On the asymptotic behaviour of the posterior distribution in hidden Markov Models

Elisabeth Gassiat, Judith Rousseau

On the asymptotic behaviour of the posterior distribution in hidden Markov Models

Elisabeth Gassiat

Laboratoire de Mathématique, Université Paris-Sud et CNRS, Orsay, France

Judith Rousseau

ENSAE-CREST and CEREMADE, Université Paris-Dauphine, Paris, France

Abstract. In this paper, we investigate the asymptotic behaviour of the posterior distribution in hidden Markov models (HMMs). We obtain a general asymptotic result, and give conditions on the prior under which we obtain a rate of convergence for the posterior distribution of the marginal distributions of the process. We then focus on the situation where the hidden Markov chain evolves on a finite state space but where the number of hidden states might be larger than the true one. It is known that the likelihood ratio test statistic for overfitted HMMs has a non standard behavior and is unbounded. Our conditions on the prior may be seen as a way to penalize parameters to avoid this phenomenon. We are then able to define a consistent Bayesian estimator of the number of hidden states. We also give a precise description of the situation when the observations are i.i.d. and we allow 2 possible hidden states. Some simulations are presented to illustrate our results.

1. Introduction

Hidden Markov models are stochastic processes $(X_j, Y_j)_{j\geq 0}$ where $(X_j)_{j\geq 0}$ is a Markov chain living in a state space $\mathcal{X}$ and conditionnally on $(X_j)_{j\geq 0}$ the Y_j 's are independent with a distribution depending only on X_j and living in $\mathcal{Y}$. The observations are $Y_{1:n} = (Y_1, \cdots, Y_n)$ and the associated states $X_{1:n} = (X_1, \cdots, X_n)$ are unobserved. Hidden Markov models are useful tools to model time series where the observed phenomenon is driven by a latent Markov chain. They may be seen as a dynamic extension of mixture models. They have been used successfully in a variety of applications such as economics (e.g. Albert and Chib (1993)), genomics (e.g. Churchill (1989)), signal processing and image analysis (e.g. Andrieu and Doucet (2000)), ecology (e.g. Guttorp (1995)), speech recognition (e.g. Rabiner (1989)) to name but a few. The books by MacDonald and Zucchini (1997) and Cappé et al. (2004) provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. When the state space $\mathcal{X}$ of the hidden Markov chain is finite, the number of hidden states induces a classification of the regimes in which the time series evolves. They often have a practical interpretation in the modelization of the underlying phenomenon. It is thus of importance to be able to infer both the number of hidden states (which we call the order of the HMM) from data, when it is not known in advance and the associated parameters.

In the frequentist literature, penalized likelihood methods have been proposed to estimate the order of a HMM, using for instance Bayesian information criteria (BIC for short). These methods were applied for instance in Leroux and Putterman (1992), Rydén et al. (1998), but without theoretical consistency results. Later, it has been observed that the likelihood ratio statistics is unbouded, in the very simple situation where one wants to test between 1 or 2 hidden states, see Gassiat and Kéribin (2000). The question whether BIC penalized likelihood methods lead to consistent order estimation stayed open. Using tools borrowed from information theory, it has been possible to calibrate heavier penalties in maximum likelihood methods to obtain consistent estimators of the order, see Gassiat and Boucheron (2003), Chambaz et al. (2009). The use of penalized marginal pseudo likelihood was also proved to lead to weakly consistent estimators by Gassiat (2002).

On the Bayesian side, various methods were proposed to deal with an unknown number of hidden states, but no theoretical result exists to validate the methods. Reversible jump methods have been built, leading to satisfactory results on simulation and real data, see Boys and Henderson

(2004), Green and Richardson (2002), Robert et al. (2000), Spezia (2010). The ideas of variational Bayesian methods were developed in McGrory and Titterington (2009).

Recently, one of the authors proposed a theoretical analysis of the posterior distribution for overfitted mixtures, see Rousseau and Mengersen (2011). In this paper, it is proved that one may choose the prior in such a way that extra components are emptied, or in such a way that extra components merge with true ones. More precisely, if a Dirichlet prior $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ is considered on the k weights of the mixture components, small values of the α_j's imply that the posterior distribution will tend to empty the extra components of the mixture when the true distribution has a smaller number, say $k_0 < k$ of true components.
One aim of our paper is to understand if such an analysis may be extended to dynamic mixtures, that is to HMMs. Since HMMs are much more complicated models regarding order estimation, with unbounded likelihood ratio statistics that are still not well understood, our results do not cover all choices of prior distributions to empty extra components or to merge them with true ones. Only this last possibility is fully understood. Consider a finite state space HMM, with k states and with independent Dirichlet prior distributions $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ on each row of the transition matrix of the latent Markov chain. We prove that if the parameters α_j's are large enough, extra components merge to true ones. We are also able to propose a Bayesian consistent estimator of the number of hidden states, without using variable dimension algorithms such as reversible jump algorithms, which are often difficult to implement efficiently. On the other hand, if the parameters α_j's are small the posterior becomes harder to implement, see Section 4. Hence from a computational viewpoint, choosing smaller α_j's leads to greater difficulty and it does not seem to be counter-balanced by a more stable asymptotic behaviour of the posterior. We are thus able to give guidelines to choose the prior in such a way that the posterior leads to interpretable results by chosing large enough parameters in the Dirichlet prior.

In Section 2, we give a general theorem on the asymptotic behaviour of the posterior distribution. To our knowledge, this is the first general theoretical result for HMM Bayesian estimation. Though Ghosal and van der Vaart (2006) give rates of convergence for the posterior in possibly dependent observation models, they cannot be applied to the order estimation problem, as explained in Section 2.2. In Section 3 we consider the case of finite state space HMMs. Using the general result of Section 2, we explain how it is possible to choose the prior in such a way that the posterior gives consistent estimation of the marginal distributions. In this case we also obtain convergence rates. We are then able to derive a consistent Bayesian estimator of the number of hidden states, which does not require a prior on the number of states nor the computation of marginal likelihoods in the different candidate models. To our knowledge, this is the first consistency result on Bayesian order estimation in the case of hidden Markov models. In the specific situation where the overfitting is by only one state and the observations are i.i.d., we are able to describe more precisely what choice of the prior leads to the merging of the two states together with convergence rates. In Section 4 we present a simulation study to illustrate our results and to investigate what happens for other choices of priors.

2. Posterior concentration rates for HMMs : a general result

Since we could not find in the literature any result on the asymptotic concentration of the posterior distribution in HMM models we first present a general theorem where the posterior concentration is described in such models. We first describe the general setting and we give some notations that are used throughout the paper.

2.1. Setting and notations

Recall that HMMs model pairs (X_i, Y_i), $i = 1, \ldots, n$, where $(X_i)_i$ is the unobserved Markov chain living on a state space $\mathcal{X}$ and the observations $(Y_i)_{i=1}^n$ are conditionally independent given the $(X_i)_{i=1}^n$ and live in $\mathcal{Y}$. The spaces $\mathcal{X}, \mathcal{Y}$ can be general and we only assume that they are Polish spaces endowed with their Borel σ-fields. The hidden Markov chain $(X_i)_{i=1}^n$ has a Markov kernel $Q_\theta, \theta \in \Theta$ where Θ is a subset of an euclidian space and the conditional distribution of Y_i given X_i has density with respect to some given measure ν on $\mathcal{Y}$ denoted by $g_\theta(y|x)$, $x \in \mathcal{X}$, $\theta \in \Theta$. With an abuse of notations we also denote ν the product measure $\nu^{\otimes l}$ on $\mathcal{Y}^l$. We assume that the Markov

kernels Q_θ admit a (not necessarily unique) stationary distribution μ_θ, for each $\theta \in \Theta$. We write $\mathbb{P}_\theta$ for the probability distribution of the stationary HMM $(X_j, Y_j)_{j\geq 1}$ with parameter θ. That is, for any integer n, any measurable set A in the Borel σ-field of $\mathcal{X}^n \times \mathcal{Y}^n$:

$$\mathbb{P}_\theta\left((X_1,\dots,X_n,Y_1,\dots,Y_n)\in A\right) = \int_A \mu_\theta(dx_1)\prod_{i=1}^{n-1} Q_\theta\left(x_i,dx_{i+1}\right)\prod_{i=1}^{n} g_\theta\left(y_i|x_i\right)\nu(dy_1)\dots\nu(dy_n). \tag{1}$$

Thus for any integer n, under $\mathbb{P}_\theta$, $Y_{1:n} = (Y_1,\dots,Y_n)$ has a probability density with respect to $\nu(dy_1)\cdots\nu(dy_n)$ equal to

$$f_{n,\theta}(y_1,\dots,y_n) = \int_{\mathcal{X}^n} \mu_\theta(dx_1)\prod_{i=1}^{n-1} Q_\theta\left(x_i,dx_{i+1}\right)\prod_{i=1}^{n} g_\theta\left(y_i|x_i\right).$$

We denote π the prior distribution on Θ. As is often the case in Bayesian analysis of HMMs, instead of computing the stationary distribution μ_θ of the hidden Markov chain X for each θ, we consider a probability distribution $\pi_{\mathcal{X}}$ on the unobserved initial state X_0. Denote $\ell_n(\theta,x)$ the log-likelihood starting from x, for all $x \in \mathcal{X}$, which is given by

$$\ell_n\left(\theta,x\right) = \log\left[\int_{\mathcal{X}^{n+1}} Q_\theta\left(x,dx_1\right)\prod_{i=1}^{n-1} Q_\theta\left(x_i,dx_{i+1}\right)\prod_{i=1}^{n} g_\theta\left(Y_i|x_i\right)\right].$$

Similarly, the log-likelihood starting from a distribution π_0 on $\mathcal{X}$ is denoted $\ell_n\left(\theta,\pi_0\right)$ i.e.

$$\ell_n\left(\theta,\pi_0\right) = \log\left[\int_{\mathcal{X}} e^{\ell_n(\theta,x)} d\pi_0(x)\right].$$

The posterior distribution can then be written as

$$\mathbb{P}^\pi\left(A|Y_{1:n}\right) = \frac{\int_{A\times\mathcal{X}} e^{\ell_n(\theta,x)}\pi\left(d\theta\right)\pi_{\mathcal{X}}\left(dx\right)}{\int_{\Theta\times\mathcal{X}} e^{\ell_n(\theta,x)}\pi\left(d\theta\right)\pi_{\mathcal{X}}\left(dx\right)} \tag{2}$$

for any Borel set $A \subset \Theta$.
We shall also use the notation $\mathbb{P}_{\theta,x}$ for the probability distribution of the HMM starting from x, that is, for any integer n, any measurable set A in the Borel σ-field of $\mathcal{X}^n \times \mathcal{Y}^n$:

$$\mathbb{P}_{\theta,x}\left((X_1,\dots,X_n,Y_1,\dots,Y_n)\in A\right) = \int_A Q_\theta(x,dx_1)\prod_{i=1}^{n-1} Q_\theta\left(x_i,dx_{i+1}\right)\prod_{i=1}^{n} g_\theta\left(y_i|x_i\right)\nu(dy_1)\dots\nu(dy_n),$$

so that for any $\theta \in \Theta$,

$$\mathbb{P}_\theta = \int_{\mathcal{X}} \mathbb{P}_{\theta,x}\mu_\theta\left(dx\right).$$

We denote by E_θ the expectation under $\mathbb{P}_\theta$ and by $E_{\theta,x}$ the expectation under $\mathbb{P}_{\theta,x}$.

We assume throughout the paper that we are given a stationary HMM $(X_j, Y_j)_{j\geq 1}$ with distribution $\mathbb{P}_{\theta_0}$ for some $\theta_0 \in \Theta$. We will be interested in the asymptotic behaviour of the posterior distribution of finite marginals of the process. Indeed, marginals (of dimension at least 2) capture the transition of the Markov chain together with the emission parameters as we shall explain below. Thus we define for any integer $l \geq 2$, and for any $\theta \in \Theta$, the probability density $f_{l,\theta}(\cdots)$ of $(Y_1,\dots,Y_l)$ under $\mathbb{P}_\theta$. That is for any parameter θ,

$$f_{l,\theta}\left(y_1,\dots,y_l\right) = \int \mu_\theta\left(dx_1\right) Q_\theta\left(x_1,dx_2\right)\cdots Q_\theta\left(x_{l-1},dx_l\right) g_\theta\left(y_1|x_1\right)\cdots g_\theta\left(y_l|x_l\right)$$

so that $f_{l,\theta}$ is a mixture in $\mathcal{Y}^l$ of product probability measures. When such mixtures are identifiable, knowledge of $f_{l,\theta}$ leads to the knowledge of the mixing measure, which itself gives the knowledge of the distribution of the hidden Markov chain. Mixtures of products of gaussian distributions are

identifiable, for instance, but many other families of mixtures are identifiable, see MacLachlan and Peel (2000), Hall and Zhou (2003), Allman et al. (2009).

Define, for any $\theta \in \Theta$, real numbers $\rho_\theta \geq 1$ and $\mathbb{R}_\theta > 0$ such that, for any integer m, any $x \in \mathcal{X}$

$$\|Q_\theta^m(x,.) - \mu_\theta\|_{TV} \leq R_\theta \rho_\theta^{-m} \tag{3}$$

where $\|\cdot\|_{TV}$ is the total variation norm. If it is possible to set $\rho_\theta > 1$, the Markov chain $(X_n)_{n\geq 1}$ is uniformly ergodic and μ_θ is its unique stationary distribution.

Throughout the paper $\nabla_\theta h$ denotes the gradient vector of the function h when considered as a function of θ, and $D^2_\theta h$ its Hessian matrix. We denote by $B_d(\gamma, \epsilon)$ the d dimensional ball centered at γ with radius ϵ, when $\gamma \in \mathbb{R}^d$. The notation $a_n \lesssim b_n$ means that a_n is larger than b_n up to a positive constant that does not depend on n.

2.2. General HMMs

We now derive posterior concentration rates in the framework of Hidden Markov models. This setup follows the ideas of Ghosal and van der Vaart (2006). The proof of Theorem 1 is given in Section 6.

THEOREM 1. *Assume*

- *(A0)* $\rho_{\theta_0} > 1$.
- *(A1). There exists* $S_n \subset \Theta \times \mathcal{X}$, $D > 0$, $A > 0$ *and* x_0 *in* $\mathcal{X}$ *such that for any integer* n,

  $$\ell_n(\theta_0) - \ell_n(\theta_0, x_0) \leq A,$$

  and for any sequence $(C_n)_{n\geq 1}$ *of real numbers tending to* $+\infty$

  $$\sup_{(\theta,x)\in S_n} \mathbb{P}_{\theta_0}\left[\ell_n(\theta, x) - \ell_n(\theta_0, x_0) < -C_n\right] = o(1), \pi \otimes \pi_{\mathcal{X}}[S_n] \gtrsim n^{-D/2}.$$

- *(A2). There exists a sequence* $(\mathcal{F}_n)_{n\geq 1}$ *of subsets of* Θ *such that*

  $$\pi(\mathcal{F}_n^c) = o(n^{-D/2}).$$

- *(A3). There exists* $\delta_0 > 0$ *and* $M > 0$ *such that for all* $\delta_0 > \delta > 0$,

  $$N(\delta, \mathcal{F}_n, d_l(.,.)) \leq \left(\frac{n}{\delta}\right)^M$$

  where $N(\delta, \mathcal{F}_n, d_l(.,.))$ *is the smallest number of* $\theta_j \in \mathcal{F}_n$ *such that for all* $\theta \in \mathcal{F}_n$ *there exists a* θ_j *with* $d_l(\theta_j, \theta) \leq \delta$. *Here* $d_l(\theta, \theta_j) = \|f_{l,\theta} - f_{l,\theta_j}\|_1 := \int_{\mathcal{Y}^l} |f_{l,\theta} - f_{l,\theta_j}|(y) d\nu(y)$.

Then there exists K *large enough such that*

$$\mathbb{P}^\pi\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq K\sqrt{\frac{\log n}{n}} | Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1).$$

Theorem 1 gives the posterior concentration rate of $\|f_{l,\theta} - f_{l,\theta_0}\|_1$ up to the parameter $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1}$. This is in sharp contrast with the results in Ghosal and van der Vaart (2006), though the proof of our theorem follows the same lines. In Ghosal and van der Vaart (2006), applications to Markov chains or to gaussian time series of their general theorem use assumptions that lead in some sense to lower bound the coefficient $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1}$. This corresponds to choosing a prior whose support in Θ is included in a set where $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1}$ is uniformly bounded from below. Even in the *simple* case of finite state space HMMs, which are extensively used in practice, this type of priors would be awkward. We investigate this case in details in Section 3. If such a prior is considered, then Theorem 1 implies that the posterior distribution concentrates on $\{\theta; f_{l,\theta} = f_{l,\theta_0}\}$ and also provides

a concentration rate of the posterior distribution of order $O((\log n/n)^{1/2})$, in terms of the L_1 norm on $f_{l,\theta} - f_{l,\theta_0}$.

