Bernoulli **20**(4), 2014, 2039–2075
DOI: 10.3150/13-BEJ550

About the posterior distribution in hidden Markov models with unknown number of states

ELISABETH GASSIAT[1] and JUDITH ROUSSEAU[2]

[1]*Laboratoire de Mathématiques d'Orsay UMR 8628, Université Paris-Sud, Bâtiment 425, 91405 Orsay-Cédex, France. E-mail: elisabeth.gassiat@math.u-psud.fr*
[2]*CREST-ENSAE, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France. E-mail: rousseau@ceremade.dauphine.fr*

We consider finite state space stationary hidden Markov models (HMMs) in the situation where the number of hidden states is unknown. We provide a frequentist asymptotic evaluation of Bayesian analysis methods. Our main result gives posterior concentration rates for the marginal densities, that is for the density of a fixed number of consecutive observations. Using conditions on the prior, we are then able to define a consistent Bayesian estimator of the number of hidden states. It is known that the likelihood ratio test statistic for overfitted HMMs has a nonstandard behaviour and is unbounded. Our conditions on the prior may be seen as a way to penalize parameters to avoid this phenomenon. Inference of parameters is a much more difficult task than inference of marginal densities, we still provide a precise description of the situation when the observations are i.i.d. and we allow for 2 possible hidden states.

Keywords: Bayesian statistics; hidden Markov models; number of components; order selection; posterior distribution

1. Introduction

Finite state space hidden Markov models (which will be shortened to HMMs throughout the paper) are stochastic processes $(X_j, Y_j)_{j\geq 1}$ where $(X_j)_{j\geq 1}$ is a Markov chain living in a finite state space $\mathcal{X}$ and conditionally on $(X_j)_{j\geq 1}$ the Y_j's are independent with a distribution depending only on X_j and living in $\mathcal{Y}$. HMMs are useful tools to model time series where the observed phenomenon is driven by a latent Markov chain. They have been used successfully in a variety of applications, the books MacDonald and Zucchini [14], Zucchini and MacDonald [23] and Cappé *et al.* [2] provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. Finite state space HMMs may also be seen as a dynamic extension of finite mixture models and may be used to do unsupervised clustering. The hidden states often have a practical interpretation in the modelling 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 the data, and the associated parameters.

The aim of this paper is to provide a frequentist asymptotic analysis of Bayesian methods used for statistical inference in finite state space HMMs when the order is unknown. Let us first review what is known on the subject and important questions that still stay unsolved.

1350-7265

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 [13], Rydén *et al.* [21], but without theoretical consistency results. Later, it has been observed that the likelihood ratio statistics is unbounded, in the very simple situation where one wants to test between 1 or 2 hidden states, see Gassiat and Kéribin [9]. 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 [7], Chambaz *et al.* [3]. The use of penalized marginal pseudo likelihood was also proved to lead to weakly consistent estimators by Gassiat [6].

On the Bayesian side, various methods were proposed to deal with an unknown number of hidden states, but no frequentist theoretical result exists for these methods. Notice though that, if the number of states is known, de Gunst and Shcherbakova [4] obtain a Bernstein–von Mises theorem for the posterior distribution, under additional (but usual) regularity conditions. When the order is unknown, reversible jump methods have been built, leading to satisfactory results on simulation and real data, see Boys and Henderson [1], Green and Richardson [12], Robert *et al.* [19], Spezia [22]. The ideas of variational Bayesian methods were developed in McGrory and Titterington [15]. Recently, one of the authors proposed a frequentist asymptotic analysis of the posterior distribution for overfitted mixtures when the observations are i.i.d., see Rousseau and Mengersen [20]. 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 $\mathcal{D}(\alpha_1, \dots, \alpha_k)$ 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 HMMs.

As is well known in the statistical analysis of overfitted finite mixtures, the difficulty of the problem comes from the non-identifiability of the parameters. But what is specific to HMMs is that the non-identifiability of the parameters leads to the fact that neighbourhoods of the "true" parameter values contain transition matrices arbitrarily close to non-ergodic transition matrices. To understand this on a simple example, just consider the case of HMMs with two hidden states, say p is the probability of going from state 1 to state 2 and q the probability of going from state 2 to state 1. If the observations are in fact independently distributed, their distribution may be seen as a HMM with two hidden states where $q = 1 - p$. Neighbourhoods of the "true" values $(p, 1 - p)$ contain parameters such that p is small or $1 - p$ is small, leading to hidden Markov chains having mixing coefficients very close to 1. Imposing a prior condition such as $\delta \le p \le 1 - \delta$ for some $\delta > 0$ is not satisfactory.

Our first main result Theorem 1 gives concentration rates for the posterior distribution of the marginal densities of a fixed number of consecutive observations. First, under mild assumptions on the densities and the prior, we obtain the asymptotic posterior concentration rate $\sqrt{n}$, n the number of observations, up to a $\log n$ factor, when the loss function is the L_1 norm between densities multiplied by some function of the ergodicity coefficient of the hidden Markov chain. Then, with more stringent assumptions on the prior, we give posterior concentration rates for the marginal densities in L_1 norm only (without the ergodicity coefficient). For instance, 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. Then our theorem says that if the sum of the parameters α_j's is large enough, the posterior distribution of the marginal densities in L_1 norm concentrates at a polynomial rate in n. These results are obtained as applications of a general theorem we prove about concentration rates for the posterior distribution of the marginal densities when the state space of the HMM is not constrained to be a finite set, see Theorem 4.

A byproduct of the non-identifiability for overfitted mixtures or HMMs is the fact that, going back from marginal densities to the parameters is not easy. The local geometry of finite mixtures has been understood by Gassiat and van Handel [8], and following their approach in the HMM context we can go back from the L_1 norm between densities to the parameters. We are then able to propose a Bayesian consistent estimator of the number of hidden states, see Theorem 2, under the same conditions on the prior as in Theorem 1. To our knowledge, this is the first consistency result on Bayesian order estimation in the case of HMMs.

Finally, obtaining posterior concentration rates for the parameters themselves seems to be very difficult, and we propose a more complete analysis in the simple situation of HMMs with 2 hidden states and independent observations. In such a case, we prove that, if all the parameters (not only the sum of them) of the prior Dirichlet distribution are large enough, then extra components merge with true ones, see Theorem 3. We believe this to be more general but have not been able to prove it.

The organization of the paper is the following. In Section 2, we first set the model and notations. In subsequent subsections, we give Theorems 1, 2 and 3. In Section 3, we give the posterior concentration theorem for general HMMs, Theorem 4, on which Theorem 1 is based. All proofs are given in Section 4.

2. Finite state space hidden Markov models

2.1. Model and notations

Recall that finite state space HMMs model pairs $(X_i, Y_i)_{i\geq 1}$ where $(X_i)_{i\geq 1}$ is the unobserved Markov chain living on a finite state space $\mathcal{X} = \{1, \ldots, k\}$ and the observations $(Y_i)_{i\geq 1}$ are conditionally independent given the $(X_i)_{i\geq 1}$. The observations take value in $\mathcal{Y}$, which is assumed to be a Polish space endowed with its σ-field. Throughout the paper, we denote $x_{1:n} = (x_1, \ldots, x_n)$.

The hidden Markov chain $(X_i)_{i\geq 1}$ has a Markov transition matrix $Q = (q_{ij})_{1\leq i,j\leq k}$. The conditional distribution of Y_i given X_i has a density with respect to some given measure ν on $\mathcal{Y}$. We denote by $g_{\gamma_j}(y)$, $j = 1, \ldots, k$, the conditional density of Y_i given $X_i = j$. Here, $\gamma_j \in \Gamma \subset \mathbb{R}^d$ for $j = 1, \ldots, k$, the γ_j's are called the emission parameters. In the following, we parametrize the transition matrices on $\{1, \ldots, k\}$ as $(q_{ij})_{1\leq i\leq k, 1\leq j\leq k-1}$ (implying that $q_{ik} = 1 - \sum_{j=1}^{k-1} q_{ij}$ for all $i \leq k$) and we denote by Δ_k the set of probability mass functions $\Delta_k = \{(u_1, \ldots, u_{k-1}) : u_1 \geq 0, \ldots, u_{k-1} \geq 0, \sum_{i=1}^{k-1} u_i \leq 1\}$. 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 denote the overall parameter by $\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1; \gamma_1, \ldots, \gamma_k) \in \Theta_k$ where $\Theta_k = \Delta_k^k \times \Gamma^k$. To alleviate notations, we will write $\theta = (Q; \gamma_1, \ldots, \gamma_k)$, where $Q = (q_{ij})_{1\leq i,j\leq k}$, $q_{ik} = 1 - \sum_{j=1}^{k-1} q_{ij}$ for all $i \leq k$.

Throughout the paper, $\nabla_\theta h$ denotes the gradient vector of the function h when considered as a function of θ, and $D^i_\theta h$ its ith derivative operator with respect to θ, for $i \geq 1$. 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 $a_n \gtrsim b_n$ means that a_n is larger than b_n up to a positive constant that is fixed throughout.

Any Markov chain on a finite state space with transition matrix Q admits a stationary distribution which we denote by μ_Q, if it admits more than one we choose one of them. Then for any finite state space Markov chain with transition matrix Q it is possible to define real numbers $\rho_Q \geq 1$ such that, for any integer m, any $j \leq k$

$$\sum_{j=1}^{k} \left| \left(Q^m\right)_{ij} - \mu_Q(j) \right| \leq \rho_Q^{-m}, \qquad \rho_Q = \left(1 - \sum_{j=1}^{k} \min_{1 \leq i \leq k} q_{ij}\right)^{-1}, \tag{1}$$

where Q^m is the m-step transition matrix of the Markov chain. If $\rho_Q > 1$, the Markov chain $(X_n)_{n\geq 1}$ is uniformly geometrically ergodic and μ_Q is its unique stationary distribution. In the following, we shall also denote μ_θ and ρ_θ in the place of μ_Q and ρ_Q when $\theta = (Q; \gamma_1, \dots, \gamma_k)$.

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 set A in the Borel σ-field of $\mathcal{X}^n \times \mathcal{Y}^n$:

$$\begin{aligned} &\mathbb{P}_\theta\big((X_1, \dots, X_n, Y_1, \dots, Y_n) \in A\big) \\ &\quad = \sum_{x_1, \dots, x_n = 1}^{k} \int_{\mathcal{Y}^n} \mathbb{1}_A(x_{1:n}, y_{1:n}) \mu_Q(x_1) \prod_{i=1}^{n-1} q_{x_i x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}}(y_i) \nu(\mathrm{d}y_1) \cdots \nu(\mathrm{d}y_n). \end{aligned} \tag{2}$$

Thus for any integer n, under $\mathbb{P}_\theta$, $Y_{1:n} = (Y_1, \dots, Y_n)$ has a probability density with respect to $\nu(\mathrm{d}y_1) \cdots \nu(\mathrm{d}y_n)$ equal to

$$f_{n,\theta}(y_1, \dots, y_n) = \sum_{x_1, \dots, x_n = 1}^{k} \mu_Q(x_1) \prod_{i=1}^{n-1} q_{x_i x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}}(y_i). \tag{3}$$

We note E_θ for the expectation under $\mathbb{P}_\theta$.

We denote Π_k the prior distribution on Θ_k. As is often the case in Bayesian analysis of HMMs, instead of computing the stationary distribution μ_Q of the hidden Markov chain with transition matrix Q, we consider a probability distribution $\pi_{\mathcal{X}}$ on the unobserved initial state X_0. Denote $\ell_n(\theta, x_0)$ the log-likelihood starting from x_0, for all $x_0 \in \{1, \dots, k\}$, we have

$$\ell_n(\theta, x_0) = \log\left[\sum_{x_1, \dots, x_n = 1}^{k} \prod_{i=0}^{n-1} q_{x_i x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}}(y_i)\right].$$

The log-likelihood starting from a probability distribution $\pi_{\mathcal{X}}$ on $\mathcal{X}$ is then given by $\log[\sum_{x_0=1}^{k} \mathrm{e}^{\ell_n(\theta, x_0)} \pi_{\mathcal{X}}(x_0)]$. This may also be interpreted as taking a prior $\Pi = \Pi_k \otimes \pi_{\mathcal{X}}$ over

$\Theta_k \times \{1, \dots, k\}$. The posterior distribution can then be written as

$$\mathbb{P}^{\Pi}(A|Y_{1:n}) = \frac{\sum_{x_0=1}^{k} \int_A \mathrm{e}^{\ell_n(\theta,x_0)} \Pi_k(\mathrm{d}\theta) \pi_{\mathcal{X}}(x_0)}{\sum_{x_0=1}^{k} \int_\Theta \mathrm{e}^{\ell_n(\theta,x_0)} \Pi_k(\mathrm{d}\theta) \pi_{\mathcal{X}}(x_0)} \tag{4}$$

for any Borel set $A \subset \Theta_k$.

