Bernoulli **22**(1), 2016, 193–212
DOI: 10.3150/14-BEJ631

Nonparametric finite translation hidden Markov models and extensions

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

In this paper, we consider nonparametric finite translation hidden Markov models, or more generally finite translation mixtures with dependent latent variables. We prove that all the parameters of the model are identifiable as soon as the matrix that defines the joint distribution of two consecutive latent variables is non-singular and the translation parameters are distinct. Under this assumption, we provide a consistent estimator of the number of populations, of the translation parameters and of the distribution of two consecutive latent variables, which we prove to be asymptotically normally distributed under mild dependency assumptions. We propose a nonparametric estimator of the unknown translated density. In case the latent variables form a Markov chain, we prove that this estimator is minimax adaptive over regularity classes of densities.

Keywords: dependent latent variable models; hidden Markov models; nonparametric estimation; translation mixtures

1. Introduction

Finite state space hidden Markov models (shortened as HMMs throughout the paper) were introduced to model data coming from heterogeneous populations when the observed phenomenon is driven by a latent Markov chain. They may also be seen as a dynamic extension of finite mixture models. Such models may be described as follows. Consider a sequence $(S_i)_{i\in\mathbb{N}}$ of latent (unobserved) random variables with finite state space $\{1,\dots,k\}$, and a sequence $(Y_i)_{i\in\mathbb{N}}$ of random variables (the observations) such that, conditionally to $(S_i)_{i\in\mathbb{N}}$, the Y_i's are independently distributed, and for each i, the distribution of Y_i depends only on the current latent variable S_i. The latent variables may be interpreted as the labelling of the population the observation comes from. Such a model may also be phrased as a mixture model with dependent regime, and in case the sequence $(S_i)_{i\in\mathbb{N}}$ forms a Markov chain, this defines a HMM.

To be able to infer about the population structures, one usually states parametric models, since in general, it is not possible to recover individual distributions from a convex combination of them without additional information. See, for instance, McLachlan and Peel [24] or Marin *et al.* [20] for a review of mixture models' methods, and Cappé *et al.* [9] for a recent state of the art concerning HMMs. But parametric modeling may lead to poor results in particular applications (see, for instance, the discussion on the Old faithfull dataset in Azzaline and Bowman [3]), and some nonparametric procedures have been considered in applied papers, for example, in climate state identification in Lambert *et al.* [19] or for copy number variants identification in DNA analysis,

for which a nonparametric hidden Markov model has been proposed in Yau *et al.* [30]. However, no theoretical result has been proved until now to validate those nonparametric procedures.

The aim of this paper is to provide a full theoretical justification of the use of nonparametric methods in the case of finite translation HMMs (which is used in Yau *et al.* [30]) or more generally finite translated mixtures with dependent latent variables. We consider location models

$$Y_i = m_{S_i} + \varepsilon_i, \qquad i \in \mathbb{N}, \tag{1}$$

where $(\varepsilon_i)_{i\in\mathbb{N}}$ is a sequence of independent identically distributed random variables taking values in $\mathbb{R}$, and $m_j \in \mathbb{R}$, $j = 1, \ldots, k$. The aim is to estimate the parameters k, $m_1, \ldots, m_k$, the distribution of the latent variables $(S_i)_{i\in\mathbb{N}}$ and the distribution F of the ε_i's. As usual for finite mixtures, one may recover the parameters only up to relabelling, and obviously, F may only be estimated up to a translation (that would be reversely reported to the m_j's).

Our most important result here is that in case the latent variables are *not* independent; model (1) is identifiable without any assumption on F. To be precise, if Q is the $k \times k$-matrix such that $Q_{i,j}$ is the probability that $S_1 = i$ and $S_2 = j$, we prove that the knowledge of the distribution of (Y_1, Y_2) allows the identification of k, $m_1, \ldots, m_k$, Q and F as soon as Q is a non-singular matrix, whatever F may be; see Theorem 1.

This identifiability result may seem surprising, since it is obvious that for independent variables, such a result does not hold. Indeed, with independent observations one has only access to the marginal distribution of Y_1 which is given by

$$P_{\mu,F}(\cdot) = \sum_{j=1}^{k} \mu(j) F(\cdot - m_j). \tag{2}$$

Here, $\mu(j) = P(S_1 = j)$, $j = 1, \ldots, k$. An equivalent representation of (2) corresponds, for instance, to $k = 1$, $m_1 = 0$ and $F = P_{\mu,F}$ the marginal distribution. Thus, to be able to infer about the mixture model in case of independent observations, one needs further restrictive assumptions. Hunter *et al.* [17] have considered model (2) with the additional assumption that F is symmetrical; see also Bordes *et al.* [7] and Butucea and Vandekerkhove [8].

To obtain our identifiability result, we take advantage of the joint distribution of (Y_1, Y_2) under model (1), which is given by

$$P_{\theta,F}(A \times B) = \sum_{i=1}^{k} \sum_{i=1}^{k} Q_{i,j} F(A - m_i) F(B - m_j), \qquad \forall A, B \in \mathcal{B}_{\mathbb{R}}, \tag{3}$$

where $\mathcal{B}_{\mathbb{R}}$ denotes the Borel σ field of $\mathbb{R}$ and $\theta = (m, (Q_{i,j})_{1\le i,j\le k,(i,j)\neq(k,k)})$, with $m = (m_1, \ldots, m_k) \in \mathbb{R}^k$. Notice that here, one cannot use recent results about mixtures with repeated measurements. Indeed, one could interpret (3) as a mixture with k^2 components in two dimensions, but the linear independence required for identifiability in Allman *et al.* [1] never holds. Independent mixtures with repeated measurements (or multivariate) have received recent interest since they were proved to be nonparametrically identifiable under some structural assumptions. See Hall and Zhou [15], Kasahara and Shimotsu [18], Allman *et al.* [1], Bonhomme *et al.* [5] and references therein; see also Henry *et al.* [16]. An extension of our identifiability Theorem 1

has recently been obtained by one of the authors in Gassiat *et al.* [13], in the restricted context of hidden Markov models with known number of states.

Building upon our identifiability result, we propose an estimator of k, and of the parametric part of the distribution, namely Q and $m_1, \dots, m_k$. Moreover, we prove that our estimator is $\sqrt{n}$-consistent, with asymptotic Gaussian distribution, under mild dependency assumptions; see Theorem 2. When the number of populations is known and if the translation parameters m_j, $j \leq k$ are known to be bounded by a given constant, we prove that the estimator (centered and at $\sqrt{n}$-scale) has a sub-Gaussian distribution; see Theorem 3.

As soon as Q and $m_1, \dots, m_k$ are consistently estimated, one may imagine various nonparametric estimation methods for F. In the context of hidden Markov models as considered in Yau *et al.* [30], we propose an estimator of the nonparametric part of the distribution, namely F, assuming that it is absolutely continuous with respect to Lebesgue measure. This estimator uses the model selection approach developed in Massart [21], with the penalized estimated pseudo likelihood contrast based on marginal densities $\sum_{j=1}^{k} \widehat{\mu}(j) f(y - \widehat{m}_j)$. We prove that our nonparametric estimator is adaptive over regular classes of densities; see Theorem 4 and Corollary 1.

The paper is organized as follows. In Section 2, we present and prove our general identifiability theorem. In Section 3, we define an estimator of the order and of the parametric part, and state the convergence results: asymptotic Gaussian distribution and deviation inequalities. In Section 4, we explain our nonparametric estimator of the density of F using model selection methods, and state an oracle inequality and adaptive convergence results. Most of the proofs are postponed either to the Appendix, for the first 3 sections or to the supplementary material Gassiat and Rousseau [14] for the last section.

2. General identifiability result

Let $\mathcal{Q}_k$ be the set of probability mass functions on $\{1, \dots, k\}^2$, that is the set of $k \times k$ matrices $Q = (Q_{i,j})_{1 \leq i,j \leq k}$ such that for all $(i, j) \in \{1, \dots, k\}^2$, $Q_{i,j} \geq 0$, and $\sum_{i=1}^{k} \sum_{j=1}^{k} Q_{i,j} = 1$. With $\theta = (m, (Q_{i,j})_{1 \leq i,j \leq k, (i,j) \neq (k,k)})$, $m = (m_1, \dots, m_k) \in \mathbb{R}^k$, recall that $P_{\theta,F}$ is the joint distribution of (Y_1, Y_2) under model (1). In this case, ordering the coefficients $m_1 \leq m_2 \leq \cdots \leq m_k$ and replacing F by $F(\cdot - m_1)$ leads to the same model so that without loss of generality we fix $0 = m_1 \leq m_2 \leq \cdots \leq m_k$. Let Θ_k be the set of parameters θ such that $m_1 = 0 \leq m_2 \leq \cdots \leq m_k$ and $Q \in \mathcal{Q}_k$, where $Q = (Q_{i,j})_{1 \leq i,j \leq k}$, $Q_{k,k} = 1 - \sum_{(i,j) \neq (k,k)} Q_{i,j}$.

Let also Θ_k^0 be the set of parameters $\theta = (m, (Q_{i,j})_{1 \leq i,j \leq k, (i,j) \neq (k,k)}) \in \Theta_k$ such that $m_1 = 0 < m_2 < \cdots < m_k$ and $\det(Q) \neq 0$. We then have the following result on the identification of F and θ from $P_{\theta,F}$.

