Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space

Elisabeth Gassiat, Sylvain Le Corff, Luc Lehéricy

Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space

Élisabeth Gassiat*, Sylvain Le Corff**, and Luc Lehéricy*

*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.
**Samovar, CNRS, Télécom SudParis, Institut Polytechnique de Paris.

Abstract

This paper considers hidden Markov models where the observations are given as the sum of a latent state which lies in a general state space and some independent noise with unknown distribution. It is shown that these fully nonparametric translation models are identifiable with respect to both the distribution of the latent variables and the distribution of the noise, under mostly a light tail assumption on the latent variables. Two nonparametric estimation methods are proposed and we prove that the corresponding estimators are consistent for the weak convergence topology. These results are illustrated with numerical experiments.

1 Introduction

This paper considers nonparametric translation hidden Markov models where, for all $i = 1, \dots, n$, the observation Y_i is

$$Y_i = X_i + \varepsilon_i \,, \tag{1}$$

where $n \geqslant 1$ is the number of observations, $(X_i)_{i=1,\dots,n}$ is a d dimensional hidden stationary Markov chain and $(\varepsilon_i)_{i=1,\dots,n}$ are independent identically distributed random variables independent of $(X_i)_{i=1,\dots,n}$. Both the distributions of the latent variables and of the noise ε_1 are unknown. The first objective of this paper is to prove that the law of the hidden states may be recovered using only the observations $(Y_i)_{i=1,\dots,n}$ when no assumption is made on the noise distribution and with only a weak nonparametric assumption on the distribution of the hidden Markov chain. In addition, consistent estimation procedures based either on a least squares or on a maximum likelihood approach are proposed. This work provides the first contribution to establish identifiability results in a fully nonparametric setting for hidden Markov models with general state space.

The use of latent data models is ubiquitous in time series analysis across a wide range of applied science and engineering domains such as signal processing [Crouse et al., 1998], genomics [Yau et al., 2011, Wang et al., 2017], target tracking [Särkkä et al., 2007], enhancement and segmentation of speech and audio signals [Rabiner, 1989], see also [Särkkä, 2013, Douc et al., 2014, Zucchini et al., 2016] and the numerous references therein. The specific setting of translation hidden Markov models described by (1) is commonly used in statistical signal processing, such as for nonlinear phase estimation, where the problem appears in many applications: detection of phase synchronization, estimation of instantaneous frequencies or in neuroscience, see [Dahlhaus et al., 2018], [Fell and Axmacher, 2011] and the references therein. In these

applications, the latent signal is modeled as $X_i = g(Z_i)$, $g : \mathbb{R}^d \to \mathbb{R}^d$ for some sequence $(Z_i)_{i\geqslant 1}$ of relevant hidden variables. In [Dumont and Le Corff, 2017b], such models are used to detect oscillation patterns in human electrocardiogam recordings and to estimate a noisy Rossler attractor.

Although parametric hidden Markov models have been widely studied and are appealing for a wide range of applications, parametric inference procedures may lead to poor results in real data and high dimensional learning problems. This explains the recent keen interest for nonparametric latent data models which have been introduced in many disciplines such as climate state identification [Lambert et al., 2003, Touron, 2019], genomics [Yau et al., 2011], statistical modelling of animal movement [Langrock et al., 2015] or biology [Volant et al., 2014]. For finite state space hidden Markov models, such nonparametric modeling has been recently validated by theoretical identifiability results and the analysis of estimation procedures with provable guarantees, see [Gassiat et al., 2016], [Alexandrovich et al., 2016], [De Castro et al., 2016], [Lehéricy, 2018]. In this setting, the parameters to be estimated are the transition matrix of the hidden chain and the emission densities. See also [Gassiat and Rousseau, 2016] and [Akakpo, 2019] for translation hidden Markov models with finite state space. While certainly of interest, the finite state space setting may be too restrictive for many applications.

The inverse problem in (1) is also known as the deconvolution problem. There is a wide range of literature on density deconvolution when the distribution of the noise ε_i is assumed to be known and the random variables $(X_i, \varepsilon_i)_{i=1,\dots,n}$ are assumed to be independent and identically distributed, see [Devroye, 1989], [Liu and Taylor, 1989], [Stefanski and Carroll, 1990], for some early nonparametric deconvolution methods, [Carroll and Hall, 1988] and [Fan, 1991] for minimax rates, see also [Dedecker et al., 2015] and references therein for a recent work. However, when the distribution of the noise is unknown and the observations are independent, model (1) can not be identified in full generality.

In this paper, we establish the identifiability of the fully nonparametric hidden translation model under the weak assumption that the Laplace transform of the latent Markov chain has an exponential growth smaller than 2, see Theorem 1. In the case of real valued hidden Markov models, identifiability is extended to latent variables having Laplace transform with exponential growth smaller than 3, see Theorem 2. Two different methods are proposed to recover the distribution of the latent variables. The first one is a least squares method arising naturally from the identifiability proof, the second one is the classical maximum likelihood method using discrete probability measures as approximation of all probability measures. Both estimators are proved to be consistent for the weak convergence topology, see Theorem 3 and Theorem 4.

The paper is organized as follows. Section 2 displays the general identifiability results. The consistency of the least squares approach and that of the maximum likelihood estimation procedures are given in Section 3. These results are supported by simulations in Section 4.

2 Identifiability theorems

Consider a sequence of random variables $(Y_i)_{i\geqslant 1}$ taking values in $\mathbb{R}^d$ and satisfying model (1) in which the hidden Markov chain $(X_i)_{i\geqslant 1}$ is stationary. Endow $\mathbb{R}^d$ with its Borel sigma-field $\mathcal{B}(\mathbb{R}^d)$. For each transition kernel $K : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ having a unique stationary distribution μ_K, define the measure R_K on $\mathbb{R}^{2d}$ as follows. For all $A \in \mathcal{B}(\mathbb{R}^{2d})$, $R_K(A) = \int \mu_K(\mathrm{d}x)K(x,\mathrm{d}y)\mathbb{1}_A(x,y)$. For any probability distribution P on $\mathbb{R}^d$, denote by $\mathbb{P}_{K,P}$ the distribution of the sequence $(Y_i)_{i\geqslant 1}$ when the stationary Markov chain $(X_i)_{i\geqslant 1}$ has transition K and ε_1 has distribution P. For any $\rho > 0$, let $\mathcal{M}_\rho$ be the set of finite measures μ on $\mathbb{R}^d$ such that there exist $A, B > 0$ satisfying, for all $\lambda \in \mathbb{R}^d$, $\int \exp\left(\lambda^T x\right) \mathrm{d}\mu(x) \leqslant A \exp\left(B\|\lambda\|^\rho\right)$, where for a vector λ in a Euclidian space, $\|\lambda\|$ denotes its euclidian norm and λ^T denotes its transpose vector. Notice that if K is such that $\mu_K \in \mathcal{M}_\rho$ for some ρ, then the function $\Phi_{R_K} : \mathbb{C}^d \times \mathbb{C}^d \longrightarrow \mathbb{C}$, defined for all

$(z_1, z_2) \in \mathbb{C}^d \times \mathbb{C}^d$ by $\Phi_R(z_1, z_2) = \int \exp\left(z_1^T x_1 + z_2^T x_2\right) \mathrm{d}R(x_1, x_2)$, is a multivariate analytic function. Consider the following assumption.

H1 For any $z_0 \in \mathbb{C}^d$, $z \mapsto \Phi_R(z_0, z)$ is not the null function or $z \mapsto \Phi_R(z, z_0)$ is not the null function.

Throughout this paper, the assertion *$R_K = R_{\widetilde{K}}$ and $P = \widetilde{P}$ up to translation* means that there exists $m \in \mathbb{R}^d$ such that if (X_1, X_2) has distribution R_K and $(\varepsilon_1, \varepsilon_2)$ has distribution $P \otimes P$, then $(X_1 - m, X_2 - m)$ has distribution $R_{\widetilde{K}}$ and $(\varepsilon_1 + m, \varepsilon_2 + m)$ has distribution $\widetilde{P} \otimes \widetilde{P}$. The following theorems state that the distribution of the observations allows to recover the kernel of the hidden Markov chain and the distribution of the noise up to translation.

Theorem 1. *Assume that K (resp. $\widetilde{K}$) is a transition kernel on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ admitting a unique stationary distribution μ_K (resp. $\mu_{\widetilde{K}}$). Assume that there exists $\rho < 2$ such that $\mu_K \in \mathcal{M}_\rho$ and $\mu_{\widetilde{K}} \in \mathcal{M}_\rho$. Assume that R_K and $R_{\widetilde{K}}$ satisfy assumption H1. Then, $\mathbb{P}_{K,P} = \mathbb{P}_{\widetilde{K},\widetilde{P}}$ implies that $R_K = R_{\widetilde{K}}$ and $P = \widetilde{P}$ up to translation.*

In the case of real valued random variables, identifiability holds for a larger class of transition kernels, including Gaussian Markov chains.

Theorem 2 (case $d = 1$)**.** *Assume that K (resp. $\widetilde{K}$) is a transition kernel on $\mathbb{R} \times \mathcal{B}(\mathbb{R})$ admitting a unique stationary distribution μ_K (resp. $\mu_{\widetilde{K}}$) and a density with respect to the Lebesgue measure. Assume that there exists $\rho < 3$ such that $\mu_K \in \mathcal{M}_\rho$ and $\mu_{\widetilde{K}} \in \mathcal{M}_\rho$. Assume that R_K and $R_{\widetilde{K}}$ satisfy assumption H1. Assume moreover that if the stationary Markov chain with transition kernel K (resp. $\widetilde{K}$) is Gaussian, it is not a sequence of independent and identically distributed variables. Then, $\mathbb{P}_{K,P} = \mathbb{P}_{\widetilde{K},\widetilde{P}}$ implies that $R_K = R_{\widetilde{K}}$ and $P = \widetilde{P}$ up to translation.*

Remark 1. *One way to fix the "up to translation" indeterminacy when the noise has a first order moment is to assume that $\mathbb{E}[\varepsilon_1] = 0$.*

Remark 2. *In nonparametric hidden regression models, $X_i = g(Z_i)$ where $g : \mathbb{R}^d \to \mathbb{R}^d$ and $(Z_i)_{i \geqslant 1}$ is a sequence of hidden variables. Under the assumption that g is one-to-one, if $(Z_i)_{i \geqslant 1}$ is a Markov chain, then $(X_i)_{i \geqslant 1}$ is also a Markov chain. Then, when H1 holds, Theorem 1 extends the identification results of [Dumont and Le Corff, 2017a, Dumont and Le Corff, 2017b] to the cases where the distribution of the additive noise is unknown. Numerical experiments in the case where $g : x \mapsto \cos x$ are given in Section 4.*

Theorems 1 and 2 are proved in Appendix A.

3 Consistent estimation

3.1 Using least squares for characteristic functions

In the following, objects related to the true (unknown) distribution $\mathbb{P}^\star$ of the observed process are denoted with the superscript $\star$. Let $\mathcal{S}$ be a compact neighborhood of 0 in $\mathbb{R}^{2d}$, and let $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a positive function on $\mathcal{S}$. Let $\phi^\star$ be the characteristic function of ε_1. For any probability distribution R on $\mathbb{R}^d \times \mathbb{R}^d$, define

$$M(R) = \int_{\mathcal{S}} \left|\Phi_{R^\star}(it_1, it_2)\Phi_R(it_1, 0)\Phi_R(0, it_2) - \Phi_R(it_1, it_2)\Phi_{R^\star}(it_1, 0)\Phi_{R^\star}(0, it_2)\right|^2 \\ |\phi^\star(t_1)\phi^\star(t_2)|^2 w(t_1, t_2) \mathrm{d}t_1 \mathrm{d}t_2 \,.$$

Under appropriate assumptions, by the proof of Theorem 1, $M(R) = 0$ if and only if $R = R^\star = R_{K^\star}$ up to translation. Using an estimator $\widehat{\Phi}_n$ of the characteristic function of (Y_1, Y_2), define an estimator of $M(\cdot)$ by

$$M_n(R) = \int_{\mathcal{S}} \left| \widehat{\Phi}_n(t_1,t_2)\Phi_R(it_1,0)\Phi_R(0,it_2) - \Phi_R(it_1,it_2)\widehat{\Phi}_n(t_1,0)\widehat{\Phi}_n(0,t_2) \right|^2 w(t_1,t_2)\mathrm{d}t_1\mathrm{d}t_2 \,.$$

Let $\mathcal{R}$ be a set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ such that for some $\rho < 2$, for all $R \in \mathcal{R}$, both marginal distributions of R are in $\mathcal{M}_\rho$ and R satisfies assumption H1. Define $\widehat{R}_n$ as an element of $\mathcal{R}$ satisfying

$$M_n(\widehat{R}_n) = \inf_{R \in \mathcal{R}} M_n(R).$$

Under the assumptions of Theorem 3, $\widehat{R}_n$ exists but may be not uniquely defined because of translation invariance. Let d be a distance that metrizes weak convergence on $\mathcal{R}$, and define $Z_n(t_1,t_2)$ by $Z_n(t_1,t_2) = \sqrt{n}(\widehat{\Phi}_n(t_1,t_2) - \Phi_{R^\star}(it_1,it_2)\phi_1^\star(t_1)\phi_2^\star(t_2))$.

Theorem 3. *Assume that $\mathcal{R}$ is compact for the weak convergence topology and that $R^\star \in \mathcal{R}$. Assume moreover that $\sup_{(t_1,t_2)\in\mathcal{S}} |Z_n(t_1,t_2)| = O_{\mathbb{P}^\star}(1)$. Then, $M(\widehat{R}_n) = O_{\mathbb{P}^\star}(n^{-1/2})$, and $d(\widehat{R}_n, \mathcal{R}^\star)$ tends to 0 in $\mathbb{P}^\star$-probability as n tends to infinity, where $\mathcal{R}^\star$ is the set of $R \in \mathcal{R}$ that are equal to $R^\star$ up to translation.*

Theorem 3 is proved in Appendix A. If $\widehat{\Phi}_n$ is the empirical estimator, then the assumption on Z_n holds as soon as the hidden Markov chain is strongly mixing, see for instance [Doukhan et al., 1994] and [Doukhan et al., 1995]. Here is an example where the other assumptions of Theorem 3 are easily verified. Consider $(R_\theta)_{\theta\in\Theta}$ a family of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ such that each R_θ admit a density r_θ with respect to the Lebesgue measure. Assume that Θ is a compact subset of a Euclidian space and let $\mathcal{P}(\Theta)$ be the set of probability distributions on Θ endowed with its Borel sigma-field. Let $\mathcal{R}$ be the following set of mixtures: $\mathcal{R} = \left\{ R_\mu = \int R_\theta \mathrm{d}\mu(\theta),\ \mu \in \mathcal{P}(\Theta) \right\}$. If $(\theta, x_1, x_2) \mapsto r_\theta(x_1,x_2)$ is a continuous and bounded function, then $\mathcal{R}$ is compact for the weak convergence topology. Also, if there exists some ρ such that the measure with density $\sup_\theta r_\theta$ with respect to the Lebesgue measure belongs to $\mathcal{M}_\rho$, then all $R \in \mathcal{R}$ belong to $\mathcal{M}_\rho$. Moreover, for any $\mu \in \mathcal{P}(\Theta)$, for any $(z_0,z_1) \in \mathbb{C}^d \times \mathbb{C}^d$, $\Phi_{R_\mu}(z_0,z_1) = \int_\Theta \Phi_{R_\theta}(z_0,z_1)\mathrm{d}\mu(\theta)$, so that as soon as for some $u_0 \in C^d$, for all $\theta \in \Theta$, $\Phi_{R_\theta}(z_0, zu_0)$ tends to $+\infty$ when $z \in \mathbb{R}$ tends to $+\infty$, then $z \mapsto \Phi_{R_\mu}(z_0,z)$ can not be the null function.

