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

Elisabeth Gassiat, Luc Lehéricy, Sylvain Le Corff

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

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

*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.
**Institut Mines-Télécom, Télécom Sud Paris, Département CITI, France.

Abstract

In this paper, we consider partially observed dynamical systems where the observations are given as the sum of latent variables lying in a general state space and some independent noise with unknown distribution. In the case of dependent latent variables such as Markov chains, it is shown that this fully nonparametric model is 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

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. In such frameworks, often referred to in the literature as *partially observed dynamical systems*, only indirect observations of the state sequence, possibly lying in a high dimensional state space, are available to perform statistical inference.

Solving inverse problems, i.e. recovering information about the hidden process using the observations, is a long standing statistical problem. In the case of hidden Markov models (HMMs), the sequence of latent states is assumed to be a Markov chain and the conditional distributions of the observations given the states are referred to as emission distributions. Although parametric HMMs have been widely studied and are appealing for a wide range of applications, inference procedures for such models 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, pear], genomics [Yau et al., 2011], statistical modelling of animal movement [Langrock et al., 2015] or biology [Volant et al., 2014]. For finite state space HMMs, 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] for translation

HMMs with finite state space. While certainly of interest, the finite state space setting may be too restrictive for many applications.

The aim of this paper is to propose a solution to fully nonparametric translation HMMs when, for all $1 \leqslant i \leqslant n$, the observation Y_i is given by

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

where X_i is the latent state and ε_i is the noise. This work can be considered as the first contribution to establish theoretical results in a general nonparametric setting for latent data models with general state space.

The inverse problem in (1) is to infer the distribution of the latent data based on $(Y_1, \dots, Y_n)$. When the observations are i.i.d., it is 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 with a nowhere vanishing Fourier transform, in the situation where the random variables $(X_i, \varepsilon_i)_{1 \leqslant i \leqslant n}$ are assumed to be independent and for all $1 \leqslant i \leqslant n$, ε_i is independent of X_i, 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 [Eckle et al., 2016] and references therein for a recent work. However, when the distribution of the noise is also unknown, model (1) can not be identified in full generality. In [Gassiat and Rousseau, 2016], the authors proved that when the latent variables take finitely many values, all the parameters of the model are identifiable as soon as the matrix that defines the joint distribution of two consecutive latent variables is nonsingular and the location parameters are distinct. When there are two possible distinct states the assumption on the matrix is equivalent to the fact that the latent variables are not independent.

The first objective of this paper is to prove that, when the state space is $\mathbb{R}^d$ for some $d \geqslant 1$, and *if one considers non independent observations*, identifiability can be obtained in model (1) without any assumption on the distribution of the noise. The second objective is to propose nonparametric estimators and to prove that they are consistent. In Section 2.1, the identifiability of the fully nonparametric hidden translation model is established under the weak assumption that the Laplace transform of the latent state has an exponential growth smaller than 2, see Theorem 1. This result is then displayed in Section 2.2 in the specific case where the latent data is a stationary Markov chain, see Corollary 1. In the case of real valued HMMs, identifiability is extended to latent variables having Laplace transform with exponential growth smaller than 3, see Theorem 2. In Section 3, 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. Simulations are presented in Section 4. The Appendices contain proofs and further examples where the theory developed in this paper could be applied.

2 Identifiability theorems

2.1 Setting and theorem

Let $X = (X_1, X_2)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2)$ be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for $i \in \{1, 2\}$, X_i and ε_i take values in $\mathbb{R}^{d_i}$ and such that ε is independent of X. For $i \in \{1, 2\}$, define

$$Y_i = X_i + \varepsilon_i \,.$$

Let $\mathbb{P}^{(2)}_{R,P}$ be the distribution of $Y = (Y_1, Y_2)$ when X has distribution R and $(\varepsilon_i)_{i\in\{1,2\}}$ are independent and such that for $i \in \{1,2\}$, ε_i has distribution P_i, with $P = (P_1, P_2)$. Let $\mathcal{A}$ be the set of distributions on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that for all $(\lambda_1, \lambda_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\int \exp\left(\lambda_1^T x_1 + \lambda_2^T x_2\right) R(\mathrm{d}x_1, \mathrm{d}x_2) < +\infty\,,$$

where for any vector λ in a Euclidian space, λ^T denotes its transpose vector. Note that if $R \in \mathcal{A}$, then the function Φ_R, defined as

$$\begin{aligned}\Phi_R : \mathbb{C}^{d_1} \times \mathbb{C}^{d_2} &\longrightarrow \mathbb{C}\\ (z_1, z_2) &\mapsto \int \exp\left(z_1^T x_1 + z_2^T x_2\right) R(\mathrm{d}x_1, \mathrm{d}x_2)\,,\end{aligned}$$

is a multivariate analytic function. For any $\rho > 0$, let $\mathcal{M}_\rho$ be the set of probability distributions μ on $\mathbb{R}^d$ for some $d \geqslant 1$ such that there exist $A, B > 0$ satisfying, for all $\lambda \in \mathbb{R}^d$,

$$\int \exp\left(\lambda^T x\right) \mu(\mathrm{d}x) \leqslant A \exp\left(B\|\lambda\|^\rho\right)\,,$$

where for a vector λ in a Euclidian space, $\|\lambda\|$ denotes its euclidian norm. For any distribution $R \in \mathcal{A}$ and any random variables (X_1, X_2) on $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution R, denote by R_1 (resp. R_2) the marginal distribution of X_1 (resp. X_2). Consider the following assumptions.

H1 The distribution $R \in \mathcal{A}$ is said to satisfy H1 if and only if for any $z_0 \in \mathbb{C}^{d_1}$, $z \mapsto \Phi_R(z_0, z)$ is not the null function and for any $z_0 \in \mathbb{C}^{d_2}$, $z \mapsto \Phi_R(z, z_0)$ is not the null function.

H2 In the case where $d_1 = d_2 = d$, $R \in \mathcal{A}$ is said to satisfy H2 if and only if $R_1 = R_2$ and 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.

In addition, the assertion *$R = \widetilde{R}$ and $P = \widetilde{P}$ up to translation* means that there exists $m = (m_1, m_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that if X has distribution R and for $i \in \{1,2\}$, ε_i has distribution P_i, then $(X_i - m_i)_{i\in\{1,2\}}$ has distribution $\widetilde{R}$ and for $i \in \{1,2\}$, $\epsilon_i + m_i$ has distribution $\widetilde{P}_i$.

Theorem 1. *Assume that $R \in \mathcal{A}$, $\widetilde{R} \in \mathcal{A}$, and that there exists $\rho < 2$ such that one of the following assumptions holds.*

- *(A1) R_1, R_2, $\widetilde{R}_1$ and $\widetilde{R}_2$ are in $\mathcal{M}_\rho$ and R and $\widetilde{R}$ satisfy assumption H1.*
- *(A2) R_1 and $\widetilde{R}_1$ are in $\mathcal{M}_\rho$ and R and $\widetilde{R}$ satisfy assumption H2.*

Then, $\mathbb{P}^{(2)}_{R,P} = \mathbb{P}^{(2)}_{\widetilde{R},\widetilde{P}}$ implies that $R = \widetilde{R}$ and $P = \widetilde{P}$ up to translation.

