Electronic Journal of Statistics
Vol. 9 (2015) 717–752
ISSN: 1935-7524
DOI: 10.1214/15-EJS1017

Posterior consistency for nonparametric hidden Markov models with finite state space

Elodie Vernet

Laboratoire de Mathématiques, Université Paris-Sud, Orsay, France
e-mail: elodie.vernet@math.u-psud.fr

Abstract: In this paper we study posterior consistency for different topologies on the parameters for hidden Markov models with finite state space. We first obtain weak and strong posterior consistency for the marginal density function of finitely many consecutive observations. We deduce posterior consistency for the different components of the parameter. We also obtain posterior consistency for marginal smoothing distributions in the discrete case. We finally apply our results to independent emission distributions, translated emission distributions and discrete HMMs, under various types of priors.

MSC 2010 subject classifications: 62G20.
Keywords and phrases: Bayesian nonparametrics, consistency, hidden Markov models.

Received July 2014.

Contents

1. Introduction

Hidden Markov models (HMMs) have been widely used in diverse fields such as speech recognition, genomics or econometrics since their introduction in Baum and Petrie [2]. The books MacDonald and Zucchini [16], MacDonald and Zuc-

chini [17], and Cappé, Moulines and Rydén [3] provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. Finite state space HMMs are stochastic processes $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $(X_t)_{t\in\mathbb{N}}$ is a Markov chain taking values in a finite set, and conditionally to $(X_t)_{t\in\mathbb{N}}$, the random variables Y_t, $t \in \mathbb{N}$, are independent, the distribution of Y_t depending only on X_t. The conditional distributions of Y_t given X_t, for all possible values of X_t, are called emission distributions. The name "hidden Markov model" comes from the fact that the observations are the Y_t's only, one cannot access to the states $(X_t)_t$ of the Markov chain. Finite state space HMMs can be used to model heterogeneous variables coming from different populations, the states of the (hidden) Markov chain defining the population the observed variable comes from. HMMs are very popular dynamical models especially because of their computational tractability since there exist efficient algorithms to compute the likelihood and to recover the posterior distribution of the hidden states given the observations.

Frequentist asymptotic properties of estimators of HMMs parameters have been studied since the 1990s. Consistency and asymptotic normality of the maximum likelihood estimator have been established in the parametric case, see Douc and Matias [6], Douc, Moulines and Rydén [7], and references in Cappé, Moulines and Rydén [3], see also Douc et al. [8] for the most general consistency result up to now. As to Bayesian asymptotic results, there are only very few and recent results, see de Gunst and Shcherbakova [5] when the number of hidden states is known, Gassiat and Rousseau [14] when the number of hidden states is unknown. All these results concern parametric HMMs.

Non parametric HMMs in the sense that the form of the emission distribution is not specified have only very recently been considered, since identifiability remained an open problem until Gassiat and Rousseau [13] and Gassiat, Cleynen and Robin [12], who prove a general identifiability result. Because parametric modeling of emission distributions may lead to poor results in practice, in particular for clustering purposes, recent interest in using non parametric HMMs appeared in applications, see Yau et al. [21], Gassiat, Cleynen and Robin [12] and references therein. Theoretical results for estimation procedures in non parametric HMMs have also been obtained only very recently: Dumont and Le Corff [10] concerns regression models with hidden (Markovian) regressors and unknown regression functions in Gaussian noise, and Gassiat and Rousseau [13] is about translated emission distributions.

In this paper, we obtain posterior consistency results for Bayesian procedures in finite state space non parametric HMMs. To our knowledge, this is the first result on posterior consistency in such models. In Section 2.2, we prove posterior consistency in terms of the weak topology and the L_1-norm on marginal densities of consecutive observations. Our main result is obtained under assumptions on the emission densities and on the prior which are very similar to the ones in the i.i.d. case, see Theorem 2.1. This result relies on a new control of the Kullback-Leibler divergence for HMMs, see Lemma 2.2. Yet estimating the distribution of consecutive observations is not the main objective of a practitioner. Classifying the observations according to their corresponding hidden states or

estimating the parameters of the model often are the questions of interest, see for instance Yau et al. [21], Whiting, Lambert and Metcalfe [20] and Couvreur and Couvreur [4]. In Section 2.3 we build upon the recent identifiability result to deduce from Theorem 2.1 posterior consistency for each component of the parameters. We obtain in general posterior consistency for the transition matrix of the Markov chain and for the emission probability distribution in the weak topology, see Theorem 2.3. Stronger results are established in particular cases, see Corollary 3.2 and Theorem 3.4. Finally, some examples of priors that fulfill the assumptions of Theorems 2.1 and 2.3 are studied in Section 3.

Particularly in Section 3.3 the discrete case is thoroughly studied with a Dirichlet process prior. Sufficient and almost necessary assumptions to apply Theorem 2.1 are given in Proposition 3.5. Moreover in this framework, posterior consistency of the marginal smoothing distributions, used in segmentation or classification, is derived in Theorem 3.4.

All proofs are given in Appendices A and B.

2. Settings and main theorem

2.1. Notations

We now precise the model and give some notations. Recall that finite state space HMMs are stochastic processes $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $(X_t)_{t\in\mathbb{N}}$ is a Markov chain taking values in a finite set, and conditionally on $(X_t)_{t\in\mathbb{N}}$, the random variables Y_t, $t \in \mathbb{N}$, are independent. The distribution of Y_t depending only on X_t is called the emission distribution. The number k of hidden states is known, so that the state space of the Markov chain is set to $\{1, \dots, k\}$. Throughout the paper, for any integer n, an n-uple $(x_1, \dots, x_n)$ is denoted $x_{1:n}$.

Let $\Delta_k = \{(x_1, \dots, x_k) : x_i \geqslant 0,\ i = 1, \dots, k;\ \sum_{i=1}^k x_i = 1\}$ denote the $(k-1)$-dimensional simplex. Let Q denote the $k \times k$ transition matrix of the Markov chain, so that identifying Q as the k-uple of transition distributions (the lines of the matrix), we write $Q \in \Delta_k^k$. We denote $\mu \in \Delta_k$ the initial probability measure, that is the distribution of X_1. For $\underline{q} \geqslant 0$, we also define

$$\Delta^k(\underline{q}) = \{Q \in \Delta_k^k \ :\ \min_{i,j\leqslant k} Q_{i,j} \geqslant \underline{q}\},$$

so that $\Delta^k(0) = \Delta_k^k$. We now recall some properties of Markov chains with transition matrix in $\Delta^k(\underline{q})$. Note that $\underline{q}$ needs to be less than $\frac{1}{k}$ for $\Delta^k(\underline{q})$ to be non empty. Then for all Q in $\Delta^k(\underline{q})$, $\max_{i,j} Q_{i,j} \leqslant 1-(k-1)\underline{q}$. Also, if $Q \in \Delta^k(\underline{q})$, then for any $i \in \{1, \dots, k\}$ and $A \subset \{1, \dots, k\}$, $\sum_{j\in A} Q_{i,j} \geqslant k\underline{q}u(A)$, with u the uniform probability on $\{1, \dots, k\}$. Besides if $Q \in \Delta^k(\underline{q})$ with $\underline{q} > 0$, the chain is irreducible, positive recurrent and admits a unique stationary probability measure denoted μ^Q for which $\underline{q} \leqslant \mu^Q(i) \leqslant 1 - (k-1)\underline{q}$, $1 \leqslant i \leqslant k$.

We assume that the observation space is $\mathbb{R}^d$ endowed with its Borel sigma field. Let $\mathcal{F}$ be the set of probability density functions with respect to a reference measure λ on $\mathbb{R}^d$. $\mathcal{F}^k$ is the set of possible emission densities, that is for $f =$

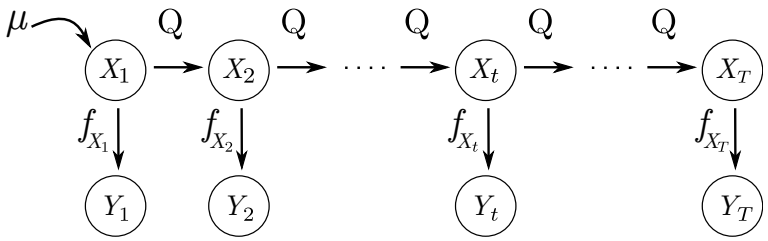

FIG 1. *The model.*

$(f_1, \dots, f_k) \in \mathcal{F}^k$, the distribution of Y_t conditionally to $X_t = i$ will be $f_i\lambda$, $i = 1, \dots, k$. See Figure 1 for a visualization of the model.

Let

$$\Theta = \{\theta = (Q, f) \ : \ Q \in \Delta_k^k, f \in \mathcal{F}^k\}$$

and

$$\Theta(\underline{q}) = \{\theta = (Q, f) \ : \ Q \in \Delta^k(\underline{q}), f \in \mathcal{F}^k\}.$$

Then $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$) denotes the probability distribution of $(X_t, Y_t)_{t\in\mathbb{N}}$ under θ and initial probability measure $\mu^\theta := \mu^Q$ (respectively μ). Let p_l^θ ($p_l^{\theta,\mu}$ resp.) denote the probability density of $Y_1, \dots, Y_l$ with respect to $\lambda^{\otimes l}$ under $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$). and P_l^θ ($P_l^{\theta,\mu}$ resp.) the marginal distribution of $Y_1, \dots, Y_l$ under $\mathbb{P}^\theta$ (resp. $\mathbb{P}^{\theta,\mu}$). So for any $\theta \in \Theta$, initial probability measure μ, and measurable set A of $\{1, \dots, k\}^l \times (\mathbb{R}^d)^l$:

$$\begin{aligned}
&\mathbb{P}^{\theta,\mu}((X_{1:l}, Y_{1:l}) \in A) \\
&\quad = \int \sum_{x_1,\dots,x_l=1}^{k} \mathbb{1}_{(x_1,\dots,x_l,y_1,\dots,y_l)\in A}\ \mu_{x_1} Q_{x_1,x_2} \dots Q_{x_{l-1},x_l} \\
&\qquad f_{x_1}(y_1) \dots f_{x_l}(y_l)\lambda(dy_1) \dots \lambda(dy_l),
\end{aligned}$$

$$p_l^{\theta,\mu}(y_1, \dots, y_l) = \sum_{x_1,\dots,x_l=1}^{k} \mu_{x_1} Q_{x_1,x_2} \dots Q_{x_{l-1},x_l} f_{x_1}(y_1) \dots f_{x_l}(y_l),$$

and $P_l^{\theta,\mu} = p_l^{\theta,\mu}\lambda^{\otimes l}$.

We denote by $\delta_\mu \otimes \pi$ the prior on $\Delta_k \times \Theta$, where $\mu \in \Delta_k$ is an initial probability measure. We assume that π is a product of probability measures on Θ, $\pi = \pi_Q \otimes \pi_f$ such that π_Q is a probability distribution on Δ_k^k and π_f is a probability distribution on $\mathcal{F}^k$.

We assume throughout the paper that the observations are distributed from $\mathbb{P}^{\theta^*}$ so that their distribution is a stationary HMM. We are interested in posterior consistency, that is to prove that with $\mathbb{P}^{\theta^*}$-probability one, for all neighborhood U of θ^*:

$$\lim_{n\to+\infty} \pi(U|Y_{1:n}) = 1.$$

The choice of a topology on the parameters arises here. For any distance or pseudometric D, we denote $N(\delta, A, D)$ the δ-covering number of the set A with

respect to D, that is the minimum number N of elements $a_1, \dots, a_N$ such that for all $a \in A$, there exists $n \leqslant N$ such that $D(a, a_n) \leqslant \delta$.

For $k \times k$ matrices M, we use

$$\|M\| = \max_{1 \leqslant i,j \leqslant k} |M_{i,j}|.$$

For probability distributions P_1 and P_2, let p_1 and p_2 be their respective densities with respect to some dominated measure ν. We use the L_1-norm:

$$\|p_1 - p_2\|_{L_1(\nu)} = \int |p_1 - p_2| d\nu$$

and the Kullback-Leibler divergence:

$$KL(P_1, P_2) = \begin{cases} \int p_1 \log(\frac{p_1}{p_2}) d\nu & \text{if } P_1 << P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

We also denote $KL(p_1, p_2)$ for $KL(p_1\nu, p_2\nu)$. On $\mathcal{F}^k$ we use the distance $d(\cdot, \cdot)$ defined for all $g = (g_1, \dots, g_k)$, $\tilde{g} = (\tilde{g_1}, \dots, \tilde{g_k})$ by

$$d(g, \tilde{g}) = \max_{1 \leqslant j \leqslant k} \|g_j - \tilde{g_j}\|_{L_1(\lambda)}.$$

On $\Theta(\underline{q})$, we use the following pseudometric for $l \geqslant 3$, $l \in \mathbb{N}$,

$$D_l(\theta, \theta') = \int |p_l^\theta(y_1, \dots, y_l) - p_l^{\theta'}(y_1, \dots, y_l)| \lambda(dy_1) \dots \lambda(dy_l) = \|p_l^\theta - p_l^{\theta'}\|_{L_1(\lambda^{\otimes l})}.$$

Then a D_l-neighborhood of θ is a set which contains a set $\{\theta' : D_l(\theta, \theta') < \varepsilon\}$ for some $\varepsilon > 0$. We also use the weak topology on marginal distributions $(P_l^\theta)_\theta$. We recall that in any neighborhood of P_l^θ in the weak topology on probability measures there is a subset which is a union of sets of the form

$$\left\{ P \ : \ \left| \int h_j dP - \int h_j p_l^\theta d\lambda^{\otimes l} \right| < \varepsilon_j, \ j = 1, \dots, N \right\},$$

where for all $1 \leqslant j \leqslant N$, $\varepsilon_j > 0$ and h_j is in the set $\mathcal{C}_b((\mathbb{R}^d)^l)$ of all bounded continuous functions from $(\mathbb{R}^d)^l$ to $\mathbb{R}$. We prove posterior consistency in this general nonparametric context using this weak topology on marginal distributions $(P_l^\theta)_\theta$ and the D_l-pseudometric in Section 2.2. We study the posterior consistency for the transition matrix and the emission distributions separately in Section 2.3.

Finally the sign $\lesssim$ is used for inequalities up to a multiplicative constant possibly depending on fixed parameters.

