log n + \tau(n,k_0,m_0) - \mathrm{pen}(n,k,m) + \mathrm{pen}(n,k_0,m_0)\right\} \times \sum_{y_1^n \in \mathcal{Y}^n} \mathbb{Q}_{k,m}(y_1^n).\end{aligned}$$

This is the expected result, since $\mathbb{Q}_{k,m}$ is a probability measure. Let us assume now that $\mathbb{Q}_{k,m} = \mathrm{ML}_{k,m}$. Similarly,

$$\begin{aligned}\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m)\right\} &\leq \mathbb{P}_0\left\{\log\frac{\mathrm{ML}_{k,m}(Y_1^n)}{\mathrm{ML}_{k_0,m_0}(Y_1^n)} \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0)\right\} \\ &\leq \mathbb{P}_0\left\{\log\frac{\mathrm{ML}_{k,m}(Y_1^n)}{\mathbb{P}_0(Y_1^n)} \geq \mathrm{pen}(n,k,m) - \mathrm{pen}(n,k_0,m_0)\right\} \\ &\leq \sum_{y_1^n \in \mathcal{Y}^n} \mathrm{ML}_{k,m}(y_1^n)\exp\left\{-\mathrm{pen}(n,k,m) + \mathrm{pen}(n,k_0,m_0)\right\}.\end{aligned}$$

Using the bound $\mathrm{ML}_{k,m}(y_1^n) \leq \mathrm{KT}_{k,m}(y_1^n)\exp\{N(k,m)/2 \cdot \log n + \tau(n,k,m)\}$ given by Lemma 3.4 yields the expected result. Thus, the proof is complete. □

The proof of Proposition 3.3 is now at hand.

Proof of Proposition 3.3. Let us denote by A_n the event $\{\widehat{(k,m)}_n \succ (k_0,m_0)\}$. By virtue of the Borel-Cantelli lemma, it is sufficient to prove that $\sum_{n\geq 1}\mathbb{P}_0(A_n)$ is finite in order to conclude that overestimation eventually does not occur, $\mathbb{P}_0$-almost surely.

Let us assume that $\mathbb{Q}_{k,m}$ =NML$_{k,m}$ or KT$_{k,m}$ (the very similar proof in the case $\mathbb{Q}_{k,m}=\mathrm{ML}_{k,m}$ is omitted). If C_0 bounds sequence $\{\tau(n,k_0,m_0)\}_n$, then

$$\begin{aligned}\mathbb{P}_0\{A_n\} = & \sum_{(k,m)\succ(k_0,m_0)} \mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m)\right\}\\ & \overset{\text{(a)}}{\leq} \sum_{(k,m)\succ(k_0,m_0)} \exp\left\{\tfrac{1}{2}N(k,m)\log n + \tau(n,k_0,m_0) - \mathrm{pen}(n,k,m) + \mathrm{pen}(n,k_0,m_0)\right\}\\ \overset{\text{(b)}}{\leq} \sum_{(k,m)\succ(k_0,m_0)} \exp & \left\{ -\left[\sum_{(k,m)\succ(k',m')\succ(k_0,m_0)} \frac{1}{2}N(k',m')\log n\right] + \tau(n,k_0,m_0) - \alpha[\varphi(k,m)-\varphi(k_0,m_0)]\log n\right\}\\ & \leq C_0 \sum_{(k,m)\succ(k_0,m_0)} \exp\{-\alpha[\varphi(k,m)-\varphi(k_0,m_0)]\log n\}.\end{aligned}$$

Here, Proposition 3.5 and $N(k,m)\geq N(k_0,m_0)$ (for all $(k,m)\not\succ(k_0,m_0)$) yield (a) and (b) follows from the definition of the penalty term (note that the second sum may be empty). Now $\varphi:\mathbb{N}^\star\times\mathbb{N}\to\mathbb{N}$ increases, hence

$$\mathbb{P}_0\{A_n\} \leq C_0 \sum_{j\geq 1}\exp\{-\alpha j\log n\} \leq C_0 n^{-\alpha}(1-n^{-\alpha})^{-1} = O(n^{-\alpha}).$$

Since $\alpha>1$, the sum $\sum_n \mathbb{P}_0\{A_n\}$ is finite, and the proof is complete. □

3.2. No underestimation

In this section, we prove that, $\mathbb{P}_0$-almost surely, $\widehat{(k,m)}_n$ does not underestimate the true order (k_0,m_0) eventually.