Remark 1. *One way to fix the "up to translation" indeterminacy when the noises have a first order moment is to assume that they are centered, that is $\mathbb{E}[\varepsilon_i] = 0$ for $i \in \{1,2\}$.*

Remark 2. *Note that there is no assumption at all on the distributions of the noises ε_1 and ε_2. For instance, it is not required that their characteristic functions do not vanish, which is usually assumed in the literature on deconvolution. The only assumptions concern the distribution of X. When (X_1, X_2) has compact support, then $R \in \mathcal{A}$ and the marginal distributions of X_1 and X_2 are in $\mathcal{M}_1$. Regarding assumption H1*

and assumption H2, using Hadamard's factorization Theorem, see [Stein and Shakarchi, 2003, Chapter 5, Theorem 5.1], if $R \in \mathcal{M}_\rho$ with $\rho < 2$, then $\Phi_R(0,\cdot)$ has no zeros if and only if X_1 is deterministic and $\Phi_R(\cdot,0)$ has no zeros if and only if X_2 is deterministic. Thus, if X_1 and X_2 are not deterministic, $\Phi_R(0,\cdot)$ and $\Phi_R(\cdot,0)$ have zeros. Moreover, if for $i \in \{1,2\}$, X_i can be decomposed as $X_i = \widetilde{X}_i + \eta_i$, with η_1 and η_2 independent variables independent of $\widetilde{X} = (\widetilde{X}_1, \widetilde{X}_2)$, and if for some z_0, $\mathbb{E}[e^{z_0^T \eta_1}] = 0$ and for some z_1, $\mathbb{E}[e^{z_1^T \eta_2}] = 0$, then H1 does not hold. In other words, H1 and H2 can hold only if all the additive noise has been removed from X.

Proof. In this section, the following result, which may be established by arguing variable by variable, is used repeatedly. 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 $\mathbb{P}^{(2)}_{R,P} = \mathbb{P}^{(2)}_{\widetilde{R},\widetilde{P}}$ and let ϕ_i (resp. $\widetilde{\phi}_i$) be the characteristic function of P_i (resp. $\widetilde{P}_i$) for $i \in \{1,2\}$. Since the distribution of Y_1 and Y_2 are the same under $\mathbb{P}^{(2)}_{R,P}$ and $\mathbb{P}^{(2)}_{\widetilde{R},\widetilde{P}}$, for any $t \in \mathbb{R}^{d_1}$,

$$\phi_1(t)\,\Phi_R(it,0) = \widetilde{\phi}_1(t)\,\Phi_{\widetilde{R}}(it,0) \tag{2}$$

and for any $t \in \mathbb{R}^{d_2}$,

$$\phi_2(t)\,\Phi_R(0,it) = \widetilde{\phi}_2(t)\,\Phi_{\widetilde{R}}(0,it)\,. \tag{3}$$

Since the distribution of (Y_1, Y_2) is the same under $\mathbb{P}^{(2)}_{R,P}$ and $\mathbb{P}^{(2)}_{\widetilde{R},\widetilde{P}}$, for any $(t_1,t_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\phi_1(t_1)\,\phi_2(t_2)\,\Phi_R(it_1,it_2) = \widetilde{\phi}_1(t_1)\,\widetilde{\phi}_2(t_2)\,\Phi_{\widetilde{R}}(it_1,it_2)\,. \tag{4}$$

There exists a neighborhood V of 0 in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that for all $t = (t_1,t_2) \in V$, $\phi_1(t_1) \neq 0$, $\phi_2(t_2) \neq 0$, $\widetilde{\phi}_1(t_1) \neq 0$, $\widetilde{\phi}_2(t_2) \neq 0$, so that (2), (3) and (4) imply that for any $(t_1,t_2) \in V^2$,

$$\Phi_R(it_1,it_2)\,\Phi_{\widetilde{R}}(it_1,0)\,\Phi_{\widetilde{R}}(0,it_2) = \Phi_{\widetilde{R}}(it_1,it_2)\,\Phi_R(it_1,0)\,\Phi_R(0,it_2)\,. \tag{5}$$

Since $(z_1,z_2) \mapsto \Phi_R(z_1,z_2)\,\Phi_{\widetilde{R}}(z_1,0)\,\Phi_{\widetilde{R}}(0,z_2) - \Phi_{\widetilde{R}}(z_1,z_2)\,\Phi_R(z_1,0)\,\Phi_R(0,z_2)$ is a multivariate analytic function of d_1+d_2 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 $z_1 \in \mathbb{C}^{d_1}$ and $z_2 \in \mathbb{C}^{d_2}$,

$$\Phi_R(z_1,z_2)\,\Phi_{\widetilde{R}}(z_1,0)\,\Phi_{\widetilde{R}}(0,z_2) = \Phi_{\widetilde{R}}(z_1,z_2)\,\Phi_R(z_1,0)\,\Phi_R(0,z_2)\,. \tag{6}$$

i). Consider first the situation where (A1) holds. Fix $(u_2,\dots,u_d) \in \mathbb{C}^{d_1-1}$ and let $\mathcal{Z}$ be the set of zeros of $u \mapsto \Phi_R(u,u_2,\dots,u_d,0)$ and $\widetilde{\mathcal{Z}}$ be the set of zeros of $u \mapsto \Phi_{\widetilde{R}}(u,u_2,\dots,u_d,0)$. Let $u_1 \in \mathcal{Z}$. Write $z_1 = (u_1,u_2,\dots,u_d)$ so that by (6), for any $z_2 \in \mathbb{C}^{d_2}$,

$$\Phi_R(z_1,z_2)\,\Phi_{\widetilde{R}}(z_1,0)\,\Phi_{\widetilde{R}}(0,z_2) = 0\,. \tag{7}$$

Using (A1), $z_2 \to \Phi_R(z_1,z_2)$ is not the null function. Thus, there exists $z_2^\star$ in $\mathbb{C}^{d_2}$ such that $\Phi_R(z_1,z_2^\star) \neq 0$ and by continuity, there exists an open neighborhood of $z_2^\star$ such that for all z_2 in this open set, $\Phi_R(z_1,z_2) \neq 0$. Since $z \mapsto \Phi_{\widetilde{R}}(0,z)$ is not the null function and is analytic on $\mathbb{C}^{d_2}$, it can not be null all over this open set, so that there exists z_2 such that simultaneously $\Phi_R(z_1,z_2) \neq 0$ and $\Phi_{\widetilde{R}}(0,z_2) \neq 0$. Then (7) leads to $\Phi_{\widetilde{R}}(z_1,0) = 0$, so that $\mathcal{Z} \subset \widetilde{\mathcal{Z}}$. A symmetric argument yields $\widetilde{\mathcal{Z}} \subset \mathcal{Z}$ so that $\mathcal{Z} = \widetilde{\mathcal{Z}}$. Moreover, the analytic functions $u \mapsto \Phi_R(u,u_2,\dots,u_d,0)$ and $u \mapsto \Phi_{\widetilde{R}}(u,u_2,\dots,u_d,0)$ 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 coefficients depending on $(u_2, \dots, u_d)$) such that for all $u \in \mathbb{C}$,