3.2 Using maximum likelihood

Using the fact that continuous distributions may be approximated by discrete distributions, we consider finite state space hidden Markov models and the associated maximum likelihood estimator. The idea is to replace the (continuous) support of the hidden process by a finite support. Increasing the number of support points reduces the approximation error (the bias) while increasing the estimation error. Thus, a careful bias-variance trade-off has to be performed to obtain consistent estimators. In this section, we present a penalized likelihood estimator that automatically selects the number of support points. Its consistency is obtained from the oracle inequality proved in [Lehéricy, 2018], Theorem 8.

Assume that the hidden process $(X_i)_{i\geqslant 1}$ takes values in a known compact set $\Lambda = [-L,L]^d \subset \mathbb{R}^d$ and that the distribution of the noise is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Denote by $K^\star$ the transition kernel of the hidden process, and by $\gamma^\star$ the density of the noise with respect to the Lebesgue measure.

Transition kernels on finite sets are described by the number of points r of their support, the vector $\mathfrak{x} = (x_1, \dots, x_r)$ of their support points and the transition matrix Q between these points: for all

$(z, z') \in \{1, \dots, r\}^2$, $Q(z, z') = \mathbb{P}(X_1 = x_{z'} | X_0 = x_z)$. For a vector $\mathfrak{X} \in \Lambda^r$, a transition matrix Q with stationary distribution μ_Q and a density γ, the log-likelihood of the parameter $(\mathfrak{X}, Q, \gamma)$ given the observations $(Y_i)_{1 \leqslant i \leqslant n}$ is

$$\ell_n(\mathfrak{X}, Q, \gamma) = \log \left(\sum_{z_1, \dots, z_n \in \{1, \dots, r\}} \mu_Q(z_1) \gamma(Y_1 - x_{z_1}) \prod_{t=2}^{n} Q(z_{t-1}, z_t) \gamma(Y_t - x_{z_t}) \right). \tag{2}$$

Let Θ be a compact subset of $\mathbb{R}^d \times GL_d(\mathbb{R})$ and $f : y \in \mathbb{R}^d \longmapsto (2\pi)^{-d/2} \exp(-\|y\|^2/2)$ be the density of a standard multivariate normal distribution. Write $\mathcal{P}(\Theta)$ the set of probability measures on Θ, let

$$\Gamma = \left\{ \gamma : y \longmapsto \int_\Theta |\det(\Sigma)| f\left(\Sigma(y - \mu)\right) \mathrm{d}p(\mu, \Sigma) : p \in \mathcal{P}(\Theta), \int_\Theta \mu \mathrm{d}p(\mu, \Sigma) = 0 \right\} \tag{3}$$

be the set of densities of location-scale mixtures of f with parameters in Θ, and assume that $\gamma^* \in \Gamma$. The condition $\int_\Theta \mu \mathrm{d}p(\mu, \Sigma) = 0$ ensures that all densities in Γ are centered. For $(\mu, \Sigma) \in \Theta$, write $\delta_{\mu,\Sigma}$ the Dirac measure centered on (μ, Σ). Let $(G_D)_{D \geqslant 1}$ be the models for the emission density, defined for all $D \geqslant 1$ by

$$G_D = \left\{ \gamma : y \longmapsto \sum_{i=1}^{D} p_i \det(\Sigma_i) f\left(\Sigma_i(y - \mu_i)\right) : \sum_{i=1}^{D} p_i \delta_{(\mu_i, \Sigma_i)} \in \mathcal{P}(\Theta), \sum_{i=1}^{D} p_i \mu_i = 0 \right\}. \tag{4}$$

G_D is the set of all densities in Γ whose mixing measure has a finite support of at most D points.

Transition kernels are understood as functions from Λ to $\mathcal{P}(\Lambda)$ endowed with the weak convergence topology, which is metrized by the Wasserstein 1 metric W_1. It is assumed that all kernels used in the proposed procedure share the same modulus of continuity ω. It is possible to assume that ω is a concave function with no loss of generality since $\mathcal{P}(\Lambda)$ has finite W_1-diameter. Let $C \geqslant 2$ be a constant.

H2 The application $x \in \Lambda \longmapsto K^\star(x, \cdot) \in (\mathcal{P}(\Lambda), W_1)$ admits the modulus of continuity $\omega/2$ and there exists a probability measure $\lambda^\star$ on Λ such that for all $x \in \Lambda$, $K^\star(x, \cdot)$ has a density with values in $[2/C, C/2]$ with respect to $\lambda^\star$.

The collection of models $(S_{r,D})_{r \geqslant 1, D \geqslant 1}$ used in the maximum likelihood estimation is defined as follows. For all $r \geqslant 1$ and $D \geqslant 1$, let $S_{r,D}$ be the set of all $(\mathfrak{X}, Q, \gamma) \in \Lambda^r \times [1/(Cr), C/r]^{r \times r} \times G_D$ such that Q is a transition matrix and the transition kernel $x_z \longmapsto \sum_{z'=1}^{r} Q(z, z') \delta_{x_{z'}}$ admits the modulus of continuity ω with respect to W_1. For each $r \geqslant 1$ and $D \geqslant 1$, the maximum likelihood estimator of model $S_{r,D}$ is defined by

$$(\widehat{\mathfrak{X}}_{r,D}, \widehat{Q}_{r,D}, \widehat{\gamma}_{r,D}) \in \operatorname*{arg\,max}_{(\mathfrak{X}, Q, \gamma) \in S_{r,D}} \frac{1}{n} \ell_n(\mathfrak{X}, Q, \gamma). \tag{5}$$

Then, select the number of states and the model dimension using the penalized likelihood:

$$(\widehat{r}_n, \widehat{D}_n) \in \operatorname*{arg\,max}_{r \leqslant \log n, D \leqslant n} \left(\frac{1}{n} \ell_n(\widehat{\mathfrak{X}}_{r,D}, \widehat{Q}_{r,D}, \widehat{\gamma}_{r,D}) - (D + r^2) \frac{(\log n)^{15}}{n} \right)$$

and define the final estimators $(\widehat{\mathfrak{X}}_n, \widehat{Q}_n, \widehat{\gamma}_n) = (\widehat{\mathfrak{X}}_{\widehat{r}_n, \widehat{D}_n}, \widehat{Q}_{\widehat{r}_n, \widehat{D}_n}, \widehat{\gamma}_{\widehat{r}_n, \widehat{D}_n})$. In order to state the consistency result, a continuous kernel associated with the discrete kernels of the models has to be introduced. For $(\mathfrak{X}, Q, \gamma) \in S_{r,D}$, denote by $K_{\mathfrak{X},Q}$ a transition kernel on Λ that admits the modulus of continuity ω with respect to the Wasserstein 1 metric, extends the kernel defined by Q on $\{x_z\}_{z=1,\dots,r}$ and such that the support of $K_{\mathfrak{X},Q}(x, \cdot)$ is in $\{x_z\}_{z=1,\dots,r}$ for all $x \in \Lambda$. Linear interpolation provides a way to construct such a kernel as soon as the modulus ω is concave.

Theorem 4. *Assume that assumptions H1 and H2 hold. Let $\lambda^\star$ be the measure defined in assumption H2 and Supp$(\lambda^\star)$ its support. Then almost surely,*

$$\sup_{x\in Supp(\lambda^\star)} W_1(K_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n}(x,\cdot),K^\star(x,\cdot)) \underset{n\to\infty}{\longrightarrow} 0$$

and $\|\widehat{\gamma}_n-\gamma^\star\|_1 \underset{n\to\infty}{\longrightarrow} 0$. In particular, almost surely under $\mathbb{P}^\star$, for all $x\in$ Supp$(\lambda^\star)$, $K_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n}(x,\cdot)\longrightarrow K^\star(x,\cdot)$ for the weak convergence topology and if $\mathbb{P}^X_K$ denotes the distribution of the stationary Markov chain with transition kernel K, $\mathbb{P}^X_{K_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n}}\longrightarrow\mathbb{P}^X_{K^\star}$ for the weak convergence topology.

This result is a special case of a theorem stated and proved in Appendix B that holds for more general sets Γ and $(G_D)_{D\geqslant 1}$.

4 Simulations

Consider the model where Z_0 is a uniform random variable on $(0,2\pi)$ and for all $k\geqslant 1$,

$$Z_k=\phi Z_{k-1}+\sigma_x\varepsilon_k\,,\quad X_k=\cos(Z_k)\quad\text{and}\quad Y_k=X_k+\sigma_y\eta_k\,,$$

where $(\phi,\sigma_x,\sigma_y)\in[-1,1]\times\mathbb{R}_+^*\times\mathbb{R}_+^*$ and where $(\varepsilon_k,\eta_k)_{k\geqslant 1}$ are independent standard Gaussian random variables independent of Z_0. The parameters $(\phi,\sigma_x,\sigma_y)=(1,0.1,0.1)$ are used to sample the observations. Assumption H2 holds: the transition kernel $K^\star$ of $(X_k)_{k\geqslant 1}$ is $1/2$-Hölder and the probability measure $\lambda^\star$ can be taken as the invariant measure of $K^\star$.

This section provides numerical illustrations of the maximum likelihood approach, additional simulations using least squares for the characteristic functions are given in Section C. The performance of the estimation procedure proposed in Section 3.2 is assessed in the case where $\Lambda=\mathbb{R}$ and Γ is as in (3) with $\Theta=\mathbb{R}\times(0,+\infty)$. Although the compacity assumptions of Section 3.2 are not satisfied, in practice, the estimator is shown to converge to the true distribution. The main reason for these assumptions is to ensure theoretical consistency by ruling out the worst case scenarios where the estimators are degenerate.

For each $n\in\{5000,10000,20000,50000,100000,200000\}$, 10 independent and identically distributed sequences $(Y_i)_{i=1,\dots,n}$ are generated. For each sample, an approximation of the maximum likelihood estimator is computed using the Estimation-Maximization algorithm [Dempster et al., 1977] for $D=2$ and $r\in\{10,20,30\}$. Then, using $N_X=5000$, $N_W=4$ and $N_X\times N_W$ independent and identically distributed pairs $(X_{1,i}^{(j)},X_{2,i}^{(j)})_{i=1,\dots,N_X,j=1,\dots,N_W}\sim R_{K^\star}$, the error criterion is the estimated Wasserstein distance between the estimated and the true distribution of (X_1,X_2):

$$\text{Error}(\widehat{\mathfrak{x}}_n,\widehat{Q}_n)=\frac{1}{N_W}\sum_{j=1}^{N_W}W_1\left(R_{K_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n}},\frac{1}{N_X}\sum_{i=1}^{N_X}\delta_{(X_{1,i}^{(j)},X_{2,i}^{(j)})}\right),\tag{6}$$

or equivalently (when written as a distance between weighted point processes)

$$\text{Error}(\widehat{\mathfrak{x}}_n,\widehat{Q}_n)=\frac{1}{N_W}\sum_{j=1}^{N_W}W_1\left(\sum_{x,x'\in\widehat{\mathfrak{x}}_n}R_{K_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n}}(x,x')\delta_{(x,x')},\frac{1}{N_X}\sum_{i=1}^{N_X}\delta_{(X_{1,i}^{(j)},X_{2,i}^{(j)})}\right).$$

The distance W_1 is computed using function `wasserstein` from R package `transport` [Schuhmacher et al., 2019, R Core Team, 2017]. The results are displayed in Figure 1.

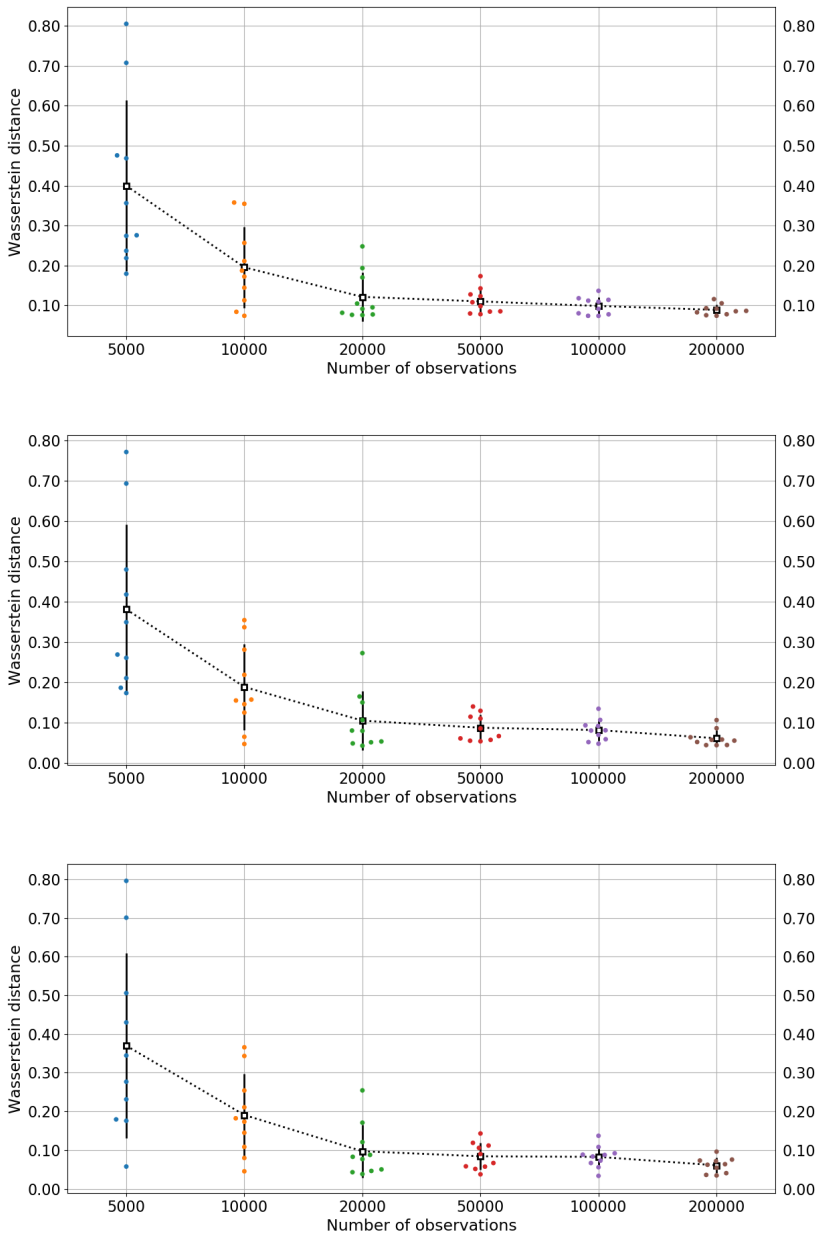

Figure 1: Wasserstein distance computed as in (6) for $r = 10$ (top), $r = 20$ (middle) and $r = 30$ (bottom). Each dot is an estimated value with the maximum likelihood approach. For each value of r, the mean value (squares) over all runs as well as the empirical standard deviation (bars) are displayed.

n	$r=10$	$r=20$	$r=30$
5000	[0.196; 0.327; 0.773]	[0.170; 0.312; 0.778]	[0.059; 0.304; 0.800]
10000	[0.075; 0.182; 0.355]	[0.047; 0.169; 0.363]	[0.045; 0.184; 0.371]
20000	[0.075; 0.097; 0.261]	[0.045; 0.082; 0.267]	[0.036; 0.079; 0.255]
50000	[0.077; 0.098; 0.166]	[0.048; 0.077; 0.155]	[0.034; 0.074; 0.160]
100000	[0.076; 0.103; 0.149]	[0.046; 0.091; 0.142]	[0.038; 0.084; 0.139]
200000	[0.076; 0.087; 0.110]	[0.045; 0.065; 0.100]	[0.037; 0.062; 0.107]

Figure 2: Summary of the Wasserstein distance computed as in (6). Each cell contains the min, median and max value of the error over the 10 simulations with corresponding r and n.

A Proof of Theorem 1, Theorem 2 and Theorem 3

A.1 Proof of Theorem 1

The following result, which may be established by arguing variable by variable, is used repeatedly in this proof. If a multivariate function is analytic on the whole multivariate complex space and is the null function in an open set of the multivariate real space or in an open set of the multivariate purely imaginary space, then it is the null function on the whole multivariate complex space.