2.2. Main theorem

In this section, we state our general theorem on posterior consistency for nonparametric hidden Markov models in the weak topology on marginal distributions $(P_l^\theta)_\theta$ and the D_l-topology. Fix $l \geqslant 3$. We consider the following assumptions:

(A0) For all $1 \leqslant i \leqslant k$, $\int f_i^*(y)|\log(f_i^*(y))|\lambda(dy) < +\infty$,

(A1) for all $\varepsilon > 0$ small enough there exists a set $\Theta_\varepsilon \subset \Theta(\underline{q})$ such that $\pi(\Theta_\varepsilon) > 0$ and for all $\theta = (Q, f) \in \Theta_\varepsilon$,

(A1a) $\|Q - Q^*\| < \varepsilon$,

(A1b) $\max_{1\leqslant i\leqslant k} \int f_i^*(y) \max_{1\leqslant j\leqslant k} \log(\frac{f_j^*(y)}{f_j(y)})\lambda(dy) < \varepsilon$,

(A1c) for all $y \in \mathbb{R}^d$ such that $\sum_{i=1}^k f_i^*(y) > 0$, $\sum_{j=1}^k f_j(y) > 0$,

(A1d) $\sup_{y:\sum_{i=1}^k f_i^*(y)>0} \max_{1\leqslant j\leqslant k} f_j(y) < +\infty$,

(A2) for all $n > 0$, for all $\delta > 0$, there exists a set $\mathcal{F}_n \subset \mathcal{F}^k$ and a real number $r_1 > 0$ such that $\pi_f\big((\mathcal{F}_n)^c\big) \lesssim e^{-nr_1}$ and such that

$$\sum_{n>0} N\left(\frac{\delta}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \exp\left(-\frac{n\delta^2 k^2 \underline{q}^2}{32l}\right) < +\infty.$$

Theorem 2.1. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leqslant i \leqslant k$, $\mu_i \geqslant \underline{q}$.*

a) If Assumptions (A0) and (A1) holds then, for all weak neighborhood U of $P_l^{\theta^*}$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(U|Y_{1:n}) = 1\right) = 1.$$

b) Moreover if Assumptions (A0), (A1) and (A2) hold then, for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\ \{\theta :\ D_l(\theta,\theta^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

Remark 2.1. We assume everywhere in the paper that the support of π_Q is included in $\Delta^k(\underline{q})$. It means the results of this paper can only be applied to priors π_Q on transition matrices which vanish close to the border of Δ_k^k. This assumption is satisfied by a product of truncated Dirichlet distribution, i.e. if the lines $Q_{i,\cdot}$ of Q are independently distributed from a law proportional to:

$$Q_{i,1}^{\alpha_1-1} \dots Q_{i,k}^{\alpha_k-1} \mathbb{1}_{\{\underline{q}\leqslant Q_{i,j}\leqslant 1,\ \forall 1\leqslant j\leqslant k\}} dQ_{i,1} \dots dQ_{i,k}$$

where $\alpha_1, \dots, \alpha_k > 0$.

The restriction on $\Delta^k(\underline{q})$ comes from the test built in Gassiat and Rousseau [14]. On this set, HMMs are geometrically ergodic. It is a common assumption in the literature see Douc and Matias [6], Douc, Moulines and Rydén [7], or Douc et al. [8] for instance. Besides Gassiat and Rousseau [14] explain the difficulty which appears when the Markov chain does not mix well. They are also able to obtain a less restrictive assumption on the support of the prior on transition matrices. In return they assume a more restrictive assumption on the log-likelihood, compare Equations (11) and (13) with their Assumption C1.

In the case of density estimation with i.i.d. observations, it is usual to control the Kullback-Leibler support of the prior to show weak posterior consistency

and to control, in addition, a metric entropy to obtain strong consistency, see Chapter 4 of Ghosh and Ramamoorthi [15]. Assumptions (A1) and (A2) are similar in spirit. Assumptions (A0) and (A1) replace the assumption on the true density function being in the Kullback-Leibler support of the prior in the i.i.d. case. (A1a) ensures that the transition matrices of Θ_ε are in a ball of radius ε around the true transition matrix. Under (A1b) the emission densities are in an ε Kullback-Leibler ball around the true one. (A0), (A1b), (A1c) and (A1d) are assumptions under which the log-likelihood converges $\mathbb{P}^{\theta^*}$-a.s. and in $L_1(\mathbb{P}^{\theta^*})$. (A2) is very similar to the assumptions of the metric entropy of Theorem 4.4.4 in Ghosh and Ramamoorthi [15].

In Appendix A, the proof of Theorem 2.1 relies on the method of Barron [1]. It consists of controlling Kullback-Leibler neighborhoods and building tests. The construction of tests is quite straightforward thanks to Rio's inequality [18] which generalizes Hoeffding's inequality. To prove a), we use the usual strategy presented in Section 4.4.1 in Ghosh and Ramamoorthi [15] together with Rio's inequality [18] and Gassiat and Rousseau [14]. To prove b), we use the tests of Gassiat and Rousseau [14]. To control the Kullback-Leibler neighborhoods, we use the following lemma whose proof is given in Appendix A.

Lemma 2.2. *Let θ^* be in $\Theta(\underline{q})$. If (A1) holds then, for all $0 < \varepsilon < 1$, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ and for all $\theta \in \Theta_\varepsilon$:*

$$\frac{1}{n} KL(\mathbb{P}_n^{\theta^*}, \mathbb{P}_n^{\theta,\mu}) \leqslant \frac{3}{\underline{q}}\, \varepsilon.$$

2.3. Consistency of each component of the parameter

In this Section we look at the consequences of Theorem 2.1 on posterior consistency for the transition matrix and the emission distributions separately. Estimating consistently the components of the parameter is of great importance. First one may want to know the proportion of each population or the probability of moving from one population to another, i.e. the transition matrix. Secondly, these components are important to recover the smoothing distribution, i.e. the distribution of a hidden state given the observations, and then to cluster the observations, see Cappé, Moulines and Rydén [3] and Theorem 3.4.

In practice, estimating the marginal density of l consecutive observations is not the first purpose. Yet estimating the parameters and the hidden states is often the goal. For instance, Whiting, Lambert and Metcalfe [20] give an algorithm to estimate the stationary probability measure of the Markov chain derived from the transition matrix. While Yau et al. [21] and Couvreur and Couvreur [4] are interested in estimating the hidden states.

The consistency for each component of the parameter, i.e. the transition matrix and the emission distributions, does not directly result from consistency of the marginal distribution of the observations, see Dumont and Le Corff [10]. Identifiability seems to be necessary to obtain this implication yet it is not sufficient. We obtain posterior consistency for the components of the parameter

thanks to the result of identifiability of Gassiat, Cleynen and Robin [12] and as usually by proving the continuity of the functional

$$\begin{cases} ((p_l^\theta)_\theta, L_1) & \to & (\Theta, \text{the topology } \mathcal{T} \text{ described in the following}) \\ p_l^\theta & \mapsto & \theta \end{cases}.$$

We use a product topology on the set of parameters. In particular we study consistency in the topology associated with the sup norm on transition matrices $\|\cdot\|$ and the weak topology on probability measures for the emission distributions up to label switching. To deal with label switching, we need the following definitions. Let $\mathcal{S}_k$ denote the symmetric group on $\{1,\dots,k\}$. Let σ be a permutation in $\mathcal{S}_k$, for all matrices $Q \in \Delta_k^k$, we denote σQ the following matrix: for all $1 \leqslant i,j \leqslant k$,

$$(\sigma Q)_{i,j} = Q_{\sigma(i),\sigma(j)}.$$

If $(X_t,Y_t)_{t\in\mathbb{N}}$ is distributed from $P^{(Q,f)}$ and $\tilde{X}_t = \sigma^{-1}(X_t)$, for $\sigma \in \mathcal{S}_k$, then $(\tilde{X}_t,Y_t)_{t\in\mathbb{N}}$ is distributed from $P^{(\sigma Q,(f_{\sigma(1)},\dots,f_{\sigma(k)}))}$, i.e the labels of the Markov chain have been switched but $(Y_t)_{t\in\mathbb{N}}$ has the same distribution. Then, in generality, from the distribution of the observations one can at most recover the parameter up to label switching. Gassiat, Cleynen and Robin [12] proved that it is possible by knowing the joint distribution of at least three consecutive observations.

In Theorem 2.3, whose proof is given in Appendix A, we prove that under the assumption of identifiability, posterior consistency in the D_l topology implies that the posterior concentrates around (Q^*,f^*) up to label switching, i.e. around $\{\sigma Q^*,(f^*_{\sigma(1)},\dots,f^*_{\sigma(k)})\}_{\sigma\in\mathcal{S}_k}$. In other words we obtain posterior consistency considering neighborhoods of the form

$$\left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)} \in U_{f_i^*}, i = 1\dots k\right\}$$

where U_{Q^*} is a neighborhood of Q^* and for all $1 \leqslant i \leqslant k$, $U_{f_i^*}$ is a weak neighborhood of $f_i^*\lambda$. That is to say we consider the product topology $\mathcal{T}$ of the sup norm topology on transition matrices and of the weak topology on the emission distributions up to label switching.

Theorem 2.3. *Let $\theta^* = (Q^*,f^*) \in \Theta$ such that $f_1^*\lambda,\dots,f_k^*\lambda$ are linearly independent and Q^* has full rank.*

If the posterior is consistent for the D_l pseudo-metric with $l \geqslant 3$, i.e. if for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\ \{\theta :\ D_l(\theta,\theta^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

then the posterior is consistent for the topology $\mathcal{T}$, i.e. for all weak neighborhood $U_{f_i^*}$ of $f_i^*\lambda$, for all $1 \leqslant i \leqslant k$ and for all neighborhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)}\lambda \in U_{f_i^*},\ 1 \leqslant i \leqslant k\right\}\middle| Y_{1:n}\right) = 1\right) = 1. \tag{1}$$

Remark 2.2. In particular, Equation (1) implies that for all $\varepsilon > 0$

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\left(\bigcup_{\sigma\in\mathcal{S}_k}\{Q : \|Q - \sigma Q^*\| < \varepsilon\} \;\middle|\; Y_{1:n}\right) = 1\right) = 1.$$

It means that under the assumptions of Theorem 2.3, the posterior concentrates around $\{\sigma Q^*, \sigma \in \mathcal{S}_k\}$. Equation (1) also implies that for all $N \in \mathbb{N}$, for all $h_i \in \mathcal{C}_b(\mathbb{R}^d)$ and for all $\varepsilon_i > 0$,

$$\mathbb{P}^{\theta^*}\Bigg(\lim_{n\to+\infty} \pi\Bigg(\bigcup_{\sigma\in\mathcal{S}_k}\Big\{f : \Big|\int h_i f_j d\lambda - \int h_i f^*_{\sigma(j)} d\lambda\Big| < \varepsilon_i, \\ \text{for all } 1 \leqslant i, j \leqslant k\Big\} \;\Big|\; Y_{1:n}\Bigg) = 1\Bigg) = 1.$$

This last result allows to consistently recover smooth functionals of the emission distributions $(f^*_j)_j$ such as $\int_K f^*_j d\lambda$ where K is compact. We obtain stronger results in Sections 3.2 and 3.3.

The uncertainty due to label switching can be removed if there is only one possible permutation σ associated to a parameter θ as in Proposition 2.4, proved in Appendix A. This Proposition 2.4 may be useful if one knows some characteristics of the hidden states which order them. The function H, in Proposition 2.4, enables to order the hidden states and then to get rid of label switching.

Proposition 2.4. *Let $\theta^* \in \Theta$ such that $f^*_1\lambda, \dots, f^*_k\lambda$ are linearly independent and Q^* has full rank. Let $H : (\Theta, \mathcal{T}_1) \to \mathbb{R}^k$ be a continuous function, where $\mathcal{T}_1$ is the product topology of the sup norm topology on transition matrices and of the weak topology on the emission distributions. Assume that for all permutation $\sigma \in \mathcal{S}_k$ and for all $\theta = (Q, f) \in \Theta$,*

$$H_i((\sigma Q, f_{\sigma(1)}, \dots, f_{\sigma(k)})) = H_{\sigma(i)}(\theta), \tag{2}$$

$$H_1(\theta^*) < \cdots < H_k(\theta^*), \tag{3}$$

$$\pi\Big(\{\theta : \; H_1(\theta) < \cdots < H_k(\theta)\}\Big) = 1. \tag{4}$$

If the posterior is consistent for the topology $\mathcal{T}$, i.e. for all weak neighborhood $U_{f^*_i}$ of $f^*_i\lambda$, for all $1 \leqslant i \leqslant k$ and for all neighborhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\Bigg(\lim_{n\to+\infty} \pi\Bigg(\Big\{\exists\sigma \in \mathcal{S}_k;\; \sigma Q \in U_{Q^*},\; f_{\sigma(i)}\lambda \in U_{f^*_i},\; 1 \leqslant i \leqslant k\Big\}\Big|Y_{1:n}\Bigg) = 1\Bigg) = 1 \tag{1}$$

then for all weak neighborhood $U_{f^*_i}$ of $f^*_i\lambda$, for all $1 \leqslant i \leqslant k$ and for all neighborhood U_{Q^*} of Q^*,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to+\infty} \pi\Bigg(\Big\{\; Q \in U_{Q^*},\; f_i\lambda \in U_{f^*_i},\; 1 \leqslant i \leqslant k\Big\} \;\Big|\; Y_{1:n}\Bigg) = 1\right) = 1.$$

Here we give some examples of possible functions H:

$$H_i(\theta) = Q_{i,i} \quad \text{or} \quad H_i(\theta) = \int \phi f_i d\lambda, \tag{5}$$

where ϕ is bounded and continuous. Even if in practice, one would often like to use $H_i(\theta) = \int y f_i(y)\lambda(dy)$, Proposition 2.4 does not allow it. Indeed, in this case H is not continuous. Yet taking a continuous truncated version of the identity for ϕ in Equation (5) may help.

3. Examples of priors on f

In this section we apply Theorems 2.1 and 2.3 for different types of priors and emission models. In Section 3.1 we deal with emission distributions which are independent mixtures of Gaussian distributions. Translated emission distributions are studied in Section 3.2. Finally we consider the discrete case with Dirichlet process priors in Section 3.3.

Assumptions (A1b) and (A2) are purposely designed to resemble the types of assumptions found in density estimation for i.i.d. observations. This allows us to use existing results on consistency in the case of i.i.d. observations. This is done in Sections 3.1 and 3.2 with a prior based on a usual prior on densities, which is a mixture of Gaussian distributions such as in Tokdar [19]. Two ways of using a prior on densities are considered. In Section 3.1, the emission distributions are independently distributed under a usual prior on densities. In Section 3.2, the emission distributions are designed from a unique density, distributed from a usual prior, which is translated. Contrariwise in the discrete case we develop a new method to deal with the Dirichlet process prior in Section 3.3.