Proposition 3.6. *Under the assumptions of Theorem 3.1, $\mathbb{P}_0$-almost surely, $\widehat{(k,m)}_n \not\succ (k_0,m_0)$ eventually.*

The first step while proving Proposition 3.6 is to relate the distribution of $\widehat{(k,m)}_n$ with the behaviour of the logarithm of the maximum likelihood ratio $[\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n)]$. This is the purpose of Lemma 3.7, whose proof is given in the appendix. From now on, "infinitely often" abbreviates to "i.o.".

Lemma 3.7. *Under the assumptions of Theorem 3.1, for every $k\geq 1$ and $m\geq 0$, there exists a sequence $\{\varepsilon_n\}$ of random variables that converges to zero $\mathbb{P}_0$-almost surely such that, for all $n\geq 1$,*

$$\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m) \text{ i.o.}\right\} \leq \mathbb{P}_0\left\{\frac{1}{n}\left[\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n)\right] \geq \varepsilon_n \text{ i.o.}\right\}.$$

Now, Proposition 3.6 essentially relies on two properties: a) the existence of a convenient Strong Law of Large Numbers for logarithms of likelihood ratios, in the spirit of the Shannon-Breiman-McMillan theorem – see Lemma 3.8; b) the existence of a finite sieve for the set of all ergodic distributions in $\Pi^{k,m}$ – see Lemma 3.9.

Let us recall that for any probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on the same measurable space $(\Omega,\mathcal{A})$ the relative entropy $\mathcal{D}(\mathbb{P}_1|\mathbb{P}_2)$ is defined by

$$\mathcal{D}(\mathbb{P}_1|\mathbb{P}_2) = \int \log\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}\mathrm{d}\mathbb{P}_1,$$

if $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$, and $+\infty$ otherwise.

Now, consider any probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on the same sequence space $(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$, with marginals onto $(\Omega^n, \mathcal{A}^n)$ denoted by $\mathbb{P}_1^n$ and $\mathbb{P}_2^n$, respectively. The asymptotic relative entropy $\mathcal{D}_\infty(\mathbb{P}_1|\mathbb{P}_2)$ (or divergence rate) is defined, when it exists, by

$$\mathcal{D}_\infty(\mathbb{P}_1|\mathbb{P}_2) = \lim_{n\to\infty} \frac{1}{n}\mathcal{D}(\mathbb{P}_1^n|\mathbb{P}_2^n).$$

Lemma 3.8 (Shannon-Breiman-McMillan)**.** *Let $\{Y_j\}_{j\geq 1}$ be an ergodic stationary process whose distribution $\mathbb{P}_0$ belongs to $\cup_{k\geq 1, m\geq 0}\Pi^{k,m}$. For all $k \geq 1$, $m \geq 0$ and any stationary ergodic $\mathbb{P}_\theta \in \Pi^{k,m}$, the divergence rate $\mathcal{D}_\infty(\mathbb{P}_0|\mathbb{P}_\theta)$ exists and is finite. Moreover, $\mathbb{P}_0$-almost surely,*

$$\lim_{n\to\infty} \frac{1}{n}\left[\log \mathbb{P}_\theta(Y_1^n) - \log \mathbb{P}_0(Y_1^n)\right] = -\mathcal{D}_\infty(\mathbb{P}_0|\mathbb{P}_\theta). \tag{8}$$

We omit the proof of Lemma 3.8, which is a generalization of a similar classical theorem that holds for hidden Markov models [2,9,11,15]. Lemma 3.8 notably ensures the existence of $\mathcal{D}_\infty(\mathbb{P}_1|\mathbb{P}_2)$ for stationary ergodic distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ belonging to $\cup_{k\geq 1, m\geq 0}\Pi^{k,m}$.

Stating the existence of a finite sieve involves two new subsets. For any $\delta > 0$, let us denote by $\Pi_\delta^{k,m}$ the subset of stationary probabilities $\mathbb{P}_\theta$ in $\Pi^{k,m}$ such that θ has all its coordinates lower bounded by δ. Moreover, let $\Pi_e^{k,m}$ stand for the subset of stationary ergodic probabilities in $\Pi^{k,m}$.