Assume that $\mathbb{P}_{K,P} = \mathbb{P}_{\widetilde{K},\widetilde{P}}$ and let ϕ (resp. $\widetilde{\phi}$) be the characteristic function of P (resp. $\widetilde{P}$). Notice that $\Phi_{R_K}(it_1, it_2)$ (resp. $\Phi_{R_{\widetilde{K}}}(it_1, it_2)$) for real numbers numbers t_1 and t_2 defines the characteristic function of (X_1, X_2) when the Markov chain has kernel K (resp. $\widetilde{K}$) and $\Phi_{R_K}(it, 0) = \Phi_{R_K}(0, it)$ (resp. $\Phi_{R_{\widetilde{K}}}(it, 0) = \Phi_{R_K}(0, it)$ for real numbers t defines the characteristic function of any X_i when the Markov chain has kernel K (resp. $\widetilde{K}$). Since the distribution of Y_1 and Y_2 are the same under $\mathbb{P}_{K,P}$ and $\mathbb{P}_{\widetilde{K},\widetilde{P}}$, for any $t \in \mathbb{R}^d$,

$$\phi(t)\,\Phi_{R_K}(it, 0) = \widetilde{\phi}(t)\,\Phi_{R_{\widetilde{K}}}(it, 0)\,. \tag{7}$$

Since the distribution of (Y_1, Y_2) is the same under $\mathbb{P}_{K,P}$ and $\mathbb{P}_{\widetilde{K},\widetilde{P}}$, for any $(t_1, t_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\phi(t_1)\,\phi(t_2)\,\Phi_{R_K}(it_1, it_2) = \widetilde{\phi}(t_1)\,\widetilde{\phi}(t_2)\,\Phi_{R_{\widetilde{K}}}(it_1, it_2)\,. \tag{8}$$

There exists a neighborhood V of 0 in $\mathbb{R}^d \times \mathbb{R}^d$ such that for all $t = (t_1, t_2) \in V$, $\phi(t_1) \neq 0$, $\phi(t_2) \neq 0$, $\widetilde{\phi}(t_1) \neq 0$, $\widetilde{\phi}(t_2) \neq 0$, so that (7) and (8) imply that for any $(t_1, t_2) \in V^2$,

$$\Phi_{R_K}(it_1, it_2)\,\Phi_{R_{\widetilde{K}}}(it_1, 0)\,\Phi_{R_{\widetilde{K}}}(it_2, 0) = \Phi_{R_{\widetilde{K}}}(it_1, it_2)\,\Phi_{R_K}(it_1, 0)\,\Phi_{R_K}(it_2, 0)\,. \tag{9}$$

Since $(z_1, z_2) \mapsto \Phi_{R_K}(z_1, z_2)\,\Phi_{R_{\widetilde{K}}}(z_1, 0)\,\Phi_{R_{\widetilde{K}}}(z_2, 0) - \Phi_{R_{\widetilde{K}}}(z_1, z_2)\,\Phi_{R_K}(z_1, 0)\,\Phi_{R_K}(z_2, 0)$ is a multivariate analytic function of $2d$ variables which is zero in a purely imaginary neighborhood of 0, then it is the null function on the whole multivariate complex space so that for any $z_1 \in \mathbb{C}^d$ and $z_2 \in \mathbb{C}^d$,

$$\Phi_{R_K}(z_1, z_2)\,\Phi_{R_{\widetilde{K}}}(z_1, 0)\,\Phi_{R_{\widetilde{K}}}(z_2, 0) = \Phi_{R_{\widetilde{K}}}(z_1, z_2)\,\Phi_{R_K}(z_1, 0)\,\Phi_{R_K}(z_2, 0)\,. \tag{10}$$

Fix $(u_2, \ldots, u_d) \in \mathbb{C}^{d-1}$ and let Z be the set of zeros of $u \mapsto \Phi_R(u, u_2, \ldots, u_d, 0)$ and $\widetilde{\mathsf{Z}}$ be the set of zeros of $u \mapsto \Phi_{R_{\widetilde{K}}}(u, u_2, \ldots, u_d, 0)$. Let $u_1 \in \mathsf{Z}$ and write $z_1 = (u_1, u_2, \ldots, u_d)$ so that by (10), for any $z_2 \in \mathbb{C}^d$,

$$\Phi_{R_K}(z_1, z_2)\,\Phi_{R_{\widetilde{K}}}(z_1, 0)\,\Phi_{R_{\widetilde{K}}}(z_2, 0) = 0 \text{ and } \Phi_{R_K}(z_2, z_1)\,\Phi_{R_{\widetilde{K}}}(z_1, 0)\,\Phi_{R_{\widetilde{K}}}(z_2, 0) = 0\,. \tag{11}$$

By assumption H1, $z_2 \to \Phi_{R_K}(z_1, z_2)$ is not the null function or $z_2 \to \Phi_{R_K}(z_2, z_1)$ is not the null function. Assume without loss of generality that $z_2 \to \Phi_{R_K}(z_1, z_2)$ is not the null function (the proof follows the same steps in the other case). Then, there exists $z_2^\star$ in $\mathbb{C}^d$ such that $\Phi_{R_K}(z_1, z_2^\star) \neq 0$ and by continuity, there exists an open neighborhood of $z_2^\star$ such that for all z_2 in this open set, $\Phi_{R_K}(z_1, z_2) \neq 0$. Since $z \mapsto \Phi_{R_{\widetilde{K}}}(z, 0)$ is not the null function and is analytic on $\mathbb{C}^d$, it can not be null all over this open set, so that there exists z_2 such that simultaneously $\Phi_{R_K}(z_1, z_2) \neq 0$ and $\Phi_{R_{\widetilde{K}}}(z_2, 0) \neq 0$. Then (11) leads to $\Phi_{R_{\widetilde{K}}}(z_1, 0) = 0$, so that $\mathsf{Z} \subset \widetilde{\mathsf{Z}}$. A symmetric argument yields $\widetilde{\mathsf{Z}} \subset \mathsf{Z}$ so that $\mathsf{Z} = \widetilde{\mathsf{Z}}$.

Moreover, the analytic functions $u \mapsto \Phi_{R_K}(u, u_2, \dots, u_d, 0)$ and $u \mapsto \Phi_{R_{\widetilde{K}}}(u, u_2, \dots, u_d, 0)$ have exponential growth order less than 2, so that using Hadamard's factorization Theorem, see [Stein and Shakarchi, 2003, Chapter 5, Theorem 5.1], there exists a polynomial function s with degree at most 1 (and with coefficients depending on $(u_2, \dots, u_d)$) such that for all $u \in \mathbb{C}$,

$$\Phi_{R_K}(u, u_2, \dots, u_d, 0) = \mathrm{e}^{s(u)} \Phi_{R_{\widetilde{K}}}(u, u_2, \dots, u_d, 0)\,.$$

Arguing similarly for all variables, there exists a polynomial function S on $\mathbb{C}^d$ with degree at most 1 in each variable such that for all $(u_1, \dots, u_d) \in \mathbb{C}^d$,

$$\Phi_{R_K}(u_1, u_2, \dots, u_d, 0) = \mathrm{e}^{S(u_1, u_2, \dots, u_d)} \Phi_{R_{\widetilde{K}}}(u_1, u_2, \dots, u_d, 0)\,. \tag{12}$$

Since $\Phi_{R_K}(0, \dots, 0) = \Phi_{R_{\widetilde{K}}}(0, \dots, 0) = 1$, the constant term of the polynomial S is 0. Assume that $\mu_{\widetilde{K}}$ is not supported by 0. Thus, there exist $a = (a_1, \dots, a_d) \in \mathbb{R}^d$, $\alpha > 0$ and $\delta > 0$ such that

$$0 \notin \prod_{j=1}^d [a_j - \alpha, a_j + \alpha] \quad \text{and} \quad \mu_{\widetilde{K}} \left(\prod_{j=1}^d [a_j - \alpha, a_j + \alpha] \right) \geqslant \delta\,,$$

which gives, for all $\lambda \in \mathbb{R}^d$,

$$\Phi_{R_{\widetilde{K}}}(\lambda, 0) \geqslant \delta \mathrm{e}^{\sum_{j=1}^d \inf_{x \in [a_j - \alpha, a_j + \alpha]} \lambda_j x}\,.$$

When $\mu_{\widetilde{K}}$ is supported by 0 then, for all $\lambda \in \mathbb{R}^d$, $\Phi_{R_{\widetilde{K}}}(\lambda, 0) = 1$. Since $\mu_{\widetilde{K}} \in \mathcal{M}_\rho$ for some $\rho < 2$, if S has degree at least 2, then $\Phi_{R_K}(\cdot, 0)$ has exponential growth of order at least 2, contradicting the assumption. Then, S has degree at most 1 and there exists $m \in \mathbb{C}^d$ such that for all $z \in \mathbb{C}^d$,

$$\Phi_{R_K}(z, 0) = \mathrm{e}^{m^T z} \Phi_{R_{\widetilde{K}}}(z, 0)\,. \tag{13}$$

As for all $z \in \mathbb{R}^d$, $\Phi_{R_K}(-iz, 0) = \overline{\Phi_{R_K}(iz, 0)}$ and $\Phi_{R_{\widetilde{K}}}(-iz, 0) = \overline{\Phi_{R_{\widetilde{K}}}(iz, 0)}$, then $m \in \mathbb{R}^d$. Combining (13) with (10) yields, for all $(t_1, t_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\Phi_{R_K}(it_1, it_2) = e^{im^T t_1 + im^T t_2} \Phi_{R_{\widetilde{K}}}(it_1, it_2)\,. \tag{14}$$

Then, using (7), for all $t \in \mathbb{R}^d$ such that $\Phi_{R_K}(it, 0) \neq 0$, $\phi(t) = \mathrm{e}^{-im^T t} \widetilde{\phi}(t)$. Since the set of zeros of $t \mapsto \Phi_{R_K}(it, 0)$ has empty interior, for each t such that $\Phi_{R_K}(it, 0) = 0$ it is possible to find a sequence $(t_n)_{n \geqslant 1}$ such that t_n tends to t and for all n, $\Phi_{R_K}(it_n, 0) \neq 0$. But ϕ and $\widetilde{\phi}$ are continuous functions, so that for all $t \in \mathbb{R}$,

$$\phi(t) = \mathrm{e}^{-im^T t} \widetilde{\phi}(t)\,. \tag{15}$$

The proof is concluded by noting that (14) and (15) imply that $R_K = R_{\widetilde{K}}$ and $P = \widetilde{P}$ up to translation.

A.2 Proof of Theorem 2

Using Hadamard's Theorem, if $\mu_K \in \mathcal{M}_\rho$ (resp. $\mu_{\widetilde{K}} \in \mathcal{M}_\rho$) with $\rho < 3$, then $z \mapsto \Phi_{R_K}(z,0)$ (resp. $z \mapsto \Phi_{R_{\widetilde{K}}}(z,0)$) has no zeros if and only if the Markov chain is Gaussian. Therefore, the assumptions of Theorem 2 imply that in all cases, the stationary Markov chains with transition kernel K (resp. $\widetilde{K}$) is not a sequence of independent and identically distributed variables.

Following the same steps as in the proof of Theorem 1, there exists a polynom S with real coefficients and degree at most 2 such that, for all $z \in \mathbb{C}$, $\Phi_{R_K}(z,0) = \mathrm{e}^{S(z)}\Phi_{R_{\widetilde{K}}}(z,0)$, and for all $(z_1,z_2) \in \mathbb{C}\times\mathbb{C}$,

$$\Phi_{R_K}(z_1,z_2) = e^{S(z_1)}e^{S(z_2)}\Phi_{R_{\widetilde{K}}}(z_1,z_2)\,. \tag{16}$$

Assume that S has degree equal to 2. Then, there exist real numbers a, b, c such that for all $z \in \mathbb{C}$, $S(z) = az^2 + bz + c$. With no loss of generality assume that $a > 0$ (otherwise, replace K by $\widetilde{K}$). Then, (16) means that there exist independent and identically distributed Gaussian variables η_i, with variance $2a$, such that, if $(X_i)_{i\geqslant 1}$ is a stationary Markov chain with transition kernel K and $(\widetilde{X}_i)_{i\geqslant 1}$ is a stationary Markov chain with transition kernel $\widetilde{K}$, $(X_i)_{i\geqslant 1}$ has the same distribution as $(\widetilde{X}_i + \eta_i)_{i\geqslant 1}$, with η_i, $i \geqslant 1$, independent of $(\widetilde{X}_i)_{i\geqslant 1}$. Using Lemma 1, this implies that the $(X_i)_{i\geqslant 1}$, are independent and identically distributed, contradicting the assumption of Theorem 2. Then, S has degree at most 1, and the end of the proof of Theorem 2 follows the same steps as the proof of Theorem 1.

Lemma 1. *Assume that $(X_i)_{i\geqslant 1}$ is a stationary real valued Markov chain with transition kernel having a density with respect to the Lebesgue measure. Assume that $(\eta_i)_{i\geqslant 1}$ is a sequence of independent and identically distributed real valued Gaussian random variables with positive variance and independent of $(X_i)_{i\geqslant 1}$. If $(X_i + \eta_i)_{i\geqslant 1}$ is Markov chain, then $(X_i)_{i\geqslant 1}$ is an independent and identically distributed sequence.*

Proof. For all $x \in \mathbb{R}$, let $x' \mapsto q(x,x')$ be the density of the transition kernel of the Markov chain $(X_i)_{i\geqslant 1}$ with respect to the Lebesgue measure and μ be its stationary density. Denote m the mean and σ^2 the variance of η_1, and let ϕ be the density of η_1 . The fact that $(X_i+\eta_i)_{i\geqslant 1}$ is a Markov chain implies that the conditional distribution of $X_3+\eta_3$, conditionally to $(X_2+\eta_2, X_1+\eta_1)$, equals the conditional distribution of $X_3+\eta_3$, conditionally to $X_2+\eta_2$ alone. This rewrites as follows. For all real numbers y_1, y_2, y_3,

$$\begin{aligned}&\int \mu(x_1)q(x_1,x_2)\phi(y_1-x_1)\phi(y_2-x_2)q(x_2,x_3)\phi(y_3-x_3)\mu(x_4)\phi(y_2-x_4)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4\\ &= \int \mu(x_1)q(x_1,x_2)\phi(y_1-x_1)\phi(y_2-x_2)\mu(x_4)q(x_4,x_3)\phi(y_3-x_3)\phi(y_2-x_4)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4. \end{aligned}\tag{17}$$

But for all real numbers x and y, $\phi(y-x) = \phi(x-y-2m)$. Since y is a complete statistic for $\phi(x-y-2m)\mathrm{d}x$, (17) implies that for all real numbers x_1, x_3, y_2,

$$\int \mu(x_1)q(x_1,x_2)\mu(x_4)[q(x_2,x_3)-q(x_4,x_3)]\phi(y_2-x_2)\phi(y_2-x_4)\mathrm{d}x_2\mathrm{d}x_4 = 0\,. \tag{18}$$

Using that $\phi(y_2-x_2)\phi(y_2-x_4) = \phi(\sqrt{2}[y_2-(x_2+x_4)/2)])\phi((x_2-x_4+m)/\sqrt{2})$, (18)) implies that for all real numbers x_1, x_3, u,

$$\int \mu(x_1)q\left(x_1,\frac{u+v}{2}\right)\mu\left(\frac{u-v}{2}\right)\left[q\left(\frac{u+v}{2},x_3\right)-q\left(\frac{u-v}{2},x_3\right)\right]\phi((v+m)/\sqrt{2})\mathrm{d}v = 0\,. \tag{19}$$