Theorem 1. *Let F and $\tilde{F}$ be any probability distributions on $\mathbb{R}$. Let k and $\tilde{k}$ be positive integers. If $\theta \in \Theta_k^0$ and $\tilde{\theta} \in \Theta_{\tilde{k}}^0$, then*

$$P_{\theta,F} = P_{\tilde{\theta},\tilde{F}} \quad \Longrightarrow \quad k = \tilde{k}, \theta = \tilde{\theta} \text{ and } F = \tilde{F}.$$

Remark 1. In the same way, it is possible to identify ℓ-marginals, for any $\ell \geq 2$, that is the distribution of $(S_1, \ldots, S_\ell)$, m and F on the basis of the distribution of $(Y_1, \ldots, Y_\ell)$.

Remark 2. An important class of models is that of hidden Markov models. In that case, if Q is the stationary distribution of two consecutive variables of the hidden Markov chain, $\det(Q) \neq 0$ if and only if the transition matrix is non-singular and the stationary distribution gives positive weights to each point. When $k = 2$, we thus have $\det(Q) \neq 0$ if and only if S_1 and S_2 are not independent.

Remark 3. The independent case is a special case where $\det(Q) = 0$. In the independent case, to get identifiable models, further assumptions are needed; see Hunter *et al.* [17], Bordes *et al.* [7], Butucea and Vandekerkhove [8] where symmetry of F is required.

Proof of Theorem 1. Denote by ϕ_F the characteristic function of F, $\phi_{\tilde{F}}$ the characteristic function of $\tilde{F}$, $\phi_{\theta,1}$ (respectively, $\phi_{\tilde{\theta},1}$) the characteristic function of the distribution of m_{S_1} under $P_{\theta,F}$ (respectively, under $P_{\tilde{\theta},\tilde{F}}$), $\phi_{\theta,2}$ (respectively, $\phi_{\tilde{\theta},2}$) the characteristic function of the distribution of m_{S_2} under $P_{\theta,F}$ (respectively, under $P_{\tilde{\theta},\tilde{F}}$), and Φ_θ (respectively, $\Phi_{\tilde{\theta}}$) the characteristic function of the distribution of (m_{S_1}, m_{S_2}) under $P_{\theta,F}$ (respectively, under $P_{\tilde{\theta},\tilde{F}}$). Then since the distribution of Y_1 is the same under $P_{\theta,F}$ and $P_{\tilde{\theta},\tilde{F}}$, one gets that for any $t \in \mathbb{R}$,

$$\phi_F(t)\phi_{\theta,1}(t) = \phi_{\tilde{F}}(t)\phi_{\tilde{\theta},1}(t). \tag{4}$$

Similarly, for any $t \in \mathbb{R}$,

$$\phi_F(t)\phi_{\theta,2}(t) = \phi_{\tilde{F}}(t)\phi_{\tilde{\theta},2}(t). \tag{5}$$

Since the distribution of (Y_1, Y_2) is the same under $P_{\theta,F}$ and $P_{\tilde{\theta},\tilde{F}}$, one gets that for any $\mathbf{t} = (t_1, t_2) \in \mathbb{R}^2$,

$$\phi_F(t_1)\phi_F(t_2)\Phi_\theta(\mathbf{t}) = \phi_{\tilde{F}}(t_1)\phi_{\tilde{F}}(t_2)\Phi_{\tilde{\theta}}(\mathbf{t}). \tag{6}$$

There exists a neighborhood V of 0 such that for all $t \in V$, $\phi_F(t) \neq 0$, so that (4), (5) and (6) imply that for any $\mathbf{t} = (t_1, t_2) \in V^2$,

$$\Phi_\theta(\mathbf{t})\phi_{\tilde{\theta},1}(t_1)\phi_{\tilde{\theta},2}(t_2) = \Phi_{\tilde{\theta}}(\mathbf{t})\phi_{\theta,1}(t_1)\phi_{\theta,2}(t_2). \tag{7}$$

Let t_1 be a fixed real number in V. $\Phi_\theta(t_1, t_2)$, $\phi_{\tilde{\theta},2}(t_2)$, $\Phi_{\tilde{\theta}}(t_1, t_2)$, $\phi_{\theta,2}(t_2)$ have analytic continuations for all complex numbers z_2, $\Phi_\theta(t_1, z_2)$, $\phi_{\tilde{\theta}}(z_2)$, $\Phi_{\tilde{\theta}}(t_1, z_2)$, $\phi_\theta(z_2)$ which are entire functions so that (7) holds with z_2 in place of t_2 for all z_2 in the complex plane $\mathbb{C}$ and any $t_1 \in V$. Again, let z_2 be a fixed complex number in $\mathbb{C}$. $\Phi_\theta(t_1, z_2)$, $\phi_{\tilde{\theta},1}(t_1)$, $\Phi_{\tilde{\theta}}(t_1, z_2)$, $\phi_{\theta,1}(t_1)$ have analytic continuations $\Phi_\theta(z_1, z_2)$, $\phi_{\tilde{\theta}}(z_1)$, $\Phi_{\tilde{\theta}}(z_1, z_2)$, $\phi_\theta(z_1)$ which are entire functions so that (7) holds with z_1 in place of t_1 and z_2 in place of t_2 for all $(z_1, z_2) \in \mathbb{C}^2$.

Let now $\mathcal{Z}$ be the set of zeros of $\phi_{\theta,1}$, $\tilde{\mathcal{Z}}$ be the set of zeros of $\phi_{\tilde{\theta},1}$ and fix $z_1 \in \mathcal{Z}$. Then, for any $z_2 \in \mathbb{C}$,

$$\Phi_\theta(z_1, z_2)\phi_{\tilde{\theta},1}(z_1)\phi_{\tilde{\theta},2}(z_2) = 0. \tag{8}$$

We now prove that $z_2 \to \Phi_\theta(z_1, \cdot)$ is not the null function. For any $z \in \mathbb{C}$,

$$\Phi_\theta(z_1, z) = \sum_{\ell=1}^{k} \left[\sum_{j=1}^{k} Q_{\ell,j} \mathrm{e}^{\mathrm{i}m_j z_1} \right] \mathrm{e}^{\mathrm{i}m_\ell z}.$$

Since $0 = m_1 < m_2 < \cdots < m_k$, if $\Phi_\theta(z_1, \cdot)$ was the null function, we would have for all $\ell = 1, \ldots, k$

$$\sum_{j=1}^{k} Q_{\ell,j} \mathrm{e}^{\mathrm{i}m_j z_1} = 0,$$

which is impossible since $\det(Q) \neq 0$. Thus, $\Phi_\theta(z_1, \cdot)$ is an entire function which has isolated zeros, $\phi_{\tilde{\theta},2}(\cdot)$ also, and it is possible to choose z_2 in $\mathbb{C}$ such that $\Phi_\theta(z_1, z_2) \neq 0$ and $\phi_{\tilde{\theta},2}(z_2) \neq 0$. Then (8) leads to $\phi_{\tilde{\theta},1}(z_1) = 0$, so that $\mathcal{Z} \subset \tilde{\mathcal{Z}}$. A symmetric argument gives $\tilde{\mathcal{Z}} \subset \mathcal{Z}$ so that $\mathcal{Z} = \tilde{\mathcal{Z}}$. Moreover, $\phi_{\theta,1}$ and $\phi_{\tilde{\theta},1}$ have growth order 1, so that using Hadamard's factorization theorem (see [27], Theorem 5.1), one gets that there exists a polynomial R of degree ≤ 1 such that for all $z \in \mathbb{C}$,

$$\phi_{\theta,1}(z) = \mathrm{e}^{R(z)} \phi_{\tilde{\theta},1}(z).$$

But using $\phi_{\theta,1}(0) = \phi_{\tilde{\theta},1}(0) = 1$, we get that there exists a complex number a such that $\phi_{\tilde{\theta},1}(z) = \mathrm{e}^{az} \phi_{\theta,1}(z)$. Since for all $z \in \mathbb{R}$, $\phi_{\theta',1}(-z) = \bar{\phi}_{\theta',1}(z)$, there exists $r \in \mathbb{R}$ such that $a = ir$, and $k = \tilde{k}$. Using $m_1 < m_2 < \cdots < m_k$ and $\tilde{m}_1 < \tilde{m}_2 < \cdots < \tilde{m}_{\tilde{k}}$, we get $\tilde{m}_j = m_j + r$, $j = 1, \ldots, k$. Using now $0 = m_1 = \tilde{m}_1$, we get $r = 0$ so that $\phi_{\theta,1} = \phi_{\tilde{\theta},1}$. Similar arguments lead to $\phi_{\theta,2} = \phi_{\tilde{\theta},2}$. Combining this with (7), we obtain $\Phi_\theta = \Phi_{\tilde{\theta}}$ which in turns implies $\theta = \tilde{\theta}$. Thus, using (4), for all $t \in \mathbb{R}$ such that $\phi_{\theta,1}(t) \neq 0$, $\phi_F(t) = \phi_{\tilde{F}}(t)$. Since $\phi_{\theta,1}$ has isolated zeros and ϕ_F, $\phi_{\tilde{F}}$ are continuous functions, one gets $\phi_F = \phi_{\tilde{F}}$ so that $F = \tilde{F}$. □

3. Estimation of the parametric part

3.1. Assumptions on the model

Hereafter, we are given a stationary sequence $(Y_i)_{i \in \mathbb{N}}$ of real random variables with distribution $\mathbb{P}^\star$. We assume that (1) holds, with $(S_i)_{i \in \mathbb{N}}$ a stationary sequence of non-observed random variables taking values in $\{1, \ldots, k^\star\}$. We denote by $F^\star$ the common probability distribution of the ε_i's, and $m^\star \in \mathbb{R}^{k^\star}$ the possible values of the m_{S_i}'s. Let $Q^\star \in \mathcal{Q}_{k^\star}$ be the distribution of (S_1, S_2), and $\theta^\star = (m^\star, (Q^\star_{i,j})_{(i,j) \neq (k^\star, k^\star)})$. We assume:

(A1) $\theta^\star \in \Theta^0_{k^\star}$, and all differences $m^\star_j - m^\star_i$, $i, j = 1, \ldots, k^\star$, $i \neq j$, are distinct.