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

Arguing similarly for all variables, there exists a polynomial function S on $\mathbb{C}^{d_1}$ with degree at most 1 in each variable such that for all $(u_1, \dots, u_d) \in \mathbb{C}^d$,

$$\Phi_R(u_1, u_2, \dots, u_d, 0) = \mathrm{e}^{S(u_1, u_2, \dots, u_d)} \Phi_{\widetilde{R}}(u_1, u_2, \dots, u_d, 0) \,. \tag{8}$$

Since $\Phi_R(0, \dots, 0) = \Phi_{\widetilde{R}}(0, \dots, 0) = 1$, the constant term of the polynomial S is 0. On the other hand, if (A1) or (A2) holds, then under $\widetilde{R}$, X_1 is not deterministic and its probability mass function is not supported by 0 (see Remark 2). Thus, there exist $a = (a_1, \dots, a_{d_1}) \in \mathbb{R}^{d_1}$, $\alpha > 0$ and $\delta > 0$ such that

$$0 \notin \prod_{j=1}^{d_1} [a_j - \alpha, a_j + \alpha] \quad \text{and} \quad \mathbb{P}^{(2)}_{\widetilde{R}, \widetilde{P}} \left(X_1 \in \prod_{j=1}^{d_1} [a_j - \alpha, a_j + \alpha] \right) \geqslant \delta \,,$$

which gives, for all $\lambda \in \mathbb{R}^{d_1}$,

$$\Phi_{\widetilde{R}}(\lambda, 0) \geqslant \delta \mathrm{e}^{\sum_{j=1}^{d_1} \inf_{x \in [a_j - \alpha, a_j + \alpha]} \lambda_j x} \,.$$

Since $\widetilde{R}_1 \in \mathcal{M}_\rho$ for some $\rho < 2$, if S has degree at least 2, then $\Phi_R(\cdot, 0)$ has exponential growth of order at least 2, contradicting assumption (A1) and (A2). Then, S has degree at most 1 and there exists $m_1 \in \mathbb{C}^{d_1}$ such that for all $z \in \mathbb{C}^{d_1}$, $\Phi_R(z, 0) = \mathrm{e}^{m_1^T z} \Phi_{\widetilde{R}}(z, 0)$. As for all $z \in \mathbb{R}^d$, $\Phi_R(-iz, 0) = \overline{\Phi_R(iz, 0)}$ and $\Phi_{\widetilde{R}}(-iz, 0) = \overline{\Phi_{\widetilde{R}}(iz, 0)}$, $m_1 \in \mathbb{R}^{d_1}$. Arguing similarly for the function $\Phi_R(0, z_2)$, there exists $m_2 \in \mathbb{R}^{d_2}$ such that for all $z \in \mathbb{C}^{d_2}$, $\Phi_R(0, z) = \mathrm{e}^{m_2^T z} \Phi_{\widetilde{R}}(0, z)$.

ii). Consider now the situation where (A2) holds. Then, for all $z \in \mathbb{C}^d$, $\Phi_R(z, 0) = \Phi_R(0, z)$ and $\Phi_{\widetilde{R}}(z, 0) = \Phi_{\widetilde{R}}(0, z)$, and following the same steps as the first part of the proof, there exists $m = (m_1, m_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that for all $z \in \mathbb{C}^d$,

$$\Phi_R(z, 0) = \mathrm{e}^{m_1^T z} \Phi_{\widetilde{R}}(z, 0) \quad \text{and} \quad \Phi_R(0, z) = \mathrm{e}^{m_2^T z} \Phi_{\widetilde{R}}(0, z) \,. \tag{9}$$

Combining (9) with (6) yields, for all $(t_1, t_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\Phi_R(it_1, it_2) = e^{i m_1^T t_1 + i m_2^T t_2} \Phi_{\widetilde{R}}(it_1, it_2) \,. \tag{10}$$

Then, using (2), for all $t \in \mathbb{R}^d$ such that $\Phi_R(it, 0) \neq 0$, $\phi_1(t) = \mathrm{e}^{-i m_1^T t} \widetilde{\phi}_1(t)$. Since the set of zeros of $t \mapsto \Phi_R(it, 0)$ has empty interior, for each t such that $\Phi_R(it, 0) = 0$ it is possible to find a sequence $(t_n)_{n \geqslant 1}$ such that t_n tends to t and for all n, $\Phi_R(it_n, 0) \neq 0$. But ϕ_1 and $\widetilde{\phi}_1$ are continuous functions, so that for all $t \in \mathbb{R}$,

$$\phi_1(t) = \mathrm{e}^{-i m_1^T t} \widetilde{\phi}_1(t) \,. \tag{11}$$

The fact that for all $t \in \mathbb{R}$,

$$\phi_2(t) = \mathrm{e}^{-i m_2^T t} \widetilde{\phi}_2(t) \tag{12}$$

follows from the same arguments. The proof is concluded by noting that (10), (11) and (12) imply that $R = \widetilde{R}$ and $P = \widetilde{P}$ up to translation. □

2.2 The case of translation hidden Markov models

Consider a sequence of random variables $(Y_n)_{n\geqslant 1}$ taking values in $\mathbb{R}^d$ such that for all $n \in \mathbb{N}$,

$$Y_n = X_n + \varepsilon_n \,, \tag{13}$$

where $(X_n)_{n\geqslant 1}$ is a stationary Markov chain, and $(\varepsilon_n)_{n\geqslant 1}$ is a sequence of independent and identically distributed (i.i.d.) random variables independent of $(X_n)_{n\geqslant 1}$. For all transition kernel $K : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ with stationary distribution μ_K, define the measure R_K on $\mathbb{R}^d$ 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 $(Y_n)_{n\geqslant 1}$ when the stationary Markov chain $(X_n)_{n\geqslant 1}$ has transition K and the ε_i's have distribution P. Theorem 1 has the following corollary.

Corollary 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 $R_K \in \mathcal{A}$, $R_{\widetilde{K}} \in \mathcal{A}$, and 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 (A2). 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 $R_K \in \mathcal{A}$, $R_{\widetilde{K}} \in \mathcal{A}$, and 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 (A2). Assume moreover that if the stationary Markov chain with transition kernel K (resp. $\widetilde{K}$) is Gaussian, it is not a sequence of i.i.d. 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.*

2.3 Further examples

Theorem 1 applies to other models such as deconvolution with repeated measurements, errors in variable regression models, or nonparametric hidden regression variables, see Section D for more details. Note that all examples given in that section are submodels of the general model displayed in Section 2.1, so that Theorem 1 not only proves that such submodels are identifiable, but also that they may be recovered even in the larger general model.