Let $\mathcal{M}_k$ be the set of all possible probability distributions $\mathbb{P}_\theta$ for all $\theta \in \Theta_k$. We say that the HMM $\mathbb{P}_\theta$ has order k_0 if the probability distribution of $(Y_n)_{n\geq 1}$ 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 $k > k_0$. Indeed, let Q^0 be a $k_0 \times k_0$ transition matrix, and $(\gamma_1^0, \dots, \gamma_{k_0}^0) \in \Gamma^{k_0}$ be parameters that define a HMM of order k_0. Then, $\theta = (Q; \gamma_1^0, \dots, \gamma_{k_0}^0, \dots, \gamma_{k_0}^0) \in \Theta_k$ with $Q = (q_{ij}, 1 \leq i, j \leq k)$ such that:

$$\begin{aligned} q_{ij} &= q_{ij}^0, \qquad i, j < k_0, \\ q_{ij} &= q_{k_0 j}^0, \qquad i \geq k_0, j < k_0, \\ \sum_{l=k_0}^{k} q_{il} &= q_{ik_0}^0, \qquad i \leq k_0, \quad \text{and} \quad \sum_{l=k_0}^{k} q_{il} = q_{k_0 k_0}^0, \qquad i \geq k_0 \end{aligned} \tag{5}$$

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

2.2. Posterior convergence rates for the finite marginal densities

Let $\theta_0 = (Q^0; \gamma_1^0, \dots, \gamma_{k_0}^0) \in \Theta_{k_0}$, $Q^0 = (q_{ij}^0)_{1\leq i\leq k_0, 1\leq j\leq k_0}$, be the parameter of a HMM of order $k_0 \leq k$. We now assume that $\mathbb{P}_{\theta_0}$ is the distribution of the observations. In this section, we fix an integer l and study the posterior distribution of the density of l consecutive observations, that is $f_{l,\theta}$, given by (3) with $n = l$. We study the posterior concentration rate around f_{l,θ_0} in terms of the L_1 loss function, when $\mathbb{P}_{\theta_0}$ is possibly of order $k_0 < k$. In this case, Theorem 2.1 of de Gunst and Shcherbakova [4] does not apply and there is no result in the literature about the frequentist asymptotic properties of the posterior distribution. The interesting and difficult feature of this case is that even though θ_0 is parameterized as an ergodic Markov chain Q^0 with k states and some identical emission parameters as described in (5), f_{l,θ_0} can be approached by marginals $f_{l,\theta}$ for which ρ_θ is arbitrarily close to 1, which deteriorates the posterior concentration rate, see Theorem 1.

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 γ_i, $i = 1, \dots, k$, which are independently distributed from ω. Hence, the prior density of Π_k (with respect to the Lebesgue measure) is equal to $\pi_k = \pi^{\otimes k} \otimes \omega^{\otimes k}$. We still denote by $\pi_{\mathcal{X}}$ a probability on $\{1, \dots, k\}$, we assume that $\pi_{\mathcal{X}}(x) > 0$ for all $x \in \{1, \dots, k\}$ and set $\Pi = \Pi_k \otimes \pi_{\mathcal{X}}$. We shall use the following assumptions.

A0 $q^0_{ij} > 0$, $1 \le i \le k_0$, $1 \le j \le k_0$.

A1 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

$$\int \sup_{\gamma' \in B_d(\gamma,\epsilon)} \left\| \nabla_\gamma \log g_{\gamma'}(y) \right\|^2 g_\gamma(y)\nu(\mathrm{d}y) < +\infty,$$

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

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

A2 There exist $a > 0$ and $b > 0$ such that

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

A3 π is continuous and positive on Δ^0_k, and there exists $C, \alpha_1 > 0, \dots, \alpha_k > 0$ such that (Dirichlet type priors):

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

and ω is continuous and positive on Γ and satisfies

$$\int_{\|x\| \ge n^b} \omega(x)\,\mathrm{d}x = \mathrm{o}\big(n^{-k(k-1+d)/2}\big), \tag{6}$$

with b defined in assumption **A2**.

We will alternatively replace **A3** by

A3bis π is continuous and positive on Δ^0_k, and there exists C such that (exponential type priors):

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

and ω is continuous and positive on Γ and satisfies (6).

Theorem 1. *Assume* **A0–A3**. *Then*, *there exists* K *large enough such that*

$$\mathbb{P}^\Pi\left[\theta : \|f_{l,\theta} - f_{l,\theta_0}\|_1(\rho_\theta - 1) \ge K\sqrt{\frac{\log n}{n}} \Big| Y_{1:n}\right] = \mathrm{o}_{\mathbb{P}_{\theta_0}}(1), \tag{7}$$

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

$$\mathbb{P}^{\Pi}\big[\theta : \| f_{l,\theta} - f_{l,\theta_0} \|_1 \ge 2K n^{-(\bar{\alpha}-k(k-1+d))/(2\bar{\alpha})}(\log n) \big| Y_{1:n}\big] = o_{\mathbb{P}_{\theta_0}}(1). \tag{8}$$

If we replace **A3** *by* **A3bis**, *then there exists* K *large enough such that*

$$\mathbb{P}^{\Pi}\big[\theta : \| f_{l,\theta} - f_{l,\theta_0} \|_1 \ge 2K n^{-1/2}(\log n)^{3/2} \big| Y_{1:n}\big] = o_{\mathbb{P}_{\theta_0}}(1). \tag{9}$$

Theorem 1 is proved in Section 4.1 as a consequence of Theorem 4 stated in Section 3, which gives posterior concentration rates for general HMMs.

Assumption **A0** is the usual ergodic condition on the finite state space Markov chain. Assumptions **A1** and **A2** are mild usual regularity conditions on the emission densities g_γ and hold for instance for multidimensional Gaussian distributions, Poisson distributions, or any regular exponential families. Assumption **A3** on the prior distribution of the transition matrix Q is satisfied for instance if each row of Q follows a Dirichlet distribution or a mixture of Dirichlet distributions, as used in Nur *et al.* [16], and assumption (6) is verified for densities ω that have at most polynomial tails.

The constraint on $\bar{\alpha} = \sum_i \alpha_i$ or condition **A3bis** are used to ensure that (8) and (9) hold respectively. The posterior concentration result (7) implies that the posterior distribution might put non-negligible mass on values of θ for which $\rho_\theta - 1$ is small and $\| f_{l,\theta} - f_{l,\theta_0} \|_1$ is not. These are parameter values associated to nearly non-ergodic latent Markov chains. Since $\rho_\theta - 1$ is small is equivalent to $\sum_j \min_i q_{ij}$ is small, the condition $\bar{\alpha} > k(k-1+d)$ prevents such pathological behaviour by ensuring that the prior mass of such sets is small enough. This condition is therefore of a different nature than Rousseau and Mengersen's [20] condition on the prior, which characterizes the asymptotic behaviour of the posterior distribution on the parameter θ. In other words, their condition allows in (static) mixture models to go from a posterior concentration result on $f_{l,\theta}$ to a posterior concentration result on θ whereas, here, the constraint on $\bar{\alpha}$ is used to obtain a posterior concentration result on $f_{l,\theta}$. Going back from $\| f_{l,\theta} - f_{l,\theta_0} \|_1$ to the parameters requires a deeper understanding of the geometry of finite HMMs, similar to the one developed in Gassiat and van Handel [8]. This will be needed to estimate the order of the HMM in Section 2.3, and fully explored when $k_0 = 1$ and $k = 2$ in Section 2.4.

For general priors, we do not know whether the $\sqrt{\log n}$ factor appearing in (7) could be replaced or not by any sequence tending to infinity. In the case where the α_i's are large enough (Dirichlet type priors), and when $k_0 = 1$ and $k = 2$, we obtain a concentration rate without the $\sqrt{\log n}$ factor, see Lemma 2 in Section 3. To do so, we prove Lemma 3 in Section 3 for which we need to compute explicitly the stationary distribution and the predictive probabilities to obtain a precise control of the likelihood, for θ's such that $\mathbb{P}_\theta$ is near $\mathbb{P}_{\theta_0}$, and to control local entropies of slices for θ's such that $\mathbb{P}_\theta$ is near $\mathbb{P}_{\theta_0}$ and where $\rho_\theta - 1$ might be small. It is not clear to us that extending such computations to the general case is possible in a similar fashion. The $\log n$ terms appearing in (8) and (9) are consequences of the $\sqrt{\log n}$ term appearing in (7).

2.3. Consistent Bayesian estimation of the number of states

To define a Bayesian estimator of the number of hidden states k_0, we need to decide how many states have enough probability mass, and are such that their emission parameters are different enough. We will be able to do it under the assumptions of Theorem 1. Set $w_n = n^{-(\bar{\alpha}-k(k+d-1))/(2\bar{\alpha})}\log n$ if **A3** holds and $\bar{\alpha} > k(k+d-1)$, and set $w_n = n^{-1/2}(\log n)^{3/2}$ if instead **A3bis** holds. Let $(u_n)_{n\geq 1}$ and $(v_n)_{n\geq 1}$ be sequences of positive real numbers tending to 0 as n tends to infinity such that $w_n = \mathrm{o}(u_n v_n)$. As in Rousseau and Mengersen [20], in the case of a misspecified model with $k_0 < k$, f_{l,θ_0} can be represented by merging components or by emptying extra components. For any $\theta \in \Theta_k$, we thus define $J(\theta)$ as

$$J(\theta) = \{ j : \mathbb{P}_\theta(X_1 = j) \geq u_n \},$$

that is, $J(\theta)$ corresponds to the set of *non-empty* components. To cluster the components that have similar emission parameters, we define for all $j \in J(\theta)$

$$A_j(\theta) = \{ i \in J(\theta) : \|\gamma_j - \gamma_i\|^2 \leq v_n \}$$

and the clusters are defined by: for all $j_1, j_2 \in J(\theta)$, j_1 and j_2 belong to the same cluster (noted $j_1 \sim j_2$) if and only if there exist $r > 1$ and $i_1, \ldots, i_r \in J(\theta)$ with $i_1 = j_1$ and $i_r = j_2$ such that for all $1 \leq l \leq r-1$, $A_{i_l}(\theta) \cap A_{i_{l+1}}(\theta) \neq \varnothing$. We then define the effective order of the HMM at θ as the number $L(\theta)$ of different clusters, that is, as the number of equivalent classes with respect to the equivalence relation $\sim$ defined above. By a good choice of u_n and v_n, we construct a consistent estimator of k_0 by considering either the posterior mode of $L(\theta)$ or its posterior median. This is presented in Theorem 2.

To prove that this gives a consistent estimator, we need an inequality that relates the L_1 distance between the l-marginals, $\|f_{l,\theta} - f_{l,\theta_0}\|_1$, to a distance between the parameter θ and parameters $\tilde{\theta}_0$ in Θ_k such that $f_{l,\tilde{\theta}_0} = f_{l,\theta_0}$. Such an inequality will be proved in Section 4.2, under the following structural assumption.

Let $T = \{\mathbf{t} = (t_1, \ldots, t_{k_0}) \in \{1, \ldots, k\}^{k_0} : t_i < t_{i+1}, i = 0, \ldots, k_0 - 1\}$. If b is a vector, b^T denotes its transpose.