We do not assume that $k^\star$ is known, so that the aim is to estimate $\theta^\star$ and $k^\star$ altogether. Stationarity implies that the marginal distributions in $Q^\star$ are identical so that we write from now on $\phi_{\theta^\star} = \phi_{\theta^\star,1} = \phi_{\theta^\star,2}$.

The idea to estimate $\theta^\star$ and k^* is to use equation (7) which holds if and only if the parameters are equal. Consider w any probability density on $\mathbb{R}^2$ with compact support $\mathcal{S}$, positive on $\mathcal{S}$ and with 0 belonging to the interior of $\mathcal{S}$; typically $\mathcal{S} = [-a, a]^2$ for some positive a. Define, for any integer k and $\theta \in \Theta_k$,

$$M(k,\theta) = \int_{\mathbb{R}^2} \left|\Phi_{\theta^\star}(t_1,t_2)\phi_{\theta,1}(t_1)\phi_{\theta,2}(t_2) - \Phi_\theta(t_1,t_2)\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\right|^2 \times \left|\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\right|^2 w(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2. \tag{9}$$

We shall use $M(k,\theta)$ as a contrast function. Indeed, thanks to Theorem 1, $\theta \in \Theta_k^0$ is such that $M(k,\theta) = 0$ if and only if $k = k^\star$ and $\theta = \theta^\star$.

We estimate $M(k,\cdot)$ by

$$M_n(k,\theta) = \int_{\mathbb{R}^2} \left|\widehat{\Phi}_n(t_1,t_2)\phi_{\theta,1}(t_1)\phi_{\theta,2}(t_2) - \Phi_\theta(t_1,t_2)\widehat{\phi}_{n,1}(t_1)\widehat{\phi}_{n,2}(t_2)\right|^2 w(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2, \tag{10}$$

where

$$\widehat{\Phi}_n(t_1,t_2) = \frac{1}{n}\sum_{j=1}^{n-1} \exp i(t_1 Y_j + t_2 Y_{j+1}), \tag{11}$$

$\widehat{\phi}_{n,1}(t) = \widehat{\Phi}_n(t,0)$ and $\widehat{\phi}_{n,2}(t) = \widehat{\Phi}_n(0,t)$. Define, for any $\mathbf{t} = (t_1,t_2) \in \mathbb{R}^2$

$$Z_n(\mathbf{t}) = \sqrt{n}\big(\widehat{\Phi}_n(\mathbf{t}) - \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\big).$$

Our main assumptions on the model are the following.

(A2) The process $(Z_n(\mathbf{t}))_{\mathbf{t}\in\mathcal{S}}$ converges weakly to a Gaussian process $(Z(\mathbf{t}))_{\mathbf{t}\in\mathcal{S}}$ in the set of complex continuous functions on $\mathcal{S}$ endowed with the uniform norm and with covariance kernel $\Gamma(\cdot,\cdot)$.

(A3) There exist positive real numbers E and c (depending on $\theta^\star$) such that for all $x \geq 0$ and $n \geq 1$,

$$\mathbb{P}^\star\Big(\sup_{\mathbf{t}\in\mathcal{S}}\big|Z_n(\mathbf{t})\big| \geq E + x\Big) \leq \exp\big(-cx^2\big).$$

(A2) will be used to obtain the asymptotic distribution of the estimator, and (A3) to obtain non-asymptotic deviation inequalities. Note that (A2) and (A3) are, for instance, verified under mixing conditions on the Y_j's. This follows applying results of Doukhan *et al.* [11,12] and Rio [26]. In particular, in the HMM situation, if $(S_i)_{i\geq 1}$ is an ergodic Markov chain, then (A2) and (A3) hold.

3.2. Definition of the estimator

In case $k^\star$ is known, one may build a $\sqrt{n}$-consistent M-estimator in the usual way. Let $\mathcal{K}$ be a compact subset of $\Theta_{k^\star}^0$, and assume that $\theta^\star$ lies in the interior of $\mathcal{K}$. Denote by $\overline{\theta}_n(\mathcal{K})$ a minimizer of $M_n(k^\star,\cdot)$ over $\mathcal{K}$. Then, as soon as $M_n(k^\star,\cdot)$ converges uniformly in probability to

$M(k^\star,\cdot)$, using the identifiability result Theorem 1, one gets the consistency of $\overline{\theta}_n(\mathcal{K})$. If moreover $\sqrt{n}\nabla M_n(k^\star,\theta^\star)$ converges in distribution ($\nabla M_n(k,\theta)$ denotes the gradient of $M_n(k,\cdot)$ at point θ), and if $D_2M_n(k^\star,\cdot)$ (the Hessian of $M_n(k^\star,\cdot)$) converges in probability to a non-singular matrix, uniformly in a neighborhood of $\theta^\star$, then $\sqrt{n}(\overline{\theta}_n(\mathcal{K})-\theta^\star)$ converges in distribution. This is stated below and proved in the Appendix.

But this requires the prior knowledge of $k^\star$ and of some compact set included in $\Theta^0_{k^\star}$ with $\theta^\star$ in its interior. To obtain an estimator using no prior information and having the same asymptotic properties as $\overline{\theta}_n(\mathcal{K})$, we use a preliminary consistent estimator $(k_n,\tilde{\theta}_n)$ of $(k^\star,\theta^\star)$, and then minimize $M_n(k_n,\cdot)$ over a compact subset of $\Theta^0_{k_n}$ such that $\tilde{\theta}_n$ lies in its interior to get the estimator $\widehat{\theta}_n$. Consistency of $\widehat{\theta}_n$ follows from that of $(k_n,\tilde{\theta}_n)$ and Theorem 1, and once consistency is obtained, since the derivation of the asymptotic properties requires only local analysis, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_n-\theta^\star)$ is the same as that of $\sqrt{n}(\overline{\theta}_n(\mathcal{K})-\theta^\star)$ (which does not depend on $\mathcal{K}$). This is stated in Theorem 2 below and proved in the Appendix.

We now explain the construction of the preliminary estimator $(k_n,\tilde{\theta}_n)$ and of the estimator $\widehat{\theta}_n$. Define $J:\mathbb{N}\to\mathbb{R}_+$ an increasing function tending to infinity at infinity, and for any integer k, I_k a positive continuous function on Θ^0_k and tending to $+\infty$ on the boundary of Θ^0_k or whenever $\|m\|$ tends to infinity. For instance, one may take

$$I_k\bigl(m,(Q_{i,j})_{(i,j)\neq(k,k)}\bigr)=-\log\det Q-\sum_{i=2}^{k}\log\frac{|m_i-m_{i-1}|}{(1+\|m\|_\infty)^2}.$$

Let $(k_n,\tilde{\theta}_n)$ be a minimizer over $\{(k,\theta):k\in\mathbb{N},\theta\in\Theta_k\}$ of

$$C_n(k,\theta)=M_n(k,\theta)+\lambda_n\bigl[J(k)+I_k(\theta)\bigr],$$

where $(\lambda_n)_{n\in\mathbb{N}}$ is a decreasing sequence of real numbers tending to 0 at infinity such that

$$\lim_{n\to+\infty}\sqrt{n}\lambda_n=+\infty. \tag{12}$$

We now define $\widehat{\theta}_n$ as a minimizer of $M_n(k_n,\cdot)$ over the compact subset of $\Theta^0_{k_n}$ (and such that $\tilde{\theta}_n$ lies in its interior)

$$\bigl\{\theta\in\Theta_{k_n}:I_{k_n}(\theta)\leq 2I_{k_n}(\tilde{\theta}_n)\bigr\}.$$

3.3. Asymptotic results

Our first result gives the asymptotic distribution of $\widehat{\theta}_n$. Let V be the variance of the Gaussian random variable

$$\int\bigl\{C(\mathbf{t})\bigl[Z(-\mathbf{t})\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2)-\Phi_{\theta^\star}(-\mathbf{t})\bigl(Z(-t_1,0)\phi_{\theta^\star}(-t_2)+Z(0,-t_2)\phi_{\theta^\star}(-t_1)\bigr)\bigr]$$
$$+C(-\mathbf{t})\bigl[Z(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)-\Phi_{\theta^\star}(\mathbf{t})\bigl(Z(t_1,0)\phi_{\theta^\star}(t_2)+Z(0,t_2)\phi_{\theta^\star}(t_1)\bigr)\bigr]\bigr\}w(\mathbf{t})\,\mathrm{d}\mathbf{t},$$

where

$$C(\mathbf{t}) = \Phi_{\theta^\star}(\mathbf{t})\nabla\big(\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\big) - \nabla\Phi_{\theta^\star}(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2).$$

Denote also $D_2 H$ the second derivative with respect to θ of any function H, whenever it makes sense and by ∇H its gradient (with respect to H).