Lemma 3.9. *Let us set $k \geq 1$ and $m \geq 0$. For every $\varepsilon > 0$, there exist $\delta > 0$, a finite set of indexes $I_\varepsilon^{k,m}$ and a finite set of stationary probabilities $\{\mathbb{P}_i\}_{i\in I_\varepsilon^{k,m}}$ included in $\Pi_\delta^{k,m}$ such that, for all stationary ergodic $\mathbb{P}_\theta \in \Pi^{k,m}$, there exists some $\mathbb{P}_i$ ($i \in I_\varepsilon^{k,m}$) which guarantees that:*

$$\sup_{n\in\mathbb{N}^\star} \max_{y_1^n \in \mathcal{Y}^n} \frac{1}{n}\left[\log \mathbb{P}_\theta(y_1^n) - \log \mathbb{P}_i(y_1^n)\right] \leq \varepsilon.$$

Lemma 3.9 is a key for replacing the term $\log \mathbb{P}_\theta$ in the left-hand side of (8) by $\log \mathrm{ML}_{k,m}$ and the right-hand side term of the same equation by $-\inf_{\mathbb{P}} \mathcal{D}_\infty(\mathbb{P}_0|\mathbb{P})$ (for $\mathbb{P}$ ranging over $\Pi_e^{k,m}$). Its proof is given in the appendix.

Proof of Proposition 3.6. Let us set $\varepsilon > 0$ such that

$$\min_{(k,m)\prec(k_0,m_0)} \inf_{\mathbb{P}\in\Pi_e^{k,m}} \mathcal{D}_\infty(\mathbb{P}_0|\mathbb{P}) > \varepsilon.$$

Such an ε exists according to a result (whose generalization is easy and omitted in our framework) first obtained by [14], Propositions 1 and 2.

Let us choose arbitrarily $(k,m)\prec(k_0,m_0)$ and prove that $\mathbb{P}_0\{\widehat{(k,m)}_n = (k,m) \text{ i.o.}\} = 0$.

According to Lemma 3.7, there exists a sequence $\{\varepsilon_n\}$ of random variables that converges to zero $\mathbb{P}_0$-almost surely such that

$$\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m) \text{ i.o.}\right\} \leq \mathbb{P}_0\left\{\frac{1}{n}\left[\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n)\right] \geq \varepsilon_n \text{ i.o.}\right\}.$$

Now, Lemma 3.9 guarantees the existence of a finite set $\{\mathbb{P}_i\}_{i\in I_\varepsilon^{k,m}}$ of stationary probability measures which belong to $\Pi_\delta^{k,m} \subset \Pi^{k,m}$ such that

$$\begin{aligned}\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k,m) \text{ i.o.}\right\} &\leq \mathbb{P}_0\left\{\frac{1}{n}\left[\max_{i\in I_\varepsilon^{k,m}} \log \mathbb{P}_i(Y_1^n) - \log \mathbb{P}_0(Y_1^n)\right] \geq (-\varepsilon + \varepsilon_n) \text{ i.o.}\right\}\\ &\leq \sum_{i\in I_\varepsilon^{k,m}} \mathbb{P}_0\left\{\frac{1}{n}\left[\log \mathbb{P}_i(Y_1^n) - \log \mathbb{P}_0(Y_1^n)\right] \geq (-\varepsilon + \varepsilon_n) \text{ i.o.}\right\}.\end{aligned}$$

Table 1. The four smallest dimensions $N(k,m)$ of MCMR of order (k,m) when $r=4$.

(m)				
1	26			
0	8	15	24	
	2	3	4	(k)

Finally, Lemma 3.8 yields the convergence of $n^{-1}[\log \mathbb{P}_i(Y_1^n) - \log \mathbb{P}_0(Y_1^n)]$ to $-\mathcal{D}_\infty(\mathbb{P}_0|\mathbb{P}_i)$, $\mathbb{P}_0$-almost surely, for all $i \in I_\varepsilon^{k,m}$. The choice of ε then ensures that

$$\mathbb{P}_0 \left\{ (\widehat{k,m})_n = (k,m) \text{ i.o.} \right\} = 0.$$

Since $(k,m) \prec (k_0,m_0)$ was chosen arbitrarily, the previous equation implies that

$$\mathbb{P}_0 \left\{ (\widehat{k,m})_n \prec (k_0,m_0) \text{ i.o.} \right\} = 0$$

or, put in other words, that $\mathbb{P}_0$-almost surely, $(\widehat{k,m})_n \succcurlyeq (k_0,m_0)$ eventually. Thus, the proof is complete. □

4. Simulation study

In this section, we choose to discard the case $k=1$. Indeed, this case corresponds to Markov models and thus, to a data dependency structure which is very different from that of the case where there are at least two regimes. This distinction does not appear in the theoretical part of this article. However, all results (and their proofs) can be easily adapted to that slightly different framework. Finally note that in practice, MCMR modelling with at least two different regimes ($k \geq 2$) is used for data with no finite memory. MCMR with one regime (Markov models) poorly fit such data.