In the case of over-fitted HMMs with finite state space, i.e. when θ_0 corresponds to a HMM associated with k_0 states while the model considers HMMs associated with $k > k_0$ states the parameter set has to contain all possible transition matrices and in any neighbourhood $\{\theta : \|f_{l,\theta} - f_{l,\theta_0}\|_1 \leq \epsilon\}$ there exist parameters θ such that $\rho(\theta) = 1$. Thus, one has to allow ρ_θ to be arbitrarily close to 1. We will see that a good choice of the prior, however, acting as soft thresholding, leads to the concentration of the posterior distribution around f_{θ_0}, in terms of $\|f_{l,\theta} - f_{l,\theta_0}\|_1$ alone, at a rate slower than $(\log n/n)^{1/2}$.

Assumption $(A0)$ implies that at θ_0 the hidden Markov chain X is uniformly ergodic. Assumptions $(A2) - (A3)$ mean that we can choose a sequence of compact subsets of Θ which behave like finite dimensional sets. These two conditions are similar in spirit to those considered in general theorems on posterior consistency or posterior convergence rates, see for instance Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2006). Condition $(A1)$ is close to the Kullback-Leibler condition as in Ghosal and van der Vaart (2006), adapted to a parametric context. A non parametric formulation could also have been provided, replacing C_n with $n\epsilon_n^2$ in (A1), $n^{-D/2}$ by $e^{-n\epsilon_n^2}$ in $(A1)$ and $(A2)$ and M by $n\epsilon_n^2/\log n$ in $(A3)$.

In the following section, we explain how condition (A1) can be verified under conditions that are classical in the HMM literature.

2.3. About condition (A1)

Here we assume that $\mathcal{X}$ is compact, and that the transition kernels Q_θ are absolutely continuous with respect to a measure μ such that $\mu(\mathcal{X}) = 1$, for all θ in a neighborhood of θ_0. We denote $q_\theta(.,.)$ the density of Q_θ with respect to μ forθ in this neighborhood, and define

$$\sigma_-(\theta) = \inf_{x,x' \in \mathcal{X}} q_\theta(x, x'), \ \sigma_+(\theta) = \sup_{x,x' \in \mathcal{X}} q_\theta(x, x').$$

Then, by Corollary 1 of Douc et al. (2004), it is possible to set $R_\theta = 1$ and $\rho_\theta = \left(1 - \frac{\sigma_-(\theta)}{\sigma_+(\theta)}\right)^{-1}$. Also, following the proof of Lemma 2 of Douc et al. (2004) we find that, if $\rho_{\theta_0} > 1$, then

$$\ell_n(\theta_0) - \ell_n(\theta_0, x_0) \leq \left(\frac{\rho_{\theta_0}}{\rho_{\theta_0} - 1}\right)^2.$$

To verify assumption (A1), assume that there exists a subset $V \subset \Theta$ containing θ_0 such that the densities $q_\theta(x, x')$ and $g_\theta(y|x)$ are smooth as functions of θ on V (i.e. satisfy assumptions (A6)-(A8) of Douc et al. (2004) on V) and such that

$$\inf_{\theta \in V} \rho_\theta > 1. \tag{4}$$

Then, for any θ, x, x_0,

$$\ell_n(\theta, x) - \ell_n(\theta_0, x_0) = \ell_n(\theta, x) - \ell_n(\theta, x_0) + \ell_n(\theta, x_0) - \ell_n(\theta_0, x_0) \tag{5}$$

and following the proof of Lemma 2 of Douc et al. (2004) gives that, if (A0) and (4) hold, $\mathbb{P}_{\theta_0}$-a.s.,

$$\sup_{\theta \in V} \sup_{x, x_0 \in \mathcal{X}} |\ell_n(\theta, x) - \ell_n(\theta, x_0)| \leq 2 \sup_{\theta \in V} \left(\frac{\rho_\theta}{\rho_\theta - 1}\right)^2.$$

Now for $\theta \in V$,

$$\ell_n(\theta, x_0) - \ell_n(\theta_0, x_0) = (\theta - \theta_0)^T \nabla_\theta \ell_n(\theta_0, x_0) + \int_0^1 (\theta - \theta_0)^T D_\theta^2 \ell_n(\theta_0 + u(\theta - \theta_0), x_0)(\theta - \theta_0)(1-u)du. \tag{6}$$

Following Theorem 2 in Douc et al. (2004), $n^{-1/2}\nabla_\theta \ell_n(\theta_0, x)$ converges in distribution under $\mathbb{P}_{\theta_0}$ to $\mathcal{N}(0, V_0)$ for some positive definite matrix V_0, and following Theorem 3 in Douc et al. (2004), we get that $\sup_{\theta \in V} n^{-1} D_\theta^2 \ell_n(\theta, x_0)$ converges $\mathbb{P}_{\theta_0}$ a.s. to V_0. Thus, we may set:

$$S_n = \left\{\theta \in V; \|\theta - \theta_0\| \leq 1/\sqrt{n}\right\} \times \mathcal{X}$$

so that

$$\sup_{(\theta,x)\in S_n} \mathbb{P}_{\theta_0}\left[\ell_n(\theta,x) - \ell_n(\theta_0,x_0) < -C_n\right] = o(1)$$

follows from (5) and (6). The second part of (A1) is then satisfied as soon as $\pi(S_n) > n^{-D/2}$ which is true for instance if V is a neighbourhood of θ_0 and if the prior has a density with respect to Lebesgue measure, which lower bounded by a positive constant on V. Note that the freedom in the choice of V implies that (A1) can be verified in situations where the true distribution can be approximated by $\mathbb{P}_\theta$ such that ρ_θ is arbitrarily close to 1, as long as it is possible to choose paths in Θ to approximate θ_0 that avoid such pathological θ's. This is illustrated in the case of finite state space HMMs in the following section.

3. Finite state space

Here we assume that $\mathcal{X} = \{1, ..., k\}$. We may take μ as the uniform probability measure on $\{1, ..., k\}$. We first describe the setting in this case, and then we prove that, under some general assumptions, Theorem 1 applies. Moreover, we prove how some choices of the prior give posterior concentration rates for the finite marginals without additional mixing coefficient. The results obtained in Section 3.1 are valid both when the true distribution has $k_0 = k$ different states and when it has a smaller number of states.

3.1. Posterior convergence rates for the finite marginals

Q_θ denotes the transition matrix $(q_{ij})_{i,j\le k}$ and $\theta = (q_{ij}, 1 \le i \le k, 1 \le j \le k-1; \gamma_1, ..., \gamma_k)$ with $\gamma_j \in \Gamma \subset \mathbb{R}^d$ such that $g_\theta(y|x) = g_{\gamma_x}(y)$, $x \in \mathcal{X} = \{1, ..., k\}$ for some family of probability densities $(g_\gamma)_{\gamma\in\Gamma}$ with respect to ν. We denote Θ_k the parameter space.

Let $\mathcal{M}_k$ be the set of all possible probability distributions of $(Y_n)_{n\ge1}$ under $\mathbb{P}_\theta$ for all $\theta \in \Theta_k$. We say that the HMM $\mathbb{P}_\theta$ has order k_0 if the probability distribution of $(Y_n)_{n\ge1}$ under $\mathbb{P}_\theta$ is in $\mathcal{M}_{k_0}$ and not in $\mathcal{M}_k$ for all $k < k_0$. Notice that a HMM of order k_0 may be represented as a HMM of order k for any $k > k_0$. Let Q^0 be a $k_0 \times k_0$ transition matrix, and $(\gamma_1^0, ..., \gamma_{k_0}^0) \in \Gamma^{k_0}$ be parameters that define a HMM of order k_0. Then, $\theta = (q_{ij}, 1 \le i, j \le k; \gamma_1^0, ..., \gamma_{k_0}^0, \ldots, \gamma_{k_0}^0) \in \Theta_k$ with $Q = (q_{ij}, 1 \le i, j \le k)$ such that :

$$\begin{aligned} &q_{ij} = q^0_{ij} \quad i, j < k_0 \\ &q_{ij} = q^0_{k_0 j} \quad i \ge k_0, \quad j < k_0, \\ &\sum_{l=k_0}^{k} q_{il} = q^0_{ik_0} \quad i \le k_0, \text{ and} \quad \sum_{l=k_0}^{k} q_{il} = q^0_{k_0k_0}, \quad i \ge k_0 \end{aligned} \tag{7}$$

gives $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$. Indeed, let $(X_n)_{n\ge1}$ be a Markov chain on $\{1, \ldots, k\}$ with transition matrix Q. Let Z be the function from $\{1, \ldots, k\}$ to $\{1, \ldots, k_0\}$ defined by $Z(x) = x$ if $x \le k_0$ and $Z(x) = k_0$ if $x \ge k_0$. Then $(Z(X_n))_{n\ge1}$ is a Markov chain on $\{1, \ldots, k_0\}$ with transition matrix Q^0.

In the following we parametrize the transition matrices on $\{1, \ldots, k\}$ as $(q_{ij})_{1\le i\le k, 1\le j\le k-1}$ (implying that $q_{ik} = 1 - \sum_{j=1}^{k-1} q_{ij}$ for all $i \le k$) and we denote by Δ_k the set of probability mass functions $\Delta_k = \{(u_1, \ldots, u_{k-1}) : u_1 \ge 0, \ldots, u_{k-1} \ge 0, \sum_{i=1}^{k-1} u_i \le 1\}$. We shall also use the set of positive probability mass functions $\Delta_k^0 = \{(u_1, \ldots, u_{k-1}) : u_1 > 0, \ldots, u_{k-1} > 0, \sum_{i=1}^{k-1} u_i < 1\}$. Thus, we may set $\Theta_k = \Delta_k^k \times \Gamma^k$.

Any Markov chain on $\{1, \ldots, k\}$ admits a stationary distributio, if Q_θ admits more than one stationary distribution, we choose one that we denote μ_θ. Besides (3) holds with $R_\theta = 1$ and

$$\rho_\theta = \left(1 - \sum_{j=1}^{k} \inf_{1\le i\le k} q_{ij}\right)^{-1},$$

so that and as soon as the transition matrix Q_θ has positive entries, $\rho_\theta < 1$. For any $\theta = (q_{ij}, 1 \le i \le k, 1 \le j \le k-1; \gamma_1, ..., \gamma_k) \in \Theta_k$, any $y = (y_1, \ldots, y_l)$ in $\mathcal{Y}^l$,

$$f_{l,\theta}(y) = \sum_{1\le i_1,\ldots,i_l\le k} \mu_\theta(i_1)\, q_{i_1 i_2} \cdots q_{i_{l-1} i_l} g_{\gamma_{i_1}}(y_1) \cdots g_{\gamma_{i_l}}(y_l). \tag{8}$$

Let $\pi(u_1,\dots,u_{k-1})$ be a prior density with respect to the Lebesgue measure on Δ_k, and let $\omega(\gamma)$ be a prior density on Γ (with respect to the Lebesgue measure on $\mathbb{R}^d$). We consider prior distributions such that the rows of the transitions matrix Q are independently distributed from π and independent of the component parameters $\gamma_i, i=1,...,k$, which are independently distributed from ω. Hence the prior density (with respect to the Lebesgue measure) is equal to $\pi_k = \pi^{\otimes k}\otimes\omega^{\otimes k}$. In this section we use a Dirichlet type prior, see assumption (F1) below, or an exponential type prior, see assumption (FE1) below, on the transition parameters $(q_{ij}, j\leq k)$.

THEOREM 2. *Let $\theta_0 = (q^0_{ij}, 1\leq i\leq k_0, 1\leq j\leq k_0-1; \gamma^0_1,....,\gamma^0_{k_0})\in\Theta_{k_0}$ be the parameter of a HMM of order $k_0\leq k$. Assume that*

- *(F0) $q^0_{ij}>0$, $1\leq i\leq k_0, 1\leq j\leq k_0$*
- *(F1) π is continuous and positive on Δ^0_k, and there exists C, $\alpha_1>0$, ..., $\alpha_k>0$ such that (Dirichlet type priors):*

$$\forall (u_1,\dots,u_{k-1})\in\Delta^0_k,\ u_k = 1-\sum_{i=1}^{k-1}u_i,\quad 0<\pi\left(u_1,\dots,u_{k-1}\right)\leq C u_1^{\alpha_1-1}\cdots u_k^{\alpha_k-1}$$

  and ω is continuous and positive on Γ.

- *(F2) The function $\gamma\mapsto g_\gamma(y)$ is twice continuously differentiable in Γ, and for any $\gamma\in\Gamma$, there exists $\epsilon>0$ such that*

$$\int \sup_{\gamma'\in B_d(\gamma,\epsilon)} \|\nabla_\gamma \log g_{\gamma'}(y)\|^2 g_\gamma(y)\,\nu(dy)<+\infty,\ \int \sup_{\gamma'\in B_d(\gamma,\epsilon)} \|D^2_\gamma \log g_{\gamma'}(y)\|^2 g_\gamma(y)\,\nu(dy)<+\infty,$$

  $\|\sup_{\gamma'\in B_d(\gamma,\epsilon)}\nabla_\gamma g_{\gamma'}(y)\|\in L_1(\nu)$ *and* $\|\sup_{\gamma'\in B_d(\gamma,\epsilon)} D^2_\gamma g_{\gamma'}(y)\|\in L_1(\nu)$

- *There exist $a>0$ and $b>0$ such that*

$$\sup_{\|\gamma\|\leq n^b}\int \|\nabla_\gamma g_\gamma(y)\|d\nu(y)\leq n^a.$$

Then, there exists K large enough such that

$$\mathbb{P}^{\pi_k}\left[\|f_{l,\theta}-f_{l,\theta_0}\|_1(\rho_\theta-1)\geq K\sqrt{\frac{\log n}{n}}|Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1)$$

where $\rho_\theta = \left(1-\sum_{j=1}^k \inf_{1\leq i\leq k} q_{ij}\right)^{-1}$. If moreover $\bar\alpha := \sum_{1\leq i\leq k}\alpha_i > k(k-1+d)$, then

$$\mathbb{P}^{\pi_k}\left[\|f_{l,\theta}-f_{l,\theta_0}\|_1\geq 2Kn^{-\frac{\bar\alpha-k(k-1+d)}{2\bar\alpha}}(\log n)|Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1).$$

If we replace (F1) by

- *(EF1) π is continuous and positive on Δ^0_k, and there exists C such that (exponential type priors):*

$$\forall (u_1,\dots,u_{k-1})\in\Delta^0_k,\ u_k = 1-\sum_{i=1}^{k-1}u_i,$$
$$0<\pi\left(u_1,\dots,u_{k-1}\right)\leq C\exp(-C/u_1)\cdots\exp(-C/u_k)$$

  and ω is continuous and positive on Γ,

then there exists K large enough such that

$$\mathbb{P}^{\pi_k}\left[\|f_{l,\theta}-f_{l,\theta_0}\|_1\geq 2Kn^{-1/2}(\log n)^{3/2}|Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1).$$

The proof is presented in Appendix 6.2.

Theorem 2 provides guidelines to choose the prior. Indeed, if a Dirichlet $\mathcal{D}(\alpha_1, ..., \alpha_k)$ prior is considered on each row of the transition matrix of the hidden Markov chain, then choosing large enough values for the α_j's ensures a consistent posterior distribution in terms of the L_1 distance on the marginals. If one chooses exponential type priors, it is possible to get, up to a $\log n$ factor, the posterior concentration rate of order $1/\sqrt{n}$ for the finite marginals. Interestingly, this is quite different from what happens in the case of overspecified mixtures as described by Rousseau and Mengersen (2011). In the case of independent mixture models, the posterior distribution concentrates at the rate $1/\sqrt{n}$ around the true density of the observations (in terms of the L_1 distance) under very general conditions on the prior. In Rousseau and Mengersen (2011), the authors prove that by choosing small values of α_i, the posterior distribution concentrates on the configuration where the extra components are emptied, which is desirable since it leads to an interpretable posterior distribution. Here the story is quite different because to favour empty components (small weights in the stationary measure $\mu(\theta)$) corresponds also to favour slow mixing Markov chains, i.e. Q_θ's such that $\rho(\theta)$ is close to 1. Then the asymptotic behaviour of the likelihood is much less stable and it is not clear that the posterior will concentrate on the correct densities f_{l,θ_0}. Hence, to be able to interpret correctly the posterior distribution it is more desirable to choose large values of α_i. We do not claim however that the threshold $k(k-1+d)$ on $\bar{\alpha} := \sum_{1\leq i\leq k} \alpha_i$ is sharp. Our intuition is that it is probably enough to assume, at least when $k_0 = k-1$, that $\bar{\alpha} > k(k_0-1+d)$, which corresponds to the number of constraints involved in the construction (7), but we have not been able to prove it, except in the case $k = 2$, see Section 3.4. Although a posterior concentration in terms of the marginals is useful, when interest lies in fitting the model or in prediction, in some applications it is also interesting to recover the parameters correctly. In the following section we use Theorem 2 to obtain a (partial) asymptotic identification of the parameters by the posterior distribution and to construct a consistent procedure to estimate the true number of states k_0 when $k \geq k_0$.