A4 For any $\mathbf{t} = (t_1, \ldots, t_{k_0}) \in T$, any $(\pi_i)_{i=1}^{k-t_{k_0}} \in (\mathbb{R}^+)^{k-t_{k_0}}$ (if $t_{k_0} < k$), any $(a_i)_{i=1}^{k_0}, (c_i)_{i=1}^{k_0} \in \mathbb{R}^{k_0}$, $(b_i)_{i=1}^{k_0} \in (\mathbb{R}^d)^{k_0}$, any $z_{i,j} \in \mathbb{R}^d$, $\alpha_{i,j} \in \mathbb{R}$, $i = 1, \ldots, k_0$, $j = 1, \ldots, t_i - t_{i-1}$ (with $t_0 = 0$), such that $\|z_{i,j}\| = 1$, $\alpha_{i,j} \geq 0$ and $\sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} = 1$, for any $(\gamma_i)_{i=1}^{k-t_{k_0}}$ which belong to $\Gamma \setminus \{\gamma_i^0, i = 1, \ldots, k_0\}$,

$$\sum_{i=1}^{k-t_{k_0}} \pi_i g_{\gamma_i} + \sum_{i=1}^{k_0} \bigl(a_i g_{\gamma_i^0} + b_i^T D^1 g_{\gamma_i^0}\bigr) + \sum_{i=1}^{k_0} c_i^2 \sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} z_{i,j}^T D^2 g_{\gamma_i^0} z_{i,j} = 0, \tag{10}$$

if and only if

$$a_i = 0, \qquad b_i = 0, \qquad c_i = 0 \quad \forall i = 1, \ldots, k_0, \qquad \pi_i = 0 \quad \forall i = 1, \ldots, k - t_{k_0}.$$

Assumption **A4** is a weak identifiability condition for situations when $k_0 < k$. Notice that **A4** is the same condition as in Rousseau and Mengersen [20], it is satisfied in particular for Poisson mixtures, location-scale Gaussian mixtures and any mixtures of regular exponential families.

The following theorem says that the posterior distribution of $L(\theta)$ concentrates on the true number k_0 of hidden states.

Theorem 2. *Assume that assumptions* **A0–A2** *and* **A4** *are verified. If either of the following two situations holds*:

- *Under assumption* **A3** (*Dirichlet type prior*), *if* $\bar{\alpha} > k(k+d-1)$ *and*

$$\frac{u_n v_n n^{(\bar{\alpha}-k(k+d-1))/(2\bar{\alpha})}}{\log n} \to +\infty.$$

- *Under assumption* **A3bis** (*exponential type prior*), *if* $u_n v_n n^{1/2}/(\log n)^{3/2} \to +\infty$,

then

$$\mathbb{P}^{\Pi}\bigl[\theta : L(\theta) \neq k_0 | Y_{1:n}\bigr] = \mathrm{o}_{\mathbb{P}_{\theta_0}}(1). \tag{11}$$

If $\hat{k}_n$ *is either the mode or the median of the posterior distribution of* $L(\theta)$, *then*

$$\hat{k}_n = k_0 + \mathrm{o}_{\mathbb{P}_{\theta_0}}(1). \tag{12}$$

One of the advantages of using such an estimate of the order of the HMM, is that we do not need to consider a prior on k and use reversible-jump methods, see Richardson and Green [17], which can be tricky to implement. In particular, we can consider a two-stage procedure where $\hat{k}_n$ is computed based on a model with k components where k is a reasonable upper bound on k_0 and then, fixing $k = \hat{k}_n$ an empirical Bayes procedure is defined on $(Q_{i,j}, i, j \le \hat{k}_n, \gamma_1, \dots, \gamma_{\hat{k}_n})$. On the event $\hat{k}_n = k_0$, which has probability going to 1 under $\mathbb{P}_{\theta_0}$ the model is regular and using the Bernstein–von Mises theorem of de Gunst and Shcherbakova [4], we obtain that with probability $\mathbb{P}_{\theta_0}$ going to 1, the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$ converges in distribution to the centered Gaussian with variance V_0, the inverse of Fisher information at parameter θ_0, where $\hat{\theta}_n$ is an efficient estimator of θ_0 when the order is known to be k_0, and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to the centered Gaussian with variance V_0 under $\mathbb{P}_{\theta_0}$.

The main point in the proof of Theorem 2 is to prove an inequality that relates the L_1 distance between the l-marginals, to a distance between the parameters of the HMM. Under condition **A4**, we prove that there exists a constant $c(\theta_0) > 0$ such that for any small enough positive ε,

$$\begin{aligned} &\frac{\|f_{l,\theta} - f_{l,\theta_0}\|_1}{c(\theta_0)} \\ &\quad \ge \sum_{1 \le j \le k : \forall i, \|\gamma_j - \gamma_i^0\| > \varepsilon} \mathbb{P}_\theta(X_1 = j) + \sum_{i=1}^{k_0} \bigl|\mathbb{P}_\theta\bigl(X_1 \in B(i)\bigr) - \mathbb{P}_{\theta_0}(X_1 = i)\bigr| \\ &\qquad + \sum_{i=1}^{k_0} \Biggl[\Bigl\| \sum_{j \in B(i)} \mathbb{P}_\theta(X_1 = j)\bigl(\gamma_j - \gamma_i^0\bigr)\Bigr\| + \frac{1}{2} \sum_{j \in B(i)} \mathbb{P}_\theta(X_1 = j)\bigl\|\gamma_j - \gamma_i^0\bigr\|^2\Biggr], \end{aligned} \tag{13}$$

where $B(i) = \{j : \|\gamma_j - \gamma_i^0\| \le \varepsilon\}$. The above lower bound essentially corresponds to a partition of $\{1, \dots, k\}$ into $k_0 + 1$ groups, where the first k_0 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 $k_0 + 1$), the second term controls the sum of the weights of the components belonging to the ith group, for $i = 1, \dots, k_0$ (components merging with the true ith component), the third term controls the distance between the mean value over the group i and the true value of the ith 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 ith component. A general inequality implying (13), obtained under a weaker condition, namely **A4bis**, holds and is stated and proved in Section 4.2.

As we have seen with Theorem 2, we can recover the *true parameter* θ_0 using a two-stage procedure where first $\hat{k}_n$ is estimated. However, it is also of interest to understand better the behaviour of the posterior distribution in the first stage procedure and see if some behaviour similar to what was observed in Rousseau and Mengersen [20] holds in the case of HMMs. From Theorem 1, it appears that HMMs present an extra difficulty due to the fact that, when the order is overestimated, the neighbourhood of θ's such that $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$ contains parameters leading to non-ergodic HMMs. To have a more refined understanding of the posterior distribution, we restrict our attention in Section 2.4 to the case where $k = 2$ and $k_0 = 1$ which is still nontrivial, see also Gassiat and Kéribin [9] for the description of pathological behaviours of the likelihood in such a case.

2.4. Posterior concentration for the parameters: The case $k_0 = 1$ and $k = 2$

In this section, we restrict our attention to the simpler case where $k_0 = 1$ and $k = 2$. In Theorem 3 below, we prove that if a Dirichlet type prior is considered on the rows of the transition matrix with parameters α_j's that are large enough the posterior distribution concentrates on the configuration where the two components (states) are merged (γ_1 and γ_2 are close to one another). When $k = 2$, we can parameterize θ as $\theta = (p, q, \gamma_1, \gamma_2)$, with $0 \le p \le 1$, $0 \le q \le 1$, so that

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

when $p \ne 0$ or $q \ne 0$. If $p = 0$ and $q = 0$, set $\mu_\theta = (\frac{1}{2}, \frac{1}{2})$, for instance. Also, we may take

$$\rho_\theta - 1 = (p+q) \wedge \bigl(2 - (p+q)\bigr).$$

When $k_0 = 1$, the observations are i.i.d. with distribution $g_{\gamma^0}\,\mathrm{d}\nu$, 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 $0 < q \le 1$ and any γ, or $\theta_0 = (p, 0, \gamma, \gamma^0)$ for any $0 < p \le 1$ 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):

$$\Pi_2(\mathrm{d}p, \mathrm{d}q) = C_{\alpha,\beta} p^{\alpha-1}(1-p)^{\beta-1} q^{\alpha-1}(1-q)^{\beta-1} \mathbb{1}_{0<p<1} \mathbb{1}_{0<q<1}\, \mathrm{d}p\, \mathrm{d}q,$$

thus satisfying **A3**. Then the following holds.

Theorem 3. *Assume that assumptions* **A0–A2** *together with assumption* **A4** *are verified and consider the prior described above with* $\omega(\cdot)$ *verifying* **A3**. *Assume moreover that for all* x, $\gamma \mapsto g_\gamma(x)$ *is four times continuously differentiable on* Γ, *and that for any* $\gamma \in \Gamma$ *there exists* $\epsilon > 0$ *such that for any* $i \leq 4$,

$$\int \sup_{\gamma' \in B_d(\gamma,\epsilon)} \left\| \frac{D^i_\gamma g_{\gamma'}}{g_{\gamma'}}(y) \right\|^4 g_\gamma(y)\nu(\mathrm{d}y) < +\infty. \tag{14}$$

Then, as soon as $\alpha > 3d/4$ *and* $\beta > 3d/4$, *for any sequence* ϵ_n *tending to* 0,

$$\mathbb{P}^\Pi\left(\frac{p}{p+q} \leq \epsilon_n \text{ or } \frac{q}{p+q} \leq \epsilon_n \Big| Y_{1:n} \right) = o_{\mathbb{P}_{\theta_0}}(1),$$

and for any sequence M_n *going to infinity*,

$$\mathbb{P}^\Pi\left(\|\gamma_1 - \gamma_0\| + \|\gamma_2 - \gamma_0\| \leq M_n n^{-1/4} \big| Y_{1:n} \right) = 1 + o_{\mathbb{P}_{\theta_0}}(1).$$

Theorem 3 says that the extra component cannot be emptied at rate ϵ_n, where the sequence ϵ_n can be chosen to converge to 0 as slowly as we want, so that asymptotically, under the posterior distribution neither $p/(p+q)$ nor $q/(p+p)$ are small, and the posterior distribution concentrates on the configuration where the components merge, with the emission parameters merging at rate $n^{-1/4}$. Similarly in Rousseau and Mengersen [20] the authors obtain that, for independent variables, under a Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$ prior on the weights of the mixture and if $\min \alpha_i > d/2$, the posterior distribution concentrates on configurations which do not empty the extra-components but merge them to true components. The threshold here is $3d/2$ instead of $d/2$. This is due to the fact that there are more parameters involved in a HMM model associated to k states than in a k-components mixture model. No result is obtained here in the case where the α_i's are small. This is due to the existence of non ergodic $\mathbb{P}_\theta$ in the vicinity of $\mathbb{P}_{\theta_0}$ that are not penalized by the prior in such cases. Our conclusion is thus to favour large values of the α_i's.

3. A general theorem

In this section, we present a general theorem which is used to prove Theorem 1 but which can be of interest in more general HMMs. We assume here that the unobserved Markov chain $(X_i)_{i\geq 1}$ lives in a Polish space $\mathcal{X}$ and the observations $(Y_i)_{i\geq 1}$ are conditionally independent given $(X_i)_{i\geq 1}$ and live in a Polish space $\mathcal{Y}$. $\mathcal{X}$, $\mathcal{Y}$ are endowed with their Borel σ-fields. We denote by $\theta \in \Theta$, where Θ is a subset of an Euclidean space, the parameter describing the distribution of the HMM, so that Q_θ, $\theta \in \Theta$ is the Markov kernel of $(X_i)_{i\geq 1}$ and the conditional distribution of Y_i given X_i has density with respect to some given measure ν on $\mathcal{Y}$ denoted by $g_\theta(y|x)$, $x \in \mathcal{X}$, $\theta \in \Theta$. We assume that the Markov kernels Q_θ admit a (not necessarily unique) stationary distribution μ_θ, for each $\theta \in \Theta$. We still 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 set A in the Borel

σ-field of $\mathcal{X}^n \times \mathcal{Y}^n$:

$$\begin{aligned}&\mathbb{P}_\theta\big((X_1,\dots,X_n,Y_1,\dots,Y_n)\in A\big)\\&\quad=\int_A \mu_\theta(\mathrm{d}x_1)\prod_{i=1}^{n-1}Q_\theta(x_i,\mathrm{d}x_{i+1})\prod_{i=1}^{n}g_\theta(y_i|x_i)\nu(\mathrm{d}y_1)\cdots\nu(\mathrm{d}y_n).\end{aligned} \tag{15}$$

Thus for any integer n, under $\mathbb{P}_\theta$, $Y_{1:n}=(Y_1,\dots,Y_n)$ has a probability density with respect to $\nu(\mathrm{d}y_1)\cdots\nu(\mathrm{d}y_n)$ equal to

$$f_{n,\theta}(y_1,\dots,y_n)=\int_{\mathcal{X}^n}\mu_\theta(\mathrm{d}x_1)\prod_{i=1}^{n-1}Q_\theta(x_i,\mathrm{d}x_{i+1})\prod_{i=1}^{n}g_\theta(y_i|x_i). \tag{16}$$

We denote by Π_Θ the prior distribution on Θ and by $\pi_{\mathcal{X}}$ the prior probability on the unobserved initial state, which might be different from the stationary distribution μ_θ. We set $\Pi=\Pi_\Theta\otimes\pi_{\mathcal{X}}$. Similarly to before, denote $\ell_n(\theta,x)$ the log-likelihood starting from x, for all $x\in\mathcal{X}$.