Our theoretical study is motivated by application to biology and more precisely, to genome analysis. Choosing a good model within a prescribed family is a very sensitive task. In [18], MCMR order selection (not identification) is performed for mining *Bacillus subtilis* chromosome heterogeneity. After fitting all models with $k \in \{2,\dots,8\}$ and $m \in \{0,1,2,3\}$, the authors select (by eyeball and using biological considerations) a MCMR of order $(k,m)=(3,2)$ for detecting atypical segments of length approximately 25 kb (1 kb equals $1,000$ nucleotides) upon the $4,200$ kb long chromosome. In this framework, $\mathcal{Y}$ stands for the nucleotides set $\{A,C,G,T\}$ $(r=4)$. In particular, the four smallest dimensions of MCMR are given in Table 1.

In order to illustrate our work, we undertake a simple simulation study in the framework described above. Evaluation of $\mathrm{ML}_{k,m}(y_1^n)$ is processed by Expectation-Maximization (EM) algorithm [4,7]. We run EM with multiple random initializations, and select the final result presenting the highest value. We use the package `SHOW` [19], where `SHOW` stands for Structured HOmogeneities Watcher. It is a set of executable programs that implements different uses of MCMR models for DNA sequences. The source code of `SHOW` is freely available. The software is protected by the GNU Public Licence.

We arbitrarily decide to consider only MCMR of dimension at most 26. The corresponding orders (k,m) appear in Table 1. Set $\mathcal{M} = \{\Pi^{k,m} : (k,m) \in \mathbb{N}^\star \times \mathbb{N}, N(k,m) \leq 26\}$. For each model $M_0 \in \mathcal{M}$ (line 1 in Fig. 1), we repeat 10 times (line 2) the following: we choose $\mathbb{P}_0 \in M_0$ (line 3), then simulate a chain y_1^n $(n = 100\,000)$ with distribution $\mathbb{P}_0$ (line 4), next for each model $M \in \mathcal{M}$ (line 5), for each $\tilde{n} \in \{25\,000; 50\,000; 100\,000\}$ (line 6), we evaluate $\sup_{\mathbb{P} \in M} \mathbb{P}(y_1^{\tilde{n}})$ (line 7). Afterwards, identifying the order boils down to applying (4) for a particular choice of penalty term. Before discussing this final step, let us go into details about the way we choose $\mathbb{P}_0 \in M_0$ (line 3). Because this simulation study is motivated by [18], we choose the final distribution obtained by fitting the same segment $[3\,450\,001; \,3\,475\,000]$ of length 25 kb of the *Bacillus Subtilis* chromosome than used in [18], Figure 1. For each repetition, a possibly slightly different distribution $\mathbb{P}_0$ is thus selected (EM is run with multiple *random* initializations).

```
foreach ($M_0 \in \mathcal{M}$) {
    repeat (10 times) {
        choice of a distribution $\mathbb{P}_0$ in model $M_0$
        simulation of a chain $y_1^n$ with distribution $\mathbb{P}_0$
        foreach ($M \in \mathcal{M}$) {
            foreach ($\tilde{n} \in \{25\,000; 50\,000; 100\,000\}$) {
                EM-evaluation of $\sup_{\mathbb{P}} \mathbb{P}(y_1^{\tilde{n}})$ for $\mathbb{P}$ ranging over $M$
            }
        }
    }
}
```

FIGURE 1. Evaluation of $\mathrm{ML}_{k,m}(y_1^{\tilde{n}})$ for various models index (k, m) and simulated observations y_1^n ($\tilde{n} \in \{25\,000; 50\,000; 100\,000\}, n = 100\,000$).