3.1. Independent mixtures of Gaussian distributions

We consider the well known location-scale mixture of Gaussian distributions as prior model for each f_i, namely each density under the prior is written as

$$g(y) = \int_{\mathbb{R}\times(0,+\infty)} \phi_\sigma(y-z)dP(z,\sigma) =: \phi * P, \tag{6}$$

where ϕ_σ is the Gaussian density with mean zero and variance σ^2, and P is a probability measure on $\mathbb{R}\times(0,+\infty)$. In this part, λ is the Lebesgue measure on $\mathbb{R}$. Let π_P be a probability measure on the set of probability measures on $\mathbb{R}\times(0,+\infty)$. Denote π_g the distribution of g expressed as (6) when $P \sim \pi_P$. Then we consider the prior distribution on $f = (f_1, \ldots, f_k)$ defined by $\pi_f = \pi_g^{\otimes k}$. We need the following assumptions to apply Theorem 2.1 and 2.3:

(B1)
$$\pi_P\left(P \ : \ \int \frac{1}{\sigma} dP(z,\sigma) < \infty\right) = 1,$$

(B2) for all $1 \leqslant j \leqslant k$, f_j^* is positive, continuous on $\mathbb{R}$ and bounded by $M < \infty$,
(B3) for all $1 \leqslant i \leqslant k$, $1 \leqslant j \leqslant k$,

$$\int_{\mathbb{R}} f_i^*(y) \log\left(\frac{f_j^*(y)}{\psi_j(y)}\right) \lambda(dy) < \infty$$

where $\psi_j(y) = \inf_{t\in[y-1,y+1]} f_j^*(t)$,
(B4) for all $1 \leqslant i \leqslant k$, there exists $\eta > 0$ such that

$$\int_{\mathbb{R}} |y|^{2(1+\eta)} f_i^*(y) \lambda(dy) < \infty,$$

(B5) for all $\beta > 0$, $\kappa > 0$, there exist a real number $\beta_0 > 0$, two increasing and positive sequences a_n and u_n tending to $+\infty$, and a sequence l_n decreasing to 0 such that

$$\pi_P\Big(P \ : \ P((-a_n, a_n] \times (l_n, u_n]) < 1 - \kappa\Big) \leqslant \exp(-n\beta_0),$$

$$\text{with} \qquad \frac{a_n}{l_n} \leqslant n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) \leqslant n\beta.$$

Proposition 3.1. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leqslant i \leqslant k$, $\mu_i \geqslant \underline{q}$. Assume that Q^* is in the support of π_Q and that the weak support of π_P contains all probability measures that are compactly supported.*

Then

- *(B1), (B2), (B3), (B4) imply (A1)*
- *and (B5) implies (A2).*

In particular in the case where π_P is the Dirichlet process $DP(\alpha G_0)$ with base measure αG_0, where G_0 is a probability measure on $\mathbb{R} \times (0, +\infty)$ and $\alpha > 0$, Assumption (B1) holds as soon as

$$\int_{\mathbb{R}\times(0,+\infty)} \frac{1}{\sigma} G_0(dz, d\sigma) < +\infty. \tag{7}$$

Indeed,

$$\begin{aligned}\int\int \frac{1}{\sigma} P(dz, d\sigma)\pi_P(dP) &= \int\int\int_{[\sigma,+\infty)} \frac{1}{t^2}\lambda(dt)P(dz, d\sigma)\pi_P(dP)\\ &= \int \frac{1}{\sigma} G_0(dz, d\sigma).\end{aligned}$$

Moreover using Remark 3.1 of Tokdar [19], Assumption (B5) easily holds as soon as for all $\beta > 0$, there exist a real number $\beta_0 > 0$, two increasing and positive sequences a_n and u_n tending to $+\infty$ and a sequence l_n decreasing to 0 such that

$$\begin{aligned} &G_0\left((-a_n, a_n] \times (l_n, u_n]\right)^c) \leqslant \exp(-n\beta_0),\\ &\frac{a_n}{l_n} \leqslant n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) \leqslant n\beta. \end{aligned} \tag{8}$$

3.2. Translated emission distributions

In this section we consider the special case of translated emission distributions, that is to say for all $1 \leqslant j \leqslant k$,

$$f_j(\cdot) = g(\cdot - m_j),$$

where g is a density function on $\mathbb{R}$ with respect to λ and for all $1 \leqslant j \leqslant k$, m_j is in $\mathbb{R}$. In this part, λ is still the Lebesgue measure on $\mathbb{R}$ and $d = 1$. This model has been in particular considered by Yau et al. [21] for the analysis of genomic copy number variation. First a corollary of Theorem 2.3 is given. Then the particular case of location-scale mixture of Gaussian distributions on g is studied.

Let

$$\Xi = \{\xi = (Q, m, g), Q \in \Delta_k^k, m \in \mathbb{R}^k, m_1 = 0 < m_2 < \cdots < m_k, g \in \mathcal{F}\}$$

and

$$\Xi(\underline{q}) = \{\xi = (Q, m, g) \in \Xi, Q \in \Delta^k(\underline{q})\}.$$

To $\xi = (Q, m, g) \in \Xi$, we associate $\theta = (Q, (g(\cdot - m_1), \ldots, g(\cdot - m_k))) \in \Theta$. We then denote $\mathbb{P}^\xi$ for $\mathbb{P}^\theta$. We assume that π_f is a product of probability measures,

$$\pi_f = \pi_m \otimes \pi_g,$$

where π_g is a probability measure on $\mathcal{F}$ and π_m is a probability measure on $\mathbb{R}^k$. Note that under Ξ, the model is completely identifiable, see Theorem 2.1 of Gassiat and Rousseau [13]. The uncertainty due to label switching is resolved here. In Corollary 3.2, additionally to posterior consistency for the transition matrices, we obtain posterior consistency for the parameters of translation m_j and for the weak convergence on the translated probability measure $g\lambda$. Under a stronger assumption, we get posterior consistency for the L_1-topology on the translated density distribution.

Fix $l \geqslant 3$. The following assumption replaces (A2) in the context of translated emission distributions:

(C2) for all $n > 0$, for all $\delta > 0$, there exists a set $\mathcal{F}_n \subset \mathbb{R}^k \times \mathcal{F}$ and a real number $r_1 > 0$ such that $\pi_f\big((\mathcal{F}_n)^c\big) \lesssim e^{-nr_1}$ and

$$\sum_{n>0} N\left(\frac{\delta}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \exp\left(-\frac{n\delta^2 k^2 \underline{q}^2}{32l}\right) < +\infty.$$

Corollary 3.2. *Let $\xi^* = (Q^*, m^*, g^*)$ be in $\Xi(\underline{q})$ such that $m_1^* = 0 < m_2^* < \cdots < m_k^*$ and Q^* has full rank.*

If the posterior is consistent for the D_l pseudometric with $l \geqslant 3$, i.e. if for all $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to\infty} \pi(\ \{\xi :\ D_l(\xi, \xi^*) < \varepsilon\}\ |Y_{1:n}) = 1\right) = 1.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi(\{Q : \|Q - Q^*\| < \varepsilon\} \mid Y_{1:n}) = 1\right) = 1,$$

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi(\{m : \forall 1 \leqslant j \leqslant k,\ |m_j - m_j^*| < \varepsilon\} \mid Y_{1:n}) = 1\right) = 1,$$

and for all $N \in \mathbb{N}$, *for all* $h_i \in \mathcal{C}_b(\mathbb{R}^d)$, *for all* $\varepsilon_i > 0,\ 1 \leqslant i \leqslant N$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi\left(\left\{g : \left|\int h_i g d\lambda - \int h_i g^* d\lambda\right| < \varepsilon_i\right\} \mid Y_{1:n}\right) = 1\right) = 1.$$

If moreover $\max_{1\leqslant j\leqslant k} \mu_j^* > 1/2$ *and* g^* *is uniformly continuous; then, for all* $\varepsilon > 0$,

$$\mathbb{P}^{\xi^*}\left(\lim_{n\to+\infty} \pi\left(\{g : \|g - g^*\|_{L_1(\lambda)} < \varepsilon\} \,|Y_{1:n}\right) = 1\right) = 1.$$

The proof of Corollary 3.2, in Appendix B, relies on the identifiability result of Gassiat and Rousseau [13] and Theorem 2.3.

In the same way as in Section 3.1, we propose to apply Theorem 2.1 and Corollary 3.2 to a prior based on location-scale mixtures of Gaussian distributions. In this part, we study a particular prior on the translated emission density g which is the location-scale mixture of Gaussian distributions. Then g is a sample drawn from π_g if

$$g(y) = \int_{\mathbb{R}\times(0,+\infty)} \phi_\sigma(y - z) dP(z, \sigma)$$

where P is a sample drawn from π_P and π_P is a probability measure on probability measures on $\mathbb{R} \times (0, +\infty)$. The following assumption help in proving (C2):

(D6) for all $\beta > 0$, $\kappa > 0$, there exist a real number $\beta_0 > 0$, three increasing sequences of positive numbers m_n, a_n and u_n tending to $+\infty$, and a sequence l_n decreasing to 0 such that

$$\pi_P\left(P \ :\ P((-a_n, a_n] \times (l_n, u_n]) < 1 - \kappa\right) \leqslant \exp(-n\beta_0),$$

$$\pi_m\left(([-m_n, m_n]^k)^c\right) \leqslant exp(-n\beta_0),$$

$$\frac{a_n}{l_n} \leqslant n\beta, \qquad \log\left(\frac{u_n}{l_n}\right) \leqslant n\beta, \qquad \log\left(\frac{m_n}{l_n}\right) \leqslant n\beta.$$

Proposition 3.3. *Let* $\underline{q} > 0$ *and* ξ^* *in* $\Xi(\underline{q})$. *Assume that the support of* π_Q *is included in* $\Delta^k(\underline{q})$ *and that for all* $1 \leqslant i \leqslant k$, $\mu_i \geqslant \underline{q}$. *Assume that* Q^* *is in the support of* π_Q, *that* m^* *is in the support of* π_m *and that the weak support of* π_P *contains all probability measures that are compactly supported.*

If (B1) is verified and (B2), (B3) and (B4) are verified with $f_j(\cdot) = g(\cdot - m_j),\ 1 \leqslant j \leqslant k$ *then (A1) holds.*

Moreover (D6) implies (C2).

The proof of Proposition 3.3 is very similar to that of Proposition 3.1 and is given in Appendix B.

Corollary 3.2 and Proposition 3.3 are less general than Theorem 2.3 and Proposition 3.1 respectively. In Corollary 3.2 and Proposition 3.3, it is assumed that the true emission distributions are translated versions of a unique density g^*. In practice, we expect priors on translated emission distributions not to be as robust as priors for which the emission distributions are i.i.d. such as priors of Section 3.1. Particularly if the true emission distributions have different tails, priors on translated emission distributions may lead to poor estimations.

3.3. Independent discrete emission distributions

Discrete emission distributions, i.e. when the support of λ is included in $\mathbb{N}$, have been successfully used, for instance in genomics in Gassiat, Cleynen and Robin [12].

Note that for discrete emission distributions, weak and l_1 topologies are the same so that weak posterior consistency implies l_1 posterior consistency. Thus Assumption (A2) becomes unnecessary in Theorems 2.1 and 2.3. Moreover posterior consistency for the emission distributions in the weak topology in Theorem 2.3 implies posterior consistency for the emission distributions in l_1.

In the discrete case, we prove in Theorem 3.4 that posterior consistency for the marginal distribution of finitely many observations, for the transition matrix and for the emission distributions in l_1 together with the restriction of the prior π_Q on $\Delta^k(\underline{q})$ imply posterior consistency for the marginal smoothing:

Theorem 3.4. *Let $\underline{q} > 0$. Assume that the support of π_Q is included in $\Delta^k(\underline{q})$ and that for all $1 \leqslant i \leqslant k$, $\mu_i \geqslant \underline{q}$. If $f_1^*\lambda, \dots, f_k^*\lambda$ are linearly independent, Q^* has full rank, and (A0) and (A1) hold; then, for all finite integer m,*

$$\lim_{n\to+\infty} \pi\Bigg(\Big\{\theta : \exists \sigma \in \mathcal{S}_k, \max_{1\leqslant a_j\leqslant k,\ 1\leqslant j\leqslant m} |P^\theta(X_i = \sigma(a_i),\ \forall\ 1 \leqslant i \leqslant m\ |Y_{1:n}) - P^{\theta^*}(X_i = a_i,\ \forall\ 1 \leqslant i \leqslant m \mid Y_{1:n})| < \varepsilon\Big\}\Big| Y_{1:n}\Bigg) = 1 \text{ in } P^{\theta^*}\text{-probability.}$$

The proof of Theorem 3.4 is given in Appendix A.

In the following we apply Theorems 2.1, 2.3 and 3.4 to a specific prior on the set of probability measures on $\mathbb{N}$ in the case of a HMM with discrete emission distributions. We consider a Dirichlet process $DP(\alpha G_0)$ with α a positive number and G_0 some probability measure on $\mathbb{N}$. We then consider a prior probability measure on Θ defined by

$$\pi = \pi_Q \otimes DP(\alpha G_0)^{\otimes k}.$$

In Proposition 3.5, we give sufficient and almost necessary conditions to obtain (A1). Proposition 3.5 is proved in Appendix A.

Proposition 3.5. *Let $\underline{q} > 0$. Assume that the support of the prior π_Q is included in $\Delta^k(\underline{q})$, that Q^* is in the support of π_Q and that for all $1 \leqslant i \leqslant k$, $\mu_i \geqslant \underline{q}$.*

If

$$\textit{(E1) for all } 1 \leqslant i \leqslant k, \ \sum_{l\in\mathbb{N}} \frac{f_i^*(l)}{G_0(l)} < +\infty$$

then (A1) holds.

Moreover if

$$\textit{(T) for all } 1 \leqslant i \leqslant k, \sum_{l\in\mathbb{N}} f_i^*(l)(-\log f_i^*(l)) < +\infty$$

then (A1b) implies (E1).

Remark 3.1. Therefore (E1) is not only sufficient to prove (A1b) but up to the weak assumption (T) it is also necessary. Assumption (E1) relies on the mutual control of the tails of the base measure G_0 and the true emission distributions f_j^*. Proposition 3.5 suggests choosing a heavy tailed probability measure G_0 with $G_0(l) > 0$, for all $l \in \mathbb{N}$.