We assume that we are given a stationary HMM $(X_j,Y_j)_{j\geq 1}$ with distribution $\mathbb{P}_{\theta_0}$ for some $\theta_0\in\Theta$.

For any $\theta\in\Theta$, it is possible to define real numbers $\rho_\theta\geq 1$ and $0<\mathbb{R}_\theta\leq 2$ such that, for any integer m, any $x\in\mathcal{X}$

$$\big\|Q_\theta^m(x,\cdot)-\mu_\theta\big\|_{\mathrm{TV}}\leq R_\theta\rho_\theta^{-m}, \tag{17}$$

where $\|\cdot\|_{\mathrm{TV}}$ 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. The following theorem provides a posterior concentration result in a general HMM setting, be it parametric or nonparametric and is an adaptation of Ghosal and van der Vaart [10] to the setup of HMMs. We present the assumptions needed to derive the posterior concentration rate.

C1 There exists $A>0$ such that for any $(x_0,x_1)\in\mathcal{X}^2$, $\mathbb{P}_{\theta_0}$ almost surely, $\forall n\in\mathbb{N}$, $|\ell_n(\theta_0,x_0)-\ell_n(\theta_0,x_1)|\leq A$, and there exist $S_n\subset\Theta\times\mathcal{X}$, $C_n>0$ and $\tilde{\epsilon}_n>0$ a sequence going to 0 with $n\tilde{\epsilon}_n^2\to+\infty$ such that

$$\sup_{(\theta,x)\in S_n}\mathbb{P}_{\theta_0}\big[\ell_n(\theta,x)-\ell_n(\theta_0,x_0)\leq -n\tilde{\epsilon}_n^2\big]=\mathrm{o}(1),\qquad \Pi[S_n]\gtrsim \mathrm{e}^{-C_n n\tilde{\epsilon}_n^2}.$$

C2 There exists a sequence $(\mathcal{F}_n)_{n\geq 1}$ of subsets of Θ

$$\Pi_\Theta\big(\mathcal{F}_n^c\big)=\mathrm{o}\big(\mathrm{e}^{-n\tilde{\epsilon}_n^2(1+C_n)}\big).$$

C3 There exists a sequence $\epsilon_n\geq\tilde{\epsilon}_n$ going to 0, such that $(n\tilde{\epsilon}_n^2(1+C_n))/(n\epsilon_n^2)$ goes to 0 and

$$N\left(\frac{\epsilon_n}{12},\mathcal{F}_n,d_l(\cdot,\cdot)\right)\leq \mathrm{e}^{(n\epsilon_n^2(\rho_{\theta_0}-1)^2)/(16l(2R_{\theta_0}+\rho_{\theta_0}-1)^2)},$$

where $N(\delta,\mathcal{F}_n,d_l(\cdot,\cdot))$ 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 $d_l(\theta,\theta_j) = \|f_{l,\theta} - f_{l,\theta_j}\|_1 := \int_{\mathcal{Y}^l} |f_{l,\theta} - f_{l,\theta_j}|(y)\,\mathrm{d}\nu^{\otimes l}(y)$.

C3bis There exists a sequence $\epsilon_n \geq \tilde{\epsilon}_n$ going to 0 such that

$$\sum_{m\geq 1} \frac{\Pi_\Theta(A_{n,m}(\epsilon_n))}{\Pi(S_n)} \mathrm{e}^{-(nm^2\epsilon_n^2)/(32l)} = \mathrm{o}\big(\mathrm{e}^{-n\tilde{\epsilon}_n^2}\big)$$

and

$$N\left(\frac{m\epsilon_n}{12}, A_{n,m}(\epsilon_n), d_l(\cdot,\cdot)\right) \leq \mathrm{e}^{(nm^2\epsilon_n^2(\rho_{\theta_0}-1)^2)/(16l(2R_{\theta_0}+\rho_{\theta_0}-1)^2)},$$

where

$$A_{n,m}(\epsilon) = \mathcal{F}_n \cap \left\{\theta : m\epsilon \leq \|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \leq (m+1)\epsilon\right\}.$$

Theorem 4. *Assume that* $\rho_{\theta_0} > 1$ *and that assumptions* **C1–C2** *are satisfied, together with either assumption* **C3** *or* **C3bis**. *Then*

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

Theorem 4 gives the posterior concentration rate of $\|f_{l,\theta} - f_{l,\theta_0}\|_1$ up to the parameter $\frac{\rho_\theta-1}{2R_\theta+\rho_\theta-1}$. In Ghosal and van der Vaart [10], for models of non independent variables, the authors consider a parameter space where the mixing coefficient term (for us $\rho_\theta - 1$) is uniformly bounded from below by a positive constant over Θ (see their assumption (4.1) for the application to Markov chains or their assumption on $\mathcal{F}$ in Theorem 7 for the application to Gaussian time series), or equivalently they consider 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, so that their posterior concentration rate is directly expressed in terms of $\|f_{l,\theta} - f_{l,\theta_0}\|_1$. Since we do not restrict ourselves to such frameworks the penalty term $\rho_\theta - 1$ is incorporated in our result. However Theorem 4, is proved along the same lines as Theorem 1 of Ghosal and van der Vaart [10].

The assumption $\rho_{\theta_0} > 1$ implies that the hidden Markov chain X is uniformly ergodic. Assumptions **C1–C2** and either **C3** or **C3bis** are similar in spirit to those considered in general theorems on posterior consistency or posterior convergence rates, see, for instance, Ghosh and Ramamoorthi [11] and Ghosal and van der Vaart [10]. Assumption **C3bis** is often used to eliminate some extra $\log n$ term which typically appear in nonparametric posterior concentration rates and is used in particular in the proof of Theorem 3.

4. Proofs

4.1. Proof of Theorem 1

The proof consists in showing that the assumptions of Theorem 4 are satisfied.

Following the proof of Lemma 2 of Douc *et al.* [5] we find that, since $\rho_{\theta_0} > 1$, for any $x_0 \in \mathcal{X}$,

$$\left|\ell_n(\theta_0, x_0) - \ell_n(\theta_0, x_1)\right| \le 2\left(\frac{\rho_{\theta_0}}{\rho_{\theta_0} - 1}\right)^2$$

so that setting $A = 2(\frac{\rho_{\theta_0}}{\rho_{\theta_0}-1})^2$ the first point of **C1** holds.

We shall verify assumption **C1** with $\tilde{\epsilon}_n = M_n/\sqrt{n}$ for some M_n tending slowly enough to infinity and that will be chosen later. Note that the assumption **A0** and the construction (5) allow to define a $\tilde{\theta}_0 \in \Theta_k$ such that, writing $\tilde{\theta}_0 = (\tilde{Q}^0, \tilde{\gamma}_1^0, \dots, \tilde{\gamma}_k^0)$ with $\tilde{Q}^0 = (\tilde{q}_{i,j}^0, i, j \le k)$, if V is a bounded subset of $\{\theta = (Q, \gamma_1, \dots, \gamma_k); |q_{i,j} - \tilde{q}_{i,j}^0| \le \tilde{\epsilon}_n\}$, then

$$\inf_{\theta \in V} \rho_\theta > 1, \tag{18}$$

for large enough n, and

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

Following the proof of Lemma 2 of Douc *et al.* [5] gives that, if **A0** and **A1** hold, for all $\theta \in V$ $\mathbb{P}_{\theta_0}$-a.s.,

$$\begin{aligned}\ell_n(\theta, x_0) - \ell_n(\theta_0, x_0) &= (\theta - \theta_0)^T \nabla_\theta \ell_n(\theta_0, x_0) \\ &\quad + \int_0^1 (\theta - \theta_0)^T D_\theta^2 \ell_n\big(\theta_0 + u(\theta - \theta_0), x_0\big)(\theta - \theta_0)(1 - u)\,\mathrm{d}u.\end{aligned} \tag{19}$$

Following Theorem 2 in Douc *et al.* [5], $n^{-1/2}\nabla_\theta \ell_n(\theta_0, x)$ 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.* [5], 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:

$$S_n = \left\{\theta \in V; \left\|\gamma_j - \gamma_j^0\right\| \le 1/\sqrt{n}\ \forall j \le k\right\} \times \mathcal{X}$$

so that

$$\sup_{(\theta, x) \in S_n} \mathbb{P}_{\theta_0}\left[\ell_n(\theta, x) - \ell_n(\theta_0, x_0) < -M_n\right] = \mathrm{o}(1). \tag{20}$$

Moreover, letting $D = k(k - 1 + d)$, we have $\Pi \otimes \Pi_{\mathcal{X}}(S_n) \gtrsim n^{-D/2}$ and **C1** is then satisfied setting $C_n = D \log n/(2M_n^2)$.

Let now $v_n = n^{-D/(2\min_{1\le i\le k}\alpha_i)}/\sqrt{\log n}$ and $u_n = n^{-D/(2\sum_{1\le i\le k}\alpha_i)}/\sqrt{\log n}$, and define

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

Now, if $\theta \in \mathcal{F}_n^c$, then there exist $1 \le i, j \le k$ such that $q_{ij} \le v_n$, or $\sum_{j=1}^k \inf_{1\le i\le k} q_{ij} \le u_n$, or there exists $1 \le i \le k$ such that $\|\gamma_i\| \ge n^b$. Using **A3** we easily obtain that for fixed i and j, $\Pi(\{\theta : q_{ij} \le v_n\}) = \mathrm{O}(v_n^{\alpha_j})$ and $\Pi(\{\theta : \|\gamma_i\| \ge n^b\}) = \mathrm{o}(n^{-D/2})$. Also, if $\sum_{j=1}^k \inf_{1\le i\le k} q_{ij} \le u_n$, then there exists a function $i(\cdot)$ from $\{1, \dots, k\}$ to $\{1, \dots, k\}$ whose image set has cardinality at least 2 such that $\sum_{j=1}^k q_{i(j)j} \le u_n$. This gives, using **A3**, $\Pi(\{\theta : \sum_{j=1}^k \inf_{1\le i\le k} q_{ij} \le u_n\}) = \mathrm{O}(u_n^{\sum_{1\le i\le k}\alpha_i})$. Thus,

$$\Pi(\mathcal{F}_n^c) = \mathrm{O}\big(v_n^{\min_{1\le i\le k}\alpha_i} + u_n^{\sum_{1\le i\le k}\alpha_i}\big) + \mathrm{o}\big(n^{-D/2}\big).$$

We may now choose M_n tending to infinity slowly enough so that $v_n^{\min_{1\le i\le k}\alpha_i} + u_n^{\sum_{1\le i\le k}\alpha_i} = \mathrm{o}(\mathrm{e}^{-M_n} n^{-D/2})$ and $\Pi(\mathcal{F}_n^c) = \mathrm{o}(\mathrm{e}^{-M_n} n^{-D/2})$. Then, **C2** holds.

Now, using the definition of $f_{l,\theta}$, we obtain that

$$\|f_{l,\theta_1} - f_{l,\theta_2}\|_1 \le \sum_{j=1}^k |\mu_{\theta_1} - \mu_{\theta_2}| + l \sum_{i,j=1}^k \big|Q_{i,j}^1 - Q_{i,j}^2\big| + l \max_{j\le k} \|g_{\gamma_j^1} - g_{\gamma_j^2}\|_1$$

so that using Lemma 1 below, **A1** and **A2** we get that for some constant B, $\forall(\theta_1, \theta_2) \in \mathcal{F}_n^2$

$$\|f_{l,\theta_1} - f_{l,\theta_2}\|_1 \le B\left(\frac{1}{v_n^{2c}} + n^a\right)\|\theta_1 - \theta_2\|.$$

Thus for some other constant $\tilde{B}$,

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

and **C3** holds when setting $\epsilon_n = K\sqrt{\frac{\log n}{n}}$ with K large enough.

We have proved that under assumptions **A0**, **A1**, **A2**, **A3**, Theorem 4 applies with $\epsilon_n = K\sqrt{\frac{\log n}{n}}$ so that

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

and the first part of Theorem 1 is proved. Now

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

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

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

and the theorem follows when **A3** holds. If now **A3bis** holds instead of **A3**, one gets, taking $u_n = v_n = h/\log n$, with $h > 2C/(k+d-1)$

$$\Pi(\mathcal{F}_n^c) = \mathrm{O}\big(v_n \exp(-C/v_n)\big) + \mathrm{o}\big(n^{-D/2}\big) = \mathrm{o}\big(\mathrm{e}^{-M_n} n^{-D/2}\big)$$

by choosing M_n increasing to infinity slowly enough so that **C2** and **C3** hold. The end of the proof follows similarly as before.