3 Consistent estimation

This section displays two different methods to obtain consistent estimators. The first method is directly inspired by the identifiability proof, and may be used in more general contexts than HMMs. The second method should be more accurate in the HMM context since it uses the full dependency properties of the observed process. In the following, objects related to the true (unknown) distribution $\mathbb{P}^\star$ of the observed process are denoted with the superscript $\star$.

3.1 Using least squares for characteristic functions

Let $\mathcal{S}$ be a compact neighborhood of 0 in $\mathbb{R}^{d_1+d_2}$, and let $W : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}_+$ be a positive function on $\mathcal{S}$. For any probability distribution R on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ define

$$M(R) = \int_{\mathcal{S}} |\Phi_{R^\star}(t_1,t_2)\Phi_R(t_1;0)\Phi_R(0;t_2) - \Phi_R(t_1,t_2)\Phi_{R^\star}(t_1;0)\Phi_{R^\star}(0;t_2)|^2 \\ |\phi_1^\star(t_1)\phi_2^\star(t_2)|^2 W(t_1,t_2)\mathrm{d}t_1\mathrm{d}t_2\,.$$

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

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

Note that if $R = \widetilde{R}$ up to translation, then $M(R) = M(\widetilde{R})$ and $M_n(R) = M_n(\widetilde{R})$. Let $\mathcal{R}$ be a set of probability distributions on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, such that for some $\rho < 2$, for all $R \in \mathcal{R}$, $R \in \mathcal{M}_\rho$ and satisfies assumption A1 (resp. A2). 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 $\mathcal{R}$. Define $Z_n(t_1,t_2)$ by

$$Z_n(t_1,t_2) = \sqrt{n}\left(\widehat{\Phi}_n(t_1,t_2) - \Phi_{R^\star}(t_1,t_2)\phi_1^\star(t_1)\phi_2^\star(t_2)\right)\,.$$

The following consistency theorem holds.

Theorem 3. *Assume that $\mathcal{R}$ is compact for the weak convergence topology and that $R^\star \in \mathcal{R}$. Assume moreover that*

$$\sup_{(t_1,t_2)\in\mathcal{S}} |Z_n(t_1,t_2)| = O_{\mathbb{P}^\star}(1)\,. \tag{14}$$

Then,

$$M(\widehat{R}_n) = O_{\mathbb{P}^\star}(n^{-1/2})\,, \tag{15}$$

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.

Here is an example where the assumptions of Theorem 3 are easily verified. Consider $(R_\theta)_{\theta\in\Theta}$ a parametric family of probability densities (with respect to the Lebesgue measure) on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, in which Θ is a compact subset of a Euclidian space, and let $\mathcal{P}(\Theta)$ be the set of probability distributions on Θ endowed with the 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 for all $\theta \in \Theta$, $R_\theta \in \mathcal{M}_\rho$, then for all $R \in \mathcal{R}$, $R \in \mathcal{M}_\rho$. Moreover, for any $\mu \in \mathcal{P}(\Theta)$, for any $(z_0, z_1) \in \mathbb{C}^{d_1} \times \mathbb{C}^{d_2}$,

$$\Phi_{R_\mu}(z_0, z_1) = \int_\Theta \Phi_{R_\theta}(z_0, z_1) \mathrm{d}\mu(\theta)\,,$$

so that as soon as for some $u_0 \in C^{d_2}$, 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. Regarding (14), using the empirical estimator for $\widehat{\Phi}_n$, then (14) holds under stationarity and mixing conditions, see for instance [Doukhan et al., 1994] and [Doukhan et al., 1995].

3.2 Using maximum likelihood

This Section focuses on the situation where the observed process is a HMM. Using the fact that continuous distributions may be approximated by discrete distributions, we consider finite state space HMMs and the associated maximum likelihood estimator (MLE). The consistency of the MLE is deduced from the oracle inequality proved in [Lehéricy, 2018] for possibly misspecified HMMs. We thus consider modelling assumptions for which Theorem 8 of [Lehéricy, 2018] can be applied. Assume that the hidden process $(X_n)_{n\geqslant 1}$ 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 on $\mathbb{R}^d$.

Transition kernels on finite sets are described by the number of points M of their support, the vector $m = (m_1, \dots, m_M)$ of their support points and the transition matrix Q between these points: for all $(i, j) \in \{1, \dots, M\}^2$, $Q(i, j) = \mathbb{P}(X_1 = m_j | X_0 = m_i)$. Then, for a vector m, a transition matrix Q with stationary distribution μ_Q, and a density γ, the log-likelihood is given by:

$$\ell_n(m, Q, \gamma) = \log \left(\sum_{x_1, \dots, x_n \in \{1, \dots, M\}} \mu_Q(x_1) \gamma(Y_1 - m_{x_1}) \prod_{i=1}^{n-1} Q(x_i, x_{i+1}) \gamma(Y_{i+1} - m_{x_{i+1}}) \right). \tag{16}$$

Let Γ a set of densities and consider the following assumptions.

H3 Γ is a set of continuous, positive and centered probability densities in the sense that

$$\forall \gamma \in \Gamma, \quad \forall i \in \{1, \dots, d\}, \quad \int_{\mathbb{R}^{i-1} \times (-\infty, 0] \times \mathbb{R}^{d-i}} \gamma(y) \mathrm{d}y = \frac{1}{2}\,, \tag{17}$$

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

$$\forall y \in \mathbb{R}^d, \quad b(y) = \sup_{\gamma \in \Gamma} \sup_{x \in 2\Lambda} \max(\gamma(y - x), \gamma(x - y)) \tag{18}$$

satisfies $b \in \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^\infty(\mathbb{R}^d)$. Moreover, $\gamma^\star \in \Gamma$.

The centering assumption (17) allows to fix the translation parameter in the identifiability results. Note that the probability density $b/\|b\|_1$ is also centered.

Example. *Assume that $d = 1$. Let $f : y \longmapsto (2\pi)^{-1/2} \exp(-y^2/2)$, Θ be a compact subset of $\mathbb{R} \times (0, +\infty)$ and*

$$\Gamma = \left\{ \gamma : y \longmapsto \int_\Theta \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right) \mathrm{d}p(\mu, \sigma) : p \in \mathcal{P}(\Theta), \gamma \text{ centered} \right\} \tag{19}$$

be the set of densities of mixtures of Gaussian distributions with parameters in Θ. Then Γ satisfies H3.

Denote by $\mathcal{P}(\Lambda)$ the set of probability measures on Λ. 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 such functions 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.

H4 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 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 $\lambda^\star$.