Theorem 2. *Assume* (A1), (A2) *and* (12). *Then $D_2 M(k^\star, \theta^\star)$ is non-singular, and $\sqrt{n}(\widehat{\theta}_n - \theta^*)$ converges in distribution to the centered Gaussian with variance*

$$\Sigma = \big[D_2 M\big(k^\star, \theta^\star\big)\big]^{-1} V \big[D_2 M\big(k^\star, \theta^\star\big)\big]^{-1}.$$

Moreover, for any compact subset $\mathcal{K}$ of $\Theta^0_{k^\star}$ such that $\theta^\star$ lies in the interior of $\mathcal{K}$, $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^*)$ converges in distribution to the centered Gaussian with variance Σ.

Theorem 2 is proved in Appendix A.

If one wants to use Theorem 2 to build confidence sets, one needs to have a consistent estimator of Σ. Since $D_2 M(k^\star, \cdot)$ is a continuous functions of θ, $D_2 M(k_n, \widehat{\theta}_n)$ is a consistent estimator of $D_2 M(k^\star, \theta^\star)$. Also, V may be viewed as a continuous function of $\Gamma(\cdot,\cdot)$ and θ, as easy but tedious computations show. One may use empirical estimators of $\Gamma(\cdot,\cdot)$ which are uniformly consistent under stationarity and mixing conditions, to get a consistent estimator of V. This leads to a plug-in consistent estimator of Σ.

Another possible way to estimate Σ is to use a boostrap method, following, for instance, Clemencon *et al.* [10] when the hidden variables form a Markov chain.

When we have deviation inequalities for the process Z_n, we are able to provide deviation inequalities for $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^*)$. Such inequalities have interest by themselves, they will also be used for proving adaptivity of our nonparametric estimator in Section 4.

Theorem 3. *Assume* (A1) *and* (A3). *Let $\mathcal{K}$ be a compact subset of $\Theta^0_{k^\star}$ such that $\theta^\star$ lies in the interior of $\mathcal{K}$. Then there exist positive real numbers $c^\star$, $M^\star$, and an integer $n^\star$ such that for all $n \geq n^\star$ and $M \geq M^\star$,*

$$\mathbb{P}^\star\big(\sqrt{n}\big\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\big\| \geq M\big) \leq 2\exp\big(-c^\star M^2\big).$$

In particular, for any integer p,

$$\sup_{n\geq 1} E_{\mathbb{P}^\star}\big[\big(\sqrt{n}\big\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\big\|\big)^p\big] < +\infty.$$

Theorem 3 is proved in Appendix B.

4. Estimation of the nonparametric part

In this section, we assume that $\mathbb{P}^\star$ is the distribution of a stationary ergodic HMM, that is, the sequence $(S_t)_{t\in\mathbb{N}}$ is a stationary ergodic Markov chain. Recall that in this case, (A2) and (A3) are verified and Theorems 2 and 3 hold.

We also assume that the unknown distribution $F^\star$ has density $f^\star$ with respect to Lebesgue measure. Thus, the density $s^\star$ of Y_1 writes

$$s^\star(y) = \sum_{j=1}^{k^\star} \mu^\star(j) f^\star\big(y - m_j^\star\big),$$

where $\mu^\star(j) = \sum_{i=1}^{k^\star} Q^\star_{i,j}$, $1 \le j \le k^\star$. Our aim in this section is to show that good estimation of f^* and s^* are also possible in such models.

4.1. General ideas

On the one hand, many methods for nonparametric estimation of the density $s^\star$ of a sequence of weakly dependent random variables are known. On the other hand, we now have estimators of the parameters $\mu^\star(j)$ and $m_j^\star$, $j = 1, \dots, k^\star$, with good properties thanks to Theorems 2 and 3. Thus, one may propose various ideas to obtain nonparametric estimators of $f^\star$. For instance:

- Given an estimator $\widehat{s}$ of $s^\star$, minimize with respect to f, $D(\widehat{s}_n, \sum_{j=1}^{k^\star} \widehat{\mu}(j) f(\cdot - \widehat{m}_j^\star))$ for some distance (or pseudo-distance) $D(\cdot,\cdot)$ between probability densities.
- Estimate the Fourier transform of $s^\star$, then divide by the estimator $\phi_{\widehat{\theta}}$ of $\phi_{\theta^\star}$, and get an estimator of $f^\star$ by Fourier inversion.
- Use model selection methods to get a nonparametric estimator of $f^\star$.

Our aim in this section is to show that nonparametric estimation using such ideas leads to estimators of $f^\star$ having classical nonparametric estimation properties. We choose to study model selection using penalized marginal likelihood and Gaussian mixtures as sieves. This allows to obtain direct computation of the estimator using the EM algorithm, to get oracle inequalities which lead to asymptotic adaptive estimation over regular classes of densities, and to be confident that the slope heuristics to choose the penalty term (see below) should lead to good practical results.

Of course, further work is needed to make precise in which situation one should use one method or the other.

4.2. A model selection method: Adaptive estimation

Our nonparametric procedure to estimate s^* and f^* is based on a penalized composite maximum likelihood estimator. More precisely, let $\widehat{\mu}(j) = \sum_{i=1}^{k^\star} \widehat{Q}_{i,j}$, $\widehat{m}_j$, be estimators of $\mu^\star(j)$, $m_j^\star$, $j = 1, \dots, k^\star$. Define for any density function f on $\mathbb{R}$

$$\ell_n(f) = \frac{1}{n} \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k^\star} \widehat{\mu}(j) f(Y_i - \widehat{m}_j) \right].$$

Let $\mathcal{F}$ be the set of probability densities on $\mathbb{R}$. We use the model collection $(\mathcal{F}_p)_{p\geq 2}$ of Gaussian mixtures with p components as approximation of $\mathcal{F}$ defined by, if $p \in \mathbb{N}$,

$$\mathcal{F}_p = \left\{ \sum_{i=1}^{p} \pi_i \varphi_{u_i}(x-\alpha_i), \alpha_i \in [-A_p, A_p], \right. \\ \left. u_i \in [b_p, B], \pi_i \geq 0, i = 1, \ldots, p, \sum_{i=1}^{p} \pi_i = 1 \right\}, \tag{13}$$

where B and A_p, b_p, $p \geq 2$, are positive real numbers, and where φ_β is the Gaussian density with variance β^2 given by $\varphi_\beta(x) = \exp(-x^2/2\beta^2)/\beta\sqrt{2\pi}$. For any $p \geq 2$, let $\widehat{f_p}$ be the maximizer of $\ell_n(f)$ over $\mathcal{F}_p$. Define

$$D_n(p) = -\ell_n(\widehat{f_p}) + \mathrm{pen}(p, n),$$

where $\mathrm{pen}(p, n)$ is some penalty term that has to be chosen. Then the estimator is defined by $\widehat{f} = \widehat{f_{\widehat{p}}}$, with $\widehat{p}$ any minimizer of D_n. In the context of independent and identically distributed observations, this estimator has been considered by Maugis-Rabusseau and Michel [23].

Before giving our main theoretical results about the nonparametric estimator, let us say a few words about the practical use of such methods. First of all, the computation of $\widehat{f_p}$ may be performed using the EM-algorithm, which is particularly simple for Gaussian mixtures. Then the choice of the penalty term $\mathrm{pen}(p, n)$ could be chosen using the slope heuristics as proposed in [2,4] for Gaussian regression models and further experimented in various other frameworks; see [6,22,28,29].

We consider $\widehat{\theta} = \overline{\theta}_n(\mathcal{K})$. We prove that $\widehat{s_{\widehat{p}}}$ is an adaptive estimator of $s^\star$, and that, if $\max_j \mu^\star(j) > \frac{1}{2}$, $\widehat{f_{\widehat{p}}}$ is an adaptive estimator of $f^\star$. Adaptivity will be proved on the following classes of regular densities.

Let $y_0 > 0$, $c > 0$, $M_1, M_2 > 0$, $\tau > 0$, $\lambda > 0$ and L a positive polynomial function on $\mathbb{R}$. Let also $\beta > 0$ and $\gamma > (3/2 - \beta)_+$. If we denote $\mathcal{P} = (y_0, c_0, M_1, \tau, M_2, \lambda, L)$, we define $\mathcal{H}_{\mathrm{loc}}(\beta, \gamma, \mathcal{P})$ as the set of probability densities f on $\mathbb{R}$ satisfying:

- f is monotone on $(-\infty, -y_0)$ and on $(y_0, +\infty)$, and $\inf_{|y|\leq y_0} f(y) \geq c_0 > 0$.
- $$\forall y \in \mathbb{R}, \qquad f(y) \leq M_1 \mathrm{e}^{-\tau|y|}. \tag{14}$$
- $\log f$ is $\lfloor \beta \rfloor$ times continuously differentiable with derivatives ℓ_j, $j \leq \beta$ satisfying for all $x \in \mathbb{R}$ and all $|y - x| \leq \lambda$,

  $$\left|\ell_{\lfloor\beta\rfloor}(y) - \ell_{\lfloor\beta\rfloor}(x)\right| \leq \lfloor\beta\rfloor! L(x)|y-x|^{\beta-\lfloor\beta\rfloor}$$

  and

  $$\int_{\mathbb{R}} \left|\ell_j(y)\right|^{(2\beta+\gamma)/j} f(y)\,\mathrm{d}y \leq M_2.$$

We set $A_p = a_0 \log p$, $b_p = b_0 (\log p)^2 / p$ in the definition of $\mathcal{F}_p$ and we consider $\widehat{s_{\widehat{p}}}$ where the penalty is set to

$$\operatorname{pen}(p,n) = \frac{3\kappa}{n}\big(k^\star p\big)\log n.$$

Here, κ is chosen large enough. We denote $h(\cdot,\cdot)$ the Hellinger distance between probability densities.