3.2. Application to the estimation of the number of hidden states

Let us now describe more precisely how to recover all the HMM dynamics, that is to recover the number of hidden states, and, in the following section, given that the number of hidden states is known, how to recover the parameters. For this, we need to understand what the posterior concentration result says about the parameters.

To recover conditions on the parameters from conditions on the l-marginals $f_{l,\theta}$, we need an inequality that relates the L_1 distance of the l-marginals to the parameters of the HMM. Such an inequality, stated below as inequality (9), is proved in Gassiat and van Handel (2012) for translation mixture models, with the strength of being uniform over the number (possibly infinite) of populations in the mixture. However for our purpose, we do not need such a general result, and it is possible to prove that (9) holds for more general situations than families of translated distributions, under a structural assumption implying, in particular, the weak identifiability of the multidimensional mixtures. This condition is presented in Appendix 6.5 together with the proof of the fact that if it is verified then (9) holds (Lemma 2). This condition is implied in particular by the strong identifiability assumption of Rousseau and Mengersen (2011), and it holds for any parametric family $(g_\gamma)_{\gamma\in\Gamma}$ which can be represented as an exponential family, including location-scale Gaussian mixtures, Poisson mixtures, exponential mixtures and so on. The inequality following Theorem 3.9 of Gassiat and van Handel (2012) says that there exists a constant $c(\theta_0) > 0$ such that for any

small enough positive ε,

$$\begin{aligned}\frac{\|f_{l,\theta}-f_{l,\theta_0}\|_1}{c(\theta_0)} \geq & \sum_{1\leq j\leq l:\forall i,\|\gamma_j-\gamma_i^0\|>\varepsilon} \mathbb{P}_\theta\left(X_1=j\right) \\ & + \sum_{1\leq i_1,\ldots,i_l\leq k_0} \left[\left|\mathbb{P}_\theta\left(X_{1:l}\in A(i_1,\ldots,i_l)\right)-\mathbb{P}_{\theta_0}\left(X_{1:l}=i_1\cdots i_l\right)\right|\right. \\ & + \left\|\sum_{(j_1,\ldots,j_l)\in A(i_1,\ldots,i_l)} \mathbb{P}_\theta\left(X_{1:l}=j_1\cdots j_l\right)\left\{\begin{pmatrix}\gamma_{j_1}\\ \cdots\\ \gamma_{j_l}\end{pmatrix}-\begin{pmatrix}\gamma_{i_1}^0\\ \cdots\\ \gamma_{i_l}^0\end{pmatrix}\right\}\right\| \\ & \left. + \frac{1}{2}\sum_{(j_1,\ldots,j_l)\in A(i_1,\ldots,i_l)} \mathbb{P}_\theta\left(X_{1:l}=j_1\cdots j_l\right)\left\|\begin{pmatrix}\gamma_{j_1}\\ \cdots\\ \gamma_{j_l}\end{pmatrix}-\begin{pmatrix}\gamma_{i_1}^0\\ \cdots\\ \gamma_{i_l}^0\end{pmatrix}\right\|^2\right] \quad (9)\end{aligned}$$

where $A(i_1,\ldots,i_l)=\{(j_1,\ldots,j_l):\|\gamma_{j_1}-\gamma_{i_1}^0\|\leq\varepsilon,\ldots,\|\gamma_{j_l}-\gamma_{i_l}^0\|\leq\varepsilon\}$. The above lower bound essentially correponds to a partition of $\{1,\ldots,k\}^l$ into k_0^l+1 groups, where the first k_0^l groups correspond to the components that are close to true distinct components in the multivariate mixture and the last corresponds to components that are emptied. The first term on the right hand side controls the weights of the components that are emptied (group k_0^l+1), the second term control the sum of the weights of the components belonging to the i-th group, for $i=1,\ldots,k_0^l$ (components merging with the true i-th component), the third controls the distance between the mean value over the group i and the true value of the i-th component in the true mixture while the last term controls the distance between each parameter value in group i and the true value of the i-th component. Notice now that for some constant $C>0$,

$$\left\|\begin{pmatrix}\gamma_{j_1}\\ \cdots\\ \gamma_{j_l}\end{pmatrix}-\begin{pmatrix}\gamma_{i_1}^0\\ \cdots\\ \gamma_{i_l}^0\end{pmatrix}\right\| \geq C\left\|\gamma_{j_1}-\gamma_{i_1}^0\right\|\vee\cdots\vee\left\|\gamma_{j_l}-\gamma_{i_l}^0\right\|$$

so that we get that for maybe some other constant $\tilde{c}(\theta_0)>0$:

$$\begin{aligned}\frac{\|f_{l,\theta}-f_{l,\theta_0}\|_1}{\tilde{c}(\theta_0)} \geq & \sum_{1\leq j\leq l:\forall i,\|\gamma_j-\gamma_i^0\|>\varepsilon} \mathbb{P}_\theta\left(X_1=j\right) \\ & \sum_{1\leq i_1,\ldots,i_l\leq k_0}\left|\sum_{(j_1,\ldots,j_l)\in A(i_1,\ldots,i_l)} \mathbb{P}_\theta\left(X_{1:l}=j_1\cdots j_l\right)-\mathbb{P}_{\theta_0}\left(X_{1:l}=i_1\cdots i_l\right)\right| \\ & + \sum_{i=1}^{k_0}\left\|\sum_{j:\|\gamma_j-\gamma_i^0\|\leq\varepsilon} \mathbb{P}_\theta\left(X_1=j\right)\gamma_j-\gamma_i^0\right\| + \frac{1}{2}\sum_{i=1}^{k_0}\sum_{j:\|\gamma_j-\gamma_i^0\|\leq\varepsilon} \mathbb{P}_\theta\left(X_1=j\right)\|\gamma_j-\gamma_i^0\|^2.\end{aligned}$$

It follows that as soon as

$$\|f_{l,\theta}-f_{l,\theta_0}\|_1 \lesssim u_n$$

for u_n tending to 0 as n tends to infinity (which is the case with u_n some negative power of n if the prior is well chosen), we get that, for any $j\in\{1,\ldots,k\}$, either $\mathbb{P}_\theta\left(X_1=j\right)\lesssim u_n$, or

$$\exists i\in\{1,\ldots,k_0\}\ \mathbb{P}_\theta\left(X_1=j\right)\|\gamma_j-\gamma_i^0\|^2\lesssim u_n.$$

This means that extra states are emptied or merge with true ones. Let c_n be a sequence increasing slowly to infinity and let J be the set (depending on θ) such that

$$J=\left\{j\ :\ \mathbb{P}_\theta\left(X_1=j\right)\geq u_n c_n\right\}.$$

For any $j\in J$, let $A_j=\left\{i\in J\ :\ \mathbb{P}_\theta\left(X_1=j\right)\|\gamma_j-\gamma_i\|^2\leq u_n\sqrt{c_n}\right\}$, we say that the elements j_1 and j_2 of J are merging if $A_{j_1}\cap A_{j_2}\neq\emptyset$, what we note $j_1\sim j_2$, then we say that two elements i and j are equivalent if there is a sequence $i_1,\ldots,i_r$ of elements such that $i_1=i\sim i_2, i_2\sim i_3,\ldots,i_{r-1}\sim i_r=j$, and we define L as the number of equivalent classes with respect to this

equivalence relationship. Then, under the assumptions of Theorem 2, it is possible to choose u_n such that

$$\mathbb{P}^{\pi_k}\left[L \neq k_0 | Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1). \tag{10}$$

Under the Dirichlet type prior assumption (F1), one may take $u_n = (n/\log n)^{-(\bar{\alpha}-k(k-1+d))/2\bar{\alpha})}$, while under the exponential type prior assumption (FE1), one may take $u_n = (n/\log n)^{-1/2}$. An alternative choice for u_n could be the posterior L_1 risk between l-marginals, say between marginals of order 2 :

$$u_n = E^{\pi}\left[\|f_{2,\theta} - f_{2,\hat{\theta}}\|_1 | Y_{1:n}\right],$$

where $\hat{\theta}$ is the posterior expectation of θ. This has the advantage of being automatically calibrated. There are various possible choices for c_n, we suggest $c_n = \sqrt{\log n}$, since in our simulation study, this lead to a satisfactory behaviour. Indeed, in the case $n = 500$ for instance, if the posterior risk above is of order 0.1 (as in some of our simulations) $u_n \log n > 0.6$ which is not a reasonable choice since it will erase all components whose weights are smaller than 0.6, whereas $u_n\sqrt{\log n} = 0.25$. Obviously, in practice, the choice of c_n should then depend on k. A heuristically good choice for c_n is $\sqrt{\log n/k}$, since if k is large, most weights should be small. An estimator of the order can then be for instance the posterior mode of L, however the whole posterior distribution of L is also of interest.

3.3. Consistent estimation of the parameters when the order is known

When the model is correctly specified, that is when the true number of states k_0 is equal to k, we are able to refine the concentration result in two ways : (i) obtain a better concentration rate (the usual $1/\sqrt{n}$ rate up to $\log n$) and (ii) obtain a posterior concentration rate in the parameter scale. We still consider the case where (9) holds. If now $k = k_0$, following from Theorem 2, if $\bar{\alpha} = \sum_{i=1}^{k} \alpha_i > k(k-1+d)$ with a Dirichlet type prior, or with an exponential type prior, it holds that

$$\|f_{l,\theta} - f_{l,\theta_0}\|_1 \lesssim u_n$$

where $u_n = O(n^{-\tau})$ for some $\tau > 0$. Since now (9) implies that

$$\|f_{l,\theta} - f_{l,\theta_0}\|_1 \geq \tilde{c}(\theta_0) \left\{ \sum_{1\leq i_1,\ldots,i_l \leq k_0} \left|\mathbb{P}_\theta\left(X_{1:l} = i_1 \cdots i_l\right) - \mathbb{P}_{\theta_0}\left(X_{1:l} = i_1 \cdots i_l\right)\right| \right.$$
$$\left. + \sum_{i=1}^{k_0} \mathbb{P}_\theta\left(X_1 = j\right) \|\gamma_j - \gamma_i^0\| + \frac{1}{2}\sum_{i=1}^{k_0} \mathbb{P}_\theta\left(X_1 = j\right) \|\gamma_j - \gamma_i^0\|^2 \right\},$$

we obtain that, for all $i_1 \cdots i_l$,

$$\left|\mathbb{P}_\theta\left(X_{1:l} = i_1 \cdots i_l\right) - \mathbb{P}_{\theta_0}\left(X_{1:l} = i_1 \cdots i_l\right)\right| \lesssim u_n,$$

which means in particular that for all i, j,

$$\left|\mathbb{P}_\theta\left(X_1 = i\right) - \mathbb{P}_{\theta_0}\left(X_1 = i\right)\right| \lesssim u_n, \quad \text{and} \quad |q_{i,j} - q_{i,j}^0| \lesssim u_n.$$

Since $q_{i,j}^0 > 0$ for all $i, j \leq k$, then there exists $a > 0$ such that $\mathbb{P}^{\pi}\left[\exists i, j,\ q_{i,j} \leq a | Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1)$ and if $q_{i,j} \geq a$ for all i, j, $\rho(\theta) - 1 \geq \sum_{j=1}^{k} \min_i q_{i,j} \geq ka > 0$, hence

$$\{(\rho(\theta)-1)\|f_{l,\theta_0} - f_{l,\theta}\|_1 \leq K\sqrt{\log n/n}\} \cap \{\rho(\theta) - 1 \geq ka\} \subset \{\|f_{l,\theta_0} - f_{l,\theta}\|_1 \leq K^{'}\sqrt{\log n/n}\}$$

for some $K^{'} > 0$. We thus have that

$$\mathbb{P}^{\pi}\left[\|f_{l,\theta_0} - f_{l,\theta}\|_1 \leq K^{'}\sqrt{\log n/n} | Y_{1:n}\right] = 1 + o_{\mathbb{P}_{\theta_0}}(1)$$

and in the above inequalities we can replace u_n by $\sqrt{\log n/n}$. We finally obtain the following result on the posterior concentration :

COROLLARY 1. *When $k = k_0$, when (9) holds, under the assumptions of Theorem 2, if the prior is of Dirichlet type with $\bar{\alpha} = \sum_{i=1}^{k} \alpha_i > k(k-1+d)$ or if the prior is of exponential type, there exists $K > 0$ such that*

$$\mathbb{P}^{\pi}\left[\|f_{l,\theta_0} - f_{l,\theta}\|_1 \leq K\sqrt{\frac{\log n}{n}}|Y_{1:n}\right] = 1 + o_{\mathbb{P}_{\theta_0}}(1)$$

and

$$\mathbb{P}^{\pi_{k_0}}\left[\forall i,j|q_{i,j} - q^0_{i,j}| \leq K\sqrt{\frac{\log n}{n}};\quad \|\gamma_i - \gamma^0_i\| \leq K\sqrt{\frac{\log n}{n}}|Y_{1:n}\right] = 1 + o_{\mathbb{P}_{\theta_0}}(1)\,.$$

Hence, we recover the $n^{-1/2}$ rate of convergence (up to a $\log n$ term), as soon as the prior vanishes quickly enough near $q_{i,j} = 0$.

3.4. Asymptotic behaviour of the posterior distribution when $k_0 = 1$ and $k = 2$

In this section we restrict our attention to the simpler case where $k_0 = 1$ and $k = 2$. We will see that despite its apparent simplicity, the asymptotic analysis of the posterior distribution leads to a guideline on the choice of the prior parameters α_i's which is (almost) opposite to that proposed in Rousseau and Mengersen (2011), in the case of mixture models. We still consider situations where (9) holds, and choose independent Dirichlet type priors for the rows of the transition matrix. We prove that the extra component is not emptied but merges with the true one, under large enough α_i's for the Dirichlet prior.

When $k = 2$, we can parameterize θ as $\theta = (p, q, \gamma_1, \gamma_2)$, with $0 \leq p \leq 1$, $0 \leq q \leq 1$, so that

$$Q_\theta = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},\ \mu_\theta = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$$

when $p \neq 0$ or $q \neq 0$. If $p = 0$ and $q = 0$, set $\mu_\theta = \left(\frac{1}{2}, \frac{1}{2}\right)$ for instance. Also,

$$\rho_\theta - 1 \geq (p+q) \wedge (2 - (p+q))\,.$$

When $k_0 = 1$, observations are i.i.d. with distribution $g_{\gamma^0}d\nu$, so that one may take $\theta_0 = (p, 1-p, \gamma^0, \gamma^0)$ for any $0 < p < 1$, or $\theta_0 = (0, q, \gamma^0, \gamma)$ for any $0 < q \leq 1$ and any γ, or $\theta_0 = (p, 0, \gamma, \gamma^0)$ for any $0 < p \leq 1$ and any γ. Also, for any $x \in \mathcal{X}$, $\mathbb{P}_{\theta_0,x} = \mathbb{P}_{\theta_0}$ and

$$\ell_n(\theta, x) - \ell_n(\theta_0, x_0) = \ell_n(\theta, x) - \ell_n(\theta_0, x).$$

We take independent Beta priors on (p, q) :

$$\pi(p,q) = C_{\alpha,\beta}p^{\alpha-1}(1-p)^{\beta-1}q^{\alpha-1}(1-q)^{\beta-1},$$

thus satisfying (F1).