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

H5 There exists an application $D \longmapsto c(D)$ such that $\log c(D) = O(D^\zeta)$ for some constant $\zeta > 0$ and for all $D \geqslant 1$, there exists an application $[-1, 1]^D \longrightarrow \mathbf{L}^\infty(\mathbb{R}^d)$ that is $c(D)$-Lipschitz and whose image contains G_D.

The collection of models $(S_{M,D,n})_{M \geqslant 1, D \geqslant 1, n \geqslant 1}$ used in the maximum likelihood estimation is defined as follows. For all $M \geqslant 1$, $D \geqslant 1$ and $n \geqslant 1$, let $S_{M,D,n}$ be the set of parameters $(m, Q, \gamma) \in \Lambda^M \times [0,1]^{M \times M} \times \mathbf{L}^1(\mathbb{R}^d)$ such that

- Q is a transition matrix such that for all $(i, j) \in \{1, \dots, M\}^2$, $Q(i, j) \in \left[\frac{1}{CM}, \frac{C}{M}\right]$;
- the application $m_i \longmapsto m_{Q(i,\cdot)}$ admits the modulus of continuity ω with respect to W_1 ;
- there exist $\alpha \in [n^{-2}, 1]$ and $\gamma' \in G_D$ such that $\gamma = (1 - \alpha)\gamma' + \alpha b / \|b\|_1$.

Let

$$(\widehat{m}_{M,D}, \widehat{Q}_{M,D}, \widehat{\gamma}_{M,D}) \in \operatorname*{arg\,max}_{(m,Q,\gamma) \in S_{M,D,n}} \frac{1}{n} \ell_n(m, Q, \gamma) \tag{20}$$

be the maximum likelihood estimator associated with each model. Then, select the number of states and the model dimension using the penalized likelihood:

$$(\widehat{M}_n, \widehat{D}_n) \in \operatorname*{arg\,max}_{M \leqslant \log n, D \leqslant n} \left(\frac{1}{n} \ell_n(\widehat{m}_{M,D}, \widehat{Q}_{M,D}, \widehat{\gamma}_{K,M}) - (D + M^2) \frac{(\log n)^{15}}{n} \right) \tag{21}$$

and define the final estimators

$$(\widehat{m}_n, \widehat{Q}_n, \widehat{\gamma}_n) = (\widehat{m}_{\widehat{M}_n, \widehat{D}_n}, \widehat{Q}_{\widehat{M}_n, \widehat{D}_n}, \widehat{\gamma}_{\widehat{M}_n, \widehat{D}_n}) \,. \tag{22}$$

In order to state the consistency result, a continuous kernel associated with the discrete kernels of the models has to be introduced. For $(m, Q, \gamma) \in S_{M,D,n}$, denote by $K_{m,Q}$ 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 $\{m_i\}_{i \in \{1,\dots,M\}}$ and such that the support of $K_{m,Q}(x, \cdot)$ is in $\{m_i\}_{i \in \{1,\dots,M\}}$ for all $x \in \Lambda$. Linear interpolation provides a way to construct such a kernel as soon as the modulus ω is concave. The following theorem is proved in Section C.

Theorem 4. *Assume that assumptions H2, H3, H4 and H5 hold. Then*

$$\sup_{x \in Supp(\lambda^\star)} W_1(K_{\widehat{m}_n, \widehat{Q}_n}(x, \cdot), K^\star(x, \cdot)) \underset{n\to\infty}{\longrightarrow} 0\,, \tag{23}$$

where $\lambda^\star$ is the measure defined in assumption H4 and

$$\|\widehat{\gamma}_n - \gamma^\star\|_1 \underset{n\to\infty}{\longrightarrow} 0\,. \tag{24}$$

In particular, for all $x \in Supp(\lambda^\star)$,

$$K_{\widehat{m}_n, \widehat{Q}_n}(x, \cdot) \xrightarrow[n\to\infty]{(d)} K^\star(x, \cdot) \tag{25}$$

and if $\mathbb{P}_K^X$ denotes the distribution of the stationary Markov chain with transition kernel K,

$$\mathbb{P}^X_{K_{\widehat{m}_n, \widehat{Q}_n}} \xrightarrow[n\to\infty]{(d)} \mathbb{P}^X_{K^\star}. \tag{26}$$

The first step of the proof of Theorem 4 is to check the compacity of the parameter space and the continuity of the total variation distance between the distributions of the observations. Then, Theorem 8 of [Lehéricy, 2018] shows that the maximum likelihood estimator converges to a zero of the total variation distance to the true distribution, given by the true parameters by Corollary 1.

4 Simulations

Consider the model given by $Z_0 \sim U(0, 2\pi)$ 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 i.i.d. standard Gaussian random variables independent of Z_0. The parameters $(\phi, \sigma_x, \sigma_y) = (1, 0.1, 0.1)$ are used to sample $n = 100,000$ observations.

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. 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 i.i.d. with distribution W. In the following experiments, N is set to $N = 5000$. 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) introduced in [Igel et al., 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 $(x_i, x_{i+1}) \times (x_j, x_{j+1})$, $1 \leqslant i,j \leqslant r$, given by $\alpha_{i,j} r^2$ and the benchmark estimation given by the empirical estimate that would be computed if the sequence $(X_k)_{1\leqslant k\leqslant n}$ were observed: $n^{-1}\sum_{k=1}^{n-1} \mathbb{1}_{(x_i,x_{i+1})\times(x_j,x_{j+1})}(X_k, X_{k+1})$. The results are displayed in Figure 1 when $r = 10$ with CMA-ES initialized at a random point and a maximum number of evaluations of $\widehat{M}_n(R)$ set to 15000. The associated estimated histograms for the estimation of the distribution of X_1 are displayed in Figure 2 with their confidence regions. The estimated values are defined as the mean of the optimization results over 50 independent Monte Carlo runs and the associated confidence regions are obtained with the empirical variance over those runs.

Penalized maximum likelihood. 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 (19) with $\Theta = \mathbb{R}\times(0,+\infty)$. This section serves as an illustration of the maximum likelihood approach and, although the compacity assumptions of Section 3.2 are not satisfied, the practical 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 even $D \geqslant 1$, the spaces G_D are chosen as the set of densities of mixtures of $D/2$ Gaussian distributions, that is when the measure p in (19) has $D/2$ support points. The models $S_{M,D,n}$ are defined as the set of parameters $(m, Q, \gamma) \in \mathbb{R}^M \times [0,1]^{M\times M} \times G_D$ such that Q is a transition matrix, with no constraint on Q or on the regularity of the kernel. Note that $\gamma \in G_D$, which corresponds to $\alpha = 0$ in the previous definition of $S_{M,D,n}$. An approximation of the maximum likelihood estimator is computed using the EM algorithm [Dempster et al., 1977] for $(M, D) = (10, 4)$ (γ is a mixture of two distributions). The results are displayed in Figure 1 and 2 where the estimated joint distribution and the benchmark distribution are assumed to be piecewise constant on the Voronoi diagram obtained with the output points of the maximum likelihood procedure (the histogram provides the probability that X_1 lies in each cell of the diagram).