Let $H : \mathbb{R}^3 \longrightarrow \mathbb{R}$ be any measurable and positive function. Define the measurable and positive function $G : (x, y, z) \mapsto H(x, y, z)\phi((x - y + 2m)/2\sqrt{2})$. Then by multiplying (19) by $H((u + v)/2, (u - v)/2, x_3)$ and integrating over x_1, x_3, u, we get by change of variable that

$$\int \mu(x_1)q(x_1, x_2)q(x_2, x_3)\mu(x_4)G(x_2, x_4, x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4 = \int \mu(x_1)q(x_1, x_2)\mu(x_4)q(x_4, x_3)G(x_2, x_4, x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4\,. \quad (20)$$

Let now $(\widetilde{X}_i)_{i\geqslant 1}$ be a Markov chain with the same distribution of $(X_i)_{i\geqslant 1}$ but independent of $(X_i)_{i\geqslant 1}$. Since the correspondance $G \leftrightarrow H$ between measurable positive functions is one-to-one, (20) means that for any measurable and positive function G, $\mathbb{E}\left[G\left(X_2, \widetilde{X}_2, X_3\right)\right] = \mathbb{E}\left[G\left(X_2, \widetilde{X}_2, \widetilde{X}_3\right)\right]$, which means that $(X_2, \widetilde{X}_2, X_3)$ and $(X_2, \widetilde{X}_2, \widetilde{X}_3)$ have the same distribution. But this implies that X_2 is independent of $(\widetilde{X}_2, X_3)$ which implies that X_2 is independent of X_3. □

A.3 Proof of Theorem 3

Using the fact that characteristic functions are bounded by 1, for all $R \in \mathcal{R}$,

$$|M_n(R) - M(R)| \leqslant \frac{3}{\sqrt{n}} \sup_{(t_1,t_2)\in\mathcal{S}} |Z_n(t_1, t_2)| + \frac{1}{n} \sup_{(t_1,t_2)\in\mathcal{S}} |Z_n(t_1, t_2)|^2\,, \quad (21)$$

and using the assumption on Z_n, $\sup_{R\in\mathcal{R}} |M_n(R) - M(R)| = O_{\mathbb{P}^\star}(n^{-1/2})$. Now, using the definition of $\widehat{R}_n$ and (21), $M(\widehat{R}_n) \leqslant M_n(\widehat{R}_n) + O_{\mathbb{P}^\star}(n^{-1/2}) \leqslant M_n(R^\star) + O_{\mathbb{P}^\star}(n^{-1/2}) \leqslant M(R^\star) + O_{\mathbb{P}^\star}(n^{-1/2})$. $M(\widehat{R}_n)$ is then upper bounded by a term of order $O_{\mathbb{P}^\star}(n^{-1/2})$ since $M(R^\star) = 0$, and the first assertion of Theorem 3 is proved. Now, $R \mapsto M(R)$ is continuous for the weak convergence topology, and for any $\epsilon > 0$, $\sup_{R\in\mathcal{R}, d(R,R^\star)\geqslant\epsilon} M(R)$ is attained by compacity of $\{R \in \mathcal{R}, d(R, R^\star) \geqslant \epsilon\}$, and positive since $M(R) = 0$ if and only if $R = R^\star$ up to translation. Thus using Theorem 5.7 in [van der Vaart, 1998], the set of limiting values of $(\widehat{R}_n)_{n\geqslant 1}$ for the weak convergence topology is the set of $R \in \mathcal{R}$ such that $R = R^\star$ up to translation.

B Proof of Theorem 4

B.1 General statement

This section provides in Theorem 5 a more general statement of the result claimed in Theorem 4. The proof of Theorem 5 is postponed to Section B.2. Let Γ be a set of probability densities on $\mathbb{R}^d$ that satisfies the following assumption.

H3 Γ is a set of continuous and positive probability densities that admit a first order moment and are centered in the sense that for all $\gamma \in \Gamma$,

$$\int_{\mathbb{R}^d} y\gamma(y)\mathrm{d}y = 0\,. \quad (22)$$

Γ is a compact subset of $\mathbf{L}^1(\mathbb{R}^d)$ and the envelope function

$$b : y \in \mathbb{R}^d \longmapsto \sup_{\gamma\in\Gamma} \sup_{x\in\Lambda} \max(\gamma(y - x), \gamma(x - y))$$

satisfies $b \in \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^\infty(\mathbb{R}^d)$, admits a first order moment, and there exists a constant $C_\Gamma > 0$ such that for all $\gamma \in \Gamma$ and $y \in \mathbb{R}^d$, the mapping $x \in \Lambda \longmapsto \gamma(y-x)/b(y)$ is C_Γ-Lipschitz. Finally, $\gamma^\star \in \Gamma$.

The centering assumption (22) allows to fix the translation parameter in the identifiability results.

Example. *Let f be a bounded and positive probability density on $\mathbb{R}^d$ that admits a first order moment and is centered. Assume that there exists $\epsilon > 0$ such that*

$$\sup_{\substack{(\mu,\Sigma)\in\mathbb{R}^d\times GL_d(\mathbb{R}) \\ \|\mu\|_2\leqslant\epsilon,\ \|\Sigma - Id_d\|_F\leqslant\epsilon}} f(\Sigma(\cdot-\mu)) \in \mathbf{L}^1(\mathbb{R}^d)$$

and let Θ be a compact subset of $\mathbb{R}^d \times GL_d(\mathbb{R})$. Finally, assume that there exists a function D_f such that for all $y, y' \in \mathbb{R}^d$, $|f(y) - f(y')| \leqslant D_f(y)|y - y'|$ and such that $(D_f/f) \in \mathbf{L}^\infty(\mathbb{R}^d)$. Then the set of translation-scale mixtures of f with parameters in Θ

$$\Gamma = \left\{ \gamma : y \longmapsto \int_\Theta |\det(\Sigma)| f\left(\Sigma(y-\mu)\right) \mathrm{d}p(\mu,\Sigma) : p \in \mathcal{P}(\Theta), \int_\Theta \mu \mathrm{d}p(\mu,\Sigma) = 0 \right\}$$

satisfies H3.

H4 Γ satisfies H3 with the envelope function b. Let m be the lower envelope function of Γ defined by

$$m : y \in \mathbb{R}^d \longmapsto \inf_{\gamma\in\Gamma} \inf_{x\in\Lambda} \gamma(y-x).$$

There exists $\epsilon > 0$ such that $\int b(y)[b(y)/m(y)]^\epsilon \mathrm{d}y < \infty$.

Example. *The set Γ of Gaussian location-scale mixtures of Section 3.2 satisfies H3 and H4.*

Then, consider $(G_D)_{D\geqslant 1}$ a family of subsets of Γ. The following assumption essentially means that each G_D is a parametric model with dimension D.

H5 Γ satisfies H3 and H4 with the functions b and m, the set $\bigcup_{D\geqslant 1} G_D$ is dense in Γ with respect to the $\mathbf{L}^1$ norm, and there exists a constant $\widetilde{c} > 0$, a mapping $(D, A) \in \mathbb{N}^* \times \mathbb{R}_+ \longmapsto c(D, A)$ and an increasing mapping $D \longmapsto \dim_D$ such that the following holds.

- For all $D \geqslant 1$ and $A \geqslant 0$, $\log c(D, A) \leqslant \widetilde{c}(\log \dim_D + A)$.
- For all $D \geqslant 1$, there exists a surjective application $\theta \in \Theta_D \subset [-1, 1]^{\dim_D} \longmapsto \gamma^\theta \in G_D$ such that for all $x \in \Lambda$, $A \geqslant 0$ and $y \in \mathbb{R}^d$ such that $\log(b(y)/m(y)) \leqslant A$, the mapping $\theta \in \Theta_D \longmapsto \gamma^\theta(y-x)/b(y)$ is $c(D, A)$-Lipschitz (with Θ_D endowed with the supremum norm).

The exact value of $\widetilde{c}$ only matters for the constants in the penalty.

Example. *The family $(G_D)_{D\geqslant 1}$ of finite Gaussian translation-scale mixtures of Section 3.2 satisfies H5 with $\dim_D = D(d^2 + d) + D - 1$ for all $D \geqslant 1$.*

Define the models $(S_{r,D})_{r\geqslant 1, D\geqslant 1}$ and their maximum likelihood estimators $(\widehat{\mathfrak{X}}_{r,D}, \widehat{Q}_{r,D}, \widehat{\gamma}_{r,D})$ as in Section 3.2. Then, select the number of states and the model dimension using the penalized likelihood. Let $\mathrm{pen}(n,r,D)$ be a penalty function such that $\mathrm{pen}(n,r,D) \underset{n\to+\infty}{\longrightarrow} 0$ for all r and D and such that there exists a sequence $(u_n)_{n\geqslant 1}$ satisfying $u_n \underset{n\to\infty}{\longrightarrow} +\infty$ and for all n, r, D,

$$\mathrm{pen}(n,r,D) \geqslant u_n(\dim_D + rd + r^2 - 1)\frac{(\log n)^{14}\log\log n}{n}\,.$$

For instance, for any constant $\mathrm{cst} > 0$, this inequality holds by choosing $\mathrm{pen} : (n,r,D) \longmapsto (\mathrm{cst}\cdot\dim_D + r^2)\frac{(\log n)^{15}}{n}$. Let

$$(\widehat{r}_n, \widehat{D}_n) \in \underset{r\leqslant \log n, D \text{ s.t. } \dim_D \leqslant n}{\arg\max} \left(\frac{1}{n}\ell_n(\widehat{\mathfrak{X}}_{r,D}, \widehat{Q}_{r,D}, \widehat{\gamma}_{r,D}) - \mathrm{pen}(n,r,D)\right) \tag{23}$$

and define the final estimators $(\widehat{\mathfrak{X}}_n, \widehat{Q}_n, \widehat{\gamma}_n) = (\widehat{\mathfrak{X}}_{\widehat{r}_n,\widehat{D}_n}, \widehat{Q}_{\widehat{r}_n,\widehat{D}_n}, \widehat{\gamma}_{\widehat{r}_n,\widehat{D}_n})$.

Theorem 5. *Assume that assumptions H1, H2, H3, H4 and H5 hold. Let $\lambda^\star$ be the measure defined in assumption H2. Then, almost surely*

$$\sup_{x\in \mathit{Supp}(\lambda^\star)} W_1(K_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n}(x,\cdot), K^\star(x,\cdot)) \underset{n\to\infty}{\longrightarrow} 0$$

and $\|\widehat{\gamma}_n - \gamma^\star\|_1 \underset{n\to\infty}{\longrightarrow} 0$. In particular, almost surely under $\mathbb{P}^\star$, for all $x \in \mathit{Supp}(\lambda^\star)$, $K_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n}(x,\cdot) \longrightarrow K^\star(x,\cdot)$ for the weak convergence topology and if $\mathbb{P}^X_K$ denotes the distribution of the stationary Markov chain with transition kernel K, $\mathbb{P}^X_{K_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n}} \longrightarrow \mathbb{P}^X_{K^\star}$ for the weak convergence topology.

The remaining sections of this paper are dedicated to the proof of Theorem 5.

B.2 Proof of Theorem 5

This section states a few intermediate results whose proofs are postponed to the following sections. These results are followed by the proof of Theorem 5, the consistency of the maximum likelihood estimator, which is the main result of this appendix. Let Ω^C_ω be the set of transition kernels K on Λ which admit the modulus of continuity ω with respect to the Wasserstein 1 metric and such that there exists a probability measure λ (which may depend on K) such that for all $x \in \Lambda$, $K(x,\cdot)$ is absolutely continuous with respect to λ with a density taking values in $[1/C, C]$. The kernel $K^\star$ as well as all kernels considered in the models $S_{r,D}$ belong to Ω^C_ω.

Lemma 2. *Assume that Ω^C_ω is endowed with the topology of the uniform convergence on the set of continuous functions with values in $(\mathcal{P}(\Lambda), W_1)$, and Γ is endowed with the $\mathbf{L}_1$ topology. Then $\Omega^C_\omega \times \Gamma$ endowed with the product topology is compact.*

For all probability measures μ and ν, the Kullback Leibler divergence between μ and ν is defined by

$$KL(\mu\|\nu) = \begin{cases} \int \log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\mu & \text{when } \mu \text{ is absolutely continuous with respect to } \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 3. *Let $(K_n, \gamma_n)_{n\geqslant 1} \in (\Omega^C_\omega \times \Gamma)^{\mathbb{N}^*}$. Then for all $n \geqslant 1$, the quantity $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{K_n,\gamma_n}) = \lim_{m\to+\infty}\frac{1}{m}KL(\mathbb{P}^{(m)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m)}_{K_n,\gamma_n})$ exists and is finite, and the following two statements are equivalent.*

1. $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{K_n,\gamma_n}) \underset{n\to\infty}{\longrightarrow} 0$.
2. *For all* $k \geqslant 1$, $d_{TV}(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{K_n,\gamma_n}) \underset{n\to\infty}{\longrightarrow} 0$.

The consistency of the maximum likelihood estimator relies on the following oracle inequality, which follows from [Lehéricy, 2018, Theorem 8]. It is proved in detail in Section B.7 how Proposition 1 is deduced from [Lehéricy, 2018, Theorem 8] in the setting of this paper.

Proposition 1. *For each* r *and* D*, let* $S_{r,D}$ *and* $(\widehat{\mathfrak{X}}_{r,D}, \widehat{Q}_{r,D}, \widehat{\gamma}_{r,D})$ *be defined as in Section 3.2 and Equation* (5) *and let* $(\widehat{r}_n, \widehat{D}_n)$ *be defined as in Equation* (23)*. There exist constants* C_{pen}*,* A *and* n_0 *such that the following holds. Assume that the penalty satisfies* $\mathrm{pen}(n,r,D) \geqslant C_{\mathrm{pen}}(\dim_D + rd + r^2 - 1)\log(n)^{14}/n$ *for all* $n \geqslant n_0$*,* r *and* D*. Then, for all* $n \geqslant n_0$*, with probability at least* $1 - 3n^{-2}$*,*

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n,\widehat{\gamma}_n}) \leqslant 2 \inf_{r\leqslant \log n, D \text{ s.t. } \dim_D \leqslant n} \left(\inf_{(\mathcal{X},Q,\gamma)\in S_{r,D}} \mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\mathfrak{X},Q,\gamma}) + 2\mathrm{pen}(n,r,D) \right) + A\frac{(\log n)^9}{n}\,.$$

Lemma 4. *Let* $(K_n,\gamma_n)_{n\geqslant 1} \in (\Omega_\omega^C \times \Gamma)^{\mathbb{N}^*}$ *be a sequence that converges to* (K,γ)*. Then, for all* $k \geqslant 1$*,* $d_{TV}(\mathbb{P}^{(k)}_{K,\gamma}, \mathbb{P}^{(k)}_{K_n,\gamma_n}) \underset{n\to\infty}{\longrightarrow} 0$.

Lemma 5. *There exists a sequence* $(\mathfrak{X}_t, Q_t, \gamma_t)_{t\geqslant 1}$ *taking values in* $\bigcup_{r\geqslant 1, D\geqslant 1} S_{r,D}$ *such that* $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\mathfrak{X}_t,Q_t,\gamma_t}) \underset{t\to\infty}{\longrightarrow} 0$.

Theorem 5 may now be proved. Proposition 1 actually gives a deterministic function $f : \mathbb{N}^* \longrightarrow \mathbb{R}_+$ such that for all $n \geqslant n_0$, with probability at least $1 - 3n^{-2}$,

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n,\widehat{\gamma}_n}) \leqslant f(n)\,.$$

By Lemma 5, the assumption that $\mathrm{pen}(n,r,D)$ goes to zero as n goes to infinity for each r and D and Borel-Cantelli Lemma, almost surely,

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n,\widehat{\gamma}_n}) \underset{n\to\infty}{\longrightarrow} 0\,.$$

Thus, by Lemma 3, almost surely, for all $k \geqslant 1$,

$$d_{\mathrm{TV}}\left(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n,\widehat{\gamma}_n}\right) \underset{n\to+\infty}{\longrightarrow} 0\,.$$

In particular, by Lemma 4, all limits (K,γ) of convergent subsequences of $(K_{\widehat{\mathfrak{X}}_n,\widehat{Q}_n}, \widehat{\gamma}_n)_n$ satisfy $\mathbb{P}^{(2)}_{K^\star,\gamma^\star} = \mathbb{P}^{(2)}_{K,\gamma}$, which means that $R_{K^\star} = R_K$ and $\gamma = \gamma^\star$ by Theorem 1 using assumption H1, the fact that $R_{K^\star}$ and R_K are in $\mathcal{M}_1$ since their support is in the compact set Λ^2 and the fact that the translation parameter is fixed by the centering condition on the densities. Therefore, using the continuity of K and $K^\star$, it follows that $K(x,\cdot) = K^\star(x,\cdot)$ for all $x \in \mathrm{Supp}(\lambda^\star)$. Since the set of parameters is compact by Lemma 2, Theorem 5 follows.