Theorem 4. *Assume* (A1) *and* (A3). *Then for any* $\mathcal{P}$, $\beta \geq 1/2$ *and* $\gamma > (3/2-\beta)_+$, *there exists* $C(\beta,\gamma,\mathcal{P}) > 0$, *such that*

$$\limsup_{n\to+\infty}\left(\frac{n}{(\log n)^3}\right)^{(2\beta)/(2\beta+1)} \sup_{f^\star \in \mathcal{H}_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\big[h^2\big(s^\star, \widehat{s_{\widehat{p}}}\big)\big] \leq C(\beta,\gamma,\mathcal{P}).$$

Thus, $\widehat{s_{\widehat{p}}}$ is adaptive on the regularity β of the density classes up to $(\log n)^{3\beta/(2\beta+1)}$, see Maugis-Rabusseau and Michel [23] for a lower bound of the asymptotic minimax risk in the case of independent and identically distributed random variables. The proof is based on an oracle inequality which is given in the supplementary material Gassiat and Rousseau [14].

Using Theorem 4, we can also derive adaptive asymptotic rates for the minimax L_1-risk for the estimation of f^*.

Corollary 1. *Assume* (A1) *and* (A3) *and that* $\max_j \mu^\star(j) > \frac{1}{2}$. *Then for any* $\mathcal{P}$, $\beta \geq 1/2$ *and* $\gamma > (3/2-\beta)_+$,

$$\limsup_{n\to+\infty}\left(\frac{n}{(\log n)^3}\right)^{\beta/(2\beta+1)} \sup_{f^\star \in \mathcal{H}_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star}\big[\big\|f_{\widehat{p}} - f^\star\big\|_1\big] \leq \frac{2\sqrt{C(\beta,\gamma,\mathcal{P})}}{(2\max_j \mu^\star(j) - 1)}.$$

It is possible that the constraint, $\max_j \mu^\star(j) > 1/2$ is not sharp, however, note that the Fourier transform of $s^\star$ is expressed as $\phi_{\theta^\star}\phi_{f^\star}$ with $\phi_{\theta^\star}(t) = \sum_{j=1}^k \mu^\star(j)\mathrm{e}^{\mathrm{i}tm_j^\star}$ and $\phi_{f^\star}$ the Fourier transform of $f^\star$, and that $|\phi_{\theta^\star}(t)| > 0$ for all $t \in \mathbb{R}$ if and only if $\max_j \mu^\star(j) > 1/2$, applying the main theorem of Moreno [25].

Proof of Corollary 1. We shall use

$$\big\|s^\star - \widehat{s_{\widehat{p}}}\big\|_1 \leq 2h\big(s^\star, \widehat{s_{\widehat{p}}}\big),$$

together with

$$\begin{aligned}\big\|s^\star - \widehat{s_{\widehat{p}}}\big\|_1 &= \left\|\sum_{j=1}^{k^\star}\mu^\star(j) f^\star\big(\cdot - m_j^\star\big) - \sum_{j=1}^{k^\star}\widehat{\mu}(j)\widehat{f_{\widehat{p}}}(\cdot - \widehat{m}_j)\right\|_1 \\ &\geq \left\|\sum_{j=1}^{k^\star}\mu^\star(j)\big(\widehat{f_{\widehat{p}}} - f^\star\big)(\cdot - \widehat{m}_j)\right\|_1 - \big\|\widehat{\theta}_n - \theta^\star\big\|\end{aligned}$$

$$
\begin{aligned}
&- \Big\| \sum_{j=1}^{k^\star} \mu^\star(j) \big(f^\star\big(\cdot - m_j^\star\big) - f^\star(\cdot - \widehat{m}_j) \big) \Big\|_1 \\
&\geq \Big(2 \max_j \mu^\star(j) - 1 \Big) \big\| f_{\widehat{p}} - f^\star \big\|_1 - \big\| \widehat{\theta}_n - \theta^\star \big\| \\
&\quad - \big\| f^\star\big(\cdot - m_j^\star\big) - f^\star(\cdot - \widehat{m}_j) \big\|_1
\end{aligned}
$$

which follows by using iteratively the triangle inequality. Using $\beta \geq 1/2$, Theorems 3 and 4, we thus get that

$$
\limsup_{n \to +\infty} \left(\frac{n}{(\log n)^3} \right)^{\beta/(2\beta+1)} \sup_{f^\star \in \mathcal{H}_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star} \big[\big\| f_{\widehat{p}} - f^\star \big\|_1 \big] \leq \frac{2\sqrt{C(\beta,\gamma,\mathcal{P})}}{(2 \max_j \mu^\star(j) - 1)}
$$

as soon as

$$
\lim_{n \to +\infty} \left(\frac{n}{(\log n)^3} \right)^{\beta/(2\beta+1)} \sup_{f^\star \in \mathcal{H}_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})} E_{\mathbb{P}^\star} \big[\big\| f^\star\big(\cdot - m_j^\star\big) - f^\star(\cdot - \widehat{m}_j) \big\|_1 \big] = 0. \tag{15}
$$

Now, since $f^\star \in H_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})$ with $\beta \geq 1/2$, if $|\widehat{m}_j - m_j^\star| \leq \lambda$,

$$
\big| \log f^\star(y - \widehat{m}_j) - \log f^\star\big(y - m_j^\star\big) \big| \leq L\big(y - m_j^\star\big) \big| \widehat{m}_j - m_j^\star \big|^{\beta \wedge 1}.
$$

Set $M \geq \frac{1}{2c^\star}$, and $a > 0$ such that, if $|y| \leq n^a$, then $L(y)|\widehat{m}_j - m_j^\star|^{\beta\wedge 1} \leq 1$. Observe also that since $\widehat{\theta}_n$ stays in a compact set, for large enough n, if $|y| \geq n^a$, then for any j, $|y - \widehat{m}_j| \geq n^a/2$ and $|y - m_j^\star| \geq n^a/2$. We obtain, using $|\mathrm{e}^u - 1| \leq 2u$ for $0 \leq u \leq 1$:

$$
\begin{aligned}
\big\| f^\star\big(\cdot - m_j^\star\big) - f^\star(\cdot - \widehat{m}_j) \big\|_1 \leq 2 \left(\frac{M \log n}{n} \right)^{-(\beta\wedge 1)/2} & \int L\big(y - m_j^\star\big) f^\star\big(y - m_j^\star\big) \, \mathrm{d}y \\
& + 2 \int_{|y| \geq n^a/2} f^\star(y) \, \mathrm{d}y + \mathbb{1}_{\|\theta^\star - \widehat{\theta}_n\| > \sqrt{M \log n}/\sqrt{n}},
\end{aligned}
$$

and (15) follows from Theorem 3, $\beta \geq 1/2$ and the fact that $f^\star \in H_{\mathrm{loc}}(\beta,\gamma,\mathcal{P})$ has exponentially decreasing tails. □

Appendix A: Proof of Theorem 2

First of all, we prove a lemma we shall use several times.

Lemma 1. *If $(k_n, \theta_n)_n$, $\theta_n \in \Theta_{k_n}$, is a random sequence such that there exists an integer $K \geq k^*$, and a compact subset $\mathcal{T}$ of $\bigcup_{k \leq K} \Theta_k^0$ such that*

$$
\mathbb{P}^*(k_n \leq K \text{ and } \theta_n \in \mathcal{T}) \to 1 \quad \text{and} \quad M_n(k_n, \theta_n) = \mathrm{o}_{\mathbb{P}^\star}(1),
$$

then

$$\mathbb{P}^*\big(k_n = k^\star\big) \to 1 \quad \textit{and} \quad \theta_n = \theta^\star + \mathrm{o}_{\mathbb{P}^\star}(1).$$

Proof. Using $||A|^2 - |B|^2| \le |A - B||A| + |B||$ and the fact that characteristic functions are uniformly upper bounded by 1, we get that for any integer k and any $\theta \in \Theta_k$:

$$\begin{aligned}
&\big|M_n(k,\theta) - M(k,\theta)\big| \\
&\quad \le 4 \int \big\{ \big|\widehat{\Phi}_n(t_1,t_2) - \Phi_{\theta^\star}(t_1,t_2)\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\big| \\
&\qquad + \big|\widehat{\phi}_{n,1}(t_1)\widehat{\phi}_{n,2}(t_2) - \phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\big| \big\} w(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2.
\end{aligned}$$

The upper bound does not depend on k and θ, $\widehat{\Phi}_n$ is uniformly upper bounded, and we get

$$\sup_{k\ge 2, \theta\in\Theta_k} \big|M_n(k,\theta) - M(k,\theta)\big| = \mathrm{O}\left(\sup_{\mathbf{t}\in\mathcal{S}} \left|\frac{Z_n(\mathbf{t})}{\sqrt{n}}\right|\right) = \mathrm{O}_{\mathbb{P}^\star}(1/\sqrt{n}) \tag{A.1}$$

which together with Theorem 1, which implies that (k^*, θ^*) is the only solution to $M(k,\theta) = 0$, terminates the proof. □

We now proceed to the proof of Theorem 2. Recall that $(k_n, \tilde{\theta}_n)$ is defined at the beginning of Section 3.2. Since $C_n(k_n, \tilde{\theta}_n) \le C_n(k^*, \theta^*)$ and M_n is a non-negative function, we get

$$\big[J(k_n) + I_{k_n}(\tilde{\theta}_n)\big] \le \big[J\big(k^\star\big) + I_{k^\star}\big(\theta^\star\big)\big] + \frac{M_n(k^\star,\theta^\star) - M(k^\star,\theta^\star)}{\lambda_n},$$

so that using (A.1), assumption (A2) and (12) we get

$$\big[J(k_n) + I_{k_n}(\tilde{\theta}_n)\big] \le \big[J\big(k^\star\big) + I_{k^\star}\big(\theta^\star\big)\big] + \mathrm{o}_{\mathbb{P}^\star}(1). \tag{A.2}$$