References

[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.

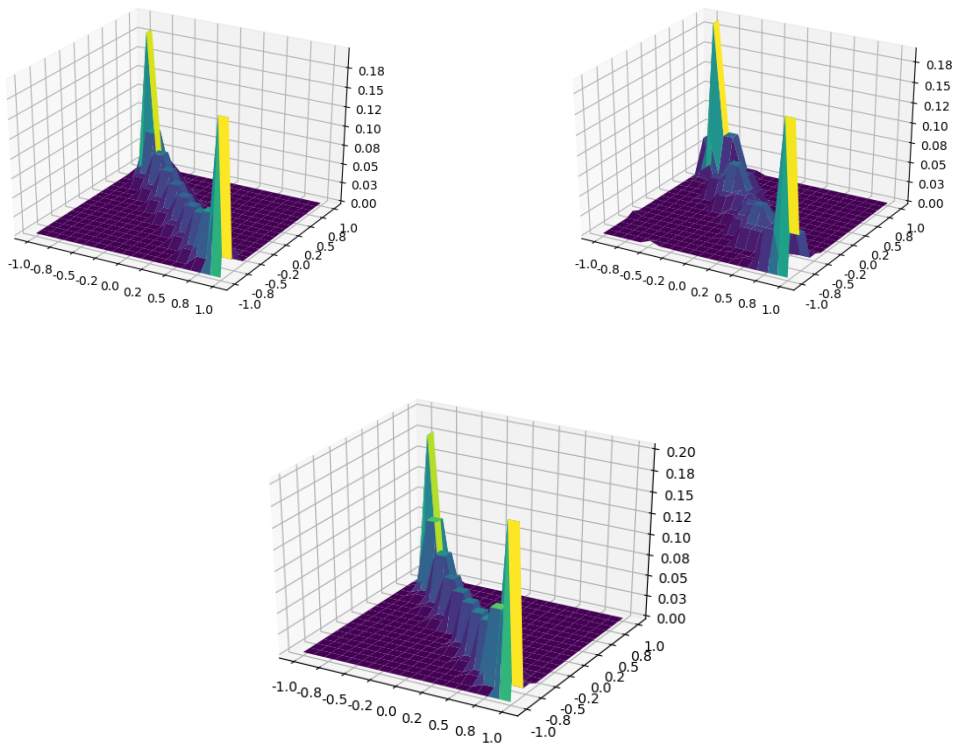

Figure 1: Estimated joint probability density of (X_1, X_2). Benchmark when the states are observed (top left), least squares estimate (top right) and maximum likelihood estimate (bottom).

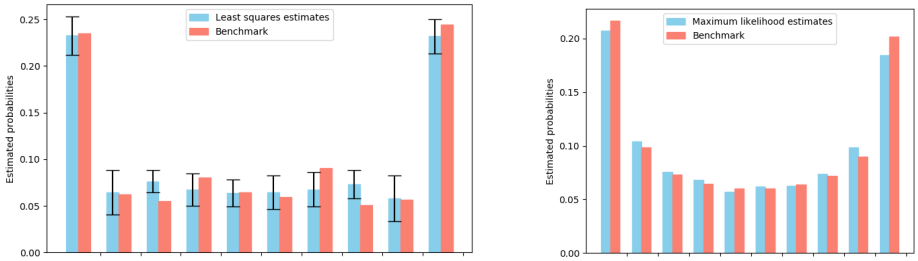

Figure 2: Estimated histograms associated with the distribution of X_1 for the least squares approach (left) and the maximum likelihood procedure (right).

[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.

[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.

[Delaigle et al., 2008] Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, 36(2):665–685.

[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.

[Eckle et al., 2016] Eckle, K., Bissantz, N., and Dette, H. (2016). Multiscale inference for multivariate deconvolution. *arXiv:1611.05201*.

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

[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.

[Igel et al., 2007] Igel, C., Hansen, N., and Roth, S. (2007). Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation*, 11:1–28.

[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. *arXiv:1807.03997*.

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

[Li and Vuong, 1998] Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.*, 65(2):139–165.

[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.

[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.

[Schennach and Hu, 2013] Schennach, S. M. and Hu, Y. (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *J. Amer. Statist. Assoc.*, 108(501):177–186.

[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, pear] Touron, A. (to appear). Consistency of the maximum likelihood estimator in seasonal hidden Markov models. *Statistics and Computing*.

[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.

A 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. $z \mapsto \Phi_{R_{\widetilde{K}}}(z,0)$) and $z \mapsto \Phi_{R_K}(0,z)$ (resp. $z \mapsto \Phi_{R_{\widetilde{K}}}(0,z)$) have no zeros if and only if the Markov chain is Gaussian. Therefore, the assumptions imply that in all cases, the stationary Markov chains with transition kernel K (resp. $\widetilde{K}$) is not a sequence of i.i.d. 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 $z \in \mathbb{C}$,

$$\Phi_{R_K}(z,0) = \mathrm{e}^{S(z)}\Phi_{R_{\widetilde{K}}}(z,0) \quad \text{and} \quad \Phi_{R_K}(0,z) = \mathrm{e}^{S(z)}\Phi_{R_{\widetilde{K}}}(0,z)\,,$$

and for all for all $(z_1, z_2) \in \mathbb{C} \times \mathbb{C}$,

$$\Phi_{R_K}(z_1,z_2) = e^{S(z_1)}e^{S(z_2)}\Phi_{R_{\widetilde{K}}}(z_1,z_2)\,. \tag{27}$$

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, (27) means that there exist i.i.d.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 i.i.d., 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 i.i.d. 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 i.i.d. sequence.*

Proof. For all $x \in \mathbb{R}$, let $x' \mapsto q(x,x')$ be the density of transition kernel $Q(x,\cdot)$ of the Markov chain $(X_i)_{i\geqslant 1}$ with respect to the Lebesgue measure and μ be its stationary density. 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. If

ϕ is the Gaussian density of η_i, for all real numbers y_1, y_2, y_3,

$$\int \mu(x_1)q(x_1,x_2)\phi(y_1-x_1)\phi(y_2-x_2)q(x_2,x_3)\phi(y_3-x_3)\mu(x_4)\phi(y_2-x_4)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4$$
$$= \int \mu(x_1)q(x_1,x_2)\phi(y_1-x_1)\phi(y_2-x_2)\mu(x_4)q(x_4,x_3)\phi(y_3-x_3)\phi(y_2-x_4)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\mathrm{d}x_4.$$

Since y is a complete statistic for $\phi(x-y)dx$, for all real numbers x_1, x_3, y_2,