Remark 3.2. We deduce from Proposition 3.5 that

$$\left\{ g^* : \ \mathbb{N} \to (0,1) \quad \text{such that} \quad \sum_{l\in\mathbb{N}} g^*(l) = 1, \right.$$
$$\left. \sum_{l\in\mathbb{N}} g^*(l)(-\log(g^*(l)) < +\infty \quad \text{and} \quad \sum_{l\in\mathbb{N}} \frac{g^*(l)}{G_0(l)} < +\infty \right\} \tag{9}$$

is a subset of the Kullback-Leibler support of the Dirichlet process $DP(\alpha G_0)$.

Acknowledgements

I want to thank Elisabeth Gassiat and Judith Rousseau for their valuable comments. I also want to thank the reviewer and the editor for their helpful comments. The research was partly supported by the grants ANR Banhdits and ANR Ipanema.

Appendix A: Proofs of key results

Proof of Lemma 2.2

For all $\theta,\ \theta^* \in \Delta^k(\underline{q})$ the Kullback-Leibler divergence between $p_n^{\theta^*}$ and p_n^{θ} is by definition equal to

$$\frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\frac{\sum_{i_1,\dots,i_n=1}^k \mu_{i_1}^* Q_{i_1,i_2}^* \dots Q_{i_{n-1},i_n}^* f_{i_1}^*(Y_1)\dots f_{i_n}^*(Y_n)}{\sum_{j_1,\dots,j_n=1}^k \mu_{j_1} Q_{j_1,j_2}\dots Q_{j_{n-1},j_n} f_{j_1}(Y_1)\dots f_{j_n}(Y_n)}\right)\right).$$

Multiplying and dividing each term of the sum in the numerator by

$$\mu_{i_1} Q_{i_1,i_2}\dots Q_{i_{n-1},i_n} f_{i_1}(Y_1)\dots f_{i_n}(Y_n),$$

we obtain

$$\frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\frac{\sum_{i_1,\cdots,i_n=1}^{k}\frac{\mu_{i_1}^*Q_{i_1,i_2}^*\cdots Q_{i_{n-1},i_n}^*f_{i_1}^*(Y_1)\cdots f_{i_n}^*(Y_n)}{\mu_{i_1}Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\cdots f_{i_n}(Y_n)}\mu_{i_1}Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\cdots f_{i_n}(Y_n)}{\sum_{j_1,\ldots,j_n=1}^{k}\mu_{j_1}Q_{j_1,j_2}\cdots Q_{j_{n-1},j_n}f_{j_1}(Y_1)\ldots f_{j_n}(Y_n)}\right)\right)$$
$$\leqslant \frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leqslant i_1,\ldots,i_n\leqslant k}\frac{\mu_{i_1}^*Q_{i_1,i_2}^*\cdots Q_{i_{n-1},i_n}^*f_{i_1}^*(Y_1)\ldots f_{i_n}^*(Y_n)}{\mu_{i_1}Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}f_{i_1}(Y_1)\ldots f_{i_n}(Y_n)}\right)\right)$$

by bounding the quotient in each term of the sum of the numerator by its maximum. Since the maximum of a product of positive factors is bounded by the product of the maxima,

$$\begin{aligned}
&\frac{1}{n}KL(p_n^{\theta^*},p_n^{\theta,\mu})\\
&\leqslant \frac{1}{n}\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leqslant i_0\leqslant k}\frac{\mu_{i_0}^*}{\mu_{i_0}}\left(\max_{1\leqslant i,j\leqslant k}\frac{Q_{i,j}^*}{Q_{i,j}}\right)^{n-1}\max_{1\leqslant i_1\leqslant k}\frac{f_{i_1}^*(Y_1)}{f_{i_1}(Y_1)}\cdots\max_{1\leqslant i_n\leqslant k}\frac{f_{i_n}^*(Y_n)}{f_{i_n}(Y_n)}\right)\right)\\
&\leqslant \frac{1}{n\underline{q}}\max_{1\leqslant i_0\leqslant k}\left|\mu_{i_0}-\mu_{i_0}^*\right|+\frac{n-1}{n\underline{q}}\max_{1\leqslant i,j\leqslant k}\left|Q_{i,j}-Q_{i,j}^*\right|\\
&\qquad\qquad+\max_{1\leqslant j\leqslant k}\int f_j^*(y)\max_{1\leqslant i\leqslant k}\log\frac{f_i^*(y)}{f_i(y)}\lambda(dy).
\end{aligned}$$

The last inequality comes from the following inequalities

$$\begin{aligned}
&\mathbb{E}_{p_n^{\theta^*}}\left(\log\left(\max_{1\leqslant i_s\leqslant k}\frac{f_{i_s}^*(Y_s)}{f_{i_s}(Y_s)}\right)\right)\\
&=\sum_{j_1,\ldots,j_n=1}^{k}\mu_{j_1}^*Q_{j_1,j_2}^*\cdots Q_{j_{n-1},j_n}^*\int f_{j_s}^*(y)\log\max_{1\leqslant i_s\leqslant k}\left(\frac{f_{i_s}^*(y)}{f_{i_s}(y)}\right)\lambda(dy)\prod_{1\leqslant t\neq s\leqslant n}\int f_{j_t}^*(y)\lambda(dy)\\
&\leqslant \max_{1\leqslant j_1\leqslant k}\int f_{j_1}^*(y)\max_{1\leqslant i_1\leqslant k}\log\frac{f_{i_1}^*(y)}{f_{i_1}(y)}\lambda(dy),
\end{aligned}$$

$$\log\left(\max_{1\leqslant i_0\leqslant k}\frac{\mu_{i_0}^*}{\mu_{i_0}}\right)\leqslant\frac{1}{\underline{q}}\max_{1\leqslant i_0\leqslant k}\left|\mu_{i_0}-\mu_{i_0}^*\right|,$$

and

$$\log\left(\max_{1\leqslant i,j\leqslant k}\frac{Q_{i,j}^*}{Q_{i,j}}\right)\leqslant\frac{1}{\underline{q}}\max_{1\leqslant i,j\leqslant k}\left|Q_{i,j}-Q_{i,j}^*\right|$$

because $\min_{1\leqslant i,j\leqslant k}(\mu_i,\mu_i^*,Q_{i,j},Q_{i,j}^*)\geqslant\underline{q}$.

Then for all $\varepsilon>0$, for n large enough, for all $\theta\in\Theta_\varepsilon$,

$$\frac{1}{n}KL(p_n^{\theta^*},p_n^{\theta,\mu})\leqslant\frac{3}{\underline{q}}\,\varepsilon.$$

Proof of Theorem 2.1

This proof relies on Theorem 5 of Barron [1]. We do not assume (A2) in the first part of the proof. First we prove that for all $a > 0$,

$$\mathbb{P}^{\theta^*}\left(\frac{\int_\Theta p_n^\theta(Y_1,\dots,Y_n)\pi(d\theta)}{p_n^{\theta^*}(Y_1,\dots,Y_n)} \leqslant \exp(-an) \text{ i.o.}\right) = 0 \tag{10}$$

that is to say

$$p_n^{\theta^*}(y_1,\dots,y_n)\lambda(dy_1)\dots\lambda(dy_n)$$

and

$$\int_\Theta p_n^\theta(y_1,\dots,y_n)\lambda(dy_1)\dots\lambda(dy_n)\pi(d\theta)$$

merge with probability one.

Let $\varepsilon > 0$. Note that Assumption (A1a) implies that $Q^* \in \Delta^k(\underline{q})$. Then by Lemma 2.2, there exists a real $\tilde{\varepsilon} > 0$ such that for n large enough, for all $\theta \in \Theta_{\tilde{\varepsilon}}$,

$$\frac{1}{n}KL(p_n^{\theta^*}, p_n^{\theta,\mu}) < \varepsilon. \tag{11}$$

Assumptions (A0), (A1b) and (A1d) imply that

$$\sum_{i=1}^k \int f_i^*(y)\left|\log\left(\sum_{j=1}^k f_j(y)\right)\right|\lambda(dy) < +\infty. \tag{12}$$

Indeed

$$\begin{aligned}&\int f_i^*(y)\left|\log\left(\sum_{j=1}^k f_j(y)\right)\right|\lambda(dy)\\ &\leqslant \int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy) + \int_{\{y:\ f_i(y)\geqslant 1\}} f_i^*(y)\log(k\max_{1\leqslant j\leqslant k} f_j(y))\lambda(dy)\end{aligned}$$

and

$$\int_{\{y:\ f_i(y)\geqslant 1\}} f_i^*(y)\log(k\max_{1\leqslant j\leqslant k} f_j(y))\lambda(dy)$$

is finite under (A1d) and

$$\int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy)$$

is finite under (A0), (A1b) and (A1d) since

$$\begin{aligned}\int f_i^*(y)\max_{1\leqslant j\leqslant k}\log\left(\frac{f_j^*(y)}{f_j(y)}\right)\lambda(dy) &\geqslant \int f_i^*(y)\log(f_i^*(y))\lambda(dy)\\ &\quad + \int_{\{y:\ f_i(y)<1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy) + \int_{\{y:\ f_i(y)\geqslant 1\}} f_i^*(y)(-\log(f_i(y)))\lambda(dy).\end{aligned}$$

Moreover by Proposition 1 of Douc, Moulines and Rydén [7], if $\theta \in \Theta(\underline{q})$ and if (A1c), (A1d) and (12) hold,

$$\frac{1}{n}\log\left(\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta,\mu}(Y_{1:n})}\right)$$

converges $\mathbb{P}^{\theta^*}$-almost surely and in $L^1(\mathbb{P}^{\theta^*})$. Let $\bar{L}(\theta)$ denote this limit:

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{p_n^{\theta^*}(Y_{1:n})}{p_n^{\theta,\mu}(Y_{1:n})}\right) =: \bar{L}(\theta),\ \mathbb{P}^{\theta^*}\text{-a.s. and in } L^1(\mathbb{P}^{\theta^*}).$$

Then using Equation (11), for all $\theta \in \Theta_{\bar{\varepsilon}}$,

$$\bar{L}(\theta) \leqslant \varepsilon. \tag{13}$$

So that for all $\varepsilon > 0$, there exists $\tilde{\varepsilon}$ such that

$$\pi\left(\theta : \bar{L}(\theta) < \varepsilon\right) \ \geqslant\ \pi(\Theta_{\tilde{\varepsilon}}) \ > 0.$$

By Lemma 10 of Barron [1], for all $a > 0$, (10) is verified.

We now have to build the tests described in Theorem 5 in Barron [1], to obtain posterior consistency first for the weak topology and secondly for the D_l-pseudometric. In the case of the weak topology, we follow the ideas of Section 4.4.1 in Ghosh and Ramamoorthi [15]. Using page 142 of Ghosh and Ramamoorthi [15], it is sufficient to consider

$$U = \left\{P \ :\ \int h dP - \int h p_l^{\theta^*} d\lambda^{\otimes l} < \varepsilon,\right\},$$

for all $\varepsilon > 0$ and $0 \leqslant h \leqslant 1$ in the set $\mathcal{C}_b((\mathbb{R}^d)^l)$. Choosing α and γ as in page 128 of Ghosh and Ramamoorthi [15], if

$$S^n = \left\{y_1, \ldots, y_n \ :\ \frac{l}{n}\sum_{j=0}^{n/l-1} h(y_{jl+1}, \ldots, y_{jl+l}) > \frac{\alpha+\gamma}{2}\right\},$$

then

$$\begin{aligned} P^{\theta^*}(S^n) &= P^{\theta^*}\left\{\sum_{j=0}^{n/l-1}\left(h(y_{jl+1}, \ldots, y_{jl+l}) - \int h p_l^{\theta^*} d\lambda^{\otimes l}\right) > \frac{n}{l}\frac{\gamma-\alpha}{2}\right\} \\ &\leqslant \exp\left(-\frac{n(\gamma-\alpha)^2(\min_{i,j} Q^*_{i,j})^2}{2l(2-k\min_{i,j} Q^*_{i,j})^2}\right) \end{aligned} \tag{14}$$

and for all $\theta \in \Theta(\underline{q})$ such that $\int h dP^\theta - \int h p_l^{\theta^*} d\lambda^{\otimes l} \geqslant \varepsilon$,

$$P^\theta((S^n)^c) \leqslant P^\theta\left\{\sum_{j=0}^{n/l-1}\left(-h(y_{jl+1}, \ldots, y_{jl+l}) + \int h p_l^{\theta} d\lambda^{\otimes l}\right) \geqslant \frac{n}{l}\frac{\gamma-\alpha}{2}\right\}$$

$$\leqslant \exp\left(-\frac{n(\gamma-\alpha)^2(\min_{i,j} Q_{i,j})^2}{2l(2-k\min_{i,j} Q_{i,j})^2}\right) \leqslant \exp\left(-\frac{n(\gamma-\alpha)^2\underline{q}^2}{2l}\right), \quad (15)$$

using the upper bound from the proof of Theorem 4 of Gassiat and Rousseau [14] based on Corollary 1 of Rio [18].