Also,

$$M_n(k_n, \tilde{\theta}_n) \le M_n\big(k^\star, \theta^\star\big) + \lambda_n\big[J\big(k^\star\big) + I_{k^\star}\big(\theta^\star\big)\big],$$

so that

$$M_n(k_n, \tilde{\theta}_n) = \mathrm{o}_{\mathbb{P}^\star}(1).$$

Thus, using (A.2) and Lemma 1

$$\mathbb{P}^*\big(k_n = k^\star\big) \to 1 \quad \text{and} \quad \tilde{\theta}_n = \theta^\star + \mathrm{o}_{\mathbb{P}^\star}(1). \tag{A.3}$$

Set now $\mathcal{K} = \{\theta \in \Theta_{k^\star} : I_{k^\star}(\theta) \le 4 I_{k^\star}(\theta^\star)\}$. $\mathcal{K}$ is a compact subset of $\Theta^0_{k^\star}$. Let E_n be the event $(k_n = k^\star$ and $\widehat{\theta}_n = \overline{\theta}_n(\mathcal{K}))$. Using Lemma 1, we get that $\overline{\theta}_n(\mathcal{K})$ is a consistent estimator of $\theta^\star$, and using (A.3) and Lemma 1, we get also that $\widehat{\theta}_n$ is a consistent estimator of $\theta^\star$, so M_n has

the same minimizer on $\mathcal{K}$ and on $\{I_{k_n}(\theta) \le 2I_{k_n}(\tilde{\theta}_n)\}$, with probability tending to 1, since they contain a neighbourhood of θ^*. Thus, $\mathbb{P}^\star(E_n) \to 1$. Now, since

$$\widehat{\theta}_n = \overline{\theta}_n(\mathcal{K})\mathbb{1}_{E_n} + \widehat{\theta}_n \mathbb{1}_{E_n^c},$$

Theorem 2 follows as soon as we prove that $\sqrt{n}(\overline{\theta}_n(\mathcal{K}) - \theta^\star)$ converges in distribution to the centered Gaussian with variance Σ. But this is a straightforward consequence of

$$D_2 M_n\big(k^\star, \theta_n\big)\big(\overline{\theta}_n(\mathcal{K}) - \theta^\star\big) = \nabla M_n\big(k^\star, \theta^\star\big),$$

for some $\theta_n \in \Theta_{k^\star}$ such that $\|\theta_n - \theta^\star\| \le \|\overline{\theta}_n(\mathcal{K}) - \theta^\star\|$, the consistency of $\overline{\theta}_n(\mathcal{K})$ and the following lemma.

Lemma 2. *Assume* (A1) *and* (A2). *Then*

- $\sqrt{n}\nabla M_n(k^\star, \theta^\star)$ *converges in distribution to a centered Gaussian with variance* V.
- $D_2 M(k^\star, \theta^\star)$ *is non-singular, and for any random variable* $\theta_n \in \Theta_{k^\star}$ *converging in* $\mathbb{P}^\star$*-probability to* $\theta^\star$, *one has*

$$D_2 M_n\big(k^\star, \theta_n\big) = D_2 M\big(k^\star, \theta^\star\big) + \mathrm{o}_{\mathbb{P}^\star}(1).$$

Proof. First notice that, in every formula, taking the conjugate of any involved function at point $\mathbf{t}$ is the same as taking the function at point $-\mathbf{t}$. This is also verified for derivatives. Write now for any $\theta \in \Theta_{k^\star}$ and any $\mathbf{t} = (t_1, t_2)$

$$G_n(\theta, \mathbf{t}) = \widehat{\Phi}_n(\mathbf{t})\phi_{\theta,1}(t_1)\phi_{\theta,2}(t_2) - \Phi_\theta(\mathbf{t})\widehat{\phi}_{n,1}(t_1)\widehat{\phi}_{n,2}(t_2)$$

so that, if $\nabla G_n(\theta, \mathbf{t})$ denotes the gradient of G_n with respect to θ at point $(\theta, \mathbf{t})$, one has

$$\nabla M_n\big(k^\star, \theta^\star\big) = \int \big[\nabla G_n\big(\theta^\star, \mathbf{t}\big)G_n\big(\theta^\star, -\mathbf{t}\big) + \nabla G_n\big(\theta^\star, -\mathbf{t}\big)G_n\big(\theta^\star, \mathbf{t}\big)\big]w(\mathbf{t})\,\mathrm{d}\mathbf{t}.$$

Now, writing $\widehat{\Phi}_n(\mathbf{t}) = \frac{Z_n(\mathbf{t})}{\sqrt{n}} + \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)$ and using (A2) one gets easily

$$\begin{aligned}
&\sqrt{n}\nabla M_n\big(k^\star, \theta^\star\big) \\
&\quad = \int \big\{\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\big[\Phi_{\theta^\star}(\mathbf{t})\nabla\big(\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\big) - \nabla\Phi_{\theta^\star}(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\big] \\
&\qquad \times \big[Z_n(-\mathbf{t})\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2) - \Phi_{\theta^\star}(-\mathbf{t})\big(Z_n(-t_1, 0)\phi_{\theta^\star}(-t_2) + Z_n(0, -t_2)\phi_{\theta^\star}(-t_1)\big)\big] \\
&\qquad + \phi_{F^\star}(-t_1)\phi_{F^\star}(-t_2)\big[\Phi_{\theta^\star}(-\mathbf{t})\nabla\big(\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2)\big) - \nabla\Phi_{\theta^\star}(-\mathbf{t})\phi_{\theta^\star}(-t_1)\phi_{\theta^\star}(-t_2)\big] \\
&\qquad\quad \times \big[Z_n(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2) - \Phi_{\theta^\star}(\mathbf{t})\big(Z_n(t_1, 0)\phi_{\theta^\star}(t_2) + Z_n(0, t_2)\phi_{\theta^\star}(t_1)\big)\big]\big\}w(\mathbf{t})\,\mathrm{d}\mathbf{t} \\
&\quad + \mathrm{O}_{\mathbb{P}^\star}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where the term $O_{\mathbb{P}^\star}(\frac{1}{\sqrt{n}})$ comes from the quadratic terms in Z_n. The convergence in distribution of $\sqrt{n}\nabla M_n(k^\star,\theta^\star)$ to a centered Gaussian with variance V follows.

Similar computation gives that for any $\theta\in\Theta_{k^\star}$

$$\begin{aligned}
&D_2M_n\big(k^\star,\theta\big)-D_2M_n\big(k^\star,\theta^\star\big)\\
&\quad=\int\big|\widehat{\Phi}_n(\mathbf{t})\big|^2\big[A_1(\mathbf{t},\theta)-A_1\big(\mathbf{t},\theta^\star\big)\big]w(\mathbf{t})\,\mathrm{d}\mathbf{t}\\
&\qquad+\int\big|\widehat{\Phi}_n(t_1,0)\big|^2\big|\widehat{\Phi}_n(0,t_2)\big|^2\big[A_2(\mathbf{t},\theta)-A_2\big(\mathbf{t},\theta^\star\big)\big]w(\mathbf{t})\,\mathrm{d}\mathbf{t}\\
&\qquad+\mathrm{Re}\bigg\{\int\widehat{\Phi}_n(-\mathbf{t})\widehat{\Phi}_n(t_1,0)\widehat{\Phi}_n(0,t_2)\big[A_3(\mathbf{t},\theta)-A_3\big(\mathbf{t},\theta^\star\big)\big]w(\mathbf{t})\,\mathrm{d}\mathbf{t}\bigg\}
\end{aligned}$$

for matrix-valued functions $A_1(\mathbf{t},\theta)$, $A_2(\mathbf{t},\theta)$, $A_3(\mathbf{t},\theta)$ that are, in a neighbourhood of $\theta^\star$, continuous in the variable θ for all $\mathbf{t}$ and uniformly upper bounded. Thus, $D_2M_n(k^\star,\theta_n)-D_2M_n(k^\star,\theta^\star)$ converges in $\mathbb{P}^\star$-probability to 0 whenever θ_n is a random variable converging in $\mathbb{P}^\star$-probability to $\theta^\star$.