B.3 Proof of Lemma 2

Let Ω_ω be the set of transition kernels on Λ which admit the modulus of continuity ω with respect to the Wasserstein 1 metric. Ω_ω is an equicontinuous family of functions from Λ to the set of probability measures $\mathcal{P}(\Lambda)$ on Λ endowed with the Wasserstein 1 metric. Since Λ is compact, convergence in Wasserstein distance is equivalent to convergence in distribution and $\mathcal{P}(\Lambda)$ is compact for the topology of the convergence in distribution, so that Arzelà-Ascoli's theorem ensures that Ω_ω is relatively compact in the class of continuous functions from Λ to $(\mathcal{P}(\Lambda), W_1)$ with respect to the uniform convergence distance. It is closed, therefore it is compact.

Recall that Ω_ω^C is the subset of Ω_ω such that $K \in \Omega_\omega^C$ if and only if there exists a probability measure λ such that for all $x \in \Lambda$, $K(x,\cdot)$ is absolutely continuous with respect to λ with a density taking values in $[1/C, C]$. Let us show that it is closed. Let $(K_n)_{n\geqslant 1}$ be a convergent sequence in Ω_ω^C and $(\lambda_n)_{n\geqslant 1}$ the associated probability measures. Write $K \in \Omega_\omega$ its limit. Without loss of generality, it is possible to assume that $\lambda_n \longrightarrow \lambda$ for some $\lambda \in \mathcal{P}(\Lambda)$ as n grows to $+\infty$. Let $\mathcal{C}_{b,+}^0$ be the set of real-valued, nonnegative, bounded and continuous function on Λ, then for all $f \in \mathcal{C}_{b,+}^0$ and all $x \in \Lambda$,

$$\int K_n(x, \mathrm{d}x') f(x') \in \left[\frac{1}{C}\int f \mathrm{d}\lambda, C\int f \mathrm{d}\lambda\right]$$

by definition of Ω_ω^C. Then, using the convergence of the sequences, for all $f \in \mathcal{C}_{b,+}^0$ and all $x \in \Lambda$,

$$\int K(x, \mathrm{d}x') f(x') \in \left[\frac{1}{C}\int f \mathrm{d}\lambda, C\int f \mathrm{d}\lambda\right] .$$

For all closed set $F \subset \Lambda$, there exists a sequence $(f_i)_{i\geqslant 1} \searrow \mathbf{1}_F$. Therefore, for all closed set $F \subset \Lambda$ and all $x \in \Lambda$,

$$K(x, F) \in \left[\frac{\lambda(F)}{C}, C\lambda(F)\right] .$$

Thus, using the regularity of Borel probability measures on polish spaces, the same holds for all measurable sets, so that $K \in \Omega_\omega^C$. Therefore, Ω_ω^C is closed, so that it is compact.

B.4 Proof of Lemma 3

The following lemma follows from the proof of Lemma 3 of [Douc et al., 2004]. In this section only, for all integers $a \leqslant b$, write Y_a^b instead of $(Y_a, \dots, Y_b)$.

Lemma 6. *Assume that assumption H3 holds. By stationarity, extend the process $(Y_t)_{t\geqslant 1}$ into a process $(Y_t)_{t\in\mathbb{Z}}$. Let $K, K' \in \Omega_\omega^C$ and $\gamma, \gamma' \in \Gamma$. Then, there exists random variables $\delta_{k,\infty}(K,\gamma)$ and $\delta_{k,\infty}(K',\gamma')$ such that almost surely, for all $k \in \mathbb{Z}$ and $m \geqslant 0$,*

$$\left|\log \frac{p_{Y_k|Y_{k-m}^{k-1},K,\gamma}(Y_k|Y_{k-m}^{k-1})}{p_{Y_k|Y_{k-m}^{k-1},K',\gamma'}(Y_k|Y_{k-m}^{k-1})} - \log\frac{\delta_{k,\infty}(K,\gamma)}{\delta_{k,\infty}(K',\gamma')}\right| \leqslant 2C^2\left(1-\frac{1}{C^2}\right)^{m-1} ,$$

and for all $k \in \mathbb{Z}$,

$$\left(\sup_{m\geqslant 0}\left|\log\frac{p_{Y_k|Y_{k-m}^{k-1},K,\gamma}(Y_k|Y_{k-m}^{k-1})}{p_{Y_k|Y_{k-m}^{k-1},K',\gamma'}(Y_k|Y_{k-m}^{k-1})}\right|\right) \vee \left|\log\frac{\delta_{k,\infty}(K,\gamma)}{\delta_{k,\infty}(K',\gamma')}\right| \in \mathbf{L}^1(\mathbb{P}^\star) .$$

Proof. Write first how the notations of this paper match those of [Douc et al., 2004]. The set $\mathcal{X}$ (resp. $\mathcal{Y}$) of [Douc et al., 2004] is Λ (resp. $\mathbb{R}^d$) and $\mathbb{R}^d$ is equiped with the measure with density $b/\|b\|_1$ with respect to the Lebesgue measure. Finally, the set Θ of [Douc et al., 2004] is $\{(K,\gamma),(K',\gamma')\}$. Contrary to the setting of [Douc et al., 2004], $\mathcal{X}$ is endowed with a measure that depends on the parameter θ. The proof of Lemma 3 of [Douc et al., 2004] holds with the following relaxed assumptions (with the notations of [Douc et al., 2004]).

(A1') For all $\theta \in \Theta$, there exists a measure μ_θ on $\mathcal{X}$ such that the transition kernel of $(X_k)_{k\geqslant 1}$ has a density q_θ with respect to μ_θ such that for all $x, x' \in \mathcal{X}$, $1/C \leqslant q_\theta(x,x') \leqslant C$.

(A3') $\bar{\mathbb{E}}_{\theta^*}[|\log b_+(Y_1, \bar{\mathbf{Y}}_0)|] < \infty$ and $\bar{\mathbb{E}}_{\theta^*}[|\log b_-(Y_1, \bar{\mathbf{Y}}_0)|] < \infty$ where $b_+(y_1, \bar{\mathbf{y}}_0) \stackrel{\Delta}{=} \sup_\theta \int_{\mathcal{X}} g_\theta(y_1|\bar{\mathbf{y}}_0, x)\mu_\theta(\mathrm{d}x)$ and $b_-(y_1, \bar{\mathbf{y}}_0) \stackrel{\Delta}{=} \inf_\theta \int_{\mathcal{X}} g_\theta(y_1|\bar{\mathbf{y}}_0, x)\mu_\theta(\mathrm{d}x)$.

These assumptions are equivalent to the following (A1") and (A3").

(A1") There exists a measure λ_K on Λ such that the transition kernel K has a density with respect to λ_K with values in $[1/C, C]$, and likewise for K'.

(A3") $\mathbb{E}^\star[|\log \int_\Lambda \|b\|_1(\gamma(Y_1 - x)/b(Y_1))\mathrm{d}\lambda_K(x)|] < \infty$, and likewise for (K',γ').

The lemma then follows from Lemma 3 of [Douc et al., 2004] applied on (K,γ) and (K',γ'). (A1") is direct by definition of Ω_ω^C. By H4, $\|b\|_1 m(y)/b(y) \leqslant \int_\Lambda g_x(y) d\lambda_K(x) \leqslant \|b\|_1$. Thus, (A3") is implied by the integrability condition of H4 since the distribution of Y_1 under $\mathbb{P}^\star$ is dominated by the distribution with density b with respect to the Lebesgue measure. □

Thus, for all $K, K' \in \Omega_\omega^C$ and $\gamma, \gamma' \in \Gamma$, the limit

$$\mathbf{K}(\mathbb{P}_{K,\gamma}\|\mathbb{P}_{K',\gamma'}) = \lim_{m\to+\infty} \frac{1}{m} KL(\mathbb{P}^{(m)}_{K,\gamma}\|\mathbb{P}^{(m)}_{K',\gamma'}) = \mathbb{E}_{K,\gamma}\left[\log \frac{\delta_{0,\infty}(K,\gamma)}{\delta_{0,\infty}(K',\gamma')}\right]$$

exists, is finite, and for all $k, m \geqslant 1$,

$$\left|k\mathbf{K}(\mathbb{P}_{K,\gamma}\|\mathbb{P}_{K',\gamma'}) - \left(KL(\mathbb{P}^{(m+k)}_{K,\gamma}\|\mathbb{P}^{(m+k)}_{K',\gamma'}) - KL(\mathbb{P}^{(m)}_{K,\gamma}\|\mathbb{P}^{(m)}_{K',\gamma'})\right)\right| \leqslant 2C^4\left(1 - \frac{1}{C^2}\right)^{m-1}.$$

Let $(K_n,\gamma_n)_{n\geqslant 1} \in (\Omega_\omega^C \times \Gamma)^{\mathbb{N}}$ be a sequence of parameters such that $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{K_n,\gamma_n}) \longrightarrow 0$. The above equation implies that for all $k \geqslant 1$, there exists sequences $(m_n)_{n\geqslant 1} \longrightarrow +\infty$ and $(l_n)_{n\geqslant 1} \longrightarrow +\infty$ such that

$$KL(\mathbb{P}^{(m_n+l_n+k)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n+l_n+k)}_{K_n,\gamma_n}) - KL(\mathbb{P}^{(m_n)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n)}_{K_n,\gamma_n}) \underset{n\to\infty}{\longrightarrow} 0\,.$$

Using the chain rule and Pinsker's inequality,

$$\begin{aligned}
&KL(\mathbb{P}^{(m_n+l_n+k)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n+l_n+k)}_{K_n,\gamma_n}) - KL(\mathbb{P}^{(m_n)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n)}_{K_n,\gamma_n}) \\
&\qquad = \mathbb{E}_{Y_1^{m_n}|K^\star,\gamma^\star}\left[KL\left(\mathbb{P}_{Y_{m_n+1}^{m_n+l_n+k}|Y_1^{m_n},K^\star,\gamma^\star}\|\mathbb{P}_{Y_{m_n+1}^{m_n+l_n+k}|Y_1^{m_n},K_n,\gamma_n}\right)\right], \\
&\qquad \geqslant \mathbb{E}_{Y_1^{m_n}|K^\star,\gamma^\star}\left[KL\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K^\star,\gamma^\star}\|\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K_n,\gamma_n}\right)\right], \\
&\qquad \geqslant 2\mathbb{E}_{Y_1^{m_n}|K^\star,\gamma^\star}\left[d^2_{\mathrm{TV}}\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K^\star,\gamma^\star},\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K_n,\gamma_n}\right)\right].
\end{aligned}$$

Since the kernels satisfy the Doeblin condition (see for instance [Cappé et al., 2005], Section 4.3.3), the resulting processes are ϕ-mixing with mixing coefficients $\phi(i) \leqslant 2(1-1/C)^i$ (see the proof of Lemma 1 of [Lehéricy, 2018] for a proof, and [Bradley, 2005] for a survey of mixing properties). In particular, for all $K \in \Omega_\omega^C$, for all positive and continuous probability density γ on $\mathbb{R}^d$ and for all $A \in \sigma(Y_1, \dots, Y_{m_n})$ such that $\mathbb{P}_{K,\gamma}(A) > 0$,

$$d_{\text{TV}}\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|A,K,\gamma}, \mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|K,\gamma}\right) \leqslant 2\left(1-\frac{1}{C}\right)^{l_n},$$

so that using the continuity and positivity of γ,

$$d_{\text{TV}}\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K,\gamma}, \mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|K,\gamma}\right) \leqslant 2\left(1-\frac{1}{C}\right)^{l_n}.$$

Finally,

$$\begin{aligned}
2\mathbb{E}_{Y_1^{m_n}|K^\star,\gamma^\star}&\left[d_{\text{TV}}^2\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K^\star,\gamma^\star}, \mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|Y_1^{m_n},K_n,\gamma_n}\right)\right] \\
&\geqslant 2\left(d_{\text{TV}}\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|K^\star,\gamma^\star}, \mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|K_n,\gamma_n}\right) - 4\left(1-\frac{1}{C}\right)^{l_n}\right)^2, \\
&\geqslant d_{\text{TV}}^2\left(\mathbb{P}_{K^\star,\gamma^\star}^{(k)}, \mathbb{P}_{K_n,\gamma_n}^{(k)}\right) - 32\left(1-\frac{1}{C}\right)^{2l_n},
\end{aligned}$$

using that $(a-b)^2 \geqslant a^2/2 - b^2$ for all $a, b \in \mathbb{R}$ and the stationarity of the distributions $\mathbb{P}_{K,\gamma}$ for all $K \in \Omega_\omega^C$ and $\gamma \in \Gamma$. Therefore, for all $k \geqslant 1$,

$$d_{\text{TV}}\left(\mathbb{P}_{K^\star,\gamma^\star}^{(k)}, \mathbb{P}_{K_n,\gamma_n}^{(k)}\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Conversely, let $(K_n, \gamma_n)_{n\geqslant 1} \in (\Omega_\omega^C \times \Gamma)^{\mathbb{N}^*}$ be a sequence of parameters such that for all $k \geqslant 1$,

$$d_{\text{TV}}\left(\mathbb{P}_{K^\star,\gamma^\star}^{(k)}, \mathbb{P}_{K_n,\gamma_n}^{(k)}\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Then by Lemma 6, for all $k, n \geqslant 1$,

$$\begin{aligned}
\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{K_n,\gamma_n}) &\leqslant \mathbb{E}KL(\mathbb{P}_{Y_k|Y_1^{k-1},K^\star,\gamma^\star} \| \mathbb{P}_{Y_k|Y_1^{k-1},K_n,\gamma_n}) + 2C^2\left(1-\frac{1}{C^2}\right)^{k-2} \\
&\leqslant KL(\mathbb{P}_{K^\star,\gamma^\star}^{(k)} \| \mathbb{P}_{K_n,\gamma_n}^{(k)}) + 2C^2\left(1-\frac{1}{C^2}\right)^{k-2}, \qquad (24)
\end{aligned}$$

by the entropy chain rule. Lemma 4 of [Shen et al., 2013] entails that there exists $\lambda_0 \in (0,1)$ such that for

all $\lambda \in (0, \lambda_0)$,

$$\begin{aligned} KL(\mathbb{P}^{(k)}_{K^\star,\gamma^\star} \| \mathbb{P}^{(k)}_{K_n,\gamma_n}) &\leqslant \left(1 + 2k \log \frac{1}{\lambda}\right) h^2(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{K_n,\gamma_n}) \\ &\quad + 2\mathbb{E}\left[\log\left(\frac{p_{Y_1^k|K^\star,\gamma^\star}}{p_{Y_1^k|K_n,\gamma_n}}\right) \mathbf{1}\left(\frac{p_{Y_1^k|K^\star,\gamma^\star}}{p_{Y_1^k|K_n,\gamma_n}} \geqslant \frac{1}{\lambda}\right)\right], \\ &\leqslant 2\left(1 + 2k \log \frac{1}{\lambda}\right) d_{\mathrm{TV}}(\mathbb{P}^{(k)}_{K^\star,\gamma^\star,}, \mathbb{P}^{(k)}_{K_n,\gamma_n}) \\ &\quad + 2\int \prod_{i=1}^k b(y_i) \log\left(\prod_{i=1}^k \frac{b(y_i)}{m(y_i)}\right) \mathbf{1}\left(\prod_{i=1}^k \frac{b(y_i)}{m(y_i)} \geqslant \frac{1}{\lambda}\right) \mathrm{d}y\,, \end{aligned}$$

using that the square of the Hellinger distance is upper bounded by the $\mathbf{L}^1$ distance, that is twice the total variation distance. The second term is finite for all λ by H4. Therefore, by carefully choosing a sequence λ that tends to zero, we obtain $\limsup_n KL(\mathbb{P}^{(k)}_{K^\star,\gamma^\star} \| \mathbb{P}^{(k)}_{K_n,\gamma_n}) = 0$ for all $k \geqslant 1$. This, together with taking k that tends to infinity in Equation (24), proves the second statement of the lemma.