This simulation study validates Theorem 3.1: when $\tilde{n} = 50\,000$ and $\tilde{n} = 100\,000$, $\widehat{(k,m)}_{\tilde{n}} = (k_0, m_0)$ ten times out of ten for each true underlying model of order (k_0, m_0). Interestingly, this numerical evidence of consistency for very large values of $\tilde{n}$ does not include the case $\tilde{n} = 25\,000$. Indeed consistency then fails: $\widehat{(k,m)}_{\tilde{n}} = (k_0, m_0)$ ten times out of ten when $(k_0, m_0) = (2, 0)$, $\widehat{(k,m)}_{\tilde{n}} = (k_0, m_0)$ eight times out of ten when $(k_0, m_0) = (3, 0)$ [$(2, 0)$ otherwise], $\widehat{(k,m)}_{\tilde{n}} = (k_0, m_0)$ two times out of ten when $(k_0, m_0) = (4, 0)$ [$(3, 0)$ otherwise], and finally $\widehat{(k,m)}_{\tilde{n}} = (3, 0) \neq (k_0, m_0)$ ten times out of ten when $(k_0, m_0) = (2, 1)$. Each time $\widehat{(k,m)}_{\tilde{n}}$ differs from (k_0, m_0), one has $N(\widehat{(k,m)}_{\tilde{n}}) \leq N(k_0, m_0)$. In other words, our penalty is too heavy for that sample size, and the asymptotic regime is arguably not reached yet when $\tilde{n} = 25\,000$ whereas it is when $\tilde{n} \geq 50\,000$.

We emphasized earlier that our penalty is heavier than the BIC penalty (*i.e.* $\frac{1}{2}N(k, m) \log n$). How does the BIC estimator behave? For every sample size $\tilde{n} \in \{25\,000; 50\,000; 100\,000\}$ and every true underlying model, the BIC estimator coincides ten times out of ten with the true order. For this estimator, the asymptotic regime is already reached when $\tilde{n} = 25\,000$. Note that a slight modification of our penalty function yields another estimator which performs as well as the BIC one: if we replace $\mathrm{pen}(\tilde{n}, k, m)$ as defined in (7) by $\frac{1}{2}\mathrm{pen}(\tilde{n}, k, m)$, then the new estimator equals the true order ten times out of ten for every sample size $\tilde{n}$ and every true underlying model. One may finally wonder for which sample size the BIC criterion reaches its asymptotic regime. If the BIC estimator behaviour is still perfect when $\tilde{n} = 25\,000$, it actually fails when $\tilde{n} = 15\,000$. Denote by $\widetilde{(k,m)}_n$ the BIC estimator: $\widetilde{(k,m)}_n = (k_0, m_0)$ ten times out of ten when $(k_0, m_0) = (2, 0)$, $\widetilde{(k,m)}_n = (k_0, m_0)$ eight times out of ten when $(k_0, m_0) = (3, 0)$ [$(2, 0)$ otherwise], $\widetilde{(k,m)}_n = (k_0, m_0)$ ten times out of ten when $(k_0, m_0) = (4, 0)$, and finally $\widetilde{(k,m)}_n = (k_0, m_0)$ nine times out of ten when $(k_0, m_0) = (2, 1)$ [$(3, 0)$ otherwise]. Again, each time $\widetilde{(k,m)}_n$ differs from (k_0, m_0), one has $N(\widetilde{(k,m)}_n) \leq N(k_0, m_0)$. It is even worse when $\tilde{n} = 10\,000$, where we obtain $\widetilde{(k,m)}_n = (k_0, m_0)$ ten times out of ten when $(k_0, m_0) = (2, 0)$, $\widetilde{(k,m)}_n = (k_0, m_0)$ eight times out of ten when $(k_0, m_0) = (3, 0)$ [$(2, 0)$ otherwise], $\widetilde{(k,m)}_n = (k_0, m_0)$ eight times out of ten when $(k_0, m_0) = (4, 0)$ [$(3, 0)$ otherwise], and finally $\widetilde{(k,m)}_n = (k_0, m_0)$ nine times out of ten when $(k_0, m_0) = (2, 1)$ [$(3, 0)$ otherwise].

In conclusion, we apply the BIC criterion to the original sequence of *Bacillus Subtilis*: the resulting order estimator equals $(2, 1)$ (results are reported in Tab. 2). Our estimator equals $(3, 0)$.