Using Theorem 5 of Barron [1] and combining Equations (14) and (15),

$$P^{\theta^*}\left(\pi\left(\left\{\theta : \int h dP^\theta - \int h p_l^{\theta^*} d\lambda^{\otimes l} < \varepsilon\right\}^c \,\middle|\, Y_{1:n}\right) \geqslant e^{-nr}, \text{ i.o. }\right) = 0$$

which implies that for all weak neighborhood U of $P_l^{\theta^*}$,

$$P^{\theta^*}\left(\pi(U^c|Y_{1:n}) \geqslant \exp(-nr) \text{ i.o. }\right) = 0,$$

so that

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(U|Y_{1:n}) = 1\right) = 1.$$

We now assume (A2) and obtain consistency for the D_l-pseudometric. Let $\varepsilon > 0$ and let

$$U = \left\{\theta \;:\; D_l(\theta,\theta^*) < \frac{2\varepsilon}{k\underline{q}}\right\} \supset \left\{\theta \;:\; D_l(\theta,\theta^*) < \varepsilon\frac{2-k\min_{1\leqslant i,j\leqslant k} Q_{i,j}}{k\min_{1\leqslant i,j\leqslant k} Q_{i,j}}\right\},$$

be a D_l-neighborhood of θ^*. Let

$$B_n^c = \Delta^k(\underline{q}) \times \mathcal{F}_n,$$

so that

$$\pi(B_n) = \pi_f(\mathcal{F}_n{}^c) \lesssim \exp(-nr_1). \quad (16)$$

In the proof of Theorem 4 of Gassiat and Rousseau [14], it is proved that for all n large enough, there exists a test ψ_n such that

$$\begin{aligned}\mathbb{E}^{\theta^*}(\psi_n) &\leqslant N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \exp\left(-\frac{n\varepsilon^2}{8l}\frac{k^2(\min_{i,j} Q^*_{i,j})^2}{(2-k\min_{i,j} Q^*_{i,j})^2}\right) \\ &\leqslant N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \exp\left(-\frac{n\varepsilon^2k^2\underline{q}^2}{32l}\right)\end{aligned} \quad (17)$$

$$\sup_{\theta\in U^c\cap B_n^c} \mathbb{P}^{\theta,\mu}(1-\psi_n) \leqslant \exp\left(-\frac{n\varepsilon^2}{32l}\right). \quad (18)$$

Note that for all $\theta, \tilde{\theta}$ in $\Theta(\underline{q})$,

$$D_l(\theta,\tilde{\theta}) \leqslant \sum_{1\leqslant i\leqslant k} |\mu_i^\theta - \mu_i^{\tilde{\theta}}| + k(l-1)\|Q-\tilde{Q}\| + l\max_{1\leqslant j\leqslant k}\|f_j - \tilde{f}_j\|_{L_1(\lambda)}.$$

The function $Q \to \mu^Q$ is continuous on the compact $\Delta^k(\underline{q})$ and thus is uniformly continuous: there exists $\alpha > 0$ such that for all $\theta, \tilde{\theta}$ in $\Theta(\underline{q})$ such that $\|Q - \tilde{Q}\| < \alpha$ then $\sum_{1 \leqslant i \leqslant k} |\mu_i^\theta - \mu_i^{\tilde{\theta}}| < \frac{\varepsilon}{36}$. This implies that

$$\begin{aligned} & N\left(\frac{\varepsilon}{12}, \Delta^k(\underline{q}) \times \mathcal{F}_n, D_l\right) \\ & \leqslant N\left(\min\left(\frac{\varepsilon}{36k(l-1)}, \alpha\right), \Delta^k(\underline{q}), \|\cdot\|\right) N\left(\frac{\varepsilon}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right) \\ & \leqslant \left(\max\left(\frac{36k(l-1)}{\varepsilon}, \frac{1}{\alpha}\right)\right)^{k(k-1)} N\left(\frac{\varepsilon}{36l}, \mathcal{F}_n, d(\cdot,\cdot)\right). \end{aligned} \tag{19}$$

Then combining Equations (16), (17), (18), (19) and using Theorem 5 of Barron [1], there exists $r > 0$ such that

$$\mathbb{P}^{\theta^*}\left(\pi\left(U^c | Y_{1:n}\right) \geqslant \exp(-nr) \text{ i.o.}\right) = 0. \tag{20}$$

And Equation (20) implies that for all $\varepsilon > 0$,

$$\mathbb{P}^{\theta^*}\left(\lim_{n\to\infty} \pi(\ \{\theta :\ D_l(\theta, \theta^*) < \varepsilon\}\ \mid Y_{1:n}) = 1\right) = 1.$$

Proof of Theorem 2.3

It is sufficient to show that for all weak neighborhood U_{f^*} of $f^*\lambda$ and neighborhood U_{Q^*} of Q^*, there exists a D_3-neighborhood U_{θ^*} of θ^* such that

$$U_{\theta^*} \subset \left\{\exists \sigma \in \mathcal{S}_k;\ \sigma Q \in U_{Q^*},\ f_{\sigma(i)} \in U_{f_i^*},\ i = 1 \dots k\right\}. \tag{21}$$

Following Gassiat, Cleynen and Robin [12], it is equivalent to show that for all sequences θ^n in $\Theta(\underline{q})$ such that $D_3(\theta^n, \theta^*) \to 0$, there exists a subsequence, that we denote again θ^n, of θ^n and $\bar{\theta} \in \Theta$ such that $\|Q^n - \bar{Q}\| \to 0$, $f_i^n\lambda$ tends to $\bar{f}_i\lambda$ in the weak topology on probability measures for all $i \leqslant k$ and $p_3^{(Q^*, f^*)} = p_3^{(\bar{Q}, \bar{f})}$.

Let θ^n in $\Theta(\underline{q})$ such that $D_3(\theta^n, \theta^*) \to 0$. As $\Delta^k(\underline{q})$ is a compact set, there exists a subsequence of Q^n that we denote again Q^n which tends to $\bar{Q} \in \Delta^k(\underline{q})$. Writing μ^n the (sub)sequence of the stationary distribution associated to Q_n, then $\mu^n \to \bar{\mu}$ where $\bar{\mu}$ is the stationary distribution associated to $\bar{Q}$. Moreover, using the reverse triangle inequality,

$$\begin{aligned} D_3(\theta^n, \theta^*) &= \|p_3^{\theta_n} - p_3^{\theta^*}\|_{L_1(\lambda^{\otimes 3})} \\ &= \int \Big| \sum_{1 \leqslant i_1, i_2, i_3 \leqslant k} \mu_{i_1}^n Q_{i_1,i_2}^n Q_{i_2,i_3}^n f_{i_1}^n(y_1) f_{i_2}^n(y_2) f_{i_3}^n(y_3) - \\ &\quad \mu_{i_1}^* Q_{i_1,i_2}^* Q_{i_2,i_3}^* f_{i_1}^*(y_1) f_{i_2}^*(y_2) f_{i_3}^*(y_3) \Big| \ \lambda(dy_1)\lambda(dy_2)\lambda(dy_3) \\ &\geqslant - \sum_{1 \leqslant i_1, i_2, i_3 \leqslant k} \left| \mu_{i_1}^n Q_{i_1,i_2}^n Q_{i_2,i_3}^n - \bar{\mu}_{i_1} \bar{Q}_{i_1,i_2} \bar{Q}_{i_2,i_3} \right| + \end{aligned}$$

$$\int \Big| \sum_{1\leqslant i_1,i_2,i_3\leqslant k} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} f^n_{i_1}(y_1)f^n_{i_2}(y_2)f^n_{i_3}(y_3) - \mu^*_{i_1}Q^*_{i_1,i_2}Q^*_{i_2,i_3}f^*_{i_1}(y_1)f^*_{i_2}(y_2)f^*_{i_3}(y_3)\Big| \lambda(dy_1)\lambda(dy_2)\lambda(dy_3),$$

since $\sum_{1\leqslant i_1,i_2,i_3\leqslant k} \Big|\mu^n_{i_1}Q^n_{i_1,i_2}Q^n_{i_2,i_3} - \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3}\Big|$ tends to zero,

$$\lim_n \int \Big| \sum_{1\leqslant i_1,i_2,i_3\leqslant k} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} f^n_{i_1}(y_1)f^n_{i_2}(y_2)f^n_{i_3}(y_3) - \mu^*_{i_1}Q^*_{i_1,i_2}Q^*_{i_2,i_3}f^*_{i_1}(y_1)f^*_{i_2}(y_2)f^*_{i_3}(y_3)\Big| \lambda(dy_1)\lambda(dy_2)\lambda(dy_3) = 0. \tag{22}$$

Let $F^n_1,\dots,F^n_k$ be the probability distribution with respective densities f^n_1, $\dots, f^n_k$ with respect to λ. Since

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} F^n_{i_1}\otimes F^n_{i_2}\otimes F^n_{i_3}$$

converges in total variation, it is tight and for all $1\leqslant i\leqslant k$, $(F^n_i)_n$ is tight. By Prohorov's theorem, for all $1\leqslant i\leqslant k$ there exists a subsequence denoted F^n_i of F^n_i which weakly converges to $\bar{F}_i$. This in turns implies that

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} F^n_{i_1}\otimes F^n_{i_2}\otimes F^n_{i_3}$$

weakly converges to

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} \bar{F}_{i_1}\otimes \bar{F}_{i_2}\otimes \bar{F}_{i_3},$$

which combined with (22), leads to

$$\sum_{i_1,i_2,i_3} \bar{\mu}_{i_1}\bar{Q}_{i_1,i_2}\bar{Q}_{i_2,i_3} \bar{F}_{i_1}\otimes \bar{F}_{i_2}\otimes \bar{F}_{i_3} = \sum_{i_1,i_2,i_3} \mu^*_{i_1}Q^*_{i_1,i_2}Q^*_{i_2,i_3} f^*_{i_1}\lambda\otimes f^*_{i_2}\lambda\otimes f^*_{i_3}\lambda.$$

By Gassiat, Cleynen and Robin [12], $\bar{Q}=Q^*$, so $\bar{\mu}=\mu^*$ and $\bar{F}_i = f^*_i\lambda$ up to a label switching, that is there exists a permutation $\sigma\in\mathcal{S}_k$ such that $\sigma\bar{Q}=Q^*$ and $\bar{F}_{\sigma(i)}=f^*_i\lambda$ so that Equation (21) holds.

In other words we have proved the continuity of the functional

$$\begin{cases} (\{p^\theta_l,\ \theta\in\Theta_I\},L_1) & \to & (\Theta_I/\mathcal{R}_\sigma,\mathcal{T}) \\ p^\theta_l & \mapsto & \theta \end{cases}$$

where $\Theta_I=\{\theta\in\Theta : Q \text{ has full rank }, f_1d\lambda\dots f_kd\lambda \text{ are linearly independent}\}$ and $\mathcal{R}_\sigma$ is the equivalence relation on Θ such that $\theta\mathcal{R}_\sigma\tilde{\theta}$ if there exists $\sigma\in\mathcal{S}_k$

such that for all $1 \leqslant i, j \leqslant k$, $Q_{i,j} = \tilde{Q}_{\sigma(i),\sigma(j)}$ and $f_i = \tilde{f}_{\sigma(i)}$; using that

$$\underset{p_l^\theta}{\left(\{p_l^\theta, \theta \in \Theta_I\}, L_1\right)} \xrightarrow[\text{continuous}]{} \underset{p_l^\theta}{\left(\{p_l^\theta, \theta \in \Theta_I\}, \begin{array}{c}\text{weak}\\ \text{topology}\end{array}\right)} \underset{\substack{\longmapsto \\ \text{continuous, bijective}}}{\longrightarrow} \overbrace{\underset{\theta}{\left(\Theta_I/\mathcal{R}_\sigma, \mathcal{T}\right)}}^{\text{compact}}.$$

Proof of Proposition 2.4

To prove Proposition 2.4, using Equation (4), it is sufficient to prove that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\begin{aligned}&\left\{\theta \in \Theta : H_1(\theta) < \cdots < H_k(\theta), \exists \sigma \in \mathcal{S}_k, \|\sigma Q - Q^*\| < \eta, \max_{1 \leqslant i \leqslant k} d_w(f_{\sigma(i)}, f_i^*) < \eta\right\}\\ &\subset \left\{\theta : H_1(\theta) < \cdots < H_k(\theta), \|Q - Q^*\| < \varepsilon, \max_{1 \leqslant i \leqslant k} d_w(f_i, f_i^*) < \varepsilon\right\}\end{aligned} \tag{23}$$

where d_w metricizes the weak topology on $\mathcal{F}$. Using Equation (3),

$$\delta := \min_{1 \leqslant i \leqslant k-1} |H_{i+1}(\theta^*) - H_i(\theta^*)| > 0 \tag{24}$$

and by continuity of H for all $\varepsilon > 0$, there exists $\eta_1 > 0$ such that for all

$$\theta \in \left\{\theta \in \Theta : \ H_1(\theta) < \cdots H_k(\theta), \ \exists \sigma \in \mathcal{S}_k, \|\sigma Q - Q^*\| < \eta_1, d_w(f_{\sigma(i)}, f_i^*) < \eta_1\right\},$$

for all $1 \leqslant i \leqslant k$, $|H_i(\theta) - H_i(\theta^*)| < \delta/2$. For such θ, using Equation (2), we obtain for all $\sigma \in \mathcal{S}_k$,

$$|H_i((\sigma Q, f_{\sigma(1)}, \ldots, f_{\sigma(k)})) - H_{\sigma(i)}(\theta^*)| < \delta/2$$

so that using Equations (3), (24) and that $H_1(\theta) < \cdots < H_k(\theta)$, the permutation σ is equal to the identity permutation. Thus Equation (23) holds with $\eta = \min(\eta_1, \varepsilon)$.