B.5 Proof of Lemma 4

The set of possible parameters $\Omega_\omega^C \times \Gamma$ is endowed with the product topology induced by the uniform convergence topology on Ω_ω^C and the $\mathbf{L}^1$ norm on Γ. It is compact for this topology. Let $(K_n, \gamma_n)_{n\geqslant 1}$ be a sequence in $\Omega_\omega \times \Gamma$ that converges to (K, γ) with respect to this topology. The aim is now to show that the distribution of $(Y_1, \dots, Y_k)$ with parameters (K_n, γ_n) converges in total variation distance to the distribution with parameters (K, γ). The transition kernel K admits a unique stationary distribution, so that Theorem 4 and the corollary of Theorem 6 of [Karr, 1975] entail that

$$\mathbb{P}^X_{K_n} \xrightarrow[n\to\infty]{(d)} \mathbb{P}^X_K\,, \tag{25}$$

where $\mathbb{P}^X_K$ denotes the distribution of a stationary Markov chain $(X_n)_{n\geqslant 1}$ with transition kernel K. This convergence holds for the distribution of the whole Markov chain, which implies in particular that the distribution of k-tuples $(X_1, \dots, X_k)$ for all $k \geqslant 1$ converges in the same way. For any $k \geqslant 1$, the total variation distance between the distributions of $(Y_1, \dots, Y_k)$ is, up to a factor 2,

$$\begin{aligned} \|p_{(Y_1,\dots,Y_k)|K,\gamma} - p_{(Y_1,\dots,Y_k)|K_n,\gamma_n}\|_1 &= \int \left| \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}^X_K(x) - \int \prod_{i=1}^k \gamma_n(y_i - x_i) \mathrm{d}\mathbb{P}^X_{K_n}(x) \right| \mathrm{d}y\,, \\ &\leqslant \int \left| \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}^X_K(x) - \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}^X_{K_n}(x) \right| \mathrm{d}y\,, \\ &\quad + \int \int \left| \prod_{i=1}^k \gamma(y_i - x_i) - \prod_{i=1}^k \gamma_n(y_i - x_i) \right| \mathrm{d}\mathbb{P}^X_{K_n}(x) \mathrm{d}y\,. \end{aligned}$$

Consider the first term of the right hand side. Since $x \longmapsto \gamma(y - x)$ is continuous and bounded for all $y \in \mathbb{R}^d$, Equation (25) yields, for all $y \in \mathbb{R}^d$,

$$\left| \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}^X_K(x) - \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}^X_{K_n}(x) \right| \xrightarrow[n\to\infty]{} 0\,.$$

Then, since $\sup_{x\in\Lambda}\gamma(y-x)\leqslant b(y)$ for all $y\in\mathbb{R}^d$, $\left|\int\prod_{i=1}^k\gamma(y_i-x_i)\mathrm{d}\mathbb{P}_K^X(x)\right|\leqslant\prod_{i=1}^k b(y_i)$, and the right hand side is integrable. The same holds for K_n, so that the dominated convergence theorem implies

$$\int\left|\int\prod_{i=1}^k\gamma(y_i-x_i)\mathrm{d}\mathbb{P}_K^X(x)-\int\prod_{i=1}^k\gamma(y_i-x_i)\mathrm{d}\mathbb{P}_{K_n}^X(x)\right|\mathrm{d}y\underset{n\to\infty}{\longrightarrow}0\,.$$

For the second term, write

$$\begin{aligned}
&\int\int\left|\prod_{i=1}^k\gamma(y_i-x_i)-\prod_{i=1}^k\gamma_n(y_i-x_i)\right|\mathrm{d}\mathbb{P}_{K_n}^X(x)\mathrm{d}y\\
&\quad\leqslant\sum_{i=1}^k\int\int\prod_{j<i}\gamma(y_j-x_j)\left|\gamma(y_i-x_i)-\gamma_n(y_i-x_i)\right|\prod_{j>i}\gamma_n(y_j-x_j)\mathrm{d}\mathbb{P}_{K_n}^X(x)\mathrm{d}y\,,\\
&\quad\leqslant\sum_{i=1}^k\int\int\left|\gamma(y_i-x_i)-\gamma_n(y_i-x_i)\right|\mathrm{d}y_i\mathrm{d}\mathbb{P}_{K_n}^X(x_i)\,,\\
&\quad=k\|\gamma-\gamma_n\|_1\,,
\end{aligned}$$

where the last term converges to 0 as $n\to\infty$. Hence, $d_{\mathrm{TV}}(\mathbb{P}_{K,\gamma}^{(k)},\mathbb{P}_{K_n,\gamma_n}^{(k)})\underset{n\to\infty}{\longrightarrow}0$ for all $k\geqslant 1$.

B.6 Proof of Lemma 5

By Lemmas 3 and 4, it suffices to show that there exists a sequence $(\mathfrak{X}_t,Q_t)_{t\geqslant1}$ such that $(\mathfrak{X}_t,Q_t,-)\in\bigcup_{r,D}S_{r,D}$ and such that the sequence of kernels $(K_t)_{t\geqslant1}=(K_{\mathfrak{X}_t,Q_t})_{t\geqslant1}$ converges to $K^\star$. The following lemma, which is a consequence of simple algebra, is stated without proof.

Lemma 7. *Let λ be a probability measure on a compact set of $\mathbb{R}^d$ which is absolutely continuous with respect to the Lebesgue measure. Then, there exists a sequence of integers $(r_t)_{t\geqslant1}\longrightarrow+\infty$ and a sequence $((A_i^t)_{1\leqslant i\leqslant r_t})_{t\geqslant1}$ of measurable partitions of the support of λ such that*

$$\begin{cases}D_t=\sup\limits_{1\leqslant i\leqslant r_t}\mathit{diam}(A_i^t)\underset{t\to+\infty}{\longrightarrow}0\,,\\ \forall t\geqslant1,\quad\forall1\leqslant i\leqslant r_t,\quad\lambda(A_i^t)\in\left[\frac{1}{2r_t},\frac{2}{r_t}\right]\,.\end{cases}$$

To address the case where $\lambda^\star$ is not absolutely continuous with respect to the Lebesgue measure, consider convolutions of the kernels. For all $\epsilon\in(0,1]$, let U_ϵ be the uniform measure on $[-\epsilon,\epsilon]^d$. For all probability measure λ on $\mathbb{R}^d$, write $\lambda*U_\epsilon$ the convolution of λ and U_ϵ, and for all transition kernel K on $\mathbb{R}^d$, write $K*U_\epsilon$ the transition kernel defined by $(K*U_\epsilon)(x,\cdot)=K(x,\cdot)*U_\epsilon$. Then $K^\star*U_\epsilon$ admit the modulus of continuity ω for all $\epsilon>0$ (since $W_1(\mu*U_\epsilon,\nu*U_\epsilon)\leqslant W_1(\mu,\nu)$ for all probability measures μ,ν) and $K^\star*U_\epsilon$ admits a density taking values in $[2/C,C/2]$ with respect to the measure $\lambda^\star*U_\epsilon$ (which is absolutely continuous with respect to the Lebesgue measure), so that it belongs to Ω_ω^C (up to enlarging Λ). Moreover, $K^\star*U_\epsilon\longrightarrow K^\star$ in Ω_ω^C as $\epsilon\longrightarrow0$. Therefore, it remains to show that for all $\epsilon>0$, the kernel $K^\star*U_\epsilon$ can be approximated by kernels in Ω_ω^C with finite support. Equivalently, assume that $\lambda^\star$ is absolutely continuous with respect to the Lebesgue measure and construct a sequence approximating $K^\star$.

Let $(r_t)_{t\geqslant1}$ and $((A_i^t)_{1\leqslant i\leqslant r_t})_{t\geqslant1}$ be the sequences obtained by applying Lemma 7 to $\lambda^\star$. For all $t\geqslant1$ and $i\in\{1,\dots,r_t\}$, let x_i^t be an element of A_i^t. For all $t\geqslant1$, the elements of the vector $\mathfrak{X}_t=(x_i^t)_{1\leqslant i\leqslant r_t}$

are distinct because $(A_i^t)_{1\leqslant i\leqslant r_t}$ is a partition of Supp$(\lambda^\star)$. Let $(\eta_t)_{t\geqslant 1} \longrightarrow 0$ be a sequence of positive numbers. Let $\widetilde{K}_t$ be the transition kernel from $\Lambda\cap(\eta_t\mathbb{Z}^d)$ to $\{x_i^t\}_{1\leqslant i\leqslant r_t}$ defined, for all $x\in\Lambda\cap(\eta_t\mathbb{Z}^d)$ and all $i\in\{1,\dots,r_t\}$, by

$$\widetilde{K}_t(x,x_i^t) = K^\star(x,A_i^t)\,.$$

By the Lemma 7 and assumption H2, $\widetilde{K}_t(x,x_i^t) \in [1/(Cr_t), C/r_t]$ for all x and i. Moreover, for all $x,x'\in\Lambda\cap(\eta_t\mathbb{Z}^d)$,

$$\begin{aligned} W_1(\widetilde{K}_t(x,\cdot),\widetilde{K}_t(x',\cdot)) \leqslant W_1(K^\star(x,\cdot),K^\star(x',\cdot)) + 2\sup_{1\leqslant i\leqslant r_t}\operatorname{diam}(A_i^t) &\leqslant \frac{\omega(|x-x'|)}{2} + 2\frac{D_t}{\eta_t}|x-x'|\,,\\ &\leqslant \omega(|x-x'|)\,, \end{aligned}$$

by choosing $\eta_t \geqslant 4D_t/\inf_{u\in(0,\operatorname{diam}(\Lambda)]}\omega(u)/u$, which is finite since ω is concave, nondecreasing and not equal to zero, so that there exists an extension $K_t\in\Omega_\omega^C$ of $\widetilde{K}_t$ such that the support of $K_t(x,\cdot)$ is $\{x_i^t\}_{1\leqslant i\leqslant r_t}$ for all $x\in\Lambda$.

For all i, j, define $Q_t(i,j) = K_t(x_i^t,x_j^t)$. All kernels considered here ($K^\star$, $\widetilde{K}_t$, K_t and $K_{\mathfrak{X}_t,Q_t}$) are kernels on the compact set Supp$(\lambda^\star)$. Therefore, we only need to show that $K_{\mathfrak{X}_t,Q_t}\longrightarrow K$ in the subset $\widetilde{\Omega}_\omega^C$ of kernels on Supp$(\lambda^\star)$ in Ω_ω^C to show that it is an approximating sequence, that is

$$\sup_{x\in\text{Supp}(\lambda^\star)} W_1(K_{\mathfrak{X}_t,Q_t}(x,\cdot),K^\star(x,\cdot)) \underset{t\to+\infty}{\longrightarrow} 0\,. \tag{26}$$

For all $x\in$ Supp$(\lambda^\star)$, let $X(x)$ (resp. $\mathfrak{X}(x)$) be one of the elements of $\Lambda\cap(\eta_t\mathbb{Z}^d)$ (resp. $\{x_i^t\}_{1\leqslant i\leqslant r_t}$) closest to x. Then $\sup_{x\in\text{Supp}(\lambda^\star)}|x-\mathfrak{X}(x)|\leqslant D_t$ and $\sup_{x\in\text{Supp}(\lambda^\star)}|x-X(x)|\leqslant\eta_t$ (with the supremum norm on $\mathbb{R}^d$) and for all $x\in$ Supp$(\lambda^\star)$,

$$\begin{aligned} W_1(K_{\mathfrak{X}_t,Q_t}(x,\cdot),K^\star(x,\cdot)) \leqslant & W_1(K_{\mathfrak{X}_t,Q_t}(x,\cdot),K_{\mathfrak{X}_t,Q_t}(\mathfrak{X}(x),\cdot))\\ &+ W_1(K_{\mathfrak{X}_t,Q_t}(\mathfrak{X}(x),\cdot),K_t(\mathfrak{X}(x),\cdot)) &(27)\\ &+ W_1(K_t(\mathfrak{X}(x),\cdot),K_t(X(\mathfrak{X}(x)),\cdot))\\ &+ W_1(K_t(X(\mathfrak{X}(x)),\cdot),K^\star(X(\mathfrak{X}(x)),\cdot)) &(28)\\ &+ W_1(K^\star(X(\mathfrak{X}(x)),\cdot),K^\star(x,\cdot))\,. \end{aligned}$$

By definition of the kernels, (27) and (28) are equal to 0. Thus, the regularity assumptions on the kernels ensure that for all $x\in$ Supp$(\lambda^\star)$,

$$W_1(K_{\mathfrak{X}_t,Q_t}(x,\cdot),K^\star(x,\cdot)) \leqslant \omega(D_t)+\omega(\eta_t)+\omega(D_t+\eta_t)/2\,,$$

which proves Equation (26).

B.7 Proof of Proposition 1

This section first states Theorem 8 of [Lehéricy, 2018] and its assumptions. It is then proved that the assumptions are satisfied and that Proposition 1 is deduced from this theorem. Let λ_b be the probability measure on $\mathbb{R}^d$ which has the density $b/\|b\|_1$ with respect to the Lebesgue measure. When necessary, the process $(Y_t)_{t\geqslant 1}$ is extended to a process $(Y_t)_{t\in\mathbb{Z}}$ by stationarity. In this section only, for all integers $a\leqslant b$, write Y_a^b instead of $(Y_a,\dots,Y_b)$.

[A⋆forgetting] There exists two constants $C_\star > 0$ and $\rho_\star \in (0,1)$ such that for all $i \in \mathbb{Z}$, for all $k, k' \in \mathbb{N}^*$ and for all $y_{i-(k\vee k')}^i \in (\mathbb{R}^d)^{(k\vee k')+1}$,

$$\left| \log\left(\frac{\mathrm{d}\mathbb{P}_{Y_i|Y_{i-k}^{i-1},K^\star,\gamma^\star}}{\mathrm{d}\lambda_b}(y_i|y_{i-k}^{i-1}) \right) - \log\left(\frac{\mathrm{d}\mathbb{P}_{Y_i|Y_{i-k'}^{i-1},K^\star,\gamma^\star}}{\mathrm{d}\lambda_b}(y_i|y_{i-k'}^{i-1}) \right) \right| \leqslant C_\star \rho_\star^{k\wedge k'-1} .$$

Let $(\Omega, \mathcal{F}, P)$ be a measured space and $\mathcal{A} \subset \mathcal{F}$ and $\mathcal{B} \subset \mathcal{F}$ be two sigma-fields. Then, the ρ-mixing coefficient between $\mathcal{A}$ and $\mathcal{B}$ is

$$\rho_{\text{mix}}(\mathcal{A}, \mathcal{B}) = \sup_{\substack{f \in \mathbf{L}^2(\Omega,\mathcal{A},P) \\ g \in \mathbf{L}^2(\Omega,\mathcal{B},P)}} |\text{Corr}(f,g)| .$$

The ρ-mixing coefficient of $(Y_t)_{t\in\mathbb{Z}}$ is

$$\rho_{\text{mix}}(n) = \rho_{\text{mix}}(\sigma(Y_i, i \geqslant n), \sigma(Y_i, i \leqslant 0)) .$$

[A⋆mixing] There exists two constants $c_\star > 0$ and $n_\star \in \mathbb{N}^*$ such that for all $n \geqslant n_\star$, $\rho_{\text{mix}}(n) \leqslant 4e^{-c_\star n}$.