A. Appendix A. Proof of Lemma 3.7

Let us set $k \geq 1$ and $m \geq 0$. The proof is straightforward when $\mathbb{Q}_{k,m} = \mathrm{ML}_{k,m}$. Indeed,

$$\mathbb{P}_0 \left\{ \widehat{(k,m)}_n = (k, m) \text{ i.o.} \right\} \leq \mathbb{P}_0 \left\{ \frac{1}{n} [\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n)] \geq -\frac{\mathrm{pen}(n, k_0, m_0)}{n} \text{ i.o.} \right\}$$

and $\mathrm{pen}(n, k_0, m_0) = o(n)$.

Table 2. EM-evaluated maximum likelihood of the original sequence (length $n = 25\,000$) for models MCMR of order (k, m) and BIC penalty $\frac{1}{2}N(k, m) \log n$. The resulting BIC order estimator equals $(2, 1)$.

(k, m)	$(2, 0)$	$(3, 0)$	$(4, 0)$	$(2, 1)$
$\log \mathrm{ML}_{k,m}(y_1^n)$	−34372.5	−34197.2	−34075.9	−33984.3
BIC penalty	40.5	75.9	121.5	131.6
$\mathrm{pen}(n, k, m)$	43.5	124.3	251.8	391.3

Let us assume that $\mathbb{Q}_{k,m} = \mathrm{NML}_{k,m}$ or $\mathrm{KT}_{k,m}$. Since $\mathrm{pen}(n, k, m)$ is non negative, the definition of $\widehat{(k,m)}_n$ readily yields that

$$\begin{aligned}\mathbb{P}_0\Big\{\widehat{(k,m)}_n = (k, m) \text{ i.o.}\Big\} \quad &\leq \quad \mathbb{P}_0\Big\{\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n) \\ &\geq \quad \log \frac{\mathrm{ML}_{k,m}(Y_1^n)}{\mathbb{Q}_{k,m}(Y_1^n)} - \log \frac{\mathbb{P}_0(Y_1^n)}{\mathbb{Q}_{k_0,m_0}(Y_1^n)} - \mathrm{pen}(n, k_0, m_0) \text{ i.o.}\Big\}.\end{aligned}$$

Then, by virtue of Lemma 3.4, it holds that:

$$\frac{1}{n}\left|\max_{y_1^n \in \mathcal{Y}^n} \left\{\log \frac{\mathrm{ML}_{k,m}(y_1^n)}{\mathbb{Q}_{k,m}(y_1^n)}\right\}\right| \xrightarrow[n\to\infty]{} 0, \tag{9}$$

$$\frac{1}{n} \max_{y_1^n \in \mathcal{Y}^n} \left\{\log \frac{\mathbb{P}_0(y_1^n)}{\mathbb{Q}_{k_0,m_0}(y_1^n)}\right\} \leq \frac{1}{n} \max_{y_1^n \in \mathcal{Y}^n} \left\{\log \frac{\mathrm{ML}_{k_0,m_0}(y_1^n)}{\mathbb{Q}_{k_0,m_0}(y_1^n)}\right\} \xrightarrow[n\to\infty]{} 0. \tag{10}$$

The final step is a variant of the so-called Barron's lemma [9], Theorem 4.4.1: a smart application of the Borel-Cantelli lemma yields that, $\mathbb{P}_0$-almost surely,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}_0(Y_1^n)}{\mathbb{Q}_{k_0,m_0}(Y_1^n)} \geq \liminf_{n\to\infty} \frac{-2\log n}{n} = 0. \tag{11}$$

Now, combining (9,10,11) with $\mathrm{pen}(n, k, m) = o(n)$ ensures the existence of a sequence $\{\varepsilon_n\}$ of random variables that converge to zero $\mathbb{P}_0$-almost surely such that

$$\mathbb{P}_0\left\{\widehat{(k,m)}_n = (k, m) \text{ i.o.}\right\} \leq \mathbb{P}_0\left\{\frac{1}{n}[\log \mathrm{ML}_{k,m}(Y_1^n) - \log \mathbb{P}_0(Y_1^n)] \geq \varepsilon_n \text{ i.o.}\right\}.$$

This concludes the proof of Lemma 3.7.

B. Appendix B. Proof of Lemma 3.9 for the existence of finite sieves

Let us set $k \geq 1$ and $m \geq 0$ and recall that the cardinality of $\mathcal{Y}$ is denoted by r. The proof of Lemma 3.9 is a straightforward consequence of the two lemmas below.