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

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

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

Let $(\widetilde{X}_i)_{i\geqslant 1}$ be a Markov chain with the same distribution of $(X_i)_{i\geqslant 1}$ but independent of $(X_i)_{i\geqslant 1}$. For any measurable and positive function H, by writing expectations for G using the previous identity when G is defined by $G:(x,y,z)\mapsto H(x,y,z)1/\phi((x-y)/2)$,

$$\mathbb{E}\left[H\left(X_2, \widetilde{X}_2, X_3\right)\right] = \mathbb{E}\left[H\left(X_2, \widetilde{X}_2, \widetilde{X}_3\right)\right]\,,$$

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. □

B Proof of Theorem 3

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

$$|M_n(R)-M(R)| \leqslant \frac{3}{\sqrt{n}}\sup_{(t_1,t_2)\in\mathcal{S}}|Z_n(t_1,t_2)| + \frac{1}{n}\sup_{(t_1,t_2)\in\mathcal{S}}|Z_n(t_1,t_2)|^2\,, \tag{28}$$

and using assumption (14), $\sup_{R\in\mathcal{R}}|M_n(R)-M(R)| = O_{\mathbb{P}^\star}(n^{-1/2})$. Now, using the definition of $\widehat{R}_n$ and (28),

$$\begin{aligned} M(\widehat{R}_n) &\leqslant M_n(\widehat{R}_n) + O_{\mathbb{P}^\star}(n^{-1/2})\,,\\ &\leqslant M_n(R^\star) + O_{\mathbb{P}^\star}(n^{-1/2})\,,\\ &\leqslant M(R^\star) + O_{\mathbb{P}^\star}(n^{-1/2})\,,\\ &= O_{\mathbb{P}^\star}(n^{-1/2})\,, \end{aligned}$$

since $M(R^\star)=0$, and (15) 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.

C Proof of Theorem 4

Compacity of the set of parameters. 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.

Let Ω_ω^C be 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 $\mathcal{C}^0_{b,+}$ be the set of real-valued, nonnegative, bounded and continuous function on Λ, then for all $f \in \mathcal{C}^0_{b,+}$ and all $x \in \Lambda$,

$$\int K_n(x, \mathrm{d}x') f(x') \in [1/C, C] \int f \mathrm{d}\lambda_n$$

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

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

For all closed set $F \subset \Lambda$, there exists a sequence $(f_i)_{i\geqslant 1} \searrow \mathbf{1}_F$, so that this implies

$$\forall F \subset \Lambda \text{ closed}, \quad \forall x \in \Lambda, \quad K(x,F) \in \left[\frac{1}{C}, C\right] \lambda(F) \,.$$

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. Finally, note that the set

$$\widetilde{\Gamma} = \left\{ \alpha \frac{b}{\|b\|_1} + (1-\alpha)\gamma : \gamma \in \Gamma, \alpha \in [0,1] \right\} \tag{29}$$

is a compact subset of $\mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^\infty(\mathbb{R}^d)$ with respect to the $\mathbf{L}^1$ norm such that all functions in $\widetilde{\Gamma}$ are positive, continuous and centered and $\sup_{\gamma\in\widetilde{\Gamma}} \sup_{x\in\Lambda} \gamma(y-x) \leqslant B(y)$ for all $y \in \mathbb{R}^d$, where the function $B \in \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^\infty(\mathbb{R}^d)$ is defined, for all $y \in \mathbb{R}^d$, by

$$B(y) = \max_{z\in\{-L,L\}^d} b(y+z) \,. \tag{30}$$

Continuity of the total variation distance between the distributions of the observed process. The set of possible parameters $\Omega_\omega^C \times \widetilde{\Gamma}$ is endowed with the product topology induced by the uniform convergence topology on Ω_ω^C and the $\mathbf{L}^1$ norm on $\widetilde{\Gamma}$. Let $(K_n, \gamma_n)_{n\geqslant 1}$ be a sequence on $\Omega_\omega \times \widetilde{\Gamma}$ that converges to (K, γ) with respect to this topology. K admits a unique stationary distribution, so that Theorem 4 and the corollary of Theorem 6 of [Karr, 1975] entail that

$$\mathbb{P}^X_{K_n} \xrightarrow[n\to\infty]{(d)} \mathbb{P}^X_K \,, \tag{31}$$

where $\mathbb{P}_K^X$ 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. Let us now show that the distribution of $(Y_1,\dots,Y_k)$ converges in total variation distance.

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

Since $x \longmapsto \gamma(y-x)$ is continuous and bounded for all $y \in \mathbb{R}^d$, Equation (31) implies that

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

Then, as $\sup_{x\in\Lambda} \gamma(y-x) \leqslant B(y)$ for all $y \in \mathbb{R}^d$ where B is defined in Equation (30),

$$\left| \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}_K^X(x) \right| \leqslant \prod_{i=1}^k B(y_i) \,,$$

and the right hand side is integrable. The same holds for K_n, so that the dominated convergence theorem implies

$$\int \left| \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}_K^X(x) - \int \prod_{i=1}^k \gamma(y_i - x_i) \mathrm{d}\mathbb{P}_{K_n}^X(x) \right| \mathrm{d}y \underset{n\to\infty}{\longrightarrow} 0 \,. \tag{32}$$

For the second term, note that

$$\begin{aligned}
&\int \int \left| \prod_{i=1}^k \gamma(y_i - x_i) - \prod_{i=1}^k \gamma_n(y_i - x_i) \right| \mathrm{d}\mathbb{P}_K^X(x) \mathrm{d}y \,, \\
&\qquad \leqslant \sum_{i=1}^k \int \int \prod_{j<i} \gamma(y_j - x_j) |\gamma(y_i - x_i) - \gamma_n(y_i - x_i)| \prod_{j>i} \gamma_n(y_i - x_i) \mathrm{d}\mathbb{P}_K^X(x) \mathrm{d}y \,, \\
&\qquad = \sum_{i=1}^k \int \int |\gamma(y_i - x_i) - \gamma_n(y_i - x_i)| \mathrm{d}y_i \mathrm{d}\mathbb{P}_K^X(x_i) \,, \\
&\qquad = k \|\gamma - \gamma_n\|_1 \,.
\end{aligned} \tag{33}$$

Let $\mathbb{P}_{K,\gamma}^{(k)}$ be the distribution of $(Y_1,\dots,Y_k)$ under the parameters (K,γ), then this equation shows that $d_{\mathrm{TV}}(P_{K,\gamma}^{(k)}, P_{K_n,\gamma_n}^{(k)}) \underset{n\to\infty}{\longrightarrow} 0$ for all $k \geqslant 1$.