Finally, note that at point $\theta^\star$ the Hessian of M simplifies into

$$D_2M\big(k^\star,\theta^\star\big)=2\int H(\mathbf{t})H(-\mathbf{t})^T\big|\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)\big|^2w(\mathbf{t})\,\mathrm{d}\mathbf{t},$$

with

$$H(\mathbf{t})=\Phi_{\theta^\star}(\mathbf{t})\big(\phi_{\theta^\star}(t_1)\nabla\phi_{\theta^\star}(t_2)+\nabla\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2)\big)-\nabla\Phi_{\theta^\star}(\mathbf{t})\phi_{\theta^\star}(t_1)\phi_{\theta^\star}(t_2).$$

Denote by $H_{m_j}(\mathbf{t})$, $j=2,\dots,k^\star$, $H_{Q_{j_1,j_2}}(\mathbf{t})$, $j_1,j_2=1,\dots,k^\star$, $(j_1,j_2)\neq(k^\star,k^\star)$ the components of the vector $H(\mathbf{t})$. Positive definiteness of $D_2M(k^\star,\theta^\star)$ can thus be established by proving that, if for all $\mathbf{t}\in\mathcal{S}$,

$$\sum_{j=2}^{k}U_{m_j}H_{m_j}(\mathbf{t})+\sum_{(j_1,j_2)\neq(k,k)}U_{j_1,j_2}H_{Q_{j_1,j_2}}(\mathbf{t})=0 \tag{A.4}$$

then

$$U_{m_j}=0,\qquad j=2,\dots,k^\star,\qquad U_{j_1,j_2}=0,\qquad j_1,j_2=1,\dots,k^\star,\qquad (j_1,j_2)\neq\big(k^\star,k^\star\big).$$

By linear independence of the functions $\mathrm{e}^{\mathrm{i}ta}$ and $t\mathrm{e}^{\mathrm{i}tb}$, this implies in particular that for all $\mathbf{t}=(t_1,t_2)$,

$$\begin{aligned}
&\sum_{j_1,\dots,j_4=1}^{k^\star}U_{m_{j_1}}\mu^\star(j_1)\mu^\star(j_2)Q^\star_{j_3,j_4}\mathrm{e}^{\mathrm{i}t_1(m^\star_{j_1}+m^\star_{j_3})+\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}\\
&\quad=\sum_{j_1,\dots,j_4=1}^{k^\star}U_{m_{j_1}}\mu^\star(j_2)\mu^\star(j_3)Q^\star_{j_1,j_4}\mathrm{e}^{\mathrm{i}t_1(m^\star_{j_1}+m^\star_{j_3})+\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}
\end{aligned} \tag{A.5}$$

with $U_{m_1} = 0$. The smallest possible term $m^\star_{j_1} + m^\star_{j_3}$ with $j_1 > 1$ is equal to $m^\star_2 = m^\star_2 + m^\star_1$ setting $j_1 = 2$ and $j_3 = 1$ only. Thus, (A.5) implies that

$$U_{m_2}\mu^\star(2)\sum_{j_2,j_4=1}^{k^\star}\mu^\star(j_2)Q^\star_{1,j_4}\mathrm{e}^{\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}$$

$$= U_{m_2}\mu^\star(1)\sum_{j_2,j_4=1}^{k^\star}\mu^\star(j_2)Q^\star_{2,j_4}\mathrm{e}^{\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}$$

for all t_2, that is,

$$U_{m_2}\mu^\star(2)\phi_{\theta^\star}(t_2)\sum_{j_4=1}^{k^\star}Q^\star_{1,j_4}\mathrm{e}^{\mathrm{i}t_2 m^\star_{j_4}} = U_{m_2}\mu^\star(1)\phi_{\theta^\star}(t_2)\sum_{j_4=1}^{k^\star}Q^\star_{2,j_4}\mathrm{e}^{\mathrm{i}t_2 m^\star_{j_4}}.$$

Since $\phi_{\theta^\star}$ has only isolated zeros, this is satisfied if and only if

$$U_{m_2}\mu^\star(2)\sum_{j_4=1}^{k^\star}Q^\star_{1,j_4}\mathrm{e}^{\mathrm{i}t_2 m^\star_{j_4}} = U_{m_2}\mu^\star(1)\sum_{j_4=1}^{k^\star}Q^\star_{2,j_4}\mathrm{e}^{\mathrm{i}t_2 m^\star_{j_4}}.$$

Thus, (A.5) is satisfied only if either $U_{m_2} = 0$ or $\mu^\star(2)Q^\star_{1,j} = \mu^\star(1)Q^\star_{2,j}$ for all j. The latter is impossible since $Q^\star$ is non-singular, thus $U_{m_2} = 0$ and (A.5) becomes

$$\sum_{j_1=3,j_2,\ldots,j_4=1}^{k^\star}U_{m_{j_1}}\mu^\star(j_1)\mu^\star(j_2)Q^\star_{j_3,j_4}\mathrm{e}^{\mathrm{i}t_1(m^\star_{j_1}+m^\star_{j_3})+\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}$$

$$= \sum_{j_1=3,j_2,\ldots,j_4=1}^{k^\star}U_{m_{j_1}}\mu^\star(j_2)\mu^\star(j_3)Q^\star_{j_1,j_4}\mathrm{e}^{\mathrm{i}t_1(m^\star_{j_1}+m^\star_{j_3})+\mathrm{i}t_2(m^\star_{j_2}+m^\star_{j_4})}.$$

The smallest possible value for $m^\star_{j_1} + m^\star_{j_3}$ is then $m^\star_3$ which is obtained with the only configuration $j_1 = 3, j_3 = 1$. The same argument as before leads to $U_{m_3} = 0$. Iteration of the argument leads to $U_{m_j} = 0$ for all $j = 1, \ldots, k^\star$. We now study the derivatives associated to Q. We write U the $k^\star \times k^\star$-matrix whose components are U_{j_1,j_2} for $(j_1, j_2) \neq (k^\star, k^\star)$ and $U_{k^\star,k^\star} = -\sum_{(j_1,j_2)\neq(k^\star,k^\star)}U_{j_1,j_2}$. Then

$$\sum_{(j_1,j_2)\neq(k^\star,k^\star)}U_{j_1,j_2}\nabla_{Q_{j_1,j_2}}\Phi_{\theta^\star}(\mathbf{t}) = V(t_1)^T U V(t_2),$$

where for any $t \in \mathbb{R}$, $V(t) = ((\mathrm{e}^{\mathrm{i}tm^\star_j})_{j=1,\ldots,k^\star})^T$, and

$$\sum_{(j_1,j_2)\neq(k^\star,k^\star)}U_{j_1,j_2}\nabla_{Q_{j_1,j_2}}\phi_{\theta^\star}(t_1) = V(t_1)^T U \mathbb{1}$$

with $\mathbb{1} = (1, \ldots, 1)^T \in \mathbb{R}^{k^\star}$, since $\phi_{\theta^\star}(t_1) = V(t_1)^T Q^\star \mathbb{1}$ and $\Phi_{\theta^\star}(\mathbf{t}) = V(t_1)^T Q^\star V(t_2)$. We can then express (A.4) as

$$\begin{aligned} V(t_1)^T \big[& Q^\star V(t_2) V(t_2)^T U \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T + Q^\star V(t_2) V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T U^T \\ & - U V(t_2) V(t_2)^T Q \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T \big] V(t_1) = 0. \end{aligned} \tag{A.6}$$

Note also that since all differences $m^\star_{j_1} - m^\star_{j_2}$, $j_1 \neq j_2$, are distinct, if A is a $k^\star \times k^\star$-matrix and $\mathcal{I}$ is an open subset of $\mathbb{R}$,

$$\big[\forall t \in \mathcal{I},\, V(t)^T A V(t) = 0\big] \Longrightarrow A + A^T = 0. \tag{A.7}$$

Then (A.6) implies

$$\begin{aligned} & Q^\star V(t_2) V(t_2)^T U \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T + Q^\star \mathbb{1}\mathbb{1}^T U^T V(t_2) V(t_2)^T \big(Q^\star\big)^T \\ & + Q^\star V(t_2) V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T U^T + U \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T V(t_2) V(t_2)^T \big(Q^\star\big)^T \\ & - U V(t_2) V(t_2)^T Q^\star \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T - Q^\star \mathbb{1}\mathbb{1}^T \big(Q^\star\big)^T V(t_2) V(t_2)^T U^T = 0. \end{aligned} \tag{A.8}$$

Recall also that $\mathbb{1}^T U \mathbb{1} = 0$ and that $Q^\star \mathbb{1} = \mu^\star$. Note that $U\mathbb{1} = \alpha \mu^\star$ with $\alpha \in \mathbb{R}$ if and only if $\alpha = 0$ since $\mathbb{1}^T U \mathbb{1} = 0$ while $\mathbb{1}^T \mu^\star = 1$. Therefore, if $U\mathbb{1} \neq 0$ there exists $w \in \mathbb{R}^{k^\star}$ such that $w^T (U\mathbb{1}) \neq 0$ while $(\mu^\star)^T w = 0$. Multiplying the above equality on the left by w^T and on the right by w leads to

$$w^T Q^\star V(t_2) V(t_2)^T \big(\mu^\star\big) (U\mathbb{1})^T w = 0$$

for all t_2 in an open set. Using (A.7) again and since $(U\mathbb{1})^T w \neq 0$, we get that

$$\mu^\star \big[\big(Q^\star\big)^T w\big]^T + \big[\big(Q^\star\big)^T w\big] \big(\mu^\star\big)^T = 0.$$

Since $\mu^\star(j) > 0$ for all j, this implies that $(Q^\star)^T w = 0$ which is impossible since $Q^\star$ has full rank. Therefore, $U\mathbb{1} = 0$ and (A.8) becomes $V(t_2)^T \mu^\star [U V(t_2) (\mu^\star)^T + \mu^\star V(t_2)^T U^T] = 0$, that is $U V(t_2) (\mu^\star)^T + \mu^\star V(t_2)^T U^T = 0$ for all t_2 in an open set. Multiplying on the left by $\mathbb{1}$ implies that $U V(t_2) = 0$ for all t_2 in an open set so that $U = 0$. □

Appendix B: Proof of Theorem 3

Define for any $\theta \in \Theta_{k^\star}$, $L_n(\theta) = M_n(k^\star, \theta) - M(k^\star, \theta)$. Then, since $M_n(k^\star, \overline{\theta}_n(\mathcal{K})) \leq M_n(k^\star, \theta^\star)$, one easily gets

$$M\big(k^\star, \overline{\theta}_n(\mathcal{K})\big) - M\big(k^\star, \theta^\star\big) \leq \big| L_n\big(\overline{\theta}_n(\mathcal{K})\big) - L_n\big(\theta^\star\big) \big|.$$