Let $u_n = K\sqrt{\frac{\log n}{n}}$ with K large enough so that, thanks to Theorem 2, $\mathbb{P}^{\pi}(B_n|Y_{1:n}) = 1 + o_P(1)$, as soon as assumption (F2) holds, and where B_n is the set

$$B_n = \left\{(p+q) \wedge (2-(p+q))\,\|f_{2,\theta} - f_{2,\theta_0}\|_1 \leq u_n\right\}.$$

Then, for any sequence of sets $(A_n)_{n\geq 1}$, for any $D > 0$, and for any sequence $(c_n)_{n\geq 1}$ of real numbers

$$\begin{aligned}
E_{\theta_0}\left[\mathbb{P}^{\pi}(A_n|Y_{1:n})\right] &= E_{\theta_0}\left[\mathbb{P}^{\pi}(A_n \cap B_n|Y_{1:n})\right] + o(1) \\
&= E_{\theta_0}\left[\frac{\int_{A_n\cap B_n\times\mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\pi(d\theta)\pi_{\mathcal{X}}(dx)}{\int_{B_n\times\mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\pi(d\theta)\pi_{\mathcal{X}}(dx)}\right] + o(1) \\
&:= E_{\theta_0}\left[\frac{N_n}{D_n}\right] + o(1) \\
&\leq \mathbb{P}_{\theta_0}\left(D_n \leq c_n n^{-D/2}\right) + \frac{n^{D/2}}{c_n}\pi_2(A_n \cap B_n) + o(1)\,.
\end{aligned}$$

Thus, if

$$\mathbb{P}_{\theta_0}\left(D_n \leq c_n n^{-D/2}\right) = o\left(1\right), \tag{11}$$

one gets

$$E_{\theta_0}\left[\mathbb{P}^{\pi}\left(A_n|Y_{1:n}\right)\right] \leq \frac{n^{D/2}}{c_n}\pi_2(A_n \cap B_n) + o\left(1\right) \tag{12}$$

Let ϵ_n decrease slowly to 0 and such that $\frac{u_n}{\epsilon_n}$ tends to 0. Consider the set

$$A_n = \left\{ \frac{p}{p+q} \leq \epsilon_n \text{ or } \frac{q}{p+q} \leq \epsilon_n \right\}.$$

Then the following holds:

COROLLARY 2. *Under the assumptions of Theorem 2, in the situation where (9) holds, and if moreover for all x, $\gamma \mapsto g_\gamma(x)$ is four times continuously differentiable on Γ, and if for any $\gamma \in \Gamma$ there exists $\epsilon > 0$ such that for any $i = 1, 2, 3, 4$, if $D^i_\gamma g_{\gamma'}$ denotes the i-th differential operator (with respect to γ) of g at point γ',*

$$\int \sup_{\gamma' \in B_d(\gamma,\epsilon)} \| \frac{D^i_\gamma g_{\gamma'}}{g_{\gamma'}}\left(y\right) \|^4 g_\gamma\left(y\right) \nu(dy) < +\infty, \tag{13}$$

the extra component cannot be emptied at rate $\epsilon_n = \frac{1}{(\log n)^3}$, that is

$$\mathbb{P}^{\pi}\left(A_n|Y_{1:n}\right) = o_P(1)$$

as soon as $\alpha > 3d/4$.

To prove Corollary 2, we prove that

$$\pi_2(A_n \cap B_n) \lesssim u_n^{2\alpha} + u_n^{\alpha+d} + u_n^{d+d/2}\epsilon_n^{\alpha-d/2}, \tag{14}$$

and that (11) holds with $D = d + d/2$ as soon as C_n tends to infinity as n tends to infinity. Thus, taking $C_n = \log\log\log n$ and using (12), Corollary 2 follows. The detailed proof is given in the appendix.

With extra work, it might be possible to obtain that Corollary 2 holds for any ϵ_n decreasing to 0 and such that $\frac{u_n}{\epsilon_n}$ tends to 0. For this, one needs to obtain the posterior concentration rate $\sqrt{\frac{1}{n}}$ instead of $\sqrt{\frac{\log n}{n}}$ in Theorem 2. This requires the use of local packing numbers instead of global packing numbers to cover the sets with respect to the distance $\|f_{l,\theta} - f_{l,\theta_0}\|_1$. It should be possible to do it by using the entropy results in Gassiat and van Handel (2012).

4. Simulations

To illustrate the results obtained in Section 3, we run a simulation study under the following setups: in all cases the model corresponds to a finite state space HMM, with $k = 2$ states and Gaussian emissions with unknown means μ_1 and μ_2 and known variance:

$$Y_i|X_i = x \sim \mathcal{N}(\mu_x, 1), \quad x \in \{1, 2\}, \quad Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

We consider independent Beta(α, α) priors on (p, q) and independent normal priors on (μ_1, μ_2) with variance 1 and mean 0.

There are various ways proposed in the literature to implement such posterior distributions. The most well known is the Gibbs sampler which simulates the hidden states (possibly after integrating out the parameters given the states), see for instance Frühwirth-Schnatter (2006). Unfortunately, we have found that such algorithms do not allow to visit well the parameter space when there are well separated multiple modes, even after combining such algorithms with parallel tempering, see Geyer (1991) or Baragatti (2011) for a review on these methods. We have therefore considered an

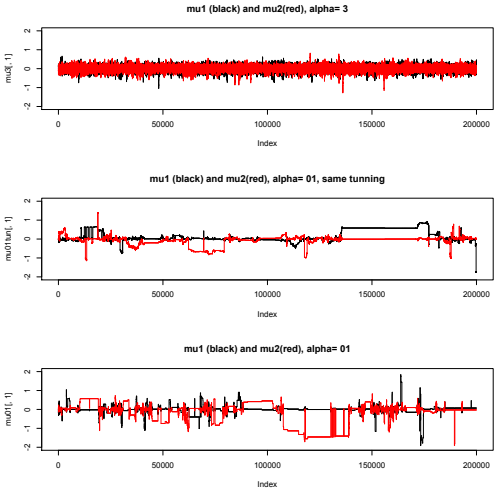

Figure 1. output of the posterior distribution on μ_1 (black) and μ (red) for $\alpha = 3$ (Left) and $\alpha = 0.1$ (Right)

Hasting-Metropolis algorithm where only the parameters are simulated and at each iteration the likelihood is computed; we have combined this algorithm with parallel tempering, as follows : For each temperature T varying in the set $1 = T_1 < T_2 < \cdots < T_L$, we propose at iteration t

$$\mu'(T) = \mu^t(T) + \sigma_\mu(T)\epsilon_\mu^t, \quad w'(T) = |w^t(T) + \sigma_w(T)\epsilon_w(T)|, \quad \theta' = (\mu', w')$$

where μ denotes the means and w the unnormalized weights. Then after the acceptance step for each T, we propose a swap between a random pair (T_i, T_j), which is accepted according to the ratio of posteriors.

Interestingly, the performance of the algorithm is strongly influenced by the prior, even for a large number of observations. Figure 1 displays the output of three Hasting-Metropolis runs, associated with the same data which are composed of $n = 1000$ independent $\mathcal{N}(0, 1)$, the top graph corresponds to $\alpha = 3$, the middle graph to $\alpha = 0.1$ based in the same tuning parameters as the top graph and the bottom graph to $\alpha = 0.1$ with tuning parameters chosen to have an acceptance rate greater than 0.12. The performance is strongly deteriorated when $\alpha = 0.1$, despite a rather large number of observations. We can see that, even decreasing significantly the variance in the proposal distributions does not prevent from being trapped in local modes, which are much more pronounced than in the case $\alpha = 3$. This might be a sign of an unstable behaviour of the likelihood in regions where p or q are small but we have no proof for this. Another aspect of the computational difficulty induced by small values of α is the label switching issue. This issue exists whatever the value of α of course, however in cases of small values of α, it is rendered even more difficult. Figure 2 shows the output of the MCMC chain of q_{11} and q_{22} before and after treatment of the label switching issue, i.e. choosing the labels, at each iteration which minimize the distance to the posterior mode (MAP) (obtained from maximizing the posterior likelihood over all the iterations). The MAP corresponds to $\hat{q}_{11} = 0.99$ and $\hat{q}_{22} = 0.9999$, in other words to a transition matrix closed to the identity matrix. The resulting posterior on (q_{11}, q_{22}) oscillates sometimes between the values $(1, 0)$ and $(0, 1)$ (or close to it), which shows that the relabelling was inefficient.To make it more efficient, we use a change of parameterization on the q_{ij}, we work with $\log q_{ij}$ instead and minimize the distance between the MAP and the values at each iteration in this new parameterization. The posterior mean of q_{11} with the relabelling based on the logarithm of the transition matrix equals 0.18, whereas it was equal to 0.48 under the other approach for the relabelling. This phenomenon does not take place when $\alpha = 3$, the same result is obtained for both parameterization, see figure 2, in particular the posterior means of q_{11} after both types of relabelling are equal to 0.48. Note that in Figure 2, the MCMC chains have been tuned to get a reasonable acceptance rate.

Hence, choosing small values for α influences strongly the behaviour of MCMC samplers, even for large values of n, at least in the above case, i.e. when the true generating process corresponds to a smaller value of the number of states than in the model. In particular we have not observed

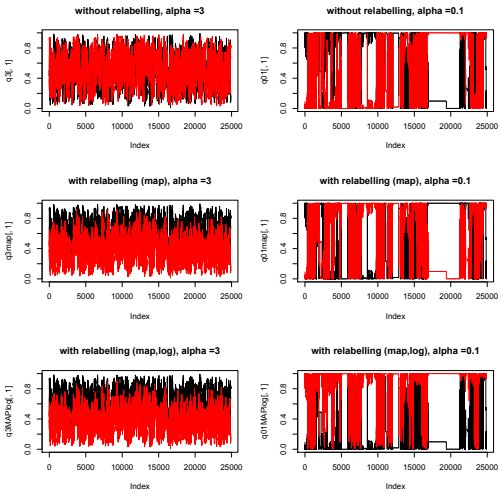

Figure 2. output of the posterior distribution on q_{11} (black) and q_{22} (red) for $\alpha = 3$ (right) and $\alpha = 0.1$ (left), without relabelling (top), with simple relabelling (middle) with relabelling in new parameterization (bottom)

this difficulty in cases where the data was generated from a HMM with two states as in the model (well specified case 1 below). We now study on simulations to what extend the behaviour of the posterior is influenced by the value of α.

The data are simulated as

- **Case 1 :** (Well specified) $k_0 = 2$, $\mu_0 = (0, 4)$, $p_0 = 0.4, q_0 = 0.7$.
- **Case 2 :** (Mis-specified) $k_0 = 1$, $\mu_0 = 0$.

In both cases the estimation will be conducted with $\alpha = 3 > k - 1 + d = 2$ (posterior concentration has been proved) and $\alpha = 0.1 < k - 1 + d$ (there is no result on posterior concentration both in terms of the l-marginals and in terms of the parameters).

Figure 3 shows the posterior expectations of the L_1 distance between the 2- marginals $f_{2,\theta}$ and f_{2,θ_0} in case 1 (left panel) and case 2 (right panel), under $n = 100, 500$. In each case the first two boxplots corresponds to $n = 100$ with $\alpha = 3$ (first) and $\alpha = 0.1$ (second) and the next two boxplots to $n = 500$. The last two boxplots of the right panel (case 2) correspond to $n = 1000$. It appears that the impact of α in case 2 is much stronger than in case 1. In the former the posterior distribution seems to have larger posterior (L_1) risk under $\alpha = 0.1$ than under $\alpha = 3$, the difference being quite significant for $n = 100$ and lessened for $n = 500$ whereas we see no real difference in case 1, even for $n = 100$. In case 2, for large n, namely $n = 1000$, the boxplot seems to indicate that the posterior L_1 risk is at least of the same order with small α ($\alpha = 0.1$) than with large α ($\alpha = 3$).

We have also computed the posterior distribution of L, for one data set, in each situation. The choice of $c_n = \sqrt{\log n}$ and $u_n = E^{\pi}[\|f_{2,\theta} - f_{2,\hat{\theta}}\|_1 | Y_{1:n}]$ leads to posterior distributions of L concentrated on the right value for each of the situations and both for $\alpha = 3$ and $\alpha = 0.1$.

5. Discussion

This paper has contributed in the understanding of the Bayesian analysis of HMMs. A positive conclusion derived from our results is that even in mis-specified cases, i.e. when the number k of states in the model is larger than the true number of states k_0, the posterior distribution will concentrate on the true values of the parameter, provided the prior forbids strongly enough small transition probabilities (for instance if the hyperparameters $\alpha_{i,j}$ in the Dirichlet priors on the rows of the transition matrix are large enough). However, contrarywise to static mixture models, there are no definite conclusions when the hyperparameters $\alpha_{i,j}$ are small. It thus gives guidelines for the choice of the prior in HMM modeling with finite state space. When the number of states is

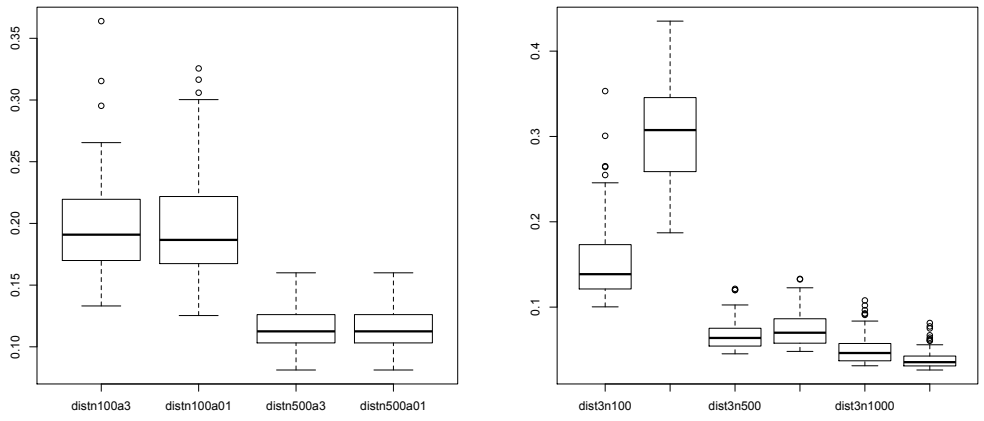

Figure 3. boxplots of the posterior expectations of $|f_{2,\theta} - f_{2,\theta_0}|$ for $\alpha = 3, 0.1$ and $n = 100, 500$, in case 1 (left) and case 2 (right)

unknown, one should preferably choose Dirichlet type priors with large enough parameters. This recommandation is supported both by theoretical and computational reasons. On the theoretical side, we have been able to prove that using such priors leads to consistent posterior distribution and we have been able to give convergence rates. On the computational side, small parameters lead to unstable chains, and heavier label switching issues, as our simulation study indicates. Thus, to have confidence in the output of our simulated posterior distribution and to obtain interpretable results one should to choose (Dirichlet) priors on the transition matrices with large enough values for the $\alpha_{i,j}$. We have been able to give a lower bound for what "large enough" means in the general finite state space situation, and sharpened this lower bound in the specific situations of 2 hidden states.

Under identifiability conditions (i.e. condition (L(l))), the posterior concentration in terms of the l-marginals can be expressed in terms of the parameters, thus enabling us to provide a consistent estimator of the number of hidden states (when the prior is correctly specified). Note that the posterior concentration rate of order $1/\sqrt{n}$ for the parameter is only obtained if the number of states is correctly specified, i.e. when $k = k_0$.

The question of the asymptotic behaviour of the posterior distribution, in case of small parameters for the Dirichlet prior remains open, though the computation of the posterior distribution becomes much more difficult in such cases. In our simulation study, the posterior distribution seems to concentrate wih n even for small values of α. Wether it really shows consistency or wether it can be extended beyong the simple case of $k_0 = 1$ and $k = 2$ is not clear to us.