Proof of Theorem 3.4

To prove Theorem 3.4 we need the following lemma:

Lemma A.1. *Let $\varepsilon > 0$, for all $0 < \varepsilon_1 < 1$, $N > 0$, $1 \leqslant j < N$ and $c > 0$ such that*

$$0 < \frac{\varepsilon_1 k^N}{c(c - \varepsilon_1)} < \frac{\varepsilon}{3} \text{ and } \frac{2(1 - \underline{q})^{N+1-j}}{\underline{q} + (1 - \underline{q})^{N+1-j}} < \frac{\varepsilon}{3}.$$

If

$$p_N^{\theta^*}(Y_{1:N}) > c \tag{25}$$

then for all $n > N$,

$$\Big\{\theta \in \Theta(\underline{q}) : \|p_N^{\theta^*} - p_N^{\theta}\|_{l_1} < \varepsilon_1,\ \exists \sigma \in \mathcal{S}_k,\ \max_{1\leqslant i\leqslant k} |\mu_{\sigma(i)}^{\theta} - \mu_i^*| < \varepsilon_1, \\ \|\sigma Q - Q^*\| < \varepsilon_1,\ \max_{1\leqslant i\leqslant k} \|f_{\sigma(i)} - f_i^*\|_{l_1} < \varepsilon_1\Big\} \\ \subset \Big\{\theta \in \Theta(\underline{q}) : \exists \sigma \in \mathcal{S}_k, \max_{1\leqslant l\leqslant k} \big|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:n})\big| < \varepsilon\Big\}.$$

Proof of Lemma A.1. Let $\theta \in \Theta(\underline{q})$ such that

$$\|p_N^{\theta^*} - p_N^{\theta}\|_{l_1} < \varepsilon_1 \tag{26}$$

and there exists $\sigma \in \mathcal{S}_k$ such that

$$\max_{1\leqslant i\leqslant k} |\mu_{\sigma(i)}^{\theta} - \mu_i^*| < \varepsilon_1,\ \|\sigma Q - Q^*\| < \varepsilon_1,\ \max_{1\leqslant i\leqslant k} \|f_{\sigma(i)} - f_i^*\|_{l_1} < \varepsilon_1. \tag{26}$$

To bound $|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = l \mid Y_{1:n})|$, we now prove that it is sufficient to bound $|P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N})|$ with $N < n$ a well chosen fixed integer thanks to the exponential forgetting of the HMM. Let $1 \leqslant a \leqslant k$,

$$\begin{aligned} &|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:n})| \\ &\quad\leqslant A_{\theta^*}^l + |P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^{\theta}(X_j = \sigma(l) \mid Y_{1:N})| + A_{\theta}^{\sigma(l)}, \end{aligned} \tag{27}$$

where for $\tilde{\theta} \in \{\theta, \theta^*\}$ and for all $1 \leqslant l \leqslant k$,

$$A_{\tilde{\theta}}^l = \left| \frac{P^{\tilde{\theta}}(Y_{1:N}, X_j = l) \sum_{1\leqslant b\leqslant k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = l, Y_{j:N})}{\sum_{1\leqslant m\leqslant k} P^{\tilde{\theta}}(Y_{1:N}, X_j = m) \sum_{1\leqslant b\leqslant k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = m, Y_{j:N})} \right. \\ \left. - \frac{P^{\tilde{\theta}}(Y_{1:N}, X_j = l) \sum_{1\leqslant b\leqslant k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = a, Y_{j:N})}{\sum_{1\leqslant m\leqslant k} P^{\tilde{\theta}}(Y_{1:N}, X_j = m) \sum_{1\leqslant b\leqslant k} P^{\tilde{\theta}}(Y_{N+1:n} \mid X_{N+1} = b) P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = a, Y_{j:N})} \right|.$$

Using Corollary 1 of Douc, Moulines and Rydén [7], i.e. the exponential forgetting of the HMM, we obtain for all $(b, \omega, m) \in \{1, \dots, k\}^3$,

$$\begin{aligned} &\big|P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = m, Y_{j:N}) - P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = \omega, Y_{j:N})\big| \\ &\quad\leqslant (1 - \underline{q})^{N+1-j} \leqslant (1 - \underline{q})^{N+1-j} \frac{P^{\tilde{\theta}}(X_{N+1} = b \mid X_j = \omega, Y_{j:N})}{\underline{q}} \end{aligned}$$

so that for $\tilde{\theta} \in \{\theta, \theta^*\}$ and for all $1 \leqslant l \leqslant k$

$$A_{\tilde{\theta}}^l \leqslant \frac{2(1 - \underline{q})^{N+1-j}}{\underline{q} + (1 - \underline{q})^{N+1-j}}. \tag{28}$$

Moreover, using (25) and (26), for all $1 \leqslant i, j \leqslant k$, $Y_{1:N} \in \mathbb{N}^N$,

$$\mu^\theta_{\sigma(i)} \geqslant \mu^*_i - \varepsilon_1, \quad Q_{\sigma(i),\sigma(j)} \geqslant Q^*_{i,j} - \varepsilon_1, \quad f_{\sigma(a_i)}(Y_i) \geqslant f^*_{a_i}(Y_i) - \varepsilon_1 \quad \text{and}$$
$$p^\theta_N(Y_{1:N}) \leqslant p^{\theta^*}_N(Y_{1:N})(1+\varepsilon_1/c),$$

we obtain

$$\begin{aligned}
&P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^\theta(X_j = \sigma(l) \mid Y_{1:N}) \\
&= \frac{\sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} Q^*_{a_1,a_2} \cdot\cdot Q^*_{a_{j-1},l} Q^*_{l,a_{j+1}} \cdot\cdot Q^*_{a_{N-1},a_N} f^*_{a_1}(Y_1) \cdot\cdot f^*_l(Y_j) \cdot\cdot f^*_{a_N}(Y_N)}{p^{\theta^*}_N(Y_{1:N})} \\
&\quad - \frac{\sum_{a_{1:j-1}, a_{j+1:N}} \mu^\theta_{\sigma(a_1)} Q_{\sigma(a_1),\sigma(a_2)} \cdots Q_{\sigma(a_{N-1}),\sigma(a_N)} f_{\sigma(a_1)}(Y_1) \cdots f_{\sigma(a_N)}(Y_N)}{p^\theta_N(Y_{1:N})} \\
&\qquad\qquad \text{where } a_j = l \text{ as in the following,} \\
&\leqslant \frac{(1+\varepsilon_1/c) \sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) - \sum_{a_{1:j-1}, a_{j+1:N}} \mu^\theta_{\sigma(a_1)} \cdots f_{\sigma(a_N)}(Y_N)}{(1+\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})} \\
&\leqslant \frac{(1+\varepsilon_1/c) \sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) - \sum_{a_{1:j-1}, a_{j+1:N}} (\mu^*_{a_1} - \varepsilon_1) \cdots (f^*_{a_N}(Y_N) - \varepsilon_1)}{c + \varepsilon_1}.
\end{aligned}$$

Expanding the product in the second sum, the numerator becomes a sum where each term is bounded by $(\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$. Indeed the first term is equal to

$$\sum_{a_{1:j-1}, a_{j+1:N}} \mu^*_{a_1} \cdots f^*_{a_N}(Y_N) = p^{\theta^*}_N(Y_{1:N})$$

which gives $(\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$ when subtracted to the first sum. The other terms are a product of a positive power of ε_1 and μ^*_i, $Q^*_{i,j}$ or $f^*_{a_i}(Y_i)$ which are all bounded by 1. Thus they are bounded by $\varepsilon_1 \leqslant (\varepsilon_1/c) p^{\theta^*}_N(Y_{1:N})$. Moreover there are k^N terms so that

$$P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^\theta(X_j = \sigma(l) \mid Y_{1:N}) \leqslant \frac{\varepsilon_1 k^N}{c(c+\varepsilon_1)}.$$

Similarly

$$P^\theta(X_j = \sigma(l) \mid Y_{1:N}) - P^{\theta^*}(X_j = l \mid Y_{1:N}) \leqslant \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)}$$

so that

$$\left|P^{\theta^*}(X_j = l \mid Y_{1:N}) - P^\theta(X_j = \sigma(l) \mid Y_{1:N})\right| \leqslant \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)}. \tag{29}$$

Combining Equations (27), (28) and (29), we obtain

$$\begin{aligned}
&|P^{\theta^*}(X_j = l \mid Y_{1:n}) - P^\theta(X_j = \sigma(l) \mid Y_{1:n})| \\
&\leqslant 2\frac{2(1-\underline{q})^{N+1-j}}{\underline{q} + (1-\underline{q})^{N+1-j}} + \frac{\varepsilon_1 k^N}{c(c-\varepsilon_1)} < \varepsilon.
\end{aligned}$$

□

We prove Theorem 3.4 for $m = 1$, one may easily generalizes the proof. Let $\beta > 0$, $j > 0$ and $\varepsilon > 0$, we fix N and $c > 0$ such that

$$\frac{2(1-\underline{q})^{N+1-j}}{\underline{q}+(1-\underline{q})^{N+1-j}} < \frac{\varepsilon}{3} \text{ and } P^{\theta^*}\left(p_N^{\theta^*}(Y_{1:N}) > c\right) > \sqrt{1-\beta} \tag{30}$$

then we choose ε_1 such that

$$0 < \frac{\varepsilon_1 2^{2N} k^N}{c(c-\varepsilon_1)} < \frac{\varepsilon}{3}. \tag{31}$$

Posterior consistency for the marginal distribution in l_1 and for all components of the parameter i.e. Theorems 2.1 and 2.3 imply that there exists M such that P^{θ^*}-a.s., for all $n \geqslant M$,

$$\pi\left(\{\theta \ : \ D_N(\theta,\theta^*) < \varepsilon_1\} \mid Y_{1:n}\right) > \frac{\sqrt{1-\beta}+1}{2} \tag{32}$$

and

$$\pi\Bigg(\{\theta \ : \ \exists\sigma\in\mathcal{S}_k, \ \max_{1\leqslant i\leqslant k}|\mu_{\sigma(i)}-\mu_i^*| < \varepsilon_1, \ \|\sigma Q - Q^*\| < \varepsilon_1, \\ \max_{1\leqslant i\leqslant k}\|f_{\sigma(i)}-f_i^*\|_{l_1}\} < \varepsilon_1 \ \Big| \ Y_{1:n}\Bigg) > \frac{\sqrt{1-\beta}+1}{2}. \tag{33}$$

Using Lemma A.1 and combining (30), (31), (32) and (33), we obtain for all $n \geqslant \max(N, M)$,

$$\begin{aligned}
&\mathbb{E}^{\theta^*}\left(\pi\left(\left\{\theta : \exists\sigma\in\mathcal{S}_k, \max_{1\leqslant l\leqslant k}\left|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j=\sigma(l)|Y_{1:n})\right| < \varepsilon\right\}|Y_{1:n}\right)\right)\\
&\geqslant \mathbb{E}^{\theta^*}\Bigg(\mathbb{1}_{p_N^{\theta^*}(Y_{1:N})>c} \\
&\qquad \pi\left(\left\{\theta\colon \exists\sigma, \max_{1\leqslant l\leqslant k}\left|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j=\sigma(l)|Y_{1:n})\right| < \varepsilon\right\}\Big|Y_{1:n}\right)\Bigg)\\
&\geqslant 1-\beta.
\end{aligned}$$

Then

$$\mathbb{E}^{\theta^*}\left(\pi\left(\left\{\theta : \exists\sigma\in\mathcal{S}_k, \max_{1\leqslant l\leqslant k}\left|P^{\theta^*}(X_j = l|Y_{1:n}) - P^{\theta}(X_j=\sigma(l)|Y_{1:n})\right| < \varepsilon\right\}|Y_{1:n}\right)\right)$$

tends to 1, which concludes the proof of Theorem 3.4.

Proof of Proposition 3.5

As under $DP(\alpha G_0)^{\otimes k}$, $f_i(l)$ is distributed from $\text{Beta}(\alpha G_0(l), \alpha\sum_{m\neq l} G_0(m))$,

$$\int_{\mathcal{F}^k}\sum_{l=1}^{+\infty} f_i^*(l)\max_{1\leqslant j\leqslant k}\left(-\log(f_j(l))\right)(DP(\alpha G_0))^{\otimes k}(df)$$

$$\leqslant \sum_{l=1}^{+\infty} f_i^*(l) \sum_{1\leqslant j\leqslant k} \int_{\mathcal{F}^k} \left(-\log(f_j(l))\right) \left(DP(\alpha G_0)\right)^{\otimes k}(df)$$
$$\leqslant \sum_{l=1}^{+\infty} \frac{f_i^*(l)\ \Gamma(\alpha)}{\Gamma(\alpha G_0(l))\Gamma\Big(\alpha \sum\limits_{m\neq l} G_0(m)\Big)} \int_0^1 -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum\limits_{m\neq l} G_0(m)-1} \lambda(dx). \tag{34}$$

On $[1/2,1]$, $-\log(x)x^{\alpha G_0(l)-1} \leqslant 2\log(2)$, so that there exists a constant C_1 which does not depend on l such that

$$\int_{1/2}^1 -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum_{m\neq l} G_0(m)-1} \lambda(dx) \leqslant C_1. \tag{35}$$

On $[0,1/2]$, $(1-x)^{\alpha \sum_{m\neq l} G_0(m))-1} \leqslant 2$, so that there exists a constant C_2 which does not depend on l such that

$$\int_0^{1/2} -\log(x) x^{\alpha G_0(l)-1}(1-x)^{\alpha \sum_{m\neq l} G_0(m)-1} \lambda(dx) \leqslant \frac{C_2}{\left(\alpha G_0(l)\right)^2}. \tag{36}$$

Moreover for all $0<\delta<1$,

$$\frac{1}{\delta} \leqslant \Gamma(\delta) = \frac{\Gamma(\delta+1)}{\delta} \leqslant \frac{2}{\delta}. \tag{37}$$

By combining Equations (34), (35), (36) and (37), for all $1 \leqslant i \leqslant k$,

$$\int_{\mathcal{F}^k} \sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leqslant j\leqslant k} \left(-\log(f_j(l))\right) \left(DP(\alpha G_0)\right)^{\otimes k}(df)$$
$$\lesssim \sum_{l=1}^{+\infty} \frac{f_i^*(l)}{\alpha G_0(l)}$$

so that using Assumption (E1),

$$\left(DP(\alpha G_0)\right)^{\otimes k}\Bigg(f_1,\ldots,f_k \ :\ \forall 1\leqslant i\leqslant k, \sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leqslant j\leqslant k}\left(-\log(f_j(l))\right) < +\infty\Bigg) = 1.$$

Note that for all $\varepsilon>0$,

$$\left\{f_1,\ldots,f_k \ :\ \forall 1\leqslant i\leqslant k,\ \sum_{l=1}^{+\infty} f_i^*(l) \max_{1\leqslant j\leqslant k}\left(-\log(f_j(l))\right) < +\infty\right\}$$
$$\subset \bigcup_{N\in\mathbb{N}} \left\{f_1,\ldots,f_k \ :\ \forall 1\leqslant i\leqslant k,\ \sum_{l=N}^{+\infty} f_i^*(l) \max_{1\leqslant j\leqslant k}\left(-\log(f_j(l))\right) < \varepsilon\right\},$$

thus arguing by contradiction, for all $\varepsilon > 0$, there exists L_ε such that

$$\big(DP(\alpha G_0)\big)^{\otimes k}\Bigg(f_1,\dots,f_k \ : \ \forall 1 \leqslant i \leqslant k, \\ \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) < \varepsilon\Bigg) > 0.$$

Using the tail free property of the Dirichlet process, for all $1 \leqslant j \leqslant k$,

$$\sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) < \varepsilon$$

and

$$\left(\frac{f_j(1)}{\sum_{l\leqslant L_\varepsilon} f_j(l)},\dots,\frac{f_j(L_\varepsilon)}{\sum_{l\leqslant L_\varepsilon} f_j(l)}\right) \tag{38}$$

are independent given $\sum_{l>L_\varepsilon} f_j(l)$ and (38) given $\sum_{l>L_\varepsilon} f_j(l)$ has a Dirichlet distribution with parameter $(\alpha G_0(1),\dots,\alpha G_0(L_\varepsilon))$. Then for all $\varepsilon > 0$, there exists L_ε such that for all $\delta \in (0,1)$,

$$\big(DP(\alpha G_0)\big)^{\otimes k}\Bigg(f_1,\dots,f_k \ : \ \forall 1 \leqslant i \leqslant k, \ \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) < \frac{\varepsilon}{2}, \\ \forall l \leqslant L_\varepsilon, \ |f_j(l) - f_j^*(l)| \leqslant c\delta\Bigg) > 0 \tag{39}$$

where $c = \min_{1\leqslant i\leqslant k} \min_{l\leqslant L_\varepsilon, f_i^*(l)>0} f_i^*(l)$.