Lemma B.1. *For all $\delta > 0$, the set of functions $\theta \mapsto \mathbb{P}_\theta(y_1^n)$ indexed by $n \in \mathbb{N}^\star$ and $y_1^n \in \mathcal{Y}^n$ is equicontinuous over $\Theta_\delta^{k,m}$.*

Lemma B.2. *For every $\theta \in \Theta_e^{k,m}$ and $\delta > 0$ small enough, there exists $\theta_\delta \in \Theta_\delta^{k,m}$ such that, for all $n \in \mathbb{N}^\star$ and $y_1^n \in \mathcal{Y}^n$, the following bound holds:*

$$\frac{1}{n}[\log \mathbb{P}_\theta(y_1^n) - \log \mathbb{P}_{\theta_\delta}(y_1^n)] \leq 2(k^2 + r^2)\delta.$$

Lemma B.1 is a simple generalization of a result of Liu and Narayan [16] (Lem. 2.6), so we omit its proof. The proof of Lemma B.2 is also adapted from [16] (see their Ex. 2). The details are postponed after the proof of Lemma 3.9.

Proof of Lemma 3.9. Let us set $\varepsilon > 0$. According to Lemma B.1, for each $\theta_\delta \in \Theta_\delta^{k,m}$, there exists an open ball $\mathcal{B}(\theta_\delta) \subset \Theta_\delta^{k,m}$ such that, for every $\theta \in \mathcal{B}(\theta_\delta)$,

$$\sup_{n\in\mathbb{N}^\star} \max_{y_1^n \in \mathcal{Y}^n} \frac{1}{n} |\log \mathbb{P}_\theta(y_1^n) - \log \mathbb{P}_{\theta_\delta}(y_1^n)| \leq \varepsilon/2.$$

Since $\Theta_\delta^{k,m}$ is a compact set, the Borel-Lebesgue property ensures the existence of a finite subset $\{\theta_\delta^i : i \in I_\varepsilon\}$ of $\Theta_\delta^{k,m}$ such that $\cup_{i\in I_\varepsilon} \mathcal{B}(\theta_\delta^i) = \Theta_\delta^{k,m}$. Let us denote by $\mathbb{P}_i$ the probability measure $\mathbb{P}_{\theta_\delta^i}$ (for each $i \in I_\varepsilon$). In summary, for all $\theta_\delta \in \Theta_\delta^{k,m}$, there exists $i \in I_\varepsilon$ such that

$$\sup_{n\in\mathbb{N}^\star} \max_{y_1^n \in \mathcal{Y}^n} \frac{1}{n} |\log \mathbb{P}_{\theta_\delta}(y_1^n) - \log \mathbb{P}_i(y_1^n)| \leq \varepsilon/2. \tag{12}$$

Let us set $\delta \leq \varepsilon/[4(k^2 + r^2)]$. By virtue of Lemma B.2, for every $\theta \in \Theta_e^{k,m}$, there exists $\theta_\delta \in \Theta_\delta^{k,m}$ such that

$$\sup_{n\in\mathbb{N}^\star} \max_{y_1^n \in \mathcal{Y}^n} \frac{1}{n} [\log \mathbb{P}_\theta(y_1^n) - \log \mathbb{P}_{\theta_\delta}(y_1^n)] \leq 2(k^2 + r^2)\delta \leq \varepsilon/2. \tag{13}$$

Combining (12,13) concludes the proof. □

Proof of Lemma B.2. Set $\theta = (A, B) \in \Theta_e^{k,m}$ (see Def. 2 for the decomposition of parameter θ) and $\delta > 0$. The parameter θ_δ is constructed in the following way.

For each row $i \in \{1, \dots, k\}$ of matrix A, replace the maximal coefficient $a(i, j_{\max})$ by $a(i, j_{\max}) - (k-1)\delta$, then add δ to the other coefficients of this row. This yields the new parameter A_δ. Moreover, for each fixed "row" $(t^m; x) \in \mathcal{Y}^m \times \mathcal{X}$, replace the maximal coefficient of matrix B, namely $b(j_{\max}|t^m; x)$, by $b(j_{\max}|t^m; x) - (r-1)\delta$, then add δ to the other coefficients.