[A⋆tail] There exists a constant $B^\star \geqslant 1$ such that for all $i \in \mathbb{Z}$, all $k \in \mathbb{N}$ and all $v \geqslant e$,

$$\mathbb{P}\left(\frac{\mathrm{d}\mathbb{P}_{Y_i|Y_{i-k}^{i-1},K^\star,\gamma^\star}}{\mathrm{d}\lambda_b}(Y_i|Y_{i-k}^{i-1}) \geqslant v^{B^\star} \right) \leqslant \frac{1}{v} .$$

[Lehéricy, 2018] considers models written $T_{r,D}$ in the following (instead of $S_{K,M,n}$ in [Lehéricy, 2018]). These models are sets of hidden Markov model parameters (not translation hidden Markov models), that is of vectors of the form (r, π, Q, g) where r is the number of values the Markov chain can take, π is the initial distribution of the Markov chain, Q is its transition matrix and $g = (g_z)_{z=1,\dots,r}$ the vector of its emission densities, that is a vector of probability densities on $\mathbb{R}^d$ with respect to the Lebesgue measure. Let $(m_{r,D})_{r\geqslant 1, D\geqslant 1}$ be a sequence of nonnegative integers. For all $n \geqslant 1$, let $\sigma_-(n) \in (0, e^{-1}]$ and let $\mathfrak{P}_n$ be a subset of $\{(r,D) \in (\mathbb{N}^*)^2 : r \leqslant 1/(2\sigma_-(n)) \text{ and } m_{r,D} \leqslant 2n\}$. This set lists the indices of the models among which the final model is selected. Let $\mathbf{T}_n = \bigcup_{(r,D)\in\mathfrak{P}_n} T_{r,D}$ be the set of all model parameters considered when n observations are available.

[Aergodic] For all $(r, \pi, Q, -) \in \mathbf{T}_n$,

$$\inf_{x,x'=1,\dots,r} Q(x,x') \geqslant \sigma_-(n) \quad \text{and} \quad \inf_{x=1,\dots,r} \pi(x) \geqslant \sigma_-(n) .$$

[Atail] There exists a constant $B(n) \geqslant 1$ such that for all $u \geqslant 1$,

$$\mathbb{P}^\star\left(\sup_{(r,-,-,g)\in\mathbf{T}_n} \left| \log \sum_{z=1}^r g_z(Y_1) \right| \geqslant B(n)u \right) \leqslant e^{-u} .$$

Finally, the assumptions **[Aentropy]** and **[Agrowth]** of [Lehéricy, 2018] are replaced by the following more general assumption, which allows to improve the penalty (the original assumptions induce a penalty proportional to $r\dim_D + rd + r^2$ instead of $\dim_D + rd + r^2$). Let $N(B, d, \epsilon)$ be the smallest number of brackets of size ϵ for the distance d needed to cover the set of functions B.

[Aentropy'] There exist a mapping $(r, D, n, A) \longmapsto C_{\text{aux}}(r, D, n, A) \geqslant 1$, a sequence of nonnegative integers $(m_{r,D})_{r\geqslant 1, D\geqslant 1}$ and a family of sets $(\mathcal{S}_{n,A})_{n\geqslant 1, A\geqslant 0} \subset \mathbb{R}^d$ such that for all $n \geqslant 1$ and $A \geqslant 0$, $\mathbb{P}^\star(Y_1 \in \mathcal{S}_{n,A}) \leqslant \exp(-2A/B(n))$ where $B(n)$ is as in **[Atail]**, for all $y \in \mathcal{S}_{n,A}$,

$$\sup_{(r',-,-,g')\in \mathbf{T}_n} \left| \log \sum_{z=1}^{r'} g'_z(y) \right| \leqslant A$$

and for all $r \geqslant 1$, $D \geqslant 1$, $n \geqslant 1$, $A \geqslant B(n)$ and $\delta \in (0,1)$,

$$N\left(\left\{ \left(y \mapsto g_z(y) \mathbf{1}_{y\in\mathcal{S}_{n,A}}\right)_{z=1,\dots,r} \right\}_{(r,-,-,g)\in T_{r,D}}, d_\infty, \delta \right) \leqslant \max\left(\frac{C_{\text{aux}}(r,D,n,A)}{\delta}, 1 \right)^{m_{r,D}}, \quad (29)$$

where d_∞ is the distance associated with the supremum norm on $(\mathbf{L}^\infty(\mathcal{Y}))^r$. Moreover, there exist an integer n_{growth} and a constant $c_{\text{growth}} > 0$ such that for all $n \geqslant n_{\text{growth}}$,

$$\sup_{(r,D)\in\mathfrak{P}_n} \log C_{\text{aux}}(r, D, n, B(n)\log n) \leqslant c_{\text{growth}} (\log n)^2 \log\log n\,.$$

Note that choosing $S_{n,A} = \{y \in \mathbb{R}^d : \sup_{(r',-,-,g')\in\mathbf{T}_n} |\log \sum_{z=1}^{r'} g'_z(y)| \leqslant A\}$ gives the original formulation of [Lehéricy, 2018]. Write $\mathbb{P}_{r,\pi,Q,g}$ the distribution of a hidden Markov model with parameter (r,π,Q,g). Lemma 4 and 5 of [Lehéricy, 2018] show that for all r, D and for all $(r,\pi,Q,g) \in T_{r,D}$, the limit $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{r,Q,g}) = \lim_m m^{-1} KL(\mathbb{P}_{Y_1^m|K^\star,\gamma^\star} \| \mathbb{P}_{Y_1^m|r,\pi,Q,g})$ exists, is finite and does not depend on π. This quantity coincides with the one defined in Lemma 3 when the hidden Markov model with parameter (r,π,Q,g) is a translation hidden Markov model with transition kernel in Ω_ω^C and emission density in Γ. Define the loglikelihood of a hidden Markov model with parameter (r,π,Q,g) by

$$\ell_n^{\text{HMM}}(r,\pi,Q,g) = \log\left(\sum_{z_1,\dots,z_n\in\{1,\dots,r\}} \pi(z_1) g_{z_1}(Y_1) \prod_{t=2}^n Q(z_{t-1},z_t) g_{z_t}(Y_t) \right).$$

Theorem 8 of [Lehéricy, 2018] may now be stated with a noteworthy modification: not all possible number of states and model indices are considered during the model selection step (30), but only the ones in $\mathfrak{P}_n$. This has no consequence on the proof.

Theorem 6. *Assume that **[A⋆forgetting]**, **[A⋆mixing]**, **[A⋆tail]**, **[Aergodic]**, **[Atail]** and **[Aentropy']** hold. Assume that $\sigma_-(n) = C_\sigma(\log n)^{-1}$ and $B(n) = C_B \log n$ for some constants $C_\sigma \geqslant 0$ and $C_B \geqslant 2$. Let $\alpha \geqslant 0$. For all r and D, let*

$$(r, \widehat{\pi}_{r,D,n}, \widehat{Q}_{r,D,n}, \widehat{g}_{r,D,n}) \in \operatorname*{arg\,max}_{(r,\pi,Q,g)\in T_{r,D}} \frac{1}{n} \ell_n^{HMM}(r,\pi,Q,g)\,,$$

$$(\widehat{r}_n, \widehat{D}_n) \in \operatorname*{arg\,max}_{(r,D)\in\mathfrak{P}_n} \left(\frac{1}{n} \ell_n^{HMM}(r, \widehat{\pi}_{r,D,n}, \widehat{Q}_{r,D,n}, \widehat{g}_{r,D,n}) - \text{pen}(n,r,D) \right), \quad (30)$$

for some function pen*, and let*

$$(\widehat{r}_n, \widehat{\pi}_n, \widehat{Q}_n, \widehat{g}_n) = (\widehat{r}_n, \widehat{\pi}_{\widehat{r}_n,\widehat{D}_n,n}, \widehat{Q}_{\widehat{r}_n,\widehat{D}_n,n}, \widehat{g}_{\widehat{r}_n,\widehat{D}_n,n})$$

be the nonparametric maximum likelihood estimator. Then, there exist constants A, C_{pen} and n_0 depending only on α, C_σ, C_B, n_*, c_* and c_{growth} such that for all

$$n \geqslant n_{growth} \vee n_0 \vee \exp\left(C_\sigma\left((1+C_*) \vee \frac{2-\rho_*}{1-\rho_*} \vee e^2\right)\right) \vee \exp\left(\frac{B^*}{C_B}\right) \vee \exp\sqrt{\frac{C_\sigma}{2}(n_*+1)}\,,$$

all $t \geqslant 1$, all $\eta \leqslant 1$, with probability at least $1 - e^{-t} - 2n^{-\alpha}$,

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{\widehat{r}_n,\widehat{Q}_n,\widehat{g}_n}) \leqslant (1+\eta) \inf_{(r,D)\in\mathfrak{P}_n}\left\{\inf_{(r,\pi,Q,g)\in T_{r,D}} \mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{r,Q,g}) + 2\mathrm{pen}(n,r,D)\right\} + \frac{A}{\eta}t\frac{(\log n)^8}{n}$$

as soon as

$$\mathrm{pen}(n,r,D) \geqslant \frac{C_{\mathrm{pen}}}{\eta}(m_{r,D}+r^2-1)\frac{(\log n)^{14}\log\log n}{n}\,.$$

Let us now check the assumptions. **[A⋆mixing]** and **[A⋆forgetting]** follow from Lemma 1 of [Lehéricy, 2018] and from H2 with $\rho_\star = 1-4/C^2$, $C_\star = C^2/4$, $n_\star = 1$ and $c_\star = -\log(1-2/C)/2$, where C is the constant from H2. **[A⋆tail]** follows from assumption H3 with $B^\star = \max(1, \log\|b\|_1)$: by definition of λ_b and b, for all $i \in \mathbb{Z}$, $k \in \mathbb{N}$, $y_{i-k}^i \in (\mathbb{R}^d)^{k+1}$ and $v \geqslant e$,

$$\frac{\mathrm{d}\mathbb{P}_{Y_i|Y_{i-k}^{i-1},K^\star,\gamma^\star}}{\mathrm{d}\lambda_b}(y_i|y_{i-k}^{i-1}) = \frac{\int \gamma^\star(y_i-x)\mathrm{d}\mathbb{P}_{X_i|Y_{i-k}^{i-1},K^\star,\gamma^\star}(x|y_{i-k}^i)}{b(y_i)/\|b\|_1} \leqslant \|b\|_1 \leqslant v^{B^\star}\,.$$

For each $r \geqslant 1$ and $D \geqslant 1$, let $m_{r,D} = \dim_D + rd$. For each $n \geqslant 1$, let $\sigma_-(n) = (2\log n)^{-1}$ and $\mathfrak{P}_n = \{(r,D) : r \leqslant \log n \text{ and } \dim_D \leqslant n\}$. For n large enough, $\mathfrak{P}_n$ is indeed a subset of $\{(r,D) \in (\mathbb{N}^*)^2 : r \leqslant 1/(2\sigma_-(n)) \text{ and } m_{r,D} \leqslant 2n\}$. For each $r \geqslant 1$ and $D \geqslant 1$, the model $T_{r,D}$ is the set of translation hidden Markov model parameters in $S_{r,D}$ seen as hidden Markov model parameters (with the dominating measure λ_b on $\mathbb{R}^d$ instead of the Lebesgue measure):

$$T_{r,D} = \left\{\left(r, \pi_Q, Q, \left(y \longmapsto \frac{\gamma(y-x_r)}{b(y)/\|b\|_1}\right)_{z=1,\dots,r}\right) : ((x_z)_{z=1,\dots,r}, Q, \gamma) \in S_{r,D}, \pi_Q Q = \pi_Q\right\}\,.$$

By definition of $S_{r,D}$, for all $(r,\pi,Q,-) \in T_{r,D}$ and $x, x' \in \{1,\dots,r\}$, $Q(x,x') \geqslant (Cr)^{-1}$ and $\pi(x) \geqslant (Cr)^{-1}$. Thus, for all $(r,\pi,Q,-) \in \mathbf{T}_n$, $Q(x,x') \geqslant (C\log n)^{-1} \geqslant \sigma_-(n)$ since $C \geqslant 2$. The same holds for π, so that **[Aergodic]** is satisfied.

By H3, for all $n \geqslant 1$ and $y \in \mathbb{R}^d$, $\sup_{(r,-,-,g)\in\mathbf{T}_n}\sum_{z=1}^r g_z(y) \leqslant \|b\|_1 \log n$, and by H4,

$$\inf_{(r,-,-,g)\in\mathbf{T}_n}\sum_{z=1}^r g_z(y) \geqslant \|b\|_1 m(y)/b(y)\,,$$

so that by Markov's inequality, for all $t > 0$, with ϵ as in H4,

$$\mathbb{P}_{K^\star,\gamma^\star}\left[\left(\inf_{(r,-,-,g)\in\mathbf{T}_n}\sum_{z=1}^r g_z(y)\right)^{-\epsilon} \geqslant t\right] \leqslant \|b\|_1^{-\epsilon}\frac{\mathbb{E}_{K^\star,\gamma^\star}[(b(Y_1)/m(Y_1))^\epsilon]}{t}\,,$$

so that there exists a constant $C_{H4} > 0$ such that

$$\mathbb{P}_{K^\star,\gamma^\star}\left[\inf_{(r,-,-,g)\in\mathbf{T}_n} \log\sum_{z=1}^{r} g_z(y) \leqslant -\frac{1}{\epsilon}u\right] \leqslant C_{H4}e^{-u}\,.$$

Thus, there exists n_{tail} such that **[Atail]** holds for any $n \geqslant n_{\text{tail}}$ and for any $B(n) \geqslant \max(2/\epsilon, \log(\|b\|_1 \log n))$. Choose $B(n) = \log n$.

Finally, **[Aentropy']** is implied by the following assumption, which follows from H3 and H5 with $c(r, D, A) = c(D, A) + C_\Gamma$.

[Aentropy"] There exists a mapping $(r, D, A) \in \mathbb{N}^* \times \mathbb{N}^* \times \mathbb{R}_+ \longmapsto c(r, D, A)$ and a constant c' such that $\log c(r, D, A) \leqslant c'(\log m_{r,D} + A)$. There exists a sequence $(\Theta_D)_{D\geqslant 1}$ of sets such that for all $D \geqslant 1$, $\Theta_D \subset [-1, 1]^{\dim_D}$ and there exists a surjective mapping $\theta \in \Theta_D \longmapsto \gamma^\theta \in G_D$. For all $r \geqslant 1$, $D \geqslant 1$, $A \geqslant 0$ and $y \in \mathbb{R}^d$ such that $\log(b(y)/m(y)) \leqslant A$, the mapping $(x, \theta) \in \Lambda^r \times \Theta_D \longmapsto (\gamma^\theta(y-x_z)/b(y))_{z\in\{1,\dots,r\}}$ is $c(r, D, A)$-Lipschitz (when Λ and Θ_D are endowed with the supremum norm).

Let us see how this implies **[Aentropy']**. Let $\mathcal{S}_{n,A} = \{y \in \mathbb{R}^d : \log(b(y)/m(y)) \leqslant A\}$. By H4 and Markov's inequality, $\mathbb{P}^\star(Y_1 \in \mathcal{S}_{n,A}) \leqslant \exp(-A\epsilon/2)$ for A large enough. Moreover, for all $A \geqslant \log(\|b\|_1 \log n)$ and $y \in \mathcal{S}_{n,A}$,

$$\sup_{(r',-,-,g')\in\mathbf{T}_n} \left|\log\sum_{z=1}^{r'} g'_z(y)\right| \leqslant \max\left(\log\frac{b(y)}{\|b\|_1 m(y)}, \log(\|b\|_1 \log n)\right) \leqslant A.$$

A bracket covering of size δ of $[-1, 1]^{rd} \times [-1, 1]^{\dim_D}$ gives a bracket covering of size δL of $\Lambda^r \times \Theta_D$, which in turn gives bracket covering of size $c(r, D, A)\delta L\|b\|_1$ of the set

$$\left\{\left(y \longmapsto \|b\|_1 \frac{\gamma(y - x_z)}{b(y)}\mathbf{1}_{y\in\mathcal{S}_{n,A}}\right)_{z=1,\dots,r} : x \in \Lambda^r, \gamma \in G_D\right\}.$$

Since there exists a bracket covering of size δ of $[-1, 1]$ with cardinality at most $\max(2/\delta, 1)$, Equation (29) of **[Aentropy']** holds with $C_{\text{aux}}(r, D, n, A) = 2c(r, D, A)L\|b\|_1$. Finally, since $\sup_{(r,D)\in\mathfrak{P}_n} \log c(r, D, A) \leqslant c'(\log n + A)$, the last part of **[Aentropy']** holds.