For all $f_1,\dots,f_k$ such that for all $1 \leqslant i,j \leqslant k$,

$$\sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) < \frac{\varepsilon}{2}$$

and for all $l \leqslant L_\varepsilon$, $|f_j(l) - f_j^*(l)| \leqslant c\delta$,

$$\begin{aligned}
&\sum_{l\in\mathbb{N}} f_i^*(l) \max_{1\leqslant j\leqslant k} \log\left(\frac{f_j^*(l)}{f_j(l)}\right) \\
&\quad = \sum_{l\leqslant L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k} \log\left(\frac{f_j^*(l)}{f_j(l)}\right) + \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k} \log(f_j^*(l)) \\
&\qquad + \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) \\
&\quad \leqslant \frac{\delta}{1-\delta} + 0 + \frac{\varepsilon}{2} \leqslant \varepsilon
\end{aligned} \tag{40}$$

for δ small enough. For such a δ denote

$$\Theta_\varepsilon = \{Q \ : \ \|Q - Q^*\| \leqslant \varepsilon\} \times \{f_1,\dots,f_k \ : \ \sum_{l>L_\varepsilon} f_i^*(l) \max_{1\leqslant j\leqslant k}(-\log(f_j(l))) < \frac{\varepsilon}{2}, \\ \forall l \leqslant L_\varepsilon, \ |f_j(l) - f_j^*(l)| \leqslant c\delta, \ \forall 1 \leqslant i,j \leqslant k\}$$

Using Equation (40), (A1b) holds. Furthermore (A1d) is obviously checked. Under Assumption (E1), $G_0(l) > 0$ when $\sum_{i=1}^k f_i^*(l) > 0$ so that (A1c) holds. Using the assumption that Q^* is in the support of π_Q, (A1a) is checked. Then using Equation (39), (A1) holds and the first part of Proposition 3.5 follows.

We now prove the second part of Proposition 3.5. We first give a representation of a discrete Dirichlet process with independent Gamma distributed random variables.

Lemma A.2 (Ferguson [11]). *Let $(Z_l)_{l\in\mathbb{N}}$ be independent random variables such that for all $l \in \mathbb{N}$,*

$$Z_l \sim \Gamma(\alpha G_0(l), 1),$$

then $\sum_{l=1}^L Z_l$ converges almost surely and its limit has a gamma distribution $\Gamma(\alpha, 1)$.

Moreover denote

$$f : \begin{cases} \mathbb{N} & \to [0,1] \\ i & \to f(i) = Z_i / (\sum_{l=1}^{+\infty} Z_l) \end{cases},$$

then f is distributed from a Dirichlet process $DP(\alpha G_0)$.

We assume (A1b) i.e. for all $\varepsilon > 0$,

$$DP(\alpha G_0)^{\otimes k}\left(\left\{f \in \mathcal{F}^k, \forall i \in \{1,\dots,k\} \ \sum_{l\in\mathbb{N}} f_i^*(l) \max_{1\leqslant j\leqslant k} \log \frac{f_j^*(l)}{f_j(l)} < \varepsilon\right\}\right) > 0.$$

Let $\varepsilon > 0$, define $\mathcal{F}_\varepsilon$ as the set of $f = (f_1,\dots,f_k) \in \mathcal{F}^k$ such that for all $1 \leqslant i \leqslant k$, for all $f \in \mathcal{F}_\varepsilon$,

$$\sum_{l\in\mathbb{N}} f_i^*(l) \log\left(\frac{f_i^*(l)}{f_i(l)}\right) < \varepsilon.$$

Then $DP(\alpha G_0)^{\otimes k}(\mathcal{F}_\varepsilon) > 0$.

Since $\sum_l f_i^*(l)(-\log f_i^*(l))$ converges, then $\sum_l f_i^*(l)(-\log f_i(l))$ converges. Using Lemma A.2, we can write f_i with independent gamma distributed random variables $(Z_l)_{l\in\mathbb{N}}$:

$$f_i(l) = \frac{Z_l}{\sum_{j\in\mathbb{N}} Z_j},$$

where $Z_l \sim \Gamma(\alpha G_0(l), 1)$. Then $\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$ converges since $\sum_{j\in\mathbb{N}} Z_j$ is finite almost surely. Since $DP(\alpha G_0)^{\otimes k}(\mathcal{F}_\varepsilon) > 0$, for all $1 \leqslant i \leqslant k$ with positive probability,

$$\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$$

converges. Using the Kolmogorov 0-1 law and the Three-Series Theorem (see Section 9.7.3 in Dudley [9]), $\sum_{l\in\mathbb{N}} f_i^*(l)(-\log(Z_l))$ converges almost surely and

$$\sum_{l\in\mathbb{N}} \mathbb{P}(|f_i^*(l)(-\log(Z_l))| > 1) < +\infty, \tag{41}$$

$$\sum_{l\in\mathbb{N}} \mathbb{E}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\leqslant 1}\big) < +\infty, \tag{42}$$

$$\sum_{l\in\mathbb{N}} \operatorname{var}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\leqslant 1}\big) < +\infty. \tag{43}$$

Equation (41) implies that

$$\begin{aligned}
+\infty &> \sum_{l\in\mathbb{N}} \mathbb{P}(|f_i^*(l)(-\log(Z_l))| > 1) \\
&\geqslant \sum_{l\in\mathbb{N}} \frac{1}{\Gamma(\alpha G_0(l))} \int_0^{\exp(-1/f_i^*(l))} x^{\alpha G_0(l)-1} e^{-x} dx \\
&\geqslant \sum_{l\in\mathbb{N}} \frac{1}{\alpha G_0(l)\Gamma(\alpha G_0(l))} \exp\left(-\exp\left(\frac{-1}{f_i^*(l)}\right) - \frac{\alpha G_0(l)}{f_i^*(l)}\right) \\
&\gtrsim \sum_{l\in\mathbb{N}} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right)
\end{aligned}$$

using Equation (37). Then

$$\lim_{l\to\infty} \frac{f_i^*(l)}{G_0(l)} = 0. \tag{44}$$

Moreover Equation (42) implies that

$$\begin{aligned}
+\infty &> \sum_{l} \mathbb{E}\big(f_i^*(l)(-\log(Z_l))\mathbb{1}_{|f_i^*(l)(-\log(Z_l))|\leqslant 1}\big) \\
&\geqslant \sum_{l} \left(\int_{\exp(-1/f_i^*(l))}^{1} \frac{1}{\Gamma(\alpha G_0(l))} f_i^*(l)(-\log(x)) x^{\alpha G_0(l)-1} e^{-x} dx \right. \\
&\qquad \left. + \int_1^{\exp(1/f_i^*(l))} \frac{1}{\Gamma(\alpha G_0(l))} f_i^*(l)(-\log(x)) x^{\alpha G_0(l)-1} e^{-x} dx \right) \\
&\geqslant \sum_{l} \left(\frac{e^{-1} f_i^*(l)}{\Gamma(\alpha G_0(l))} \int_{\exp(-1/f_i^*(l))}^{1} (-\log(x)) x^{\alpha G_0(l)-1} dx \right. \\
&\qquad \left. - \frac{1}{\Gamma(\alpha G_0(l))} \int_1^{\exp(1/f_i^*(l))} e^{-x} dx \right) \\
&\gtrsim -\alpha + \sum_{l} \frac{e^{-1} f_i^*(l)}{\alpha^2 G_0^2(l)\Gamma(\alpha G_0(l))} \\
&\qquad \left(1 - \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right) - \frac{\alpha G_0(l)}{f_i^*(l)} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right)\right) \\
&\gtrsim -\alpha + \sum_{l} \frac{f_i^*(l)}{G_0(l)}
\end{aligned}$$

using Equation (37) and that

$$\lim_{l\to\infty} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right) + \frac{\alpha G_0(l)}{f_i^*(l)} \exp\left(-\frac{\alpha G_0(l)}{f_i^*(l)}\right) = 0$$

using Equation (44). Then

$$\sum_{l\in\mathbb{N}} \frac{f_i^*(l)}{G_0(l)} < +\infty.$$

Appendix B: Other proofs

Proof of Proposition 3.1

The proof uses many ideas of Tokdar [19].

We now prove that Assumptions (B1), (B2), (B3) and (B4) imply (A1). A reproduction of the proof of Theorem 3.2. and Lemma 3.1 of Tokdar [19] shows that Assumptions (B2), (B3) and (B4) imply that for all $\varepsilon > 0$, for all $1 \leqslant j \leqslant k$ there exists a weak neighborhood V_j of a compactly supported probability measure $\tilde{P}_j$ such that for all $f_j = \phi * P_j$, $P_j \in V_j$,

$$\int_{\mathbb{R}} f_i^*(y) \max_{1\leqslant j\leqslant k} \log\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy) < \varepsilon. \tag{45}$$

Let $0 < \underline{\sigma} < \bar{\sigma}$ and $\zeta > 0$ be such that for all $1 \leqslant j \leqslant k$

$$\tilde{P}_j([-\zeta,\zeta]\times[\underline{\sigma},\bar{\sigma}]) = 1.$$

Let $\delta = \underline{\sigma}/2$. For all $1 \leqslant j \leqslant k$ define

$$U_j = \left\{ P \ : \ \left| \int_{\mathbb{R}\times(0,+\infty)} h dP - \int_{\mathbb{R}\times(0,+\infty)} h d\tilde{P}_j \right| < \varepsilon \right\},$$

where $h : \mathbb{R}\times(0,+\infty) \to [0,1]$ is a piecewise affine continuous function such that $h(z,\sigma) = 1$ for all $z \in [-\zeta,\zeta]$ and $\sigma \in [\underline{\sigma},\bar{\sigma}]$ and $h(z,\sigma) = 0$ for all $z \in [-\zeta-\delta,\zeta+\delta]^c$ and $\sigma \in [\underline{\sigma}-\delta,\bar{\sigma}+\delta]^c$. For all $\varepsilon > 0$, define

$$\Theta_\varepsilon = \{Q \ : \ \|Q - Q^*\| < \varepsilon\} \times (V_1 \cap U_1) \times \cdots \times (V_k \cap U_k).$$

Then for all $(Q, \phi*P_1, \ldots, \phi*P_k) \in \Theta_\varepsilon$, (A1b) is true according to Equation (45). In addition, for all $y \in \mathbb{R}$,

$$\begin{aligned} f_j(y) &\geqslant \int_{[-\zeta-\delta,\zeta+\delta]\times[\underline{\sigma}-\delta,\bar{\sigma}+\delta]} \phi_\sigma(y-z) P_j(dz, d\sigma) \\ &\geqslant \frac{1}{\bar{\sigma}+\delta}\, \phi_{\underline{\sigma}-\delta}\big(\max(|y-\zeta-\delta|, |y+\zeta+\delta|)\big)\ (1-\varepsilon) \end{aligned}$$

which implies (A1c). Moreover using assumption (B1), Π_P-a.s. there exists $C > 0$ such that for all $1 \leqslant j \leqslant k$,

$$f_j(y) \leqslant \int \frac{1}{\sigma} P_j(dz, d\sigma) \leqslant C$$

so that (A1d) holds. As Θ_ε is a product of neighborhoods of elements in the support of their respective prior, $\pi(\Theta_\varepsilon) > 0$, so (A1) is checked.

Now we prove that Assumption (B5) implies Assumption (A2). Let $\delta > 0$. For all $a, l, u, \kappa > 0$, such that $l < u$ denote $\mathcal{F}^{\kappa}_{a,l,u} = \{\phi * P : P((-a, a] \times (l, u]) > 1 - \kappa\}$. Using Section 4 of Tokdar [19], there exist b_0, b_1, b_2 only depending on κ such that

$$\begin{aligned} \log(N(3\kappa, (\mathcal{F}^{\kappa}_{a,l,u})^k, d)) &\leqslant k \log(N(3\kappa, \mathcal{F}^{\kappa}_{a,l,u}, \|\cdot\|_{L_1(\lambda)})) \\ &\leqslant k b_0 \left(b_1 \frac{a}{l} + b_2 \log\left(\frac{u}{l}\right) + 1 \right). \end{aligned} \tag{46}$$

Choosing $\kappa = \frac{\delta}{3*36l}$ and $\beta < \frac{\delta^2 k \underline{q}^2}{32 l b_0 (b_1 + b_2)}$, Assumption (B5) implies that Assumption (A2) holds.