6. Appendix

6.1. Proof of Theorem 1

The proof follows the same lines as in Ghosal and van der Vaart (2006). Let $(\epsilon_n)_{n\geq 1}$ be a sequence of positive real numbers. We write

$$\begin{aligned}\mathbb{P}^\pi\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n | Y_{1:n}\right] &= \frac{\int_{A_n\times\mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\pi(d\theta)\pi_{\mathcal{X}}(dx)}{\int_{\Theta\times\mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\pi(d\theta)\pi_{\mathcal{X}}(dx)} \\ &:= \frac{N_n}{D_n},\end{aligned}$$

where $A_n = \{\theta : \|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n\}$. A lower bound on D_n is obtained in the following usual way. Set for any real number C, $\Omega_n(C) = \{(\theta, x); \ell_n(\theta, x) - \ell_n(\theta_0, x_0) \geq -C\}$, which is a

random subset of $\Theta \times \mathcal{X}$ (depending on $Y_{1:n}$).

$$\begin{aligned} D_n &\geq \int_{S_n} \mathbb{1}_{\Omega_n(C)} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \pi(d\theta)\pi_{\mathcal{X}}(dx) \\ &\geq e^{-C} \pi \otimes \pi_{\mathcal{X}}(S_n \cap \Omega_n(C)), \end{aligned}$$

therefore using (A1), for any sequence $(C_n)_{n\geq 1}$ of real numbers tending to $+\infty$, there exists $c_1 > 0$ such that

$$\begin{aligned} \mathbb{P}_{\theta_0}\left[D_n < c_1 e^{-C_n} n^{-D/2}\right] &\leq \mathbb{P}_{\theta_0}\left[D_n < e^{-C_n} \pi \otimes \pi_{\mathcal{X}}(S_n)/2\right] \\ &\leq \mathbb{P}_{\theta_0}\left[\pi \otimes \pi_{\mathcal{X}}(S_n \cap \Omega_n(C_n)^c) \geq \pi \otimes \pi_{\mathcal{X}}(S_n)/2\right] \\ &\leq 2\frac{\int_{S_n} \mathbb{P}_{\theta_0}\left[\ell_n(\theta,x) - \ell_n(\theta_0,x_0) \leq -C_n\right] \pi(d\theta)\pi_{\mathcal{X}}(dx)}{\pi \otimes \pi_{\mathcal{X}}(S_n)} \\ &= o(1), \end{aligned}$$

by assumption (A1) again.
Thus, for any sequence $(C_n)_{n\geq 1}$ of real numbers tending to $+\infty$,

$$\mathbb{P}^{\pi}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n | Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1) + \frac{N_n}{D_n} \mathbb{1}_{D_n \geq c_1 n^{-D/2} e^{-C_n}}.$$

But

$$N_n = \int_{(A_n \cap \mathcal{F}_n)\times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \pi(d\theta)\pi_{\mathcal{X}}(dx) + \int_{(A_n \cap \mathcal{F}_n^c)\times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \pi(d\theta)\pi_{\mathcal{X}}(dx)$$

and

$$E_{\theta_0}\left[\int_{(A_n \cap \mathcal{F}_n^c)\times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \pi(d\theta)\pi_{\mathcal{X}}(dx)\right] = O\left[\pi\left(A_n \cap \mathcal{F}_n^c\right)\right] = o\left(n^{-D/2}\right)$$

by Fubini's theorem and the fact that, by $(A1)$, $\ell_n(\theta_0) - \ell_n(\theta_0, x_0)$ is uniformly upper bounded, so that by taking C_n tending to $+\infty$ slowly enough,

$$\mathbb{P}^{\pi}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n | Y_{1:n}\right] = o_{\mathbb{P}_{\theta_0}}(1) + \frac{\tilde{N}_n}{D_n} \mathbb{1}_{D_n \geq c_1 n^{-D/2} e^{-C_n}}$$

where $\tilde{N}_n = \int_{(A_n \cap \mathcal{F}_n)\times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \pi(d\theta)\pi_{\mathcal{X}}(dx)$. Let now $(\theta_j)_{j=1,\ldots,N}$, $N = N(\delta, \mathcal{F}_n, d_l(.,.))$, be the sequence of θ_j's in $\mathcal{F}_n$ such for all $\theta \in \mathcal{F}_n$ there exists a θ_j with $d_l(\theta_j, \theta) \leq \delta$ (for some δ to be fixed later). Assume for simplicity's sake and without loss of generality that n is a multiple of the integer l, and define

$$\phi_j = \mathbb{1}_{\sum_{i=1}^{n/l}\left(\mathbb{1}_{(Y_{li-l+1},\ldots,Y_{li})\in A_j} - \mathbb{P}_{\theta_0}((Y_1,\ldots,Y_l)\in A_j)\right) > t_j}$$

where

$$A_j = \{(y_1,\ldots,y_l) \in \mathcal{Y}\times\mathcal{Y} \,:\, f_{l,\theta_0}(y_1,\ldots,y_l) \leq f_{l,\theta_j}(y_1,\ldots,y_l)\}$$

for some positive real number t_j to be fixed later also. Note that

$$\mathbb{P}_{\theta_j}((Y_1,\ldots,Y_l)\in A_j) - \mathbb{P}_{\theta_0}((Y_1,\ldots,Y_l)\in A_j) = \frac{1}{2}\|f_{l,\theta_j} - f_{l,\theta_0}\|_1$$

Define also

$$\psi_n = \max_{1\leq j\leq N : \theta_j \in A_n} \phi_j.$$

Then

$$E_{\theta_0}\left(\frac{\tilde{N}_n}{D_n}\psi_n\right) \leq E_{\theta_0}\psi_n \leq N(\delta, \mathcal{F}_n, d(.,.)) \max_{1\leq j\leq N : \theta_j \in A_n} E_{\theta_0}\phi_j$$

and using the usual equality

$$E_{\theta_0}\left(\tilde{N}_n(1-\psi_n)\right) = \int_{\mathcal{X}} E_{\theta_0,x_0}\left(\tilde{N}_n(1-\psi_n)\right)\mu_{\theta_0}(dx_0) = \int_{(A_n\cap\mathcal{F}_n)\times\mathcal{X}} E_{\theta,x}\left((1-\psi_n)\right)\pi(d\theta)\,\pi_{\mathcal{X}}(dx)$$

so that:

$$\mathbb{P}_{\theta_0}\left(\mathbb{P}^{\pi}\left[\|f_{l,\theta}-f_{l,\theta_0}\|_1\frac{\rho_\theta-1}{2R_\theta+\rho_\theta-1}\geq\epsilon_n|Y_{1:n}\right]\right) \leq o(1)+\left(\frac{n}{\delta}\right)^M \max_{1\leq j\leq N:\theta_j\in A_n} E_{\theta_0}\phi_j + O\left[n^{D/2}e^{C_n}\int_{(A_n\cap\mathcal{F}_n)\times\mathcal{X}} E_{\theta,x}\left((1-\psi_n)\right)\pi(d\theta)\,\pi_{\mathcal{X}}(dx)\right]. \tag{15}$$

Now

$$E_{\theta_0}[\phi_j] = \mathbb{P}_{\theta_0}\left[\sum_{i=1}^{n/l}\left(\mathbb{1}_{(Y_{li-l+1},\ldots,Y_{li})\in A_j}-\mathbb{P}_{\theta_0}((Y_1,\ldots,Y_l)\in A_j)\right)>t_j\right]$$

and

$$E_{\theta,x}(1-\phi_j) = \mathbb{P}_{\theta,x}\left[\sum_{i=1}^{n/l}\left(-\mathbb{1}_{(Y_{li-l+1},\ldots,Y_{li})\in A_j}+\mathbb{P}_{\theta,x}((Y_{li-l+1},\ldots,Y_{li})\in A_j)\right) > -t_j+\sum_{i=1}^{n/l}\left(\mathbb{P}_{\theta,x}((Y_{li-l+1},\ldots,Y_{li})\in A_j)-\mathbb{P}_{\theta_0}((Y_1,\ldots,Y_l)\in A_j)\right)\right].$$

Consider the sequence $(Z_i)_{i\geq 1}$ with for all $i\geq 1$, $Z_i=(X_{li-l+1},\ldots,X_{li},Y_{li-l+1},\ldots,Y_{li})$, which is, under $\mathbb{P}_\theta$, a Markov chain with transition kernel $\bar{Q}_\theta$ given by

$$\bar{Q}_\theta(z,dz') = g_\theta(y_1'|x_1')\cdots g_\theta(y_l'|x_l')Q_\theta(x_l,dx_1')Q_\theta(x_1',dx_2')\cdots Q_\theta(x_{l-1}',dx_l')\mu(dy_1')\cdots\mu(dy_l').$$

This kernel satisfies the same uniform ergodic property as Q_θ, with the same coefficients, that is condition (3) holds with the coefficients R_θ and ρ_θ with the replacment of Q_θ by $\bar{Q}_\theta$, and we may use Rio (2000)'s exponential inequality (corollary 1) with uniform mixing coeficients (as defined in Rio (2000)) satifying $\phi(k)\leq R_\theta\rho_\theta^{-k}$, to obtain that, for any positive real number u,

$$\mathbb{P}_{\theta_0}\left[\sum_{i=1}^{n/l}\left(\mathbb{1}_{(Y_{li-l+1},\ldots,Y_{li})\in A_j}-\mathbb{P}_{\theta_0}((Y_1,\ldots,Y_l)\in A_j)\right)>u\right]\leq\exp\left\{\frac{-2lu^2(\rho_{\theta_0}-1)^2}{n(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\right\} \tag{16}$$

and

$$\mathbb{P}_{\theta,x}\left[\sum_{i=1}^{n/l}\left(-\mathbb{1}_{(Y_{li-l+1},\ldots,Y_{li})\in A_j}+\mathbb{P}_{\theta,x}((Y_{li-l+1},\ldots,Y_{li})\in A_j)\right)>u\right]\leq\exp\left\{\frac{-2lu^2(\rho_\theta-1)^2}{n(2R_\theta+\rho_\theta-1)^2}\right\}. \tag{17}$$

Set now

$$t_j=\frac{n\|f_{l,\theta_j}-f_{l,\theta_0}\|_1}{4l},\ \delta=\frac{\epsilon_n}{4l}.$$

Since for any θ, $\frac{\rho_\theta-1}{2R_\theta+\rho_\theta-1}\leq 1$ and since consequently for $\theta_j\in A_n$, $\|f_{l,\theta_j}-f_{l,\theta_0}\|_1\geq\epsilon_n$, we first get, using (16),

$$E_{\theta_0}[\phi_j]\leq\exp\left\{\frac{-n\epsilon_n^2(\rho_{\theta_0}-1)^2}{8l(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\right\}. \tag{18}$$

Now, for any $\theta \in A_n$,

$$
\begin{aligned}
&- t_j + \sum_{i=1}^{n/l} \left(\mathbb{P}_{\theta,x}((Y_{li-l+1},\dots,Y_{li}) \in A_j) - \mathbb{P}_{\theta_0}((Y_1,\dots,Y_l) \in A_j)\right) \\
&= -\frac{n\|f_{l,\theta_j} - f_{l,\theta_0}\|_1}{4l} + \frac{n}{l}\left\{\mathbb{P}_{\theta_j}((Y_1,\dots,Y_l) \in A_j) - \mathbb{P}_{\theta_0}((Y_1,\dots;Y_l) \in A_j)\right\} \\
&\quad + \frac{n}{l}\left\{\mathbb{P}_{\theta}((Y_1,\dots,Y_l) \in A_j) - \mathbb{P}_{\theta_j}((Y_1,\dots,Y_l) \in A_j)\right\} \\
&+ \sum_{i=1}^{n/l} \left(\mathbb{P}_{\theta,x}((Y_{li-l+1},\dots,Y_{li}) \in A_j) - \mathbb{P}_{\theta}((Y_1,\dots,Y_l) \in A_j)\right) \\
&\geq \frac{n\|f_{l,\theta_j} - f_{l,\theta_0}\|_1}{4l} - \frac{n\|f_{l,\theta_j} - f_{l,\theta}\|_1}{l} - \sum_{i=1}^{n/l} R_\theta \rho_\theta^{-i} \\
&\geq \frac{n\|f_{l,\theta} - f_{l,\theta_0}\|_1}{4l} - \frac{5n\|f_{l,\theta_j} - f_{l,\theta}\|_1}{4l} - \frac{R_\theta\rho_\theta}{\rho_\theta - 1} \\
&\geq \frac{n}{4l}\left(1 - \frac{5}{4l}\right)\|f_{l,\theta} - f_{l,\theta_0}\|_1 - \frac{R_\theta\rho_\theta}{\rho_\theta - 1} \geq \left(\frac{n}{4l}\left(1 - \frac{5}{4l}\right) - \frac{1}{2\epsilon_n}\right)\|f_{l,\theta} - f_{l,\theta_0}\|_1
\end{aligned}
$$

using the triangular inequality and the fact that $\|f_{l,\theta_j} - f_{l,\theta}\|_1 \leq \frac{\epsilon_n}{4l} \leq \frac{\|f_{l,\theta} - f_{l,\theta_0}\|_1}{4l}$ since $\theta \in A_n$ and $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \leq 1$. As soon as $\frac{n}{4l}\left(1 - \frac{5}{4l}\right) - \frac{1}{2\epsilon_n} > 0$, (17) gives, for $\theta \in A_n$,

$$
E_{\theta,x}\left(1 - \phi_j\right) \leq \exp\left\{-\frac{2l}{n}\left(\frac{n}{4l}\left(1 - \frac{5}{4l}\right) - \frac{1}{2\epsilon_n}\right)^2 \epsilon_n^2\right\}. \tag{19}
$$

We finally get, using (15), (18) and (19)

$$
\begin{aligned}
\mathbb{P}_{\theta_0}\left(\mathbb{P}^\pi\left[\|f_{l,\theta_j} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n | Y_{1:n}\right]\right) \quad &\leq \quad o\left(1\right) + c_2 n^{D/2} e^{C_n} \exp\left\{-\frac{2l}{n}\left(\frac{n}{4l}\left(1 - \frac{5}{4l}\right) - \frac{1}{2\epsilon_n}\right)^2 \epsilon_n^2\right\} \\
&\quad + \left(\frac{n}{\delta}\right)^M \exp\left\{\frac{-n\epsilon_n^2\left(\rho_{\theta_0} - 1\right)^2}{8l\left(2R_{\theta_0} + \rho_{\theta_0} - 1\right)^2}\right\}
\end{aligned}
$$

for some $c_2 > 0$. Taking $\epsilon_n = K\sqrt{\frac{\log n}{n}}$ and C_n tending to $+\infty$ slowly enough, it is easy to see that

$$
\mathbb{P}_{\theta_0}\left(\mathbb{P}^\pi\left[\|f_{l,\theta_j} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n | Y_{1:n}\right]\right) = o\left(1\right)
$$

as soon as K is large enough, and Theorem 1 is proved.

6.2. Proof of Theorem 2

The proof consists in showing that assumptions (A0)-(A3) of Theorem 1 are satisfied. Assumption (F0) and the construction (7) allow to define a $\tilde{\theta}_0 \in \Theta_k$ such that (A0) holds with $D = k(k-1)+kd$. Then using (F1), (F2) and the computations of Section 2.3, (A1) holds. To prove that (A2) and (A3) hold, recall that if $\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1; \gamma_1, ..., \gamma_k)$ is such that $(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1) \in \Delta_0^k$, then μ_θ is uniquely defined. Let us now define

$$
\begin{aligned}
\mathcal{F}_n = \{\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1; \gamma_1, ..., \gamma_k) \ : \ & q_{ij} \geq v_n, 1 \leq i \leq k, 1 \leq j \leq k, \\
& \sum_{j=1}^k \inf_{1 \leq i \leq k} q_{ij} \geq u_n, \ \|\gamma_i\| \leq n^b, 1 \leq i \leq k\Bigg\}.
\end{aligned}
$$

Then, using (F2) together with (8) and Lemma 1 in Appendix 6.3, we obtain that for some constant B,

$$
\forall (\theta_1, \theta_2) \in \mathcal{F}_n^2, \ \|f_{\theta_1} - f_{\theta_2}\|_1 \leq B\left(\frac{1}{v_n^{2c}} + n^a\right)\|\theta_1 - \theta_2\|
$$

so that for some other constant B,

$$N(\delta, \mathcal{F}_n, d(.,.)) \leq \left[\frac{B}{\delta}\left(\frac{1}{v_n^{2c}} + n^a\right)\right]^{k(k-1)+kd}$$

and (A3) holds if v_n is larger than some negative power of n. Now, (F1) gives

$$\pi(\mathcal{F}_n^c) = O(v_n^{\min_{1\leq i\leq k}\alpha_i} + u_n^{\sum_{1\leq i\leq k}\alpha_i}).$$

Let then $v_n = n^{-D/2\min_{1\leq i\leq k}\alpha_i}/\sqrt{\log n}$ and $u_n = n^{-D/2\sum_{1\leq i\leq k}\alpha_i}/\sqrt{\log n}$. Then, (A2) and (A3) hold. Thus, Theorem 1 implies that

$$\begin{aligned} o_{\mathbb{P}_{\theta_0}}(1) &= \mathbb{P}^{\pi_k}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1(\rho_\theta - 1) \geq K\sqrt{\frac{\log n}{n}}\middle| Y_{1:n}\right] \\ &= \mathbb{P}^{\pi_k}\left[\theta \in \mathcal{F}_n \text{ and } \|f_{l,\theta} - f_{l,\theta_0}\|_1(\rho_\theta - 1) \geq K\sqrt{\frac{\log n}{n}}\middle| Y_{1:n}\right] + o_{\mathbb{P}_{\theta_0}}(1) \end{aligned}$$

Since $\rho_\theta - 1 \geq \sum_{j=1}^k \min_{1\leq i\leq k} q_{ij}$, for all $\theta \in \mathcal{F}_n$, $\rho_\theta - 1 \geq u_n$,

$$\mathbb{P}^{\pi_k}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1(\rho_\theta - 1) \geq K\sqrt{\frac{\log n}{n}}|Y_{1:n}\right] \geq \mathbb{P}^{\pi_k}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \geq 2K\frac{1}{u_n}\sqrt{\frac{\log n}{n}}|Y_{1:n}\right],$$

and the theorem follows when (F1) holds. If now (FE1) holds instead of (F1), one gets, taking $u_n = v_n$,

$$\pi(\mathcal{F}_n^c) = O(v_n \exp(-C/v_n)).$$

Then, taking $v_n = 1/h\log n$ with small enough h gives that (A2) and (A3) hold. The end of the proof follows similarly as before.