To finish the proof of Theorem 1, we need to prove the following lemma.

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 $m = 1, \dots, k$,*

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

One may take $c = k-1$.

Let $\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1; \gamma_1, \dots, \gamma_k)$ 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

$$\begin{aligned} \mu_\theta(m) &= \frac{P_m(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k-1)}, \qquad m = 1, \dots, k-1, \\ \mu_\theta(k) &= 1 - \sum_{m=1}^{k-1} \mu_\theta(m), \end{aligned} \tag{21}$$

where $P_m, l = 1, \dots, k-1$ 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 nonzero 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}, $1 \le i \le k$, $1 \le j \le k-1$ 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 $m = 1, \dots, k$, $i = 1, \dots, k$, $j = 1, \dots, k-1$,

$$\left| \frac{\partial \mu_\theta(m)}{\partial q_{ij}} \right| \le \frac{C}{R(q_{ij}, 1 \le i \le k, 1 \le j \le k-1)^2}. \tag{22}$$

We shall now prove that

$$R(q_{ij}, 1 \le i \le k, 1 \le j \le k-1) \ge \Bigl(\inf_{1 \le i \le k, 1 \le j \le k, i \ne j} q_{ij} \Bigr)^{k-1}, \tag{23}$$

which combined with (22) and (23) implies Lemma 1. Note that we can express R as a polynomial function of $Q = q_{ij}$, $1 \le i \le k$, $1 \le j \le k$, $i \ne j$. Indeed, $\mu := (\mu_\theta(i))_{1 \le i \le k-1}$ is solution of

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

where V is the $(k-1)$-dimensional vector $(q_{kj})_{1 \le j \le k-1}$, 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 $k \ge 2$:

$$R = \sum_{\sigma \in \mathcal{S}_{k-1}} \varepsilon(\sigma) \prod_{1 \le i \le k-1, \sigma(i)=i} \Bigl(q_{ki} + \sum_{1 \le j \le k-1, j \ne i} q_{ij} \Bigr) \prod_{1 \le i \le k-1, \sigma(i) \ne i} (q_{ki} - q_{\sigma(i)i}), \tag{24}$$

where for any integer n, $\mathcal{S}_n$ is the set of permutations of $\{1, \dots, n\}$, 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 $\{1, \dots, k-1\}$, and for all $i \in B$, $j(i) \in \{1, \dots, k-1\}$ and $j(i) \ne i$. In case $B = \varnothing$, 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 $B \ne \varnothing$. Fix such a monomial with non-null 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 $A = \varnothing$). Fix also $q_{ij(i)} = 1$ for all $i \in B$. Then, if (A', B') is another partition of $\{1, \dots, k-1\}$ with $B' \ne \varnothing$, 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 $R(Q) \ge 0$, so that this coefficient is a positive integer and (23) follows.

4.2. Proof of Theorem 2

Applying Theorem 1, we get that under the assumptions of Theorem 2, there exists K such that

$$\mathbb{P}_{\theta_0}\bigl(\| f_{l,\theta} - f_{l,\theta_0} \|_1 \le 2K w_n | Y_{1:n} \bigr) = 1 + o_{\mathbb{P}_{\theta_0}}(1).$$

But if inequality (13) holds, then as soon as

$$\|f_{l,\theta} - f_{l,\theta_0}\|_1 \lesssim w_n \tag{25}$$

we get that, for any $j \in \{1, \dots, k\}$, either $\mathbb{P}_\theta(X_1 = j) \lesssim w_n$, or

$$\exists i \in \{1, \dots, k_0\}, \qquad \mathbb{P}_\theta(X_1 = j)\big\|\gamma_j - \gamma_i^0\big\|^2 \lesssim w_n.$$

Let us choose $\epsilon \le \min_{i \ne j} \|\gamma_i^0 - \gamma_j^0\|/4$ in the definition of $B(i)$ in (13). We then obtain that for large enough n, all $j_1, j_2 \in J(\theta)$, we have $j_1 \sim j_2$ if and only if they belong to the same $B(i)$, $i = 1, \dots, k_0$, so that $L(\theta) \le k_0$. On the other hand, $L(\theta) < k_0$ would mean that at least one $B(i)$ would be empty which contradicts the fact that

$$\big|\mathbb{P}_\theta\big(X_1 \in B(i)\big) - \mathbb{P}_{\theta_0}(X_1 = i)\big| \le w_n.$$

Thus, for large enough n, if (25) holds, then $L(\theta) = k_0$, so that

$$P^\Pi\big[L(\theta) = k_0 | Y^n\big] = 1 + o_{\mathbb{P}_{\theta_0}}(1).$$

To finish the proof, we now prove that (13) holds under the assumptions of Theorem 2. This will follow from Proposition 1 below which is slightly more general.

An inequality that relates the L_1 distance of the l-marginals to the parameters of the HMM is proved in Gassiat and van Handel [8] 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 obtain it for more general situations than families of translated distributions, under the structural assumption **A4**. The inequality following Theorem 3.10 of Gassiat and van Handel [8] says that there exists a constant $c(\theta_0) > 0$ such that for any small enough positive ε,

$$\begin{aligned}
&\frac{\|f_{l,\theta} - f_{l,\theta_0}\|_1}{c(\theta_0)} \\
&\quad\ge \sum_{1 \le j \le k : \forall i, \|\gamma_j - \gamma_i^0\| > \varepsilon} \mathbb{P}_\theta(X_1 = j) \\
&\qquad + \sum_{1 \le i_1, \dots, i_l \le k_0} \Bigg[\big|\mathbb{P}_\theta\big(X_{1:l} \in A(i_1, \dots, i_l)\big) - \mathbb{P}_{\theta_0}(X_{1:l} = i_1 \cdots i_l)\big| \\
&\qquad\qquad + \Bigg\| \sum_{(j_1, \dots, j_l) \in A(i_1, \dots, i_l)} \mathbb{P}_\theta(X_{1:l} = j_1 \cdots j_l) \left\{ \begin{pmatrix} \gamma_{j_1} \\ \cdots \\ \gamma_{j_l} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \cdots \\ \gamma_{i_l}^0 \end{pmatrix} \right\} \Bigg\| \\
&\qquad\qquad + \frac{1}{2} \sum_{(j_1, \dots, j_l) \in A(i_1, \dots, i_l)} \mathbb{P}_\theta(X_{1:l} = j_1 \cdots j_l) \left\| \begin{pmatrix} \gamma_{j_1} \\ \cdots \\ \gamma_{j_l} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \cdots \\ \gamma_{i_l}^0 \end{pmatrix} \right\|^2 \Bigg],
\end{aligned} \tag{26}$$

where $A(i_1,\dots,i_l)=\{(j_1,\dots,j_l):\|\gamma_{j_1}-\gamma^0_{i_1}\|\le\varepsilon,\dots,\|\gamma_{j_l}-\gamma^0_{i_l}\|\le\varepsilon\}$. The above lower bound essentially corresponds to a partition of $\{1,\dots,k\}^l$ 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 controls the sum of the weights of the components belonging to the ith group, for $i=1,\dots,k_0^l$ (components merging with the true ith component), the third term controls the distance between the mean value over the group i and the true value of the ith 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 ith component.

Notice that (13) is a consequence of (26). We shall prove that (26) holds under an assumption slightly more general than **A4**. For this, we need to introduce some notations. For all $I=(i_1,\dots,i_l)\in\{1,\dots,k\}^l$, define $\gamma_I=(\gamma_{i_1},\dots,\gamma_{i_l})$, $G_{\gamma_I}=\prod_{t=1}^l g_{\gamma_{i_t}}(y_t)$, $D^1G_{\gamma_I}$ 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 $D^2G_{\gamma_I}$ the symmetric 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. For any $\mathbf{t}=(t_1,\dots,t_{k_0})\in T$, define for all $i\in\{1,\dots,k_0\}$ the set $J(i)=\{t_{i-1}+1,\dots,t_i\}$, using $t_0=0$.

We then consider the following condition:

A4bis For any $\mathbf{t}=(t_1,\dots,t_{k_0})\in T$, for all collections $(\pi_I)_I,(\gamma_I)_I, I\notin\{1,\dots,t_{k_0}\}^l$ satisfying $\pi_I\ge 0$, $\gamma_I=(\gamma_{i_1},\dots,\gamma_{i_l})$ such that $\gamma_{i_j}=\gamma_i^0$ when $i_j\in J(i)$ for some $i\le k_0$ and $\gamma_{i_j}\in\Gamma\setminus\{\gamma_i^0, i=1,\dots,k_0\}$ when $i_j\notin\{1,\dots,t_{k_0}\}$, for all collections $(a_I)_I$, $(c_I)_I$, $(b_I)_I$, $I\in\{1,\dots,k_0\}^l$, $a_I\in\mathbb{R}$, $c_I\ge 0$ and $b_I\in\mathbb{R}^{d|I|}$, for all collection of vectors $z_{I,J}\in\mathbb{R}^{d|I|}$ with $I\in\{1,\dots,k_0\}^l$ and $J\in J(i_1)\times\cdots\times J(i_l)$ satisfying $\|z_{I,J}\|=1$, and all sequences $(\alpha_{I,J})$, satisfying $\alpha_{I,J}\ge 0$ and $\sum_{J\in J(i_1)\times\cdots\times J(i_l)}\alpha_{I,J}=1$,

$$\begin{aligned}
&\sum_{I\notin\{1,\dots,t_{k_0}\}^l}\pi_I G_{\gamma_I}+\sum_{I\in\{1,\dots,k_0\}^l}\bigl(a_I G_{\gamma_I^0}+b_I^T D^1 G_{\gamma_I^0}\bigr)\\
&\quad+\sum_{I\in\{1,\dots,k_0\}^l}c_I\sum_{J\in J(i_1)\times\cdots\times J(i_l)}\alpha_{I,J}z_{I,J}^T D^2 G_{\gamma_I^0} z_{I,J}=0\\
&\Leftrightarrow\\
&\sum_{I\notin\{1,\dots,t_{k_0}\}^l}\pi_I+\sum_{I\in\{1,\dots,k_0\}^l}\bigl(|a_I|+\|b_I\|+c_I\bigr)=0.
\end{aligned}\tag{27}$$

We have the following proposition.

Proposition 1. *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 assumption* **A4bis** *is verified,* (26) *holds. Moreover, condition* **A4bis** *is verified as soon as condition* **A4** (*corresponding to $l=1$*) *is verified.*

Let us now prove Proposition 1. To prove the first part of the proposition, we follow the ideas of the beginning of the proof of Theorem 5.11 in Gassiat and van Handel [8]. If (26) 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, $\|f_{l,\theta^n} - f_{l,\theta_0}\|_1/N_n(\theta^n)$ tends to 0 as n tends to infinity, with

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

with $A_n(i_1,\dots,i_l) = \{(j_1,\dots,j_l) : \|\gamma_{j_1} - \gamma_{i_1}^0\| \leq \varepsilon_n, \dots, \|\gamma_{j_l} - \gamma_{i_l}^0\| \leq \varepsilon_n\}$.