It is easily checked that the constructed parameter $\theta_\delta = (A_\delta, B_\delta)$ belongs to $\Theta_\delta^{k,m}$ for $\delta \leq 1/\max(k^2, r^2)$. Besides, it is also readily seen that, for all $i, j \in \{1, \dots, k\}$ and $(t^m; x) \in \mathcal{Y}^m \times \mathcal{X}$,

$$a(i;j) \leq \frac{a_\delta(i;j)}{(1 - k^2\delta)} \quad \text{and} \quad b(j|t^m; x) \leq \frac{b_\delta(j|t^m; x)}{(1 - r^2\delta)}.$$

Therefore, for all $n \in \mathbb{N}^\star$ and $y_1^n \in \mathcal{Y}^n$,

$$\mathbb{P}_\theta(y_1^n) \leq \mathbb{P}_{\theta_\delta}(y_1^n)(1 - k^2\delta)^{-n}(1 - r^2\delta)^{-n},$$

hence

$$\frac{1}{n} [\log \mathbb{P}_\theta(y_1^n) - \log \mathbb{P}_{\theta_\delta}(y_1^n)] \leq -\log(1 - k^2\delta) - \log(1 - r^2\delta).$$

This concludes the proof, because $-\log(1 - u) \leq 2u$ for any u small enough. □

Acknowledgements. We want to thank the associate editor and a referee, whose comments led to important improvements of this paper.

References

[1] D. Blackwell and L. Koopmans, On the identifiability problem for functions of finite Markov chains. *Ann. Math. Stat.* **28** (1957) 1011–1015.

[2] S. Boucheron and E. Gassiat, Order estimation and model selection, in *Inference in hidden Markov models*, Olivier Cappé, Eric Moulines, and Tobias Rydén (Eds.), Springer Series in Statistics. New York, NY: Springer (2005).

[3] R.J. Boys and D.A. Henderson, A Bayesian approach to DNA sequence segmentation. *Biometrics* **60** (2004) 573–588.

[4] O. Cappé, E. Moulines and T. Rydén (Eds.), *Inference in hidden Markov models.* Springer Series in Statistics (2005).

[5] I. Csiszár and Z. Talata, Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Info. Theory* **52** (2006) 1007–1016.

[6] L.D. Davisson, R.J. McEliece, M.B. Pursley and M.S. Wallace, Efficient universal noiseless source codes. *IEEE Trans. Inf. Theory* **27** (1981) 269–279.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B* **39** (1977) 1–38. With discussion.

[8] Y. Ephraim and N. Merhav, Hidden Markov processes. *IEEE Trans. Inform. Theory, special issue in memory of Aaron D. Wyner* **48** (2002) 1518–1569.

[9] L. Finesso, *Consistent estimation of the order for Markov and hidden Markov chains.* Ph.D. Thesis, University of Maryland, ISR, USA (1991).

[10] C-D. Fuh, Efficient likelihood estimation in state space models. *Ann. Stat.* **34** (2006) 2026–2068.

[11] E. Gassiat and S. Boucheron, Optimal error exponents in hidden Markov model order estimation. *IEEE Trans. Info. Theory* **48** (2003) 964–980.

[12] E.J. Hannan, The estimation of the order of an ARMA process. *Ann. Stat.* **8** (1980) 1071–1081.

[13] H. Ito, S.I. Amari and K. Kobayashi, Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inf. Theory* **38** (1992) 324–333.

[14] J.C. Kieffer, Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inf. Theory* **39** (1993) 893–902.

[15] B.G. Leroux, Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** (1992) 127–143.

[16] C.C. Liu and P. Narayan, Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Trans. Inf. Theory* **40** (1994) 1167–1180.

[17] R.J. MacKay, Estimating the order of a hidden markov model. *Canadian J. Stat.* **30** (2002) 573–589.

[18] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum and P. Bessières, Mining bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.* **30** (2002) 1418–1426.

[19] P. Nicolas, A.S. Tocquet and F. Muri-Majoube, `SHOW` *User Manual.* URL: `www-mig.jouy.inra.fr/ssb/SHOW/show_doc.pdf` (2004). Software available at URL: `http://www-mig.jouy.inra.fr/ssb/SHOW/`.

[20] B.M. Pötscher, Estimation of autoregressive moving-average order given an infinite number of models and approximation of spectral densities. *J. Time Ser. Anal.* **11** (1990) 165–179.

[21] C.P. Robert, T. Rydén and D.M. Titterington, Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc., Ser. B, Stat. Methodol.* **62** (2000) 57–75.

[22] T. Rydén, Estimating the order of hidden Markov models. *Statistics* **26** (1995) 345–354.

[23] Y.M. Shtar'kov, Universal sequential coding of single messages. *Probl. Inf. Trans.* **23** (1988) 175–186.