Thus, Theorem 6 holds and ensures that there exists n_0, C_{pen} and A such that if $\text{pen}(n, r, D) \geqslant C_{\text{pen}}(m_{r,D} + r^2 - 1)(\log n)^{14}/n$, then for all $n \geqslant n_0$ and $t \geqslant 1$, with probability at least $1 - e^{-t} - 2n^{-2}$,

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{\widehat{\mathfrak{x}}_n,\widehat{Q}_n,\widehat{\gamma}_n}) \leqslant 2\inf_{(r,D)\in\mathfrak{P}_n}\left\{\inf_{(\mathfrak{x},Q,\gamma)\in S_{r,D}} \mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{\mathfrak{x},Q,\gamma}) + 2\text{pen}(n, r, D)\right\} + At\frac{(\log n)^8}{n}$$

and Proposition 1 follows by taking $t = 2\log n$ and recalling that $m_{r,D} = \dim_D + rd$ and $\mathfrak{P}_n = \{(r, D) : r \leqslant \log n \text{ and } \dim_D \leqslant n\}$.

C Additional simulations based on least squares for characteristic functions

In this section, the empirical least squares criterion $M_n(R)$ introduced in Section 3.1 is approximated to obtain a practical estimate of R using the same model as in Section 4. The estimate $\widehat{\Phi}_n$ of the characteristic function of the observations (Y_1, Y_2) is given for all $(t_1, t_2) \in \mathbb{R}^2$ by

$$\widehat{\Phi}_n(t_1, t_2) = \frac{1}{n} \sum_{j=1}^{n-1} \mathrm{e}^{it_1 Y_j + it_2 Y_{j+1}} .$$

The function w is set as the probability density function of a Gaussian random variable with standard deviation $\sigma = 3$ and M_n is estimated by the Monte Carlo estimate:

$$\widehat{M}_n(R) = \frac{1}{N} \sum_{\ell=1}^{N} \left| \widehat{\Phi}_n(U_1^\ell, U_2^\ell) \Phi_R(U_1^\ell; 0) \Phi_R(0; U_2^\ell) - \Phi_R(U_1^\ell, U_2^\ell) \widehat{\Phi}_n(U_1^\ell; 0) \widehat{\Phi}_n(0; U_2^\ell) \right|^2 ,$$

where $(U_1^\ell, U_2^\ell)_{1 \leqslant \ell \leqslant N}$ are independent and identically distributed with distribution w. In the following experiments, N is set to 5000. This estimated criterion is minimized over the set $\mathcal{D}_r$ of piecewise constant probability densities on $(-1, 1) \times (-1, 1)$ with r^2 uniformly spaced cells:

$$\mathcal{D}_r = \left\{ R : \mathbb{R}^2 \to \mathbb{R}_+ \, ; \, R = \sum_{i,j=1}^{r} \alpha_{i,j} \mathbb{1}_{(x_i, x_{i+1}) \times (x_j, x_{j+1})} \right\} ,$$

where for all $1 \leqslant i, j \leqslant r$, $x_i = -1 + 2(i-1)/r$, $\alpha_{i,j} \geqslant 0$ and $\sum_{i,j=1}^{r} \alpha_{i,j} = r^{-2}$. In this setting where the support of the law of (X_1, X_2) is compact and known, the up to translation indeterminacy is ruled out. The optimization is performed using the Covariance Matrix Adaptation Evolutionary Strategy [**?**, CMA-ES,]]igel:hansen:roth:2007 which optimizes iteratively all parameters using (μ, λ)-selection. At each iteration, the best offsprings of the current parameter estimate are combined to form the population of the following iteration and the other offsprings are discarded.

The performance of the least squares approach is assessed by comparing the estimated probability that (X_1, X_2) lies in each cell $(x_i, x_{i+1}) \times (x_j, x_{j+1})$, $1 \leqslant i, j \leqslant r$, which is $\widehat{\alpha}_{i,j}^n r^2$ and the benchmark estimation $\widetilde{\alpha}_{i,j}^{\mathsf{n,emp}}$ that would be computed if the sequence $(X_k)_{1 \leqslant k \leqslant n}$ were observed: $\widetilde{p}_{i,j}^{\mathsf{n,emp}} = n^{-1} \sum_{k=1}^{n-1} \mathbb{1}_{(x_i, x_{i+1}) \times (x_j, x_{j+1})}(X_k, X_{k+1})$. The results are displayed in Figure 3 over 10 independent runs, when the order r is in $\{10, 20, 30\}$, with CMA-ES initialized at a random point, and a maximum number of evaluations of $\widehat{M}_n(R)$ set to 75000. Each estimate is obtained with a sequence of $n = 100000$ observations and the L_1 score is

$$\varepsilon_{1,n}^r = \frac{1}{r^2} \sum_{i,j=1}^{r} \left| r^2 \widehat{\alpha}_{i,j}^n - \widetilde{p}_{i,j}^{\mathsf{n,emp}} \right| . \tag{31}$$

The associated estimated probabilities for the distribution of X_1 are displayed in Figure 4 with their confidence regions.

References

[Akakpo, 2019] Akakpo, N. (2019). Inference in a hidden Markov model with multivariate log-concave emission densities. *Working paper*.

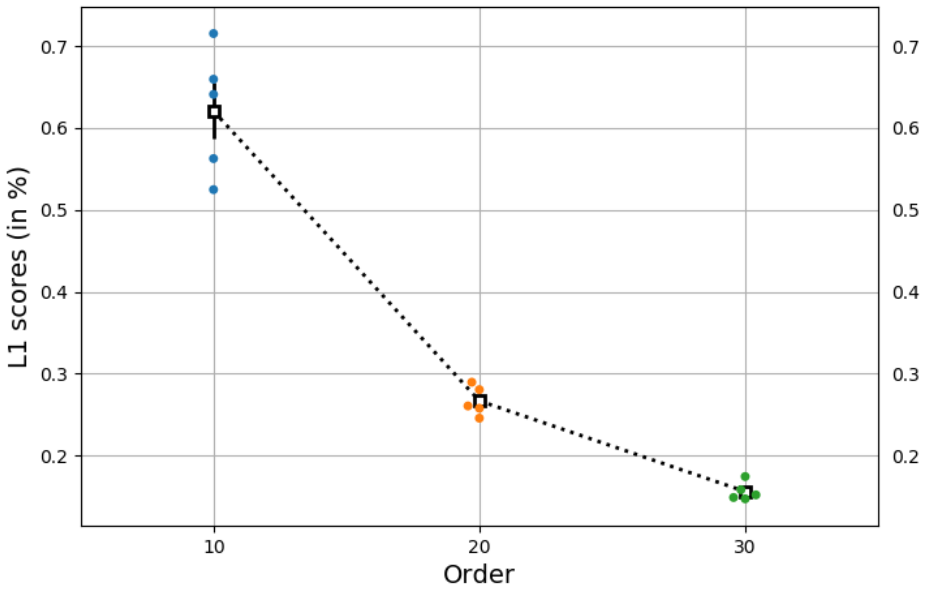

Figure 3: L_1 scores computed according to (31). Each dot is an estimated value with the least squares approach. For each value of r, the mean value (squares) over all runs as long as the empirical standard deviation (bars) are displayed.

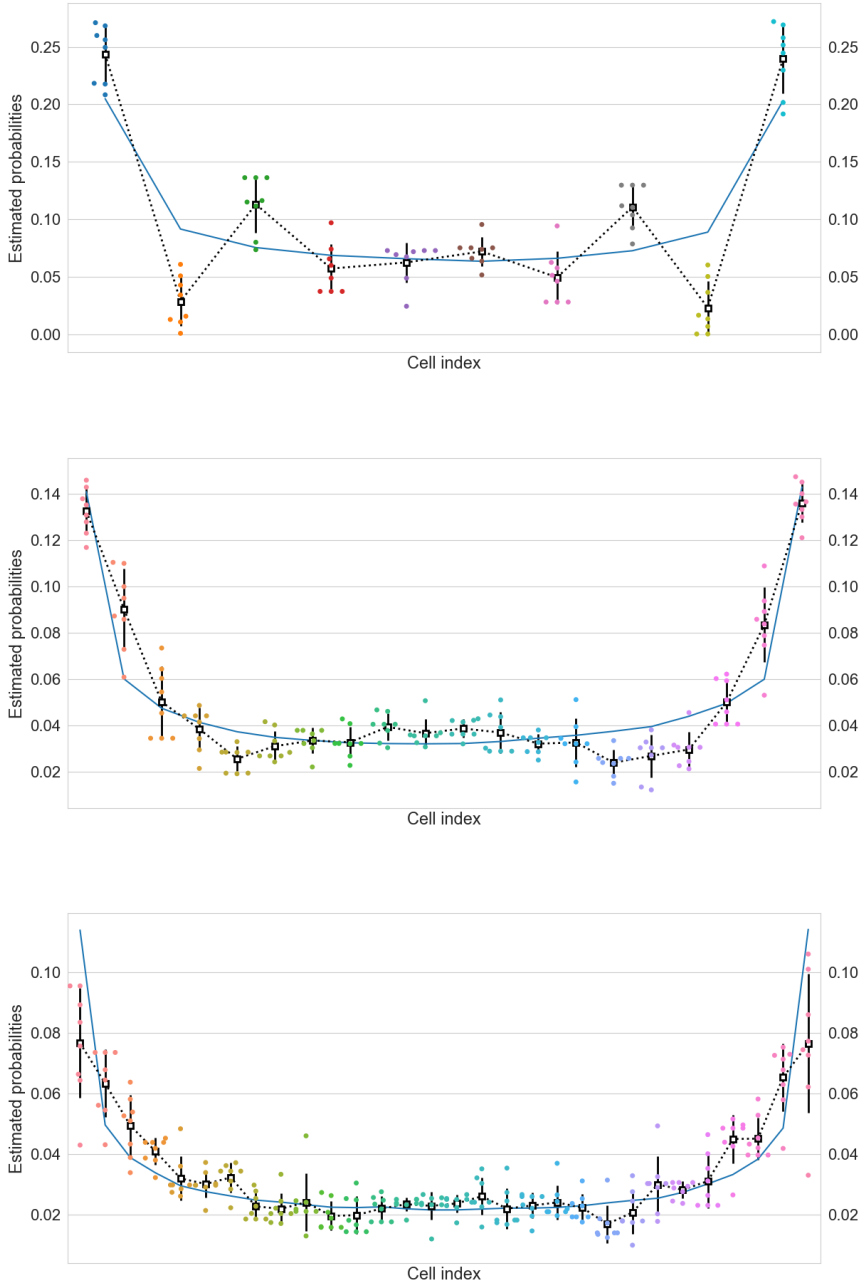

Figure 4: Estimated probabilities associated with the marginal distribution of X_1 for $r = 10$ (top), $r = 20$ (middle) and $r = 30$ (bottom). The blue line is the empirical estimate when the sequence $(X_k)_{1 \leqslant k \leqslant n}$ is observed (mean estimate over the 10 Monte Carlo runs). Each dot is an estimated value with the least squares approach. For each value of r, the mean value (squares) over all runs as long as the empirical standard deviation (bars) are displayed.

[Alexandrovich et al., 2016] Alexandrovich, G., Holzmann, H., and Leister, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434.

[Bradley, 2005] Bradley, R. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys*, 2:107–144.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

[Carroll and Hall, 1988] Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186.

[Crouse et al., 1998] Crouse, M., Nowak, R., and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902.

[Dahlhaus et al., 2018] Dahlhaus, R., Kiss, I., and Neddermeyer, J. (2018). On the relationship between the theory of cointegration and the theory of phase synchronization. *ArXiv:1201.0651*.

[De Castro et al., 2016] De Castro, Y., Gassiat, E., and Lacour, C. (2016). Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 17(111):1–43.

[Dedecker et al., 2015] Dedecker, J., Fischer, A., and Michel, B. (2015). Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Stat.*, 9(1):234–265.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society Series B*, 39:1–38.

[Devroye, 1989] Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.*, 17(2):235–239.

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

[Douc et al., 2014] Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear time series: theory, methods and applications with R examples*. CRC Press.

[Doukhan et al., 1994] Doukhan, P., Massart, P., and Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Annales de l'I.H.P.*, 30:63–82.

[Doukhan et al., 1995] Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de l'I.H.P.*, 31:393–427.

[Dumont and Le Corff, 2017a] Dumont, T. and Le Corff, S. (2017a). Nonparametric regression on hidden ϕ-mixing variables: Identifiability and consistency of a pseudo-likelihood based estimation procedure. *Bernoulli*, 23(2):990–1021.

[Dumont and Le Corff, 2017b] Dumont, T. and Le Corff, S. (2017b). Statistical inference for oscillation processes. *Statistics*, 51:61–83.

[Fan, 1991] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272.

[Fell and Axmacher, 2011] Fell, J. and Axmacher, N. (2011). The role of phase synchronization in memory processes. *Nat. Rev. Neurosci.*, 12(2):105–118.

[Gassiat et al., 2016] Gassiat, E., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71.

[Gassiat and Rousseau, 2016] Gassiat, E. and Rousseau, J. (2016). Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212.

[Karr, 1975] Karr, A. F. (1975). Weak convergence of a sequence of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 33(1):41–48.

[Lambert et al., 2003] Lambert, M. F., Whiting, J. P., and Metcalfe, A. V. (2003). A non-parametric hidden Markov model for climate state identification. *Hydrology and earth system sciences*, 7:652–667.

[Langrock et al., 2015] Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, 71(2):520–528.

[Lehéricy, 2018] Lehéricy, L. (2018). Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. *Submitted and available online at arXiv:1807.03997*.

[Lehéricy, 2018] Lehéricy, L. (2018). State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *J. Mach. Learn. Res.*

[Liu and Taylor, 1989] Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canad. J. Statist.*, 17(4):427–438.

[R Core Team, 2017] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

[Särkkä, 2013] Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press, New York, NY, USA.

[Särkkä et al., 2007] Särkkä, S., Vehtari, A., and Lampinen, J. (2007). Rao-Blackwellized particle filter for multiple target tracking. *Inofrmation Fusion*, 8(1):2–15.

[Schuhmacher et al., 2019] Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., and Schmitzer, B. (2019). *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. R package version 0.11-1.

[Shen et al., 2013] Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

[Stefanski and Carroll, 1990] Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21(2):169–184.

[Stein and Shakarchi, 2003] Stein, E. and Shakarchi, R. (2003). *Complex Analysis*. Princeton University Press, Princeton.

[Touron, 2019] Touron, A. (2019). Consistency of the maximum likelihood estimator in seasonal hidden Markov models. *Statistics and Computing*, pages 1–21.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[Volant et al., 2014] Volant, S., Bérard, C., Martin-Magniette, M.-L., and Robin, S. (2014). Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, 24(4):493–504.

[Wang et al., 2017] Wang, X., Lebarbier, E., Aubert, J., and Robin, S. (2017). Variational inference for coupled hidden Markov models applied to the joint detection of copy number variations. *arXiv preprint arXiv:1706.06742*.

[Yau et al., 2011] Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. Royal Statist. Society Series B*, 73:1–21.

[Zucchini et al., 2016] Zucchini, W., Mac Donald, I., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R*. CRC Press.