Proof of Corollary 3.2

To prove the first part of Corollary 3.2, we use Theorem 2.3 because $m_1^* < \cdots < m_k^*$ implies the linear independence of $g^*(\cdot - m_1^*)\lambda, \ldots, g^*(\cdot - m_k^*)\lambda$. Then it is sufficient to prove that for all $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\begin{aligned} &\left\{ \theta : \ \exists \sigma \in \mathcal{S}_k, \ \max_{1 \leqslant i \leqslant k} d_w\big(g(\cdot - m_{\sigma(i)}), g^*(\cdot - m_i^*)\big) < \eta, \ \|\sigma Q - Q^*\| < \eta \right\} \\ &\subset \left\{ \theta : \ d_w(g, g^*) < \varepsilon, \ \max_{1 \leqslant j \leqslant k} |m_j - m_j^*| < \varepsilon, \ \|Q - Q^*\| < \varepsilon \right\}, \end{aligned} \tag{47}$$

where d_w metricizes the weak topology on $\mathcal{F}$. Let ξ^n be a sequence of $\Theta(\underline{q})$ such that for all n there exists $\sigma_n \in \mathcal{S}_k$ such that for all $1 \leqslant i \leqslant k$,

$$d_w\big(g^n(\cdot - m^n_{\sigma_n(i)}), g^*(\cdot - m_i^*)\big) \to 0 \text{ and } \|\sigma_n Q^n - Q^*\| \to 0.$$

As there exists a finite number of permutation in $\mathcal{S}_k$, there exists a subsequence, that we denote again ξ^n, of ξ^n such that there exists a permutation σ not depending on n such that for all n and for all $1 \leqslant i \leqslant k$,

$$d_w\big(g^n(\cdot - m^n_{\sigma(i)}), g^*(\cdot - m_i^*)\big) \to 0 \text{ and } \|\sigma Q^n - Q^*\| \to 0.$$

Particularly $g^n(\cdot)\lambda$ weakly tends to $g^*(\cdot - m^*_{\sigma^{-1}(1)})\lambda$. As weak convergence implies pointwise convergence of the characteristic functions and for all $t \in \mathbb{R}$,

$$\int e^{ity} g^n(y - m^n_{\sigma(j)})\lambda(dy) = e^{itm^n_{\sigma(j)}} \int e^{ity} g^n(y)\lambda(dy)$$

then

$$\lim_{n \to \infty} e^{itm^n_{\sigma(j)}} = e^{it(m_j^* - m^*_{\sigma^{-1}(1)})}$$

for all t such that $\int e^{ity} g^*(y)\lambda(dy) \neq 0$. As any characteristic function is uniformly continuous and equal to 1 at 0, there exists $\alpha > 0$ such that $\int e^{ity} g^*(y - m^*_{\sigma^{-1}(1)})d\lambda(y) \neq 0$ for all $|t| < \alpha$. Thus for all $1 \leqslant j \leqslant k$,

$$\lim_{n \to \infty} m^n_{\sigma(j)} = m_j^* - m^*_{\sigma^{-1}(1)}.$$

Since

$$0 = m_1^* < m_2^* < \cdots < m_k^* \text{ and } 0 = m_1^n < m_2^n < \cdots < m_k^n$$

then the permutation σ is equal to the identity permutation. Then Equation (47) holds and this implies the first part of Corollary 3.2. In fact we have proved the continuity of

$$\left\{ \begin{array}{lll} (\{p_l^\xi,\ \xi \in \Xi(\underline{q}), \operatorname{rank}(Q) = k\}, L_1) & \to & (\Delta^k(0), \|\|) \times (\mathbb{R}, |)^k \times (\mathcal{F}, d_w) \\ p_l^\xi & \mapsto & \xi \end{array} \right. . \tag{48}$$

If moreover $\max_{1 \leqslant j \leqslant k} \mu_j^* > \frac{1}{2}$ and g^* is uniformly continuous, if

$$\lim_{n \to \infty} D_3(\xi^n, \xi^*) = 0$$

then

$$\lim_{n \to \infty} D_1(\xi^n, \xi^*) = 0$$

and by continuity of the functional defined in (48),

$$\lim_{n \to \infty} \max_{1 \leqslant j \leqslant k} |\mu_j^n - \mu_j^*| = 0$$

and

$$\lim_{n \to \infty} \max_{1 \leqslant j \leqslant k} |m_j^n - m_j^*| = 0$$

so that

$$\lim_{n \to \infty} \max_{1 \leqslant j \leqslant k} \|g^*(\cdot - m_j^n) - g^*(\cdot - m_j^*)\|_{L_1(\lambda)} = 0$$

since g^* is uniformly continuous. Using the following inequality proved in the proof of Corollary 1 in Gassiat and Rousseau [13]

$$\begin{aligned} \|D_1(\xi^n, \xi^*)\|_{L_1} \geqslant & \left(2 \max_{1 \leqslant j \leqslant k} \mu_j^* - 1 \right) \|g^n - g^*\|_{L_1(\lambda)} \\ & - \max_{1 \leqslant j \leqslant k} |\mu_j^n - \mu_i^*| - \max_{1 \leqslant j \leqslant k} \|g^*(\cdot - m_j^n) - g^*(\cdot - m_j^*)\|_{L_1(\lambda)} \end{aligned}$$

we obtain that $\lim_{n \to \infty} \|g^n - g^*\|_{L_1(\lambda)} = 0$ which implies the last part of Corollary 3.2.

Proof of Proposition 3.3

As in the proof of Proposition 3.1, many ideas come from Tokdar [19]. We first prove (A1) assuming that (B1), (B2), (B3) and (B4) are verified with $f_j(\cdot) = g(\cdot - m_j),\ 1 \leqslant j \leqslant k$. With the same ideas of the proof of Theorem 3.2 in Tokdar [19], for all $\varepsilon > 0$ there exists a probability measure $\tilde{P}$ on $\mathbb{R} \times (0, +\infty)$ such that there exists $0 < \underline{\sigma} < \bar{\sigma}$ and $a > 0$ satisfying

$$\tilde{P}((-a, a] \times (\underline{\sigma}, \bar{\sigma}]) = 1$$

and

$$\int g^*(y-m_i^*)\max_{1\leqslant j\leqslant k}\log\frac{g^*(y-m_j^*)}{\phi*\tilde{P}(y-m_j^*)}\lambda(dy)\leqslant\frac{\varepsilon}{3},$$

using Assumptions (B2), (B3) and (B4).

Let $G=[-a,a]\times[\underline{\sigma},\bar{\sigma}]$. Using the proof of Lemma 3.1 in Tokdar [19] for all $C>\max_{1\leqslant j\leqslant k}|m_j^*|+a+\bar{\sigma}$, for all $m_j\in[m_j^*-a,m_j^*+a]$, and for all P such that $P(G)>\frac{\underline{\sigma}}{\bar{\sigma}}$,

$$\begin{aligned}\int_{|y|>C}g^*(y-m_i^*)\max_{1\leqslant j\leqslant k}\log\frac{\phi*\tilde{P}(y-m_j^*)}{\phi*P(y-m_j)}\lambda(dy)\\ \leqslant\int_{|y|>C}g^*(y-m_i)\max_{1\leqslant j\leqslant k}\frac{1}{2}\left(\frac{|y|+|m_j^*|+2a}{\underline{\sigma}}\right)^2\lambda(dy)<\infty.\end{aligned}\tag{49}$$

Using assumption (B4) and Equation (49), we fix C such that

$$\int_{|y|>C}g^*(y-m_i^*)\max_{1\leqslant j\leqslant k}\log\frac{\phi*\tilde{P}(y-m_j^*)}{\phi*P(y-m_j)}\lambda(dy)\leqslant\frac{\varepsilon}{3}.$$

Let $G_\delta=[-a-\delta,a+\delta]\times[\underline{\sigma}-\delta,\bar{\sigma}+\delta]$, with δ chosen in $(0,\min(\frac{\underline{\sigma}}{2},\frac{a}{2})]$. Let $h:\mathbb{R}\times(0,+\infty)\to[0,1]$ be a piecewise affine continuous function such that $h(z,\sigma)=1$ on G and $h(z,\sigma)=0$ on G_δ^c. Let

$$c=\inf_{\substack{\underline{\sigma}-\delta\leqslant\sigma\leqslant\bar{\sigma}+\delta,\\ |y|\leqslant C,\\ |\theta|\leqslant a+\max_j|m_j^*|+\delta}}\phi_\sigma\left(y-\theta\right).$$

By Arzelà-Ascoli theorem there exists $y_1,\ldots,y_I$ such that for all $y\in[-C,C]$ and $1\leqslant j\leqslant k$, there exists $1\leqslant i\leqslant I$ such that

$$\sup_{(z,\sigma)\in G_\delta}\left|\phi_\sigma\left(y-m_j^*-z\right)-\phi_\sigma\left(y_i-m_j^*-z\right)\right|<c\delta.$$

Let

$$\begin{aligned}V_\delta=\Big\{P\ :\ \Big|\int h(z,\sigma)\phi_\sigma(y_i-m_j^*-z)dP(z,\sigma)-\\ \int h(z,\sigma)\phi_\sigma(y_i-m_j^*-z)d\tilde{P}(z,\sigma)\Big|<c\delta\Big\}.\end{aligned}$$

For all $P\in V_\delta$, for all $m_j\in[m_j^*-\frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}},m_j^*+\frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}}]$ and for all $1\leqslant j\leqslant k$, we get

$$\left|\frac{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)dP(z,\sigma)}{\int h(z,\sigma)\phi_\sigma(y-m_j-z)d\tilde{P}(z,\sigma)}-1\right|\leqslant4\delta$$

thus

$$\begin{aligned}
&\int_{|y|\leqslant C} g^*(y-m_i^*) \max_{1\leqslant j\leqslant k} \log \frac{\phi * \tilde{P}(y-m_j^*)}{\phi * P(y-m_j^*)} \lambda(dy) \\
&\quad \leqslant \int_{|y|\leqslant C} g^*(y-m_i^*) \max_{1\leqslant j\leqslant k} \log \frac{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)d\tilde{P}(z,\sigma)}{\int h(z,\sigma)\phi_\sigma(y-m_j^*-z)dP(z,\sigma)} \lambda(dy) \\
&\quad \leqslant \frac{4\delta}{1-4\delta}.
\end{aligned}$$

Then for δ small enough, for all $g = \phi * P$ such that $P \in V_\delta \cap \{P : P(G) > \frac{\underline{\sigma}}{\bar{\sigma}}\} = \tilde{V}_\delta$, for all $m_j \in [m_j^* - \frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}}, m_j^* + \frac{c\underline{\sigma}\delta\sqrt{2}}{\sqrt{\pi}}] = M_j^\delta$ and for all $1 \leqslant i \leqslant k$,

$$\max_{1\leqslant i\leqslant k} \int g^*(y-m_i^*) \max_{1\leqslant j\leqslant k} \log\left(\frac{g^*(y-m_j^*)}{g(y-m_j)}\right) dy < \varepsilon, \tag{50}$$

moreover,

$$\begin{aligned}
g(y-m_i) &\geqslant \int_G \phi_\sigma(y-m_i-z)P(dz,d\sigma) \\
&\geqslant \frac{\underline{\sigma}}{\bar{\sigma}} \phi_{\underline{\sigma}}(\max(|y-m_i-a|, |y-m_i+a|))P(G) \\
&\geqslant \frac{\underline{\sigma}}{\bar{\sigma}} \phi_{\underline{\sigma}}(\max(|y-m_i-a|, |y-m_i+a|)) \frac{\underline{\sigma}}{\bar{\sigma}} > 0.
\end{aligned} \tag{51}$$

Assumption (B1) ensures that (A1d) holds. Finally for all $\varepsilon > 0$, there exists $\delta > 0$ such that (A1) holds with $\Theta_\varepsilon = \{Q : \|Q - Q^*\| < \min(\varepsilon, \underline{a}/2)\} \times M_1^\delta \times \cdots \times M_k^\delta \times \tilde{V}_\delta$ using Equations (50) and (51).

We now prove (C2) thanks to Assumption (D6). Let

$$\mathcal{F}_{a,l,u,\underline{m}} = [-\underline{m}, \underline{m}]^k \times \mathcal{F}_{a,l,u},$$

where $\mathcal{F}_{a,l,u} = \mathcal{F}_{a,l,u}^2$ is defined in the proof of Proposition 3.1. Note that for all $(m, \phi * P), (\tilde{m}, \phi * \tilde{P}) \in \mathcal{F}_{a,l,u,\underline{m}}$, for all $1 \leqslant i \leqslant k$,

$$\begin{aligned}
&\|\phi * P(\cdot - m_i) - \phi * \tilde{P}(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} \\
&\leqslant \|\phi * P(\cdot - m_i) - \phi * P(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} + \|\phi * P(\cdot) - \phi * \tilde{P}(\cdot)\|_{L_1(\lambda)}.
\end{aligned}$$

The second term is dealt with in the proof of Proposition 3.1. As to the first part,

$$\|\phi * P(\cdot - m_i) - \phi * P(\cdot - \tilde{m}_i)\|_{L_1(\lambda)} \leqslant \frac{1}{l}\sqrt{\frac{2}{\pi}}|m_i - \tilde{m}_i|$$

then for all $\kappa > 0$, $a, l, u, \underline{m} > 0$ such that $l < u$,

$$N(3\kappa, \mathcal{F}_{a,l,u,\underline{m}}, d) \leqslant \left(\frac{2\underline{m}}{l\kappa} + 1\right)^k N(2\kappa, \mathcal{F}_{a,l,u}, \|\cdot\|_{L_1(\lambda)}).$$

For all $\kappa > 0$, let

$$\mathcal{F}^{\kappa}_{a,l,u,\underline{m}} = [-\underline{m}, \underline{m}]^k \times \mathcal{F}^{\kappa}_{a,l,u}.$$

Following the ideas of Lemmas 4.1 and 4.2 in Tokdar [19], there exist c_0, c_1, c_2, c_3 only depending on κ such that

$$\log\left(N(\kappa, \mathcal{F}^{\kappa}_{a,l,u,\underline{m}}), d\right) \leqslant c_0\left(c_1 k \log\frac{m}{l} + c_2\frac{a}{l} + c_3 \log\frac{u}{l} + 1\right),$$

so that (D6) implies (C2) with suitable choices of κ and β.

References

[1] Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report.

[2] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics* **37** 1554–1563. MR0202264

[3] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer. MR2159833

[4] Couvreur, L. and Couvreur, C. (2000). Wavelet-based non-parametric HMM's: Theory and applications. In *ICASSP'00* **1** 604–607.

[5] de Gunst, M. C. and Shcherbakova, O. (2008). Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics* **17** 342–356. MR2483462

[6] Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7** 381–420. MR1836737

[7] Douc, R., Moulines, E. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics* **32** 2254–2304. MR2102510

[8] Douc, R., Moulines, E., Olsson, J. and van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics* **39** 474–513. MR2797854

[9] Dudley, R. M. (2002). *Real Analysis and Probability* **74**. Cambridge University Press. MR1932358

[10] Dumont, T. and Le Corff, S. (2014). Nonparametric regression on hidden phi-mixing variables: Identifiability and consistency of a pseudo-likelihood based estimation procedure. *arXiv:1209.0633*.

[11] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2** 615–629. MR0438568

[12] Gassiat, E., Cleynen, A. and Robin, S. (2015). Inference in finite state space non parametric Hidden Markov Models and applications. *Statistics and Computing* 1–11.

[13] Gassiat, E. and Rousseau, J. (2013). Non parametric finite translation hidden Markov models and extensions. *Bernoulli*, to appear.

[14] Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075. MR3263098

[15] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer. MR1992245

[16] MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall/CRC, London, UK. MR1692202

[17] MacDonald, I. L. and Zucchini, W. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall/CRC, London, UK. MR2523850

[18] Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* **330** 905–908. MR1771956

[19] Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā* **68** 90–110. MR2301566

[20] Whiting, J. P., Lambert, M. F. and Metcalfe, A. V. (2003). Modelling persistence in annual Australian point rainfall. *Hydrology and Earth System Sciences* **7** 197–211.

[21] Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society* **73** 37–57. MR2797735