Define for any $\mathbf{t} = (t_1, t_2)$ and any θ

$$G(\theta, \mathbf{t}) = \big\{ \Phi_{\theta^\star}(\mathbf{t}) \phi_{\theta,1}(t_1) \phi_{\theta,2}(t_2) - \Phi_\theta(\mathbf{t}) \phi_{\theta^\star,1}(t_1) \phi_{\theta^\star,2}(t_2) \big\} \phi_{F^\star}(t_1) \phi_{F^\star}(t_2)$$

and

$$B_n(\theta,\mathbf{t}) = \phi_{F^\star}(t_1)\phi_{F^\star}(t_2) \times \Bigg\{ \frac{Z_n(\mathbf{t})}{\sqrt{n}}\phi_{\theta,1}(t_1)\phi_{\theta,2}(t_2) - \Phi_\theta(\mathbf{t})\Bigg[\frac{Z_n(t_1,0)}{\sqrt{n}}\phi_{\theta,2}(t_2) + \frac{Z_n(0,t_2)}{\sqrt{n}}\phi_{\theta,1}(t_1) + \frac{Z_n(t_1,0)Z_n(0,t_2)}{n}\Bigg]\Bigg\}.$$

Writing $\widehat{\Phi}_n(\mathbf{t}) = \frac{Z_n(\mathbf{t})}{\sqrt{n}} + \Phi_{\theta^\star}(\mathbf{t})\phi_{F^\star}(t_1)\phi_{F^\star}(t_2)$, one gets

$$L_n(\theta) = \int \big(\big[B_n(\theta,\mathbf{t}) + G(\theta,\mathbf{t})\big]\big[B_n(\theta,-\mathbf{t}) + G(\theta,-\mathbf{t})\big] - \big|G(\theta,\mathbf{t})\big|^2\big)w(\mathbf{t})\,\mathrm{d}\mathbf{t}.$$

Since $G(\theta^\star,\mathbf{t}) = 0$ for all $\mathbf{t}$, we obtain

$$L_n(\theta) - L_n\big(\theta^\star\big) = \int \big\{\big|B_n(\theta,\mathbf{t})\big|^2 - \big|B_n\big(\theta^\star,\mathbf{t}\big)\big|^2 + B_n(\theta,\mathbf{t})G(\theta,-\mathbf{t}) + B_n(\theta,-\mathbf{t})G(\theta,\mathbf{t})\big\}w(\mathbf{t})\,\mathrm{d}\mathbf{t}$$

which gives

$$\big|L_n(\theta) - L_n\big(\theta^\star\big)\big| \le \int \big\{\big|B_n(\theta,\mathbf{t}) - B_n\big(\theta^\star,\mathbf{t}\big)\big|\big|B_n(\theta,\mathbf{t}) + B_n\big(\theta^\star,\mathbf{t}\big)\big| + 2\big|B_n(\theta,\mathbf{t})\big|\big|G(\theta,\mathbf{t}) - G\big(\theta^\star,\mathbf{t}\big)\big|\big\}w(\mathbf{t})\,\mathrm{d}\mathbf{t}$$

which leads to

$$M\big(k^\star,\overline{\theta}_n(\mathcal{K})\big) - M\big(k^\star,\theta^\star\big) \le C W_n\big\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\big\| \tag{B.1}$$

for some constant C and any integer n, and with

$$W_n = \Big\{\frac{V_n}{\sqrt{n}} + \frac{V_n^2}{n} + \frac{V_n^3}{n^{3/2}} + \frac{V_n^4}{n^2}\Big\}, \qquad V_n = \sup_{\mathbf{t}\in\mathcal{S}}\big|Z_n(\mathbf{t})\big|.$$

Observe now that, since $D_2M(k^\star,\cdot)$ is continuous and $D_2M(k^\star,\theta^\star)$ is non-singular, there exists $\lambda > 0$ and $\alpha > 0$ such that, if $\|\theta - \theta^\star\| \le \alpha$, then $M(k^\star,\theta) - M(k^\star,\theta^\star) \ge \frac{\lambda}{2}\|\theta - \theta^\star\|^2$. Moreover, there exists $\delta > 0$ such that, if $\theta \in \mathcal{K}$ is such that $\|\theta - \theta^\star\| \ge \alpha$, then $M(k^\star,\theta) - M(k^\star,\theta^\star) \ge \delta$. Using (B.1), we obtain that for any real number M large enough,

$$\mathbb{P}^\star\big(\sqrt{n}\big\|\overline{\theta}_n(\mathcal{K}) - \theta^\star\big\| \ge M\big) \le \mathbb{P}^\star\Big(W_n \ge \frac{\delta}{2CM(\mathcal{K})}\Big) + \mathbb{P}^\star\Big(\sqrt{n}W_n \ge \frac{M\lambda}{2C}\Big),$$

where $M(\mathcal{K}) = \sup_{\theta\in\mathcal{K}}\|\theta\|$. Since, for n large enough

$$\mathbb{P}^\star\Big(W_n \ge \frac{\delta}{2CM(\mathcal{K})}\Big) \le \mathbb{P}^\star\Big(V_n \ge E + \frac{\delta\sqrt{n}}{8CM(\mathcal{K})}\Big)$$

and

$$\mathbb{P}^\star\left(\sqrt{n}W_n \geq \frac{M\lambda}{2C}\right) \leq \mathbb{P}^\star\left(V_n \geq \frac{M\lambda}{8C}\right),$$

using assumption (A3), this terminates the proof of Theorem 3.

Supplementary Material

Supplement to "Nonparametric finite translation hidden Markov models and extensions" (DOI: 10.3150/14-BEJ631SUPP; .pdf). In the supplementary material, we provide an oracle inequality which is used to prove Theorem 4, together with the proofs of the oracle inequality and of Theorem 4. We also give a concentration inequality which is used in various parts of these proofs.

Acknowledgements

The authors would like to thank the anonymous Associate Editor and the referees for valuable comments and suggestions.

This work was partly supported by the 2010–2014 grant ANR Banhdits AAP Blanc SIMI 1. Judith Rousseau is also affiliated to the CEREMADE, Université Paris Dauphine.

References

[1] Allman, E.S., Matias, C. and Rhodes, J.A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

[2] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** 245–279.

[3] Azzaline, A. and Bowman, A.W. (1990). A look at some data in the Old Faithful geyser. *Appl. Statist.* **39** 357–365.

[4] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

[5] Bonhomme, S., Jochman, K. and Robin, J. (2011). Nonparametric estimation of finite mixtures. Technical report.

[6] Bontemps, D. and Toussile, W. (2013). Clustering and variable selection for categorical multivariate data. *Electron. J. Stat.* **7** 2344–2371. MR3108816

[7] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356

[8] Butucea, C. and Vandekerkhove, P. (2014). Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.* **41** 227–239.

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

[10] Clemencon, S., Garivier, A. and Tressou, J. (2009). Pseudo-regenerative block-bootstrap for hidden Markov chains. In *Statistical Signal Processing*, 2009. IEEE.

[11] Doukhan, P., Massart, P. and Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **30** 63–82. MR1262892
[12] Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **31** 393–427. MR1324814
[13] Gassiat, E., Cleynen, A. and Robin, S. (2013). Finite state space nonparametric hidden Markov models are in general identifiable. Available at arXiv:1306.4657.
[14] Gassiat, E. and Rousseau, J. (2015). Supplement to "Nonparametric finite translation hidden Markov models and extensions". DOI:10.3150/14-BEJ631SUPP.
[15] Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. MR1962504
[16] Henry, M., Kitamura, Y. and Salanié, B. (2014). Partial identification of finite mixtures in econometric models. *Quant. Econ.* **5** 123–144.
[17] Hunter, D.R., Wang, S. and Hettmansperger, T.P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275
[18] Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica* **77** 135–175. MR2477846
[19] Lambert, M.F., Whiting, J.P. and Metcalfe, A.V. (2003). A non-parametric hidden Markov model for climate state identification. *Hydrol. Earth Syst. Sci.* **7** 652–667.
[20] Marin, J.-M., Mengersen, K. and Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking*: *Modeling and Computation* (C. Rao and D. Dey, eds.). *Handbook of Statist.* **25** 459–507. Amsterdam: Elsevier/North-Holland. MR2490536
[21] Massart, P. (2007). *Concentration Inequalities and Model Selection*: *Ecole d'Eté de Probabilités de Saint-Flour XXXIII* – 2003. Berlin: Springer. MR2319879
[22] Maugis, C. and Michel, B. (2011). Data-driven penalty calibration: A case study for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15** 320–339. MR2870518
[23] Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.* **17** 698–724. MR3126158
[24] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. MR1789474
[25] Moreno, C.J. (1973). The zeros of exponential polynomials. I. *Compos. Math.* **26** 69–78. MR0318460
[26] 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
[27] Stein, E.M. and Shakarchi, R. (2003). *Complex Analysis*. Princeton, NJ: Princeton Univ. Press. MR1976398
[28] Verzelen, N. (2009). Adaptative estimation to regular Gaussian Markov random fields. Ph.D. thesis, Univ. Paris-Sud.
[29] Villers, F. (2007). Tests et sélection de modèles pour l'analyse de donnŕes protéomiques et transcriptomiques. Ph.D. thesis, Univ. Paris-Sud.
[30] Yau, C., Papaspiliopoulos, O., Roberts, G.O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 37–57. MR2797735

Received June 2013 and revised December 2013