Convergence in Kullback-Leibler divergence entails convergence in total variation distance of the distribution of the observed process. For all probability measures μ and ν, the Kullback Leibler divergence between μ and ν is defined by

$$KL(\mu\|\nu) := \begin{cases} \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathrm{d}\mu & \text{when } \mu \text{ is absolutely continuous with respect to } \nu, \\ +\infty & \text{otherwise.} \end{cases} \tag{34}$$

Lemma 3 of [Douc et al., 2004] implies that for all $K, K' \in \Omega_\omega^C$ and for all $\gamma, \gamma' \in \widetilde{\Gamma}$, the limit

$$\mathbf{K}(\mathbb{P}_{K,\gamma}\|\mathbb{P}_{K',\gamma'}) = \lim_{m\to+\infty} \frac{1}{m} KL(\mathbb{P}^{(m)}_{K,\gamma}\|\mathbb{P}^{(m)}_{K',\gamma'}) \tag{35}$$

exists, is in $\mathbb{R}\cup\{+\infty\}$, and if it is finite, then for all $k, m \geqslant 1$,

$$\left|k\mathbf{K}(\mathbb{P}_{K,\gamma}\|\mathbb{P}_{K',\gamma'}) - \left(KL(\mathbb{P}^{(m+k)}_{K,\gamma}\|\mathbb{P}^{(m+k)}_{K',\gamma'}) - KL(\mathbb{P}^{(m)}_{K,\gamma}\|\mathbb{P}^{(m)}_{K',\gamma'})\right)\right| \leqslant C^4\left(1-\frac{1}{C^2}\right)^{m-1}. \tag{36}$$

Let $(K_n,\gamma_n)_{n\geqslant 1} \in (\Omega_\omega^C\times\widetilde{\Gamma})^{\mathbb{N}}$ 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

$$KL(\mathbb{P}^{(m_n+l_n+k)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n+l_n+k)}_{K_n,\gamma_n}) - KL(\mathbb{P}^{(m_n)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(m_n)}_{K_n,\gamma_n}) \underset{n\to\infty}{\longrightarrow} 0\,. \tag{37}$$

Note that

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

using the chain rule and Pinsker's inequality. 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 $\phi(i)\leqslant 2(1-\frac{1}{C})^i$ (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 probability density γ and for all $A\in\sigma(Y_1,\dots,Y_{m_n})$ such that $\mathbb{P}_{K,\gamma}(A)>0$,

$$d_{\mathrm{TV}}\left(\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|A,K,\gamma},\mathbb{P}_{Y_{m_n+l_n+1}^{m_n+l_n+k}|K,\gamma}\right) \leqslant 2\left(1-\frac{1}{C}\right)^{l_n}, \tag{38}$$

so that using the continuity of γ_n and γ,

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

using that $(a-b)^2 \geqslant \frac{a^2}{2} - b^2$ 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 $\gamma \in \widetilde{\Gamma}$. Therefore, one has for all $k \geqslant 1$,

$$d_{\text{TV}}\left(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{K_n,\gamma_n}\right) \underset{n\to+\infty}{\longrightarrow} 0\,. \tag{39}$$

Consistency. Let us check the assumptions of [Lehéricy, 2018]. We endow $\mathbb{R}^d$ with the probability measure with density $b/\|b\|_1$ with respect to the Lebesgue measure. Then **[A⋆mixing]** and **[A⋆forgetting]** follow from his [Lehéricy, 2018, Lemma 1] and H4. **[A⋆tail]** follows from H3 since

$$p^\star(Y_i|Y_1^{i-1}) \leqslant \frac{\sup_{x\in\Lambda}\gamma^\star(Y_i - x)}{b(Y_i)/\|b\|_1} \leqslant \|b\|_1\,.$$

[Aergodic] and **[Atail]** follow for the same reason and because $\alpha \geqslant 1/n^2$ in the construction of $S_{M,D,n}$. Finally, **[Aentropy]** and **[Agrowth]** follow from H5.

Note that our penalty has a dimension term of the form $(D + M^2)$ instead of $(DM + M^2)$. This comes from the fact that there is a single emission density instead of M densities. A careful reading of the proof shows that [Lehéricy, 2018, Theorem 8] holds with this penalty in our setting. Therefore, Theorem 8 implies that almost surely

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{\widehat{K}_n,\widehat{\gamma}_n}) \underset{n\to\infty}{\longrightarrow} 0 \tag{40}$$

as soon as there exists an approximating sequence $(m_t, Q_t, \gamma_t)_{t\geqslant 1}$ in $\bigcup_{M,D,n} S_{M,D,n}$ such that $\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{K_{m_t,Q_t},\gamma_t}) \underset{t\to+\infty}{\longrightarrow} 0$. Let us assume for now that it exists, then almost surely for all $k \geqslant 1$,

$$d_{\text{TV}}\left(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{\widehat{K}_n,\widehat{\gamma}_n}\right) \underset{n\to+\infty}{\longrightarrow} 0. \tag{41}$$

Therefore, all limits (K, γ) of convergent subsequences of $(\widehat{K}_n, \widehat{\gamma}_n)_n$ 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 Corollary 1 using assumption H2, 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 $x \in \text{Supp}(\lambda^\star)$. Since the set of parameters is compact, Theorem 4 follows.

Existence of an approximating sequence. Assume that there exists a sequence $(K_t)_{t\geqslant 1}$ in Ω_ω^C of kernels of the form K_{m_t,Q_t} with $(m_t, Q_t, -) \in \bigcup_{M,D,n} S_{M,D,n}$ such that $K_t \longrightarrow K^\star$. Let $(\delta_t)_{t\geqslant 1}$ be a sequence with values in $(0,1]$ such that $\delta_t \longrightarrow 0$ and let $\gamma_t = \delta_t \frac{b}{\|b\|_1} + (1-\delta_t)\gamma^\star$ for all $t \geqslant 1$ (note that we may replace $\gamma^\star$ by a sequence with values in $\bigcup_D G_D$ that converges to $\gamma^\star$ without changing the following proof). Then by Equations (32) and (33), there exists a sequence $(k_t)_{t\geqslant 1}$ such that $k_t \longrightarrow \infty$, $k_t^2\delta_t \log\frac{1}{\delta_t} \longrightarrow 0$ and such that for all t large enough,

$$d_{\text{TV}}(P^{(k_t)}_{K^\star,\gamma^\star}, P^{(k_t)}_{K_t,\gamma_t}) \leqslant 3k_t\delta_t\,. \tag{42}$$

Lemma 3 of [Douc et al., 2004] implies that for all $k \geqslant 1$ and for all $t \geqslant 1$,

$$\begin{aligned}\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star}\|\mathbb{P}_{K_t,\gamma_t}) &\leqslant \mathbb{E}KL(\mathbb{P}_{Y_k|Y_1,\dots,Y_{k-1},K^\star,\gamma^\star}\|\mathbb{P}_{Y_k|Y_0,\dots,Y_{k-1},K_t,\gamma_t}) + C^2\left(1-\frac{1}{C^2}\right)^{k-2}\,,\\ &\leqslant KL(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}\|\mathbb{P}^{(k)}_{K_t,\gamma_t}) + C^2\left(1-\frac{1}{C^2}\right)^{k-2}\end{aligned} \tag