6.3. Derivatives of the stationary distribution : Lemma 1

LEMMA 1. *The function $\theta \mapsto \mu_\theta$ is continuously differentiable in $(\Delta_k^0)^k \times \Gamma^k$ and there exists an integer $c > 0$ and a constant $C > 0$ such that for any $1 \leq i \leq k$, $1 \leq j \leq k-1$, any $m = 1, \ldots, k$,*

$$\left|\frac{\partial \mu_\theta(m)}{\partial q_{ij}}\right| \leq \frac{C}{(\inf_{i'\neq j'} q_{i'j'})^{2c}}.$$

One may take $c = k-1$.

Let $\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1; \gamma_1, ..., \gamma_k)$ be such that $(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1) \in \Delta_0^k$, $Q_\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k)$ is a $k \times k$ stochastic matrix with positive entries, and μ_θ is uniquely defined by the equation

$$\mu_\theta^T Q_\theta = \mu_\theta^T$$

if μ_θ is the vector $(\mu_\theta(m))_{1\leq m\leq k}$. This equation is solved by linear algebra as

$$\mu_\theta(m) = \frac{P_m(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)},\ m = 1, \ldots, k-1,\ \mu_\theta(k) = 1 - \sum_{m=1}^{k-1}\mu_\theta(m), \quad (20)$$

where P_m, $l = 1, \ldots, k-1$ and R are polynomials where the coefficients are integers (bounded by k) and the monomials are all of degree $k-1$, each variable $q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1$ appearing with power 0 or 1. Now, since the equation has a unique solution as soon as $(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1) \in \Delta_0^k$, then R is never 0 on Δ_0^k, so it may be 0 only at the boundary. Thus, as a fraction of polynomials with non zero denominator, $\theta \mapsto \mu_\theta$ is infinitely differentiable in $(\Delta_k^0)^k \times \Gamma^k$, and the derivative has components all of form

$$\frac{P(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)^2}$$

where again P is a polynomial where the coefficients are integers (bounded by $2k$) and the monomials are all of degree $k-1$, each variable $q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1$ appearing with power 0 or 1. Thus, since all q_{ij}'s are bounded by 1 there exists a constant C such that for all $m = 1, \ldots, k$, $i = 1, \ldots, k$, $j = 1, \ldots, k-1$,

$$\left| \frac{\partial \mu_\theta(m)}{\partial q_{ij}} \right| \leq \frac{C}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)^2}. \tag{21}$$

We shall now prove that

$$R\left(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1\right) \geq \left(\inf_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} q_{ij}\right)^{k-1}, \tag{22}$$

which combined with (21) and (22) implies Lemma 1. Note that we can express R as a polynomial function of $Q = q_{ij}$, $1 \leq i \leq k, 1 \leq j \leq k, i \neq j$. Indeed, $\mu := (\mu_\theta(i))_{1 \leq i \leq k-1}$ is solution of

$$\mu^T \cdot M = V^T$$

where V is the $(k-1)$-dimensional vector $(q_{kj})_{1 \leq j \leq k-1}$, and M is the $(k-1) \times (k-1)$-matrix with components $M_{i,j} = q_{kj} - q_{ij} + \mathbb{1}_{i=j}$. Since R is the determinant of M, this leads to, for any $k \geq 2$:

$$R = \sum_{\sigma \in \mathcal{S}_{k-1}} \varepsilon(\sigma) \prod_{1 \leq i \leq k-1, \sigma(i) = i} \left(q_{ki} + \sum_{1 \leq j \leq k-1, j \neq i} q_{ij} \right) \prod_{1 \leq i \leq k-1, \sigma(i) \neq i} \left(q_{ki} - q_{\sigma(i)i} \right) \tag{23}$$

where for any integer n, $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$, and for each permutation σ, $\varepsilon(\sigma)$ is its signature. Thus R is a polynomial in the components of Q where each monomial has integer coefficient and has $k-1$ different factors. The possible monomials are of form

$$\beta \prod_{i \in A} q_{ki} \prod_{i \in B} q_{ij(i)}$$

where (A, B) is a partition of $\{1, \ldots, k-1\}$, and for all $i \in B$, $j(i) \in \{1, \ldots, k-1\}$ and $j(i) \neq i$. In case $B = \emptyset$, the coefficient β of the monomial is $\sum_{\sigma \in \mathcal{S}_{k-1}} \varepsilon(\sigma) = 0$, so that we only consider partitions such that $B \neq \emptyset$. Fix such a monomial with non nul coefficient, let (A, B) be the associated partition. Let Q be such that, for all $i \in A$, $q_{ki} > 0$, for all $i \notin A$, $q_{ki} = 0$ and $q_{kk} > 0$ (used to handle the case $A = \emptyset$). Fix also $q_{ij(i)} = 1$ for all $i \in B$. Then, if (A', B') is another partition of $\{1, \ldots, k-1\}$ with $B' \neq \emptyset$, the monomial $\prod_{i \in A'} q_{ki} \prod_{i \in B'} q_{ij(i)} = 0$. Thus, $R(Q)$ equals $\prod_{i \in A} q_{ki} \prod_{i \in B} q_{ij(i)}$ times the coefficient of the monomial. But $R(Q) \geq 0$, so that this coefficient is a positive integer and (22) follows.

6.4. Proof of Corollary 2

We first prove that (14) holds. Following (9) we get that there exists $c(\theta_0) > 0$ and $\eta > 0$ such that:

- If $\|\gamma_1 - \gamma^0\| \leq \eta$ and $\|\gamma_2 - \gamma^0\| \leq \eta$,

$$\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0) \frac{1}{p+q} \left[\left\| q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \right\| + q \left\| \gamma_1 - \gamma^0 \right\|^2 + p \left\| \gamma_2 - \gamma^0 \right\|^2 \right],$$

- If $\|\gamma_1 - \gamma^0\| \leq \eta$ and $\|\gamma_1 - \gamma^0\| + \|\gamma_2 - \gamma^0\| > 2\eta$,

$$\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0) \left[\frac{p}{p+q} + \frac{q}{p+q} \left\| \gamma_1 - \gamma^0 \right\| \right],$$

- If $\|\gamma_2 - \gamma^0\| \leq \eta$ and $\|\gamma_1 - \gamma^0\| + \|\gamma_2 - \gamma^0\| > 2\eta$,

$$\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0) \left[\frac{q}{p+q} + \frac{p}{p+q} \left\| \gamma_2 - \gamma^0 \right\| \right],$$

- If $\|\gamma_1 - \gamma^0\| > \eta$ and $\|\gamma_2 - \gamma^0\| > \eta$,

$$\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0).$$

Define

$$B_n^1 = \left\{ \frac{(p+q) \wedge (2-(p+q))}{p+q} \left[\left\| q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \right\| + q \left\| \gamma_1 - \gamma^0 \right\|^2 + p \left\| \gamma_2 - \gamma^0 \right\|^2 \right] \leq u_n \right\},$$

$$B_n^2 = \left\{ \| \frac{(p+q) \wedge (2-(p+q))}{p+q} \left[p + q \left\| \gamma_1 - \gamma^0 \right\| \right] \| \leq u_n \right\},$$

$$B_n^3 = \left\{ \frac{(p+q) \wedge (2-(p+q))}{p+q} \left[q + p \left\| \gamma_2 - \gamma^0 \right\| \right] \leq u_n \right\}$$

and

$$B_n^4 = \left\{ (p+q) \wedge (2-(p+q)) \leq u_n \right\}.$$

Then,

$$\pi_2(A_n \cap B_n) \leq \pi_2(A_n \cap B_n^1) + \pi_2(A_n \cap B_n^2) + \pi_2(A_n \cap B_n^3) + \pi_2(A_n \cap B_n^4).$$

Notice that on A_n, if $p+q \geq 1$, then $p \leq \epsilon_n$ and $q \geq 1 - \epsilon_n$, or $q \leq \epsilon_n$ and $p \geq 1 - \epsilon_n$, so that also $2 - (p+q) \geq 1 - \epsilon_n$.

- On $A_n \cap B_n^1$, $\left\| q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \right\| \lesssim u_n$, $q \left\| \gamma_1 - \gamma^0 \right\|^2 \lesssim u_n$, $p \left\| \gamma_2 - \gamma^0 \right\|^2 \lesssim u_n$, and $p \lesssim \epsilon_n$ or $q \lesssim \epsilon_n$. This gives $\pi_2(A_n \cap B_n^1) \lesssim u_n^{d+d/2} \epsilon_n^{\alpha - d/2}$.
- On $A_n \cap B_n^2$, $p \lesssim u_n$ and $q \left\| \gamma_1 - \gamma^0 \right\| \lesssim u_n$ in case $p+q \leq 1$, and $p \lesssim u_n$, $1-q \lesssim u_n$ and $q \left\| \gamma_1 - \gamma^0 \right\| \lesssim u_n$ in case $p+q \geq 1$, leading to $\pi_2(A_n \cap B_n^2) \lesssim u_n^{\alpha+d} + u_n^{\alpha+\beta+d}$.
- For symmetry reasons, $\pi_2(A_n \cap B_n^3) = \pi_2(A_n \cap B_n^2)$.
- On $A_n \cap B_n^4$, $p \lesssim u_n$ and $q \lesssim u_n$, so that $\pi_2(A_n \cap B_n^4) \lesssim u_n^{2\alpha}$.

Keeping only the leading terms, we see that (14) holds.

We now prove that (11) holds with $D = d + d/2$ and c_n tending to infinity, which will finish the proof of corollary 2. Let us introduce the set, for small but fixed ϵ:

$$U_n = \Big\{ \theta = (p, q, \gamma_1, \gamma_2) \ : \ \|\gamma_1 - \gamma^0\|^2 \leq \frac{1}{\sqrt{n}}, \ \|\gamma_2 - \gamma^0\|^2 \leq \frac{1}{\sqrt{n}}, \ \|q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)\| \leq \frac{1}{\sqrt{n}}, \\ |q - \frac{1}{2}| \leq \epsilon, \ |p - \frac{1}{2}| \leq \epsilon \Big\}$$

so that $U_n \subset B_n$, and $\pi_2(U_n \gtrsim n^{-3d/4}$. Thus

$$D_n \geq \int_{U_n \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \pi(d\theta) \pi_{\mathcal{X}}(dx).$$

Let us now study $\ell_n(\theta, x) - \ell_n(\theta_0, x)$. First, following the proof of Lemma 2 of Douc et al. (2004) we find that, for any $\theta \in U_n$, for any x,

$$|\ell_n(\theta) - \ell_n(\theta, x)| \leq \left(\frac{1+2\epsilon}{1-2\epsilon} \right)^2.$$

Thus, for any $\theta \in U_n$ and any x, and since $\ell_n(\theta_0, x)$ does not depend on x,

$$\ell_n(\theta, x) - \ell_n(\theta_0, x) \geq \ell_n(\theta) - \ell_n(\theta_0) - \left(\frac{1+2\epsilon}{1-2\epsilon} \right)^2. \tag{24}$$

Let us now study $\ell_n(\theta) - \ell_n(\theta_0)$.

$$\ell_n(\theta) - \ell_n(\theta_0) = \sum_{k=1}^{n} \log \left[\mathbb{P}_\theta \left(X_k = 1 | Y_{1:k-1} \right) \frac{g_{\gamma_1}}{g_{\gamma^0}} \left(Y_k \right) + \mathbb{P}_\theta \left(X_k = 2 | Y_{1:k-1} \right) \frac{g_{\gamma_2}}{g_{\gamma^0}} \left(Y_k \right) \right]$$

and we set for $k = 1$

$$\mathbb{P}_\theta \left(X_k = 1 | Y_{1:k-1} \right) = \mathbb{P}_\theta \left(X_1 = 1 \right) = \frac{q}{p+q},$$

$$\mathbb{P}_\theta \left(X_k = 2 | Y_{1:k-1} \right) = \mathbb{P}_\theta \left(X_1 = 2 \right) = \frac{p}{p+q}.$$

Denote $p_k(\theta)$ the random variable $\mathbb{P}_\theta \left(X_k = 1 | Y_{1:k-1} \right)$, which is a function of $Y_{1:k-1}$ and thus independent of Y_k. We have the recursion

$$p_{k+1}\left(\theta\right) = \frac{(1-p)p_k(\theta)g_{\gamma_1}(Y_k) + q(1-p_k(\theta))g_{\gamma_2}(Y_k)}{p_k(\theta)g_{\gamma_1}(Y_k) + (1-p_k(\theta))g_{\gamma_2}(Y_k)}. \tag{25}$$

Note that, for any p, q in $]0,1[$, for any $k \geq 1$,

$$p_k(p, q, \gamma^0, \gamma^0) = \frac{q}{p+q}.$$

We shall denote by $D^i_{(\gamma_1)^j,(\gamma_2)^{i-j}}$ the i-th partial derivative operator j times with respect to γ_1 and $i-j$ times with respect to γ_2 ($0 \leq j \leq i$, the order in which derivatives are taken does not matter). Fix $\theta = (p, q, \gamma_1, \gamma_2) \in U_n$. When derivatives are taken at point $(p, q, \gamma^0, \gamma^0)$, they are written with 0 as superscript.
Using Taylor expansion till order 4, there exists $t \in [0,1]$ such that denoting $\theta_t = t\theta + (1-t)(p, q, \gamma^0, \gamma^0)$:

$$\ell_n(\theta) - \ell_n(\theta_0) = (\gamma_1 - \gamma^0)D^1_{\gamma_1}\ell^0_n + (\gamma_2 - \gamma^0)D^1_{\gamma_2}\ell^0_n + S_n(\theta) + T_n(\theta) + R_n(\theta, t) \tag{26}$$

where $S_n(\theta)$ denotes the term of order 2, $T_n(\theta)$ denotes the term of order 3, and $R_n(\theta, t)$ the remainder, that is

$$S_n(\theta) = (\gamma_1 - \gamma^0)^2 D^2_{(\gamma_1)^2}\ell^0_n + 2(\gamma_1 - \gamma^0)(\gamma_2 - \gamma^0)D^2_{\gamma_1,\gamma_2}\ell^0_n + (\gamma_2 - \gamma^0)^2 D^2_{(\gamma_2)^2}\ell^0_n,$$

$$\begin{aligned} T_n(\theta) = (\gamma_1 - \gamma^0)^3 D^3_{(\gamma_1)^3}\ell^0_n + 3(\gamma_1 - \gamma^0)^2(\gamma_2 - \gamma^0)D^3_{(\gamma_1)^2,\gamma_2}\ell^0_n \\ + 3(\gamma_1 - \gamma^0)(\gamma_2 - \gamma^0)^2 D^3_{\gamma_1,(\gamma_2)^2}\ell^0_n + (\gamma_2 - \gamma^0)^3 D^3_{(\gamma_2)^3}\ell^0_n \end{aligned}$$

and

$$R_n(\theta, t) = \sum_{k=0}^{4} \left(\begin{array}{c} k \\ 4 \end{array} \right) (\gamma_1 - \gamma^0)^k (\gamma_2 - \gamma^0)^{4-k} D^4_{(\gamma_1)^k,(\gamma_2)^{4-k}} \ell_n(\theta_t).$$

Easy but tedious computations lead to the following results.

$$\begin{aligned} &(\gamma_1 - \gamma^0)D^1_{\gamma_1}\ell^0_n + (\gamma_2 - \gamma^0)D^1_{\gamma_2}\ell^0_n \\ &\qquad = \left[\sum_{k=1}^{n} \frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}} \left(Y_k \right) \right] \left[\frac{q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)}{p+q} \right] \\ &\qquad = \left[\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}} \left(Y_k \right) \right] \left[\sqrt{n} \frac{q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)}{p+q} \right] \end{aligned}$$

so that

$$\sup_{\theta \in U_n} \left| (\gamma_1 - \gamma^0)D^1_{\gamma_1}\ell^0_n + (\gamma_2 - \gamma^0)D^1_{\gamma_2}\ell^0_n \right| = O_{\mathbb{P}_{\theta_0}}\left(1\right). \tag{27}$$

Also,

$$S_n(\theta) = -\left[\frac{1}{n}\sum_{k=1}^{n}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2\right]\left[\sqrt{n}\frac{q(\gamma_1-\gamma^0)+p(\gamma_2-\gamma^0)}{p+q}\right]^2$$
$$+\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right]\left[\frac{q}{p+q}\left(n^{1/4}(\gamma_1-\gamma^0)\right)^2+\frac{p}{p+q}\left(n^{1/4}(\gamma_2-\gamma^0)\right)^2\right]$$
$$+2\left(n^{1/4}(\gamma_1-\gamma^0)\right)^2\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(D^1_{\gamma_1}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right]-2\left(n^{1/4}(\gamma_2-\gamma^0)\right)^2\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(D^1_{\gamma_2}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right]$$
$$+2\left(n^{1/4}(\gamma_1-\gamma^0)(\gamma_2-\gamma^0)\right)\left[\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(D^1_{\gamma_2}p^0_k-D^1_{\gamma_1}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right].$$