Now, $f_{l,\theta^n} = \sum_{I\in\{1,\dots,k\}^l} \mathbb{P}_{\theta^n}(X_1,\dots,X_l = I)G_{\gamma_I^n}$ where $\theta^n = (Q^n, (\gamma_1^n,\dots,\gamma_k^n))$, 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_i^n converges to some limit γ_i or $\|\gamma_i^n\|$ tends to infinity. Choose now the indexation such that for $i = 1,\dots,t_1$, γ_i^n converges to γ_1^0, for $i = t_1+1,\dots,t_2$, γ_i^n converges to γ_2^0, and so on, for $i = t_{k_0-1}+1,\dots,t_{k_0}$, γ_i^n converges to $\gamma_{k_0}^0$, and if $t_{k_0} < k$, for some $\tilde{k} \leq k$, for $i = t_{k_0}+1,\dots,\tilde{k}$, γ_i^n converges to some $\gamma_i \notin \{\gamma_1^0,\dots,\gamma_{k_0}^0\}$, and for $i = \tilde{k}+1,\dots,k$, $\|\gamma_i^n\|$ tends to infinity. It is possible that $\tilde{k} = t_{k_0}$ in which case no γ_i^n converges to some $\gamma_i \notin \{\gamma_1^0,\dots,\gamma_{k_0}^0\}$. Such a $\mathbf{t} = (t_1,\dots,t_{k_0}) \in T$ exists, because if $\|f_{l,\theta^n} - f_{l,\theta_0}\|_1/N_n(\theta^n)$ tends to 0 as n tends to infinity, $\|f_{l,\theta^n} - f_{l,\theta_0}\|_1$, and $N_n(\theta^n)$ tends to 0 as n tends to infinity (if it was not the case, using the regularity of $\theta \mapsto f_{l,\theta}$ we would have a contradiction). Now along the subsequence we may write, for large enough n:

$$\begin{aligned}
N_n(\theta^n) = & \sum_{I\notin\{1,\dots,t_{k_0}\}^l} \mathbb{P}_\theta(X_{1:l} = I) \\
& + \sum_{I\in\{1,\dots,k_0\}^l}\Biggl[\Biggl|\sum_{J\in J(i_1)\times\cdots\times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) - \mathbb{P}_{\theta_0}(X_{1:l} = I)\Biggr| \\
& \qquad + \Biggl\|\sum_{J\in J(i_1)\times\cdots\times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J)(\gamma_J - \gamma_I^0)\Biggr\| \\
& \qquad + \frac{1}{2}\sum_{J\in J(i_1)\times\cdots\times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J)\|\gamma_J - \gamma_I^0\|^2\Biggr].
\end{aligned}$$

We shall use a Taylor expansion till order 2. To be perfectly rigorous in the following, we need to express explicitly I in terms of its distinct indices, $(\tilde{i}_1, \dots, \tilde{i}_{|I|})$, so that $G_{\gamma_I} = \prod_{t=1}^{|I|} \prod_{j:i_j=\tilde{i}_t} g_{\gamma_{\tilde{i}_t}}(y_j)$, but to keep notations concise we do not make such a distinction and for instance $(\gamma_J^n - \gamma_I^0)^T D^1 G_{\gamma_I^0}$ means

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

and similarly for the second derivatives. We have

$$\begin{aligned} f_{l,\theta^n} - f_{l,\theta_0} = & \sum_{I \notin \{1,\dots,t_{k_0}\}^l} \mathbb{P}_\theta(X_{1:l} = I) G_{\gamma_I^n} \\ & + \sum_{I \in \{1,\dots,k_0\}^l} \Biggl\{ \Biggl[\sum_{J \in J(i_1) \times \cdots \times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) - \mathbb{P}_{\theta_0}(X_{1:l} = I) \Biggr] G_{\gamma_I^0} \\ & \qquad + \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) \bigl(\gamma_J - \gamma_I^0\bigr)^T D^1 G_{\gamma_I^0} \\ & \qquad + \frac{1}{2} \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) \bigl(\gamma_J - \gamma_I^0\bigr)^T D^2 G_{\gamma_I^*} \bigl(\gamma_J - \gamma_I^0\bigr) \Biggr\} \end{aligned}$$

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

$$\begin{aligned} h = & \sum_{I \notin \{1,\dots,t_{k_0}\}^l} \pi_I G_{\gamma_I} + \sum_{I \in \{1,\dots,k_0\}^l} \bigl(a_I G_{\gamma_I^0} + b_I^T D^1 G_{\gamma_I^0}\bigr) \\ & + \sum_{I \in \{1,\dots,k_0\}^l} c_I \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \alpha_{I,J} z_{I,J}^T D^2 G_{\gamma_I^0} z_{I,J} \end{aligned}$$

as in condition $L(l)$, with $\sum_{I \notin \{1,\dots,t_{k_0}\}^l} \pi_I + \sum_{I \in \{1,\dots,k_0\}^l} (|a_I| + \|b_I\| + c_I) = 1$. But as $\|f_{l,\theta^n} - f_{l,\theta_0}\|_1 / N_n(\theta^n)$ 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 **A4** implies **A4bis**. Let

$$\begin{aligned} & \sum_{I \notin \{1,\dots,t_{k_0}\}^l} \pi_I G_{\gamma_I} + \sum_{I \in \{1,\dots,k_0\}^l} \bigl(a_I G_{\gamma_I^0} + b_I^T D^1 G_{\gamma_I^0}\bigr) \\ & \quad + \sum_{I \in \{1,\dots,k_0\}^l} c_I \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \alpha_{I,J} z_{I,J}^T D^2 G_{\gamma_I^0} z_{I,J} = 0 \\ & \Leftrightarrow \\ & \sum_{I \notin \{1,\dots,t_{k_0}\}^l} \pi_I + \sum_{I \in \{1,\dots,k_0\}^l} \bigl(|a_I| + \|b_I\| + c_I\bigr) = 0 \end{aligned}$$

with $\pi_I, a_I, b_I, \alpha_{I,J}$ and $z_{I,J}$ be as in assumption **A4bis**. We group the terms depending only on y_1 and we can rewrite the equation as

$$\begin{aligned}\sum_{i=t_{k_0}+1}^{k} \pi_i'(y_2,\dots,y_l) g_{\gamma_i}(y_1) + \sum_{i=1}^{k_0} \bigl(a_i'(y_2,\dots,y_l) g_{\gamma_i^0}(y_1) + b_i'^T(y_2,\dots,y_l) D^1 g_{\gamma_i^0}(y_1)\bigr) \\ + \sum_{i=1}^{k_0} \sum_{j=1}^{t_i-t_{i-1}} \sum_{i_2,\dots,i_l=1}^{k_0} c_I' \sum_{(j_2,\dots,j_l)\in J(i_2)\times\cdots\times J(i_l)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i^0}(y_1) z_{I,J}(i) = 0,\end{aligned} \tag{28}$$

where we have written

$$z_{I,J} = \bigl(z_{I,J}(i_1),\dots,z_{I,J}(i_l)\bigr), \qquad \text{with } I=(i,i_2,\dots,i_l),\ J=(j_1,\dots,j_l),\ z_{I,J}(i)\in\mathbb{R}^d$$

and

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

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

$$\sum_{i_2,\dots,i_l=1}^{k_0} c_I' \sum_{(j_2,\dots,j_l)\in J(i_2)\times\cdots\times J(i_l)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i^0}(y_1) z_{I,J}(i) = w_{i,j}^T D^2 g_{\gamma_i^0}(y_1) w_{i,j},$$

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 $c_i' = \sum_{i_2,\dots,i_l} c_I' \sum_{j=1}^{t_i-t_{i-1}} \|w_{i,j}\|^2$, then

$$\begin{aligned}&\sum_{j=1}^{t_i-t_{i-1}} \sum_{i_2,\dots,i_l=1}^{k_0} c_I' \sum_{(j_2,\dots,j_l)\in J(i_2)\times\cdots\times J(i_l)} \alpha_{I,J} z_{I,J}(i)^T D^2 g_{\gamma_i^0}(y_1) z_{I,J}(i) \\ &\quad = c_i' \sum_{j=1}^{t_i-t_{i-1}} \alpha_{i,j} w_{i,j}^T D^2 g_{\gamma_i^0}(y_1) w_{i,j}\end{aligned}$$

and (10) implies that

$$a_i' = c_i' = 0, \qquad b_i' = 0, \qquad i=1,\dots,k_0, \qquad \pi_i' = 0, \qquad i = t_{k_0}+1,\dots,k.$$

Simple calculations imply that

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

and similarly if i is such that there exists $j = 1, \dots, t_i - t_{i-1}$, $I = (i, i_2, \dots, i_l)$ and $J = (j, j_2, \dots, j_l) \in J(i) \times \cdots \times J(i_l)$ such that $c_I > 0$, $\alpha_J > 0$ and $\|z_{I,J}(i)\| > 0$, then $c_{i,i_2,\dots,i_l} = 0$ for all $i_2, \dots, i_l$. Else, by considering y_t for some other t, we obtain that (28) implies that

$$\pi_I = 0 \qquad \forall I \notin \{1, \dots, t_{k_0}\}^l, \qquad c_I = 0 \qquad \forall I \in \{1, \dots, t_{k_0}\}^l.$$

This leads to

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

A simple recursive argument implies that $b_I = 0$ for all $I \in \{1, \dots, t_{k_0}\}^l$ which in turns implies that $a_I = 0$ for all $I \in \{1, \dots, t_{k_0}\}^l$ and condition **A4bis** is verified.

4.3. Proof of Theorem 3

First, we obtain the following lemma.

Lemma 2. *Under the assumptions of Theorem 3, for any sequence M_n tending to infinity,*

$$\mathbb{P}^\Pi\left((p+q) \wedge \big(2-(p+q)\big)\|f_{2,\theta} - f_{2,\theta_0}\|_1 \leq \frac{M_n}{\sqrt{n}}\right) = 1 + o_{\mathbb{P}_{\theta_0}}(1).$$

We prove Lemma 2 by applying Theorem 4, using some of the computations of the proof of Theorem 1 but verifying assumption **C3bis** instead of **C3**. Set $S_n = U_n \times \mathcal{X}$ with

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

for small but fixed ϵ. We shall prove later the following lemma.

Lemma 3. *Let M_n tend to infinity. Then*

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

and

$$\Pi(S_n) \gtrsim n^{-3d/4}. \tag{29}$$

Now we prove that assumption **C3bis** holds with $\epsilon_n = M_n/\sqrt{n}$, which will finish the proof of Lemma 2. By Proposition 1, we obtain that there exists $c(\theta_0) > 0$ and $\eta > 0$ such that:

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

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

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

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

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

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

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

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

Similar upper bounds hold also by Taylor expansion. Thus, for any m, $A_{n,m}(\epsilon_n)$ is a subset of the set of θ's such that

$$\min\bigg\{\frac{(p+q)\wedge(2-(p+q))}{p+q}\big[\big\|q(\gamma_1-\gamma^0)+p(\gamma_2-\gamma^0)\big\|+q\|\gamma_1-\gamma^0\|^2+p\|\gamma_2-\gamma^0\|^2\big]; \frac{(p+q)\wedge(2-(p+q))}{p+q}\big[p+q\|\gamma_1-\gamma^0\|\big]; \frac{(p+q)\wedge(2-(p+q))}{p+q}\big[q+p\|\gamma_2-\gamma^0\|\big]; (p+q)\wedge\big(2-(p+q)\big)\bigg\}\lesssim(m+1)\epsilon_n.$$

This leads to

$$\Pi_2\big(A_{n,m}(\epsilon_n)\big)\lesssim\big[(m+1)\epsilon_n\big]^{2\alpha}+\big[(m+1)\epsilon_n\big]^{2\beta}+\big[(m+1)\epsilon_n\big]^{\alpha+d}$$

so that if $\alpha,\beta>3d/4$ and (29) holds, there exists $\delta>0$ such that

$$\frac{\Pi_2(A_{n,m}(\epsilon_n))\mathrm{e}^{-(nm^2\epsilon_n^2)/(32l)}}{\Pi(S_n)}\lesssim n^{-\delta}\big[(M_nm)^{2\alpha}+(M_nm)^{2\beta}+(M_nm)^{\alpha+d}\big]\mathrm{e}^{-(M_n^2m^2)/(32l)}.$$

Also for all $\epsilon>0$ small enough $A_{n,m}(\epsilon)$ contains the set of θ's such that

$$\max\bigg\{\frac{(p+q)\wedge(2-(p+q))}{p+q} \times\big[\big\|q(\gamma_1-\gamma^0)+p(\gamma_2-\gamma^0)\big\|+q\|\gamma_1-\gamma^0\|^2+p\|\gamma_2-\gamma^0\|^2\big];$$

$$\left.\begin{aligned}&\frac{(p+q)\wedge(2-(p+q))}{p+q}\big[p+q\|\gamma_1-\gamma^0\|\big];\\&\frac{(p+q)\wedge(2-(p+q))}{p+q}\big[q+p\|\gamma_2-\gamma^0\|\big];\,(p+q)\wedge\big(2-(p+q)\big)\end{aligned}\right\}\lesssim(m+1)\epsilon$$

therefore

$$N\left(\frac{m\epsilon_n}{12}, A_{n,m}(\epsilon_n), d_l(\cdot,\cdot)\right)\lesssim m^{2+2d}\lesssim \mathrm{e}^{(n\epsilon_n^2 m^2(\rho_{\theta_0}-1)^2)/(16l(2+\rho_{\theta_0}-1)^2)},$$

so that assumption **C3bis** is verified.

We now prove Theorem 3. Notice first that, by setting

$$D_n=\int_{\Theta\times\mathcal{X}}\mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\Pi_2(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x),$$

as in the proof of Theorem 4 we get that for any sequence C_n tending to infinity,

$$\mathbb{P}_{\theta_0}\big(D_n\le C_n n^{-D/2}\big)=\mathrm{o}(1)\tag{30}$$

with $D=d+d/2$.