Using (25) one gets that for all integer $k\geq 2$, ($D^1_{\gamma_1}p^0_1=0$ and $D^1_{\gamma_2}p^0_1=0$):

$$D^1_{\gamma_1}p^0_k = \frac{pq}{(p+q)^2}\sum_{l=1}^{k-1}(1-p-q)^{k-l}\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_l)$$

and

$$D^1_{\gamma_2}p^0_k = -D^1_{\gamma_1}p^0_k$$

which leads to

$$E_{\theta_0}\left[\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(D^1_{\gamma_1}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2\right]\leq\left(E_{\theta_0}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^2\right)^2.$$

and

$$E_{\theta_0}\left[\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(D^1_{\gamma_2}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2\right]\leq\left(E_{\theta_0}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^2\right)^2.$$

Thus, we obtain

$$\sup_{\theta\in U_n}|S_n(\theta)| = O_{\mathbb{P}_{\theta_0}}(1). \tag{28}$$

For the order 3 term, as soon as $\theta\in U_n$:

$$T_n(\theta) = -\left[\sum_{k=1}^{n}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^3\right]\left[\frac{q(\gamma_1-\gamma^0)+p(\gamma_2-\gamma^0)}{p+q}\right]^3$$
$$+\left[\sum_{k=1}^{n}\frac{D^3_{\gamma^3} g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right]\left[\frac{q}{p+q}\left(\gamma_1-\gamma^0\right)^3+\frac{p}{p+q}\left(\gamma_2-\gamma^0\right)^3\right]$$
$$-3\left[\sum_{k=1}^{n}\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right]\left[\frac{q(\gamma_1-\gamma^0)+p(\gamma_2-\gamma^0)}{p+q}\right]\left[\frac{q}{(p+q)^2}\left(\gamma_1-\gamma^0\right)^2+\frac{p}{(p+q)^2}\left(\gamma_2-\gamma^0\right)^2\right]$$
$$+O\left(n^{-3/4}\right)\left\{\sum_{k=1}^{n}(D^1_{\gamma_1}p^0_k)\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2+\sum_{k=1}^{n}(D^1_{\gamma_1}p^0_k)\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right.$$
$$\sum_{k=1}^{n}(D^2_{(\gamma_1)^2}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)+\sum_{k=1}^{n}(D^2_{(\gamma_2)^2}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)$$
$$\left.+\sum_{k=1}^{n}(D^2_{(\gamma_1,\gamma_2)}p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right\}$$

so that using assumptions (13)

$$\sup_{\theta\in U_n}|T_n(\theta)| = O_{\mathbb{P}_{\theta_0}}\left(n^{-1/4}\right) + O_{\mathbb{P}_{\theta_0}}(1) + O\left(n^{-1/4}\right) Z_n$$

with

$$Z_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\left\{\left[\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2 + \frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right] D^1_{\gamma_1} p^0_k + \frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\left[D^2_{(\gamma_1)^2} p^0_k + D^2_{(\gamma_2)^2} p^0_k + D^2_{(\gamma_1,\gamma_2)} p^0_k\right]\right\}.$$

Now using (25) one gets that for all integer $k \geq 1$,

$$\frac{1}{1-p-q} D^2_{(\gamma_1)^2} p^0_{k+1} = -2\frac{pq^2}{(p+q)^3}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2 + 2(D^1_{\gamma_1} p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k) + \frac{pq}{(p+q)^2}\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k) + D^2_{(\gamma_1)^2} p^0_k,$$

$$\frac{1}{1-p-q} D^2_{(\gamma_2)^2} p^0_{k+1} = 2\frac{p^2 q}{(p+q)^3}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2 - 2(D^1_{\gamma_1} p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k) - \frac{pq}{(p+q)^2}\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_k) + D^2_{(\gamma_2)^2} p^0_k,$$

$$\frac{1}{1-p-q} D^2_{(\gamma_1,\gamma_2)} p^0_{k+1} = 2\frac{pq(q-p)}{(p+q)^3}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k)\right)^2 + 2(D^1_{\gamma_1} p^0_k)\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_k) + D^2_{(\gamma_1,\gamma_2)} p^0_k,$$

and using $D^2_{(\gamma_1)^2} p^0_1 = 0, D^2_{(\gamma_2)^2} p^0_1 = 0, D^2_{(\gamma_1,\gamma_2)} p^0_1 = 0$ and easy but tedious computations one gets that for some finite $C > 0$,

$$E_{\theta_0}\left(Z_n^2\right) \leq C E_{\theta_0}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^2 \left[E_{\theta_0}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^4 + E_{\theta_0}\left(\frac{D^2_{\gamma^2} g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^2 + \left(E_{\theta_0}\left(\frac{D^1_\gamma g_{\gamma_0}}{g_{\gamma_0}}(Y_1)\right)^2\right)^2\right]$$

so that we finally obtain

$$\sup_{\theta\in U_n}|T_n(\theta)| = O_{\mathbb{P}_{\theta_0}}(1). \tag{29}$$

Let us finally study the fourth order remainder $R_n(\theta, t)$. We have

$$\sup_{\theta\in U_n}|R_n(\theta,t)| \leq \frac{1}{n}\sum_{k=1}^{n} A_{k,n} B_{k,n},$$

where, for big enough n, $A_{k,n}$ is a polynomial of degree at most 4 in $\sup_{\gamma'\in B_d(\gamma^0,\epsilon)} \| \frac{D^i_\gamma g_{\gamma'}}{g_{\gamma'}}(Y_k) \|$, and $B_{k,n}$ is a sum of terms of form

$$\sup_{\theta\in U_n}\left|\prod_{i=1}^{4}\prod_{j=0}^{i}\left(D^i_{(\gamma_1)^j,(\gamma_2)^{i-j}} p_k(\theta_t)\right)^{a_{i,j}}\right| \tag{30}$$

where the $a_{i,j}$ are non negative integers such that $\sum_{i=1}^{4}\sum_{j=0} i a_{i,j} \leq 4$.
To prove that

$$\sup_{\theta\in U_n}|R_n(\theta,t)| = O_{\mathbb{P}_{\theta_0}}(1) \tag{31}$$

holds, it is enough to prove that $E_{\theta_0}|\sum_{k=1}^{n} A_{k,n}B_{k,n}| = O(n)$. But for each k, $p_k(\theta)$ and its derivatives depend on $Y_1, \ldots, Y_{k-1}$ only, so that $A_{k,n}$ and $B_{k,n}$ are independent random variables, and

$$\begin{aligned} E_{\theta_0}\left|\sum_{k=1}^{n} A_{k,n}B_{k,n}\right| &\leq \sum_{k=1}^{n} E_{\theta_0}|A_{k,n}|\, E_{\theta_0}|B_{k,n}| \\ &\leq C \max_{i=1,2,3,4} E_{\theta_0}\left(\sup_{\gamma'\in B_d(\gamma^0,\epsilon)} \| \frac{D^i_\gamma g_{\gamma'}}{g_{\gamma'}}(Y_1) \|^4\right)\sum_{k=1}^{n} E_{\theta_0}|B_{k,n}| \end{aligned}$$

for some finite $C > 0$. Now, using (25) one gets that for all integer $k \geq 1$ and for any θ,

$$D^1_{\gamma_1} p_{k+1}(\theta) = (1-p-q)\left\{ \frac{p_k(\theta)(1-p_k(\theta))g_{\gamma_2}(Y_k)D^1_{\gamma} g_{\gamma_1}(Y_k) + g_{\gamma_1}(Y_k)g_{\gamma_2}(Y_k)D^1_{\gamma_1} p_k(\theta)}{\left(p_k(\theta)g_{\gamma_1}(Y_k) + (1-p_k(\theta))g_{\gamma_2}(Y_k)\right)^2} \right\},$$

$$D^1_{\gamma_2} p_{k+1}(\theta) = (1-p-q)\left\{ \frac{-p_k(\theta)(1-p_k(\theta))g_{\gamma_1}(Y_k)D^1_{\gamma} g_{\gamma_2}(Y_k) + g_{\gamma_1}(Y_k)g_{\gamma_2}(Y_k)D^1_{\gamma_2} p_k(\theta)}{\left(p_k(\theta)g_{\gamma_1}(Y_k) + (1-p_k(\theta))g_{\gamma_2}(Y_k)\right)^2} \right\}.$$

Notice that for any θ, any $k \geq 2$, $p_k(\theta) \in (1-p, q)$ so that for any $\theta \in U_n$, any $k \geq 2$, $p_k(\theta) \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon]$. We obtain easily that for $i = 1, 2$, $k \geq 2$,

$$\sup_{\theta \in U_n} \left| D^1_{\gamma_i} p_{k+1}(\theta) \right| \leq \left(\frac{2\epsilon}{1-8\epsilon} \right) \left\{ \sup_{\gamma' \in B_d(\gamma^0, \epsilon)} \| \frac{D^1_{\gamma} g_{\gamma'}}{g_{\gamma'}}(Y_k) \| + \sup_{\theta \in U_n} \left| D^1_{\gamma_i} p_k(\theta) \right| \right\}.$$

Using similar tricks, it is possible to get that there exists a finite constant $C > 0$ such that for any $i = 1, 2, 3, 4$, any $j = 0, \ldots, i$, any $k \geq 2$,

$$\sup_{\theta \in U_n} \left| D^i_{(\gamma_1)^j, (\gamma_2)^{i-j}} p_{k+1}(\theta) \right| \leq C\epsilon \left\{ \sup_{\gamma' \in B_d(\gamma^0, \epsilon)} \| \sum_{l=1}^{i} \frac{D^l_{\gamma^l} g_{\gamma'}}{g_{\gamma'}}(Y_k) \|^{i+1-l} + \sum_{l=1}^{i} \sum_{m=0}^{l} \sup_{\theta \in U_n} \left| D^l_{(\gamma_1)^j, (\gamma_2)^{l-j}} p_k(\theta) \right|^{i+1-l} \right\}.$$

By recursion, we obtain that there exists a finite $C > 0$ such that any term of form (30) has expectation uniformly bounded:

$$E_{\theta_0} \left[\sup_{\theta \in U_n} \left| \prod_{i=1}^{4} \prod_{j=0}^{i} \left(D^i_{(\gamma_1)^j, (\gamma_2)^{i-j}} p_k(\theta_t) \right)^{a_{i,j}} \right| \right] \leq C \max_{m=1,2,3,4} \max_{r=1,2,3,4} E_{\theta_0} \left(\sup_{\gamma' \in B_d(\gamma^0, \epsilon)} \| \frac{D^m_{\gamma} g_{\gamma'}}{g_{\gamma'}}(Y_1) \|^r \right)$$

which concludes the proof of (31).
Now, using (24), (26), (27), (28), (29) and (31), we get

$$D_n \geq e^{-O_{\mathbb{P}_{\theta_0}}(1)} \pi_2(U_n)$$

so that (11) holds with $D = d + d/2$ and any c_n tending to infinity.

6.5. General identifiability condition for the parameters of a finite state space HMM

In this Section we prove that (9) holds under some general assumption. Let us first introduce some notations. For all $\ell \leq n$, For all $I = (i_1, \ldots, i_\ell) \in \{1, \cdots, k\}^\ell$, define $\gamma_I = (\gamma_{i_1}, \ldots, \gamma_{i_\ell})$, $G_{\gamma_I} = \prod_{t=1}^{\ell} g_{\gamma_{i_t}}(y_t)$, $D^1 G_{\gamma_I}$ the vector of first derivatives of G_{γ_I} with respect to each of the distinct elements in γ_I, note that it has dimension $d \times |I|$, where $|I|$ denotes the number of distinct indices in I, and similarly define $D^2 G_{\gamma_I}$ the symetric matrix in $R^{d|I| \times d|I|}$ made of second derivatives of G_{γ_I} with respect to the distinct elements (indices) in γ_I. If b is a vector, b^T denotes the transpose vector.
Let $T = \{\mathbf{t} = (t_1, \ldots, t_{k_0}) \in \{1, \ldots, k\}^{k_0} : t_i < t_{i+1}, i = 0, \ldots, k_0 - 1\}$. For any $\mathbf{t} = (t_1, \ldots, t_{k_0}) \in T$, define for all $i \in \{1, \ldots, k_0\}$ the set $J(i) = \{t_{i-1} + 1, \ldots, t_i\}$, using $t_0 = 0$.

We then consider the following condition :

- **Condition** $(L(\ell))$ For any $\mathbf{t} = (t_1, \ldots, t_{k_0}) \in T$, for all collections $(\pi_I)_I$, $(\gamma_I)_I$, $I \notin \{1, \ldots, t_{k_0}\}^\ell$ satisfying $\pi_I \geq 0$, $\gamma_I = (\gamma_{i_1}, \ldots, \gamma_{i_\ell})$ such that $\gamma_{i_j} = \gamma^0_i$ when $i_j \in J(i)$ for some $i \leq k_0$ and $\gamma_{i_j} \in \Gamma \setminus \{\gamma^0_i, i = 1, \ldots, k_0\}$ when $i_j \notin \{1, \ldots, t_{k_0}\}$, for all collections $(a_I)_I, (c_I)_I, (b_I)_I$, $I \in \{1, \ldots, k_0\}^\ell$, $a_I \in \mathbb{R}$, $c_I \geq 0$ and $b_I \in \mathbb{R}^{d|I|}$, for all collection of vectors $z_{I,J} \in \mathbb{R}^{d|I|}$ with $I \in \{1, \ldots, k_0\}^\ell$ and $J \in J(i_1) \times \cdots \times J(i_\ell)$ satisfying $\|z_{I,J}\| = 1$, and all

sequences $(\alpha_{I,J})$, satisfying $\alpha_{I,J} \geq 0$ and $\sum_{J \in J(i_1) \times \cdots \times J(i_\ell)} \alpha_{I,J} = 1$,

$$\begin{aligned} &\sum_{I \notin \{1,\ldots,t_{k_0}\}^\ell} \pi_I G_{\gamma_I} + \sum_{I \in \{1,\ldots,k_0\}^\ell} \left(a_I G_{\gamma_I^0} + b_I^T D^1 G_{\gamma_I^0} \right) \\ &+ \sum_{I \in \{1,\ldots,k_0\}^\ell} c_I \sum_{J \in J(i_1) \times \cdots \times J(i_\ell)} \alpha_{I,J} z_{I,J}^T D^2 G_{\gamma_I^0} z_{I,J} = 0 \\ &\Leftrightarrow \\ &\sum_{I \notin \{1,\ldots,t_{k_0}\}^\ell} \pi_I + \sum_{I \in \{1,\ldots,k_0\}^\ell} (|a_I| + \|b_I\| + c_I) = 0 \end{aligned} \tag{32}$$

This condition is a multivariate version of the condition that would be required if only $\ell = 1$ was considered. In this case the condition can be written as :

- **Condition (L(1))** For any $\mathbf{t} = (t_1, \ldots, t_{k_0}) \in T$, consider $(\pi_i)_{i=1}^{k-t_{k_0}}$ (if $t_{k_0} < k$) a set of nonnegative reals, $(a_i)_{i=1}^{k_0}$ and $(b_i)_{i=1}^{k_0}$ with $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}^d$ and $(c_i)_{i=1}^{k_0}$ and $z_{i,j}, \alpha_{i,j}, i = 1, \ldots, k_0, j = 1, \ldots, t_i - t_{i-1}$, with $t_0 = 0$ and $z_{i,j} \in \mathbb{R}^d$ satisfying $\|z_{i,j}\| = 1$ and $\alpha_{i,j} \geq 0$ and $\sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} = 1$, for any $(\gamma_i)_{i=1}^{k-t_{k_0}}$ which belongs to $\Gamma \setminus \{\gamma_i^0, i = 1, \ldots, k_0\}$,

$$\sum_{i=1}^{k-t_{k_0}} \pi_i g_{\gamma_i} + \sum_{i=1}^{k_0} \left(a_i g_{\gamma_i^0} + b_i^T D^1 g_{\gamma_i^0} \right) + \sum_{i=1}^{k_0} c_i^2 \sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} z_{i,j}^T D^2 g_{\gamma_i^0} z_{i,j} = 0, \tag{33}$$

if and only if

$$a_i = 0, b_i = 0, z_{i,j} = 0 \quad \forall i = 1, \ldots, k_0, \quad \forall j = 1, \ldots, t_i - t_{i-1}, \quad \pi_i = 0 \quad \forall i = 1, \ldots, k - t_{k_0}$$

Note that the partition represents the clustering structure of the extra components, up to a permutation of the labels.

Condition (L(1)) is the same condition as in Rousseau and Mengersen (2011), so that it is satisfied in particular for Poisson mixtures, location-scale Gaussian mixtures and any mixtures of regular exponential families.

LEMMA 2. *Assume that the function $\gamma \mapsto g_\gamma(y)$ is twice continuously differentiable in Γ and that for all y, $g_\gamma(y)$ vanishes as $\|\gamma\|$ tends to infinity. Then, if condition $(L(\ell))$ is verified,* (9) *holds. Moreover, condition $(L(\ell))$ $(\ell \geq 1)$ is verified as soon as condition (L(1)) is verified.*

To prove the first part of the Lemma we follow the ideas of the beginning of the proof of Theorem 5.11 in Gassiat and van Handel (2012). If (9) does not hold, there exist a sequence of l-marginals $(f_{l,\theta^n})_{n \geq 1}$ with parameters $(\theta^n)_{n \geq 1}$ such that for some positive sequence ε_n tending to 0, $\|f_{\ell,\theta^n} - f_{\ell,\theta_0}\|_1 / N_n(\theta^n)$ tends to 0 as n tends to infinity, with

$$\begin{aligned} N_n(\theta) = &\sum_{1 \leq j \leq \ell : \forall i, \|\gamma_j - \gamma_i^0\| > \varepsilon_n} \mathbb{P}_\theta(X_1 = j) \\ &+ \sum_{1 \leq i_1, \ldots, i_\ell \leq k_0} \Bigg[\Big| \sum_{(j_1,\ldots,j_\ell) \in A_n(i_1,\ldots,i_\ell)} \mathbb{P}_\theta(X_{1:l} = j_1 \cdots j_\ell) - \mathbb{P}_{\theta_0}(X_{1:l} = i_1 \cdots i_\ell) \Big| \\ &+ \Bigg\| \sum_{(j_1,\ldots,j_\ell) \in A_n(i_1,\ldots,i_\ell)} \mathbb{P}_\theta(X_{1:\ell} = j_1 \ldots j_\ell) \left\{ \begin{pmatrix} \gamma_{j_1} \\ \cdots \\ \gamma_{j_\ell} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \cdots \\ \gamma_{i_\ell}^0 \end{pmatrix} \right\} \Bigg\| \\ &+ \frac{1}{2} \sum_{(j_1,\ldots,j_\ell) \in A_n(i_1,\ldots,i_\ell)} \mathbb{P}_\theta(X_{1:\ell} = j_1 \cdots j_l) \left\| \begin{pmatrix} \gamma_{j_1} \\ \cdots \\ \gamma_{j_\ell} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \cdots \\ \gamma_{i_\ell}^0 \end{pmatrix} \right\|^2 \Bigg] \end{aligned}$$

with $A_n(i_1,\dots,i_\ell)=\{(j_1,\dots,j_\ell):\|\gamma_{j_1}-\gamma^0_{i_1}\|\le\varepsilon_n,\dots,\|\gamma_{j_\ell}-\gamma^0_{i_\ell}\|\le\varepsilon_n\}$.
Now, $f_{\ell,\theta^n}=\sum_{I\in\{1,\dots,k\}^\ell}\mathbb{P}_{\theta^n}\left((X_1,\dots,X_\ell=I\right)G_{\gamma^n_I}$ where $\theta^n=(Q^n,(\gamma^n_1,\dots,\gamma^n_k))$, Q^n a transition matrix on $\{1,\dots,k\}$. It is possible to extract a subsequence along which, for all $i=1,\dots,k$, either γ^n_i converges to some limit γ_i or $\|\gamma^n_i\|$ tends to infinity. Choose now the indexation such that for $i=1,\dots,t_1$, γ^n_i converges to γ^0_1, for $i=t_1+1,\dots,t_2$, γ^n_i converges to γ^0_2, and so on, for $i=t_{k_0-1}+1,\dots,t_{k_0}$, γ^n_i converges to $\gamma^0_{k_0}$, and if $t_{k_0}<k$, for some $\tilde{k}\le k$, for $i=t_{k_0}+1,\dots,\tilde{k}$, γ^n_i converges to some $\gamma_i\notin\{\gamma^0_1,\dots,\gamma^0_{k_0}\}$, and for $i=\tilde{k}+1,\dots,k$, $\|\gamma^n_i\|$ tends to infinity. It is possible that $\tilde{k}=t_{k_0}$ in wich case no γ^n_i converges to some $\gamma_i\notin\{\gamma^0_1,\dots,\gamma^0_{k_0}\}$. Such a $\mathbf{t}=(t_1,\dots,t_{k_0})\in T$ exists, because if $\|f_{\ell,\theta^n}-f_{\ell,\theta_0}\|_1/N_n(\theta^n)$ tends to 0 as n tends to infinity, $\|f_{\ell,\theta^n}-f_{\ell,\theta_0}\|_1$, and $N_n(\theta^n)$ tends to 0 as n tends to infinity (if it was not the case, using the regularity of $\theta\mapsto f_{\ell,\theta}$ we would have a contradiction). Now along the subsequence we may write, for large enough n:

$$N_n(\theta^n)=\sum_{I\notin\{1,\dots,t_{k_0}\}^\ell}\mathbb{P}_\theta\left(X_{1:\ell}=I\right)+\sum_{I\in\{1,\dots,k_0\}^\ell}\left[\left|\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)-\mathbb{P}_{\theta_0}\left(X_{1:\ell}=I\right)\right|\right.$$
$$\left.+\left\|\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)\gamma_J-\gamma^0_I\right\|+\frac{1}{2}\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)\left\|\gamma_J-\gamma^0_I\right\|^2\right].$$

We shall use Taylor expansion till order 2. To be perfectly rigourous in the following, we need write I in terms of its distinct indices, $(\tilde{i}_1,\cdots,\tilde{i}_{|I|})$, and $G_{\gamma_I}=\prod_{t=1}^{|I|}\prod_{j:i_j=\tilde{i}_t}g_{\gamma_{\tilde{i}_t}}(y_j)$, however we shall not make such a distinction, so that unless otherwise stated, in such a case $(\gamma^n_J-\gamma^0_I)^TD^1G_{\gamma^0_I}$ can denote

$$\sum_{t=1}^{|I|}(\gamma_{\tilde{i}_t}-\gamma^0_{\tilde{i}_t})^T\frac{\partial G_{\gamma_I}}{\partial\gamma_{\tilde{i}_t}},$$

and similarly for the second derivatives. We have

$$f_{\ell,\theta^n}-f_{\ell,\theta_0}=\sum_{I\notin\{1,\dots,t_{k_0}\}^\ell}\mathbb{P}_\theta\left(X_{1:\ell}=I\right)G_{\gamma^n_I}+\sum_{I\in\{1,\dots,k_0\}^\ell}\left\{\left[\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)-\mathbb{P}_{\theta_0}\left(X_{1:\ell}=I\right)\right]G_{\gamma^0_I}\right.$$
$$\left.+\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)\left(\gamma_J-\gamma^0_I\right)^TD^1G_{\gamma^0_I}+\frac{1}{2}\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\mathbb{P}_\theta\left(X_{1:\ell}=J\right)\left(\gamma_J-\gamma^0_I\right)^TD^2G_{\gamma^*_I}\left(\gamma_J-\gamma^0_I\right)\right\}$$

with $\gamma^*_I\in(\gamma^n_I,\gamma^0_I)$. Thus, using the fact that for all y, $g_\gamma(y)$ vanishes as $\|\gamma\|$ tends to infinity, $f_{\ell,\theta^n}-f_{\ell,\theta_0}/N_n(\theta^n)$ converges pointwise along a subsequence to a function h of form

$$h=\sum_{I\notin\{1,\dots,t_{k_0}\}^\ell}\pi_IG_{\gamma_I}+\sum_{I\in\{1,\dots,k_0\}^\ell}\left(a_IG_{\gamma^0_I}+b_I^TD^1G_{\gamma^0_I}\right)$$
$$+\sum_{I\in\{1,\dots,k_0\}^\ell}c_I\sum_{J\in J(i_1)\times\cdots\times J(i_\ell)}\alpha_{I,J}z_{I,J}^TD^2G_{\gamma^0_I}z_{I,J}$$

as in condition $L(\ell)$, with $\sum_{I\notin\{1,\dots,t_{k_0}\}^\ell}\pi_I+\sum_{I\in\{1,\dots,k_0\}^\ell}(|a_I|+\|b_I\|+c_I)=1$. But as $\|f_{\ell,\theta^n}-f_{\ell,\theta_0}\|_1/N_n(\theta^n)$ tends to 0 as n tends to infinity, we have $\|h\|_1=0$ by Fatou's lemma, and thus $h=0$, contradicting the assumption.

Let us now prove that (L(1)) implies $(L(\ell))$. Let

$$\sum_{i=1}^{k-t_{k_0}}\pi_ig_{\gamma_i}+\sum_{i=1}^{k_0}\left(a_ig_{\gamma^0_i}+b_i^TD^1g_{\gamma^0_i}\right)+\sum_{i=1}^{k_0}c_i^2\sum_{j=1}^{t_i-t_{i-1}}\alpha_{i,j}z_{i,j}'^TD^2g_{\gamma^0_i}z_{i,j}$$

be a function as in (32). If it equals 0, by grouping the terms depending only on y_1 , we can rewrite

the equation as

$$\sum_{i=t_{k_0}+1}^{k} \pi_i'(y_2,\cdots,y_\ell) g_{\gamma_i}(y_1) + \sum_{i=1}^{k_0} \Big(a_i'(y_2,\cdots,y_\ell) g_{\gamma_i^0}(y_1) + b_i'^T(y_2,\cdots,y_\ell) D^1 g_{\gamma_i^0}(y_1) \Big) \\ + \sum_{i=1}^{k_0} \sum_{j=1}^{t_i - t_{i-1}} \sum_{i_2,\cdots,i_l=1}^{k_0} c_I' \sum_{(j_2,\cdots,j_\ell)\in J(i_2)\times\cdots\times J(i_\ell)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i}(y_1) z_{I,J}(i) = 0 \tag{34}$$

where we have written

$$z_{I,J} = (z_{I,J}(i_1),\cdots,z_{I,J}(i_\ell)), \quad \text{with} \quad I=(i_1,\cdots,i_\ell) \quad J=(j_1,,j_\ell), \quad z_{I,J}(i)\in\mathbb{R}^d$$

and

$$c_I' = c_I \prod_{t=2}^{l} g_{\gamma_{i_t}^0}(y_t)$$

Note that if for $i=1,\cdots,k_0$ and $j=1,\cdots,t_i-t_{i-1}$, there exists $w_{i,j}\in\mathbb{R}^d$ such that

$$\sum_{i_2,\cdots,i_\ell=1}^{k_0} c_I' \sum_{(j_2,\cdots,j_\ell)\in J(i_2)\times\cdots\times J(i_\ell)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i}(y_1) z_{I,J}(i) = w_{i,j}^T D^2 g_{\gamma_i}(y_1) w_{i,j}$$

where possibly $w_{i,j}=0$. Let $\alpha_{i,j} = \|w_{i,j}\|^2/(\sum_{j=1}^{t_i-t_{i-1}}\|w_{i,j}\|^2)$ if there exists j such that $\|w_{i,j}\|^2 > 0$ and $c_i' = \sum_{i_2,i_\ell} c_I' \sum_{j=1}^{t_i-t_{i-1}}\|w_{i,j}\|^2$, then

$$\sum_{j=1}^{t_i-t_{i-1}} \sum_{i_2,\cdots,i_\ell=1}^{k_0} c_I' \sum_{(j_2,\cdots,j_\ell)\in J(i_2)\times\cdots\times J(i_\ell)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i}(y_1) z_{I,J}(i) = c_i' \sum_{j=1}^{t_i-t_{i-1}} \alpha_{i,j} w_{i,j}^T D^2 g_{\gamma_i}(y_1) w_{i,j}.$$

and (33) implies that

$$a_i' = c_i' = 0, b_i' = 0 \quad i=1,\cdots,k_0, \quad \pi_i' = 0, \quad i = t_{k_0}+1,\cdots k$$

Simple calculations imply that

$$\pi_i' = \sum_{i_2,\cdots,i_l=1}^{k} \pi_I \prod_{t=2}^{\ell} g_{\gamma_{i_t}}(y_t) = 0 \quad \Leftrightarrow \forall (i_2,,i_\ell)\in\{1,\cdots,k\}^{\ell-2} \pi_{i,i_2,\cdots,i_\ell} = 0$$

and similarly if i is such that there exists $j=1,\cdots,t_i-t_{i-1}$, $I=(i,i_2,\cdots,i_\ell)$ and $J=(j,j_2,\cdots,j_\ell)\in J(i)\times\cdots\times J(i_\ell)$ such that $c_I>0$, $\alpha_J>0$ and $\|z_{I,J}(i)\|>0$, then $c_{i,i_2,\cdots,i_\ell}=0$ for all $i_2,\cdots,i_\ell$. Else, by considering y_t for some other t, we obtain that (34) implies that

$$\pi_I = 0 \quad \forall I \notin \{1,\ldots,t_{k_0}\}^\ell, \quad c_I = 0 \quad \forall I\in\{1,\ldots,t_{k_0}\}^\ell.$$

This leads to

$$b_i' = \sum_{i_2,\cdots,i_l=1}^{k_0} b_I \prod_{t\geq 2} g_{\gamma_{i_t}}(y_t) = 0 \quad \forall i=1,\cdots,k_0.$$

A simple recursive argument implies that $b_I=0$ for all $I\in\{1,\ldots,t_{k_0}\}^\ell$ which in turns implies that $a_I=0$ for all $I\in\{1,\ldots,t_{k_0}\}^\ell$ and condition $(L(\ell))$ is verified.

References

Albert, J. and S. Chib (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Business Economic Statistics 1*, 1–15.

Allman, E. S., C. Matias, and J. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist. 37*, 3099–3132.

Andrieu, C. and A. Doucet (2000). Simulated annealing for Bayesian estimation of hidden Markov models. *IEEE Trans. Information Theory 46*(3), 994–1004.

Baragatti, M. (2011). Sélection bayésienne de variables et méthodes de type parallel tempering avec et sans vraisemblance. Technical report, Université Aix-Marseille II, Marseille. Diploma-thesis.

Boys, R. and D. Henderson (2004). A baysian approach to DNA sequence segmentation (with discussion). *Biometrics 60*, 573–588.

Cappé, O., E. Moulines, and T. Rydén (2004). *Hidden Markov Models*. Springer-Verlag, New York.

Chambaz, A., A. Garivier, and E. Gassiat (2009). A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification. *Journal of Stat. Planning and Inf. 139*, 962–977.

Churchill, G. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol. 51*, 79–94.

Douc, R., E. Moulines, and T. Ryden (2004). Asymptotic properties of the maximum likelihood estimator in the autoregressive models with Markov regime. *Ann. Statist. 32*, 2254–2304.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag, New York.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. H. Poincaré Probab. Statist. 38*, 897–906.

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

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

Gassiat, E. and R. van Handel (2012). The local geometry of finite mixtures. http://arxiv.org/abs/1202.3482.

Geyer, C. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the interface*, pp. 156– 161.

Ghosal, S. and A. van der Vaart (2006). Convergence rates of posterior distributions for non iid observations. *Ann. Statist. 35*(1), 192–223.

Ghosh, J. and R. Ramamoorthi (2003). *Bayesian non parametrics*. Springer-Verlag, New York.

Green, P. and S. Richardson (2002). Hidden Markov models and disease mapping. *J. American Statist. Assoc. 97*(460), 1055–1070.

Guttorp, P. (1995). *Stochastic modelling of scientific data.* Chapman and Hall.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist. 31*, 201–224.

Leroux, B. and M. Putterman (1992). Maximum-penalised-likelihood estimation for independent and Markov dependent mixture models. *Biometrics 48*, 545–558.

MacDonald, I. L. and W. Zucchini (1997). *Hidden Markov and other models for discrete-valued time series*. London, UK: Chapman and Hall/CRC.

MacLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley.

McGrory, C. and D. Titterington (2009). Variational bayesian analysis for hidden Markov models. *Australian and New Zeland Journal of Statistics 51*, 227–244.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pp. 257–286.

Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus Acad. Sciences Paris 330*, 905–908.

Robert, C., T. Rydén, and D. Titterington (2000). Jump Markov chain Monte Carlo algorithms for Bayesian inference in hidden Markov models. *J. Royal Statist. Society Series B 62*, 57–75.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted models. *Journal of Royal Stat. Soc. B*, to appear.

Rydén, T., T. Terasvirta, and S. Asbrink (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appli. Eeconometr. 13*, 217–244.

Spezia, L. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis 31*, 1–11.