Let now ϵ_n be any sequence going to 0 and let $A_n=\{\frac{p}{p+q}\le\epsilon_n \text{ or } \frac{q}{p+q}\le\epsilon_n\}$. For some sequence M_n going to infinity and $\delta_n=M_n/\sqrt{n}$, let $B_n=\{(p+q)\wedge(2-(p+q))\|f_{2,\theta}-f_{2,\theta_0}\|_1\le\delta_n\}$. We then control with $D=d+d/2$, using Lemma 2

$$\begin{aligned}E_{\theta_0}\big[\mathbb{P}^\Pi(A_n|Y_{1:n})\big]&=E_{\theta_0}\big[\mathbb{P}^\Pi(A_n\cap B_n|Y_{1:n})\big]+\mathrm{o}(1)\\&=E_{\theta_0}\left[\frac{\int_{A_n\cap B_n\times\mathcal{X}}\mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\Pi_2(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)}{\int_{\Theta\times\mathcal{X}}\mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\Pi_2(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)}\right]+\mathrm{o}(1)\\&:=E_{\theta_0}\left[\frac{N_n}{D_n}\right]+\mathrm{o}(1)\\&\le\mathbb{P}_{\theta_0}\big(D_n\le C_n n^{-D/2}\big)+\frac{n^{D/2}}{C_n}\Pi_2(A_n\cap B_n)+\mathrm{o}(1).\end{aligned}$$

Thus using (30), the first part of Theorem 3 is proved by showing that

$$\Pi_2(A_n\cap B_n)\lesssim\delta_n^{2\alpha}+\delta_n^{\alpha+d}+\delta_n^{d+d/2}\epsilon_n^{\alpha-d/2}.\tag{31}$$

Then, the second part of Theorem 3 follows from its first part and Lemma 2.

We now prove that (31) holds. Define

$$\begin{aligned}B_n^1=\bigg\{&\frac{(p+q)\wedge(2-(p+q))}{p+q}\\&\times\big[\big\|q\big(\gamma_1-\gamma^0\big)+p\big(\gamma_2-\gamma^0\big)\big\|+q\|\gamma_1-\gamma^0\|^2+p\|\gamma_2-\gamma^0\|^2\big]\le\delta_n\bigg\},\end{aligned}$$

$$B_n^2 = \left\{ \frac{(p+q) \wedge (2-(p+q))}{p+q} \left[p + q \|\gamma_1 - \gamma^0\| \right] \le \delta_n \right\},$$
$$B_n^3 = \left\{ \frac{(p+q) \wedge (2-(p+q))}{p+q} \left[q + p \|\gamma_2 - \gamma^0\| \right] \le \delta_n \right\}$$

and

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

Then

$$\Pi_2(A_n \cap B_n) \le \Pi_2\left(A_n \cap B_n^1\right) + \Pi_2\left(A_n \cap B_n^2\right) + \Pi_2\left(A_n \cap B_n^3\right) + \Pi_2\left(A_n \cap B_n^4\right).$$

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

- On $A_n \cap B_n^1$, $\|q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)\| \lesssim \delta_n$, $q\|\gamma_1 - \gamma^0\|^2 \lesssim \delta_n$, $p\|\gamma_2 - \gamma^0\|^2 \lesssim \delta_n$, and $p \lesssim \epsilon_n$ or $q \lesssim \epsilon_n$. This gives $\Pi_2(A_n \cap B_n^1) \lesssim \delta_n^{d+d/2} \epsilon_n^{\alpha - d/2}$.
- On $A_n \cap B_n^2$, $p \lesssim \delta_n$ and $q\|\gamma_1 - \gamma^0\| \lesssim \delta_n$ in case $p+q \le 1$, and $p \lesssim \delta_n$, $1-q \lesssim \delta_n$ and $q\|\gamma_1 - \gamma^0\| \lesssim \delta_n$ in case $p+q \ge 1$, leading to $\Pi_2(A_n \cap B_n^2) \lesssim \delta_n^{\alpha+d} + \delta_n^{\alpha+\beta+d}$.
- For symmetry reasons, $\Pi_2(A_n \cap B_n^3) = \Pi_2(A_n \cap B_n^2)$.
- On $A_n \cap B_n^4$, $p \lesssim \delta_n$ and $q \lesssim \delta_n$, so that $\Pi_2(A_n \cap B_n^4) \lesssim \delta_n^{2\alpha}$.

Keeping only the leading terms, we see that (31) holds and this terminates the proof Theorem 3.

We now prove Lemma 3. We easily get $\Pi_2(U_n) \gtrsim n^{-3d/4}$, and

$$D_n \ge \int_{U_n \times \mathcal{X}} \mathrm{e}^{\ell_n(\theta,x) - \ell_n(\theta_0,x)} \Pi_2(\mathrm{d}\theta) \pi_{\mathcal{X}}(\mathrm{d}x).$$

Let us now study $\ell_n(\theta, x) - \ell_n(\theta_0, x)$. First, following the proof of Lemma 2 of Douc *et al.* [5] we find that, for any $\theta \in U_n$, for any x,

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

where $\ell_n(\theta) = \sum_{x=1}^k \mu_\theta(x) \ell_n(\theta, x)$. Thus, for any $\theta \in U_n$ and any x, and since $\ell_n(\theta_0, x)$ does not depend on x,

$$\ell_n(\theta, x) - \ell_n(\theta_0, x) \ge \ell_n(\theta) - \ell_n(\theta_0) - \left(\frac{1+2\epsilon}{1-2\epsilon}\right)^2. \tag{32}$$

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

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

and we set for $k = 1$

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

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

$$p_{k+1}(\theta) = \frac{(1-p)p_k(\theta)g_{\gamma_1}(Y_k) + q(1-p_k(\theta))g_{\gamma_2}(Y_k)}{p_k(\theta)g_{\gamma_1}(Y_k) + (1-p_k(\theta))g_{\gamma_2}(Y_k)}. \tag{33}$$

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 ith 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 $\theta_t = t\theta + (1-t)(p, q, \gamma^0, \gamma^0)$:

$$\ell_n(\theta) - \ell_n(\theta_0) = (\gamma_1 - \gamma^0)D^1_{\gamma_1}\ell^0_n + (\gamma_2 - \gamma^0)D^1_{\gamma_2}\ell^0_n + S_n(\theta) + T_n(\theta) + R_n(\theta, t), \tag{34}$$

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

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

and

$$R_n(\theta, t) = \sum_{k=0}^{4} \binom{k}{4} (\gamma_1 - \gamma^0)^k (\gamma_2 - \gamma^0)^{4-k} D^4_{(\gamma_1)^k,(\gamma_2)^{4-k}}\ell_n(\theta_t).$$

Easy but tedious computations lead to the following results.

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

so that

$$\sup_{\theta\in U_n}\big|(\gamma_1-\gamma^0)D^1_{\gamma_1}\ell^0_n+(\gamma_2-\gamma^0)D^1_{\gamma_2}\ell^0_n\big| = \mathrm{O}_{\mathbb{P}_{\theta_0}}(1). \tag{35}$$

Also,

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

Using (33) 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

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

and

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

Thus, we obtain

$$\sup_{\theta\in U_n}\big|S_n(\theta)\big| = \mathrm{O}_{\mathbb{P}_{\theta_0}}(1). \tag{36}$$

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

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

so that using assumptions (14)

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

with

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

Now using (33) one gets that for all integer $k \geq 1$,

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

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

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

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 $C > 0$,

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

so that we finally obtain

$$\sup_{\theta\in U_n}\bigl|T_n(\theta)\bigr| = \mathrm{O}_{\mathbb{P}_{\theta_0}}(1). \tag{37}$$

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

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

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

$$\sup_{\theta\in U_n}\left|\prod_{i=1}^4\prod_{j=0}^i \bigl(D^i_{(\gamma_1)^j,(\gamma_2)^{i-j}} p_k(\theta_t)\bigr)^{a_{i,j}}\right|, \tag{38}$$

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

To prove that

$$\sup_{\theta\in U_n}\bigl|R_n(\theta,t)\bigr| = \mathrm{O}_{\mathbb{P}_{\theta_0}}(1) \tag{39}$$

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

and

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

for some finite $C > 0$. Now, using (33) one gets that for all integer $k \ge 1$ and for any θ,

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

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

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

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 $j = 0, \dots, i$, any $k \ge 2$,

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

By recursion, we obtain that there exists a finite $C > 0$ such that any term of form (38) has expectation uniformly bounded:

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

which concludes the proof of (39). Now, using (32), (34), (35), (36), (37) and (39), we get

$$D_n \geq \mathrm{e}^{-\mathrm{O}_{\mathbb{P}_{\theta_0}}(1)} \Pi_2(U_n)$$

so that (20) holds with S_n satisfying (29).

4.4. Proof of Theorem 4

The proof follows the same lines as in Ghosal and van der Vaart [10]. We write

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

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 $\Omega_n = \{(\theta, x); \ell_n(\theta, x) - \ell_n(\theta_0, x_0) \geq -n\tilde{\epsilon}_n^2\}$, which is a random subset of $\Theta \times \mathcal{X}$ (depending on $Y_{1:n}$),

$$\begin{aligned}D_n &\geq \int_{S_n} \mathbb{1}_{\Omega_n} \mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x) \\ &\geq \mathrm{e}^{-n\tilde{\epsilon}_n^2} \Pi(S_n \cap \Omega_n),\end{aligned}$$

therefore

$$\begin{aligned}\mathbb{P}_{\theta_0}\left[D_n < \mathrm{e}^{-n\tilde{\epsilon}_n^2} \Pi(S_n)/2\right] &\leq \mathbb{P}_{\theta_0}\left[\Pi\left(S_n \cap \Omega_n^c\right) \geq \Pi(S_n)/2\right] \\ &\leq 2 \frac{\int_{S_n} \mathbb{P}_{\theta_0}[\ell_n(\theta, x) - \ell_n(\theta_0, x_0) \leq -n\tilde{\epsilon}_n^2] \Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)}{\Pi(S_n)} \\ &= \mathrm{o}(1)\end{aligned}$$

and

$$\mathbb{P}^{\Pi}\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n \middle| Y_{1:n}\right] = \mathrm{o}_{\mathbb{P}_{\theta_0}}(1) + \frac{N_n}{D_n} \mathbb{1}_{2D_n \geq \mathrm{e}^{-n\tilde{\epsilon}_n^2}\Pi(S_n)}.$$

But

$$\begin{aligned}N_n = &\int_{(A_n \cap \mathcal{F}_n) \times \mathcal{X}} \mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x) \\ &+ \int_{(A_n \cap \mathcal{F}_n^c) \times \mathcal{X}} \mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)\end{aligned}$$

and

$$E_{\theta_0}\left[\int_{(A_n\cap\mathcal{F}_n^c)\times\mathcal{X}} \mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)\right]$$
$$= \mathrm{O}\big[\Pi_\Theta\big(A_n\cap\mathcal{F}_n^c\big)\big] = \mathrm{o}\big(\mathrm{e}^{-n\tilde{\epsilon}_n^2(C_n+1)}\big)$$

by Fubini's theorem and assumption **C2** together with the fact that $\ell_n(\theta_0) - \ell_n(\theta_0, x_0)$ is uniformly upper bounded. This implies using assumption **C1** that

$$\mathbb{P}^\Pi\left[\|f_{l,\theta} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n \middle| Y_{1:n}\right] = \mathrm{o}_{\mathbb{P}_{\theta_0}}(1) + \frac{\tilde{N}_n}{D_n}\mathbb{1}_{2D_n \geq \mathrm{e}^{-n\tilde{\epsilon}_n^2}\Pi(S_n)}, \tag{40}$$

where $\tilde{N}_n = \int_{(A_n\cap\mathcal{F}_n)\times\mathcal{X}} \mathrm{e}^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)}\Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x)$. Let now $(\theta_j)_{j=1,\dots,N}$, $N = N(\delta, \mathcal{F}_n, d_l(\cdot,\cdot))$, 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 $d_l(\theta_j, \theta) \leq \delta$ with $\delta = \epsilon_n/12$. Assume for simplicity's sake and without loss of generality that n is a multiple of the integer l, and define

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

where

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

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

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

Define also

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

Then

$$E_{\theta_0}\left(\frac{\tilde{N}_n}{D_n}\psi_n\right) \leq E_{\theta_0}\psi_n \leq N\big(\delta, \mathcal{F}_n, d(\cdot,\cdot)\big) \max_{1\leq j\leq N:\theta_j\in A_n} E_{\theta_0}\phi_j \tag{41}$$

and

$$\begin{aligned} E_{\theta_0}\big(\tilde{N}_n(1-\psi_n)\big) &= \int_{\mathcal{X}} E_{\theta_0,x_0}\big(\tilde{N}_n(1-\psi_n)\big)\mu_{\theta_0}(\mathrm{d}x_0) \\ &= \int_{(A_n\cap\mathcal{F}_n)\times\mathcal{X}} E_{\theta,x}\big((1-\psi_n)\big)\Pi_\Theta(\mathrm{d}\theta)\pi_{\mathcal{X}}(\mathrm{d}x). \end{aligned} \tag{42}$$

Now

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

and

$$
\begin{aligned}
&E_{\theta,x}(1-\phi_j) \\
&\quad = \mathbb{P}_{\theta,x}\Biggl[\sum_{i=1}^{n/l}\bigl(-\mathbb{1}_{(Y_{li-l+1},\dots,Y_{li})\in A_j} + \mathbb{P}_{\theta,x}\bigl((Y_{li-l+1},\dots,Y_{li})\in A_j\bigr)\bigr) \\
&\qquad\qquad > -t_j + \sum_{i=1}^{n/l}\bigl(\mathbb{P}_{\theta,x}\bigl((Y_{li-l+1},\dots,Y_{li})\in A_j\bigr) - \mathbb{P}_{\theta_0}\bigl((Y_1,\dots,Y_l)\in A_j\bigr)\bigr)\Biggr].
\end{aligned}
$$

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

$$
\begin{aligned}
&\bar{Q}_\theta\bigl(z,\mathrm{d}z'\bigr) \\
&\quad = g_\theta\bigl(y_1'|x_1'\bigr)\cdots g_\theta\bigl(y_l'|x_l'\bigr)Q_\theta\bigl(x_l,\mathrm{d}x_1'\bigr)Q_\theta\bigl(x_1',\mathrm{d}x_2'\bigr)\cdots Q_\theta\bigl(x_{l-1}',\mathrm{d}x_l'\bigr)\mu\bigl(\mathrm{d}y_1'\bigr)\cdots\mu\bigl(\mathrm{d}y_l'\bigr).
\end{aligned}
$$

This kernel satisfies the same uniform ergodic property as Q_θ, with the same coefficients, that is condition (17) holds with the coefficients R_θ and ρ_θ with the replacement of Q_θ by $\bar{Q}_\theta$, and we may use Rio's [18] exponential inequality (Corollary 1) with uniform mixing coefficients (as defined in Rio [18]) satisfying $\phi(m) \le R_\theta\rho_\theta^{-m}$. Indeed, by the Markov property,

$$
\begin{aligned}
\phi(m) &= \sup_{A\in\sigma(Z_1),B\in\sigma(Z_{m+1})}\bigl(\mathbb{P}_\theta(B)-\mathbb{P}_\theta(B|A)\bigr) \\
&\le \sup_z\bigl|\mathbb{P}_\theta(Z_{m+1}\in B)-\mathbb{P}_\theta(Z_{m+1}\in B|Z_1=z)\bigr| \\
&\le R_\theta\rho_\theta^{-m}.
\end{aligned}
$$

We thus obtain that, for any positive real number u,

$$
\begin{aligned}
&\mathbb{P}_{\theta_0}\Biggl[\sum_{i=1}^{n/l}\bigl(\mathbb{1}_{(Y_{li-l+1},\dots,Y_{li})\in A_j} - \mathbb{P}_{\theta_0}\bigl((Y_1,\dots,Y_l)\in A_j\bigr)\bigr) > u\Biggr] \\
&\quad \le \exp\biggl\{\frac{-2lu^2(\rho_{\theta_0}-1)^2}{n(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\biggr\}
\end{aligned}
\tag{43}
$$

and

$$
\begin{aligned}
&\mathbb{P}_{\theta,x}\Biggl[\sum_{i=1}^{n/l}\bigl(-\mathbb{1}_{(Y_{li-l+1},\dots,Y_{li})\in A_j} + \mathbb{P}_{\theta,x}\bigl((Y_{li-l+1},\dots,Y_{li})\in A_j\bigr)\bigr) > u\Biggr] \\
&\quad \le \exp\biggl\{\frac{-2lu^2(\rho_\theta-1)^2}{n(2R_\theta+\rho_\theta-1)^2}\biggr\}.
\end{aligned}
\tag{44}
$$

Set now

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

Since for any θ, $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \le 1$ and since consequently for $\theta_j \in A_n$, $\|f_{l,\theta_j} - f_{l,\theta_0}\|_1 \ge \epsilon_n$, we first get, using (43),

$$E_{\theta_0}[\phi_j] \le \exp\left\{\frac{-n\epsilon_n^2(\rho_{\theta_0} - 1)^2}{8l(2R_{\theta_0} + \rho_{\theta_0} - 1)^2}\right\}. \tag{45}$$

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

$$\begin{aligned}
&-t_j + \sum_{i=1}^{n/l}\bigl(\mathbb{P}_{\theta,x}\bigl((Y_{li-l+1},\dots,Y_{li}) \in A_j\bigr) - \mathbb{P}_{\theta_0}\bigl((Y_1,\dots,Y_l) \in A_j\bigr)\bigr) \\
&\quad = -\frac{n\|f_{l,\theta_j} - f_{l,\theta_0}\|_1}{4l} + \frac{n}{l}\bigl\{\mathbb{P}_{\theta_j}\bigl((Y_1,\dots,Y_l) \in A_j\bigr) - \mathbb{P}_{\theta_0}\bigl((Y_1,\dots,Y_l) \in A_j\bigr)\bigr\} \\
&\qquad + \frac{n}{l}\bigl\{\mathbb{P}_{\theta}\bigl((Y_1,\dots,Y_l) \in A_j\bigr) - \mathbb{P}_{\theta_j}\bigl((Y_1,\dots,Y_l) \in A_j\bigr)\bigr\} \\
&\qquad + \sum_{i=1}^{n/l}\bigl(\mathbb{P}_{\theta,x}\bigl((Y_{li-l+1},\dots,Y_{li}) \in A_j\bigr) - \mathbb{P}_{\theta}\bigl((Y_1,\dots,Y_l) \in A_j\bigr)\bigr) \\
&\quad \ge \frac{n\|f_{l,\theta_j} - f_{l,\theta_0}\|_1}{4l} - \frac{n\|f_{l,\theta_j} - f_{l,\theta}\|_1}{l} - \sum_{i=1}^{n/l} R_\theta \rho_\theta^{-i} \\
&\quad \ge \frac{n\|f_{l,\theta_j} - f_{l,\theta_0}\|_1}{4l} - \frac{n\|f_{l,\theta_j} - f_{l,\theta}\|_1}{l} - \frac{R_\theta \rho_\theta}{\rho_\theta - 1} \\
&\quad \ge \frac{n}{4l}\left(1 - \frac{5}{12} - \frac{4l}{12n\epsilon_n}\right)\|f_{l,\theta} - f_{l,\theta_0}\|_1 \ge \frac{n}{8l}\|f_{l,\theta} - f_{l,\theta_0}\|_1
\end{aligned}$$

for large enough n, using the triangular inequality and the fact that $\|f_{l,\theta_j} - f_{l,\theta}\|_1 \le \frac{\epsilon_n}{12} \le \frac{\|f_{l,\theta} - f_{l,\theta_0}\|_1}{12}\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1}$ since $\theta \in A_n$ and $\frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \le 1$. Then for $\theta \in A_n$ and large enough n,

$$E_{\theta,x}(1 - \phi_j) \le \exp\left\{-\frac{n\epsilon_n^2}{32l}\right\}. \tag{46}$$

Combining (40), with (41), (45), (42), (46) and using assumptions **C1** and **C3** we finally obtain for large enough n

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\left(\mathbb{P}^{\Pi}\left[\|f_{l,\theta_j} - f_{l,\theta_0}\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \ge \epsilon_n \middle| Y_{1:n}\right]\right) \\
&\quad \le \mathrm{o}(1) + \mathrm{O}\bigl(\mathrm{e}^{n\tilde{\epsilon}_n^2(1+C_n)}\bigr)\exp\left\{-\frac{n\epsilon_n^2}{32l}\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\exp\left\{\frac{-n\epsilon_n^2(\rho_{\theta_0}-1)^2}{8l(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\right\}\exp\left\{\frac{n\epsilon_n^2(\rho_{\theta_0}-1)^2}{16l(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\right\}\\
&=\mathrm{o}(1).
\end{aligned}
$$

Assume now that assumption **C3bis** holds. By writing $A_n \cap \mathcal{F}_n = \bigcup_{m\geq 1} A_{n,m}(\epsilon_n)$ and using same reasoning, one gets, for some positive constant c:

$$
\begin{aligned}
&\mathbb{P}_{\theta_0}\left(\mathbb{P}^{\Pi}\left[\|f_{l,\theta_j}-f_{l,\theta_0}\|_1\frac{\rho_\theta-1}{2R_\theta+\rho_\theta-1}\geq\epsilon_n\middle|Y_{1:n}\right]\right)\\
&\quad=\mathrm{o}(1)+\mathrm{e}^{n\tilde{\epsilon}_n^2}\sum_{m\geq 1}\frac{\Pi_\Theta(A_{n,m}(\epsilon_n))}{\Pi(S_n)}\exp\left\{-\frac{nm^2\epsilon_n^2}{32l}\right\}\\
&\qquad+\sum_{m\geq 1}N\left(\frac{m\epsilon_n}{12},A_{n,m}(\epsilon_n),d_l(\cdot,\cdot)\right)\exp\left\{-\frac{nm^2\epsilon_n^2(\rho_{\theta_0}-1)^2}{8l(2R_{\theta_0}+\rho_{\theta_0}-1)^2}\right\}\\
&\quad=\mathrm{o}(1)
\end{aligned}
$$

and the second part of Theorem 4 is proved.

Acknowledgement

The authors would like to thank the anonymous referees for providing constructive comments that where very helpful to improve the readability of the paper. This work was partly supported by the 2010–2014 grant ANR Banhdits AAP Blanc SIMI 1.

References

[1] Boys, R.J. and Henderson, D.A. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics* **60** 573–588. With discussions and a reply by the author. MR2089432

[2] Cappé, O., Moulines, E. and Rydén, T. (2004). *Hidden Markov Models*. New York: Springer.

[3] Chambaz, A., Garivier, A. and Gassiat, E. (2009). A minimum description length approach to hidden Markov models with Poisson and Gaussian emissions. Application to order identification. *J. Statist. Plann. Inference* **139** 962–977. MR2479841

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

[5] Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32** 2254–2304. MR2102510

[6] Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 897–906. MR1955343

[7] Gassiat, E. and Boucheron, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory* **49** 964–980. MR1984482

[8] Gassiat, E. and van Handel, R. (2014). The local geometry of finite mixtures. *Trans. Amer. Math. Soc.* **366** 1047–1072. MR3130325

[9] Gassiat, E. and Keribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM Probab. Statist.* **4** 25–52. MR1780964
[10] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274
[11] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. New York: Springer. MR1992245
[12] Green, P.J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.* **97** 1055–1070. MR1951259
[13] Leroux, B. and Putterman, M. (1992). Maximum-penalised-likelihood estimation for independent and Markov dependent mixture models. *Biometrics* **48** 545–558.
[14] MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. *Monographs on Statistics and Applied Probability* **70**. London: Chapman & Hall. MR1692202
[15] McGrory, C.A. and Titterington, D.M. (2009). Variational Bayesian analysis for hidden Markov models. *Aust. N. Z. J. Stat.* **51** 227–244. MR2531988
[16] Nur, D., Allingham, D., Rousseau, J., Mengersen, K.L. and McVinish, R. (2009). Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis. *Comput. Statist. Data Anal.* **53** 1873–1882. MR2649552
[17] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213
[18] Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.* **330** 905–908. MR1771956
[19] Robert, C.P., Rydén, T. and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 57–75. MR1747395
[20] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 689–710. MR2867454
[21] Rydén, T., Terasvirta, T. and Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometrics* **13** 217–244.
[22] Spezia, L. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *J. Time Series Anal.* **31** 1–11. MR2640842
[23] Zucchini, W. and MacDonald, I.L. (2009). *Hidden Markov Models for Time Series*: *An Introduction Using R*. *Monographs on Statistics and Applied Probability* **110**. Boca Raton, FL: CRC Press. MR2523850

Received March 2013 and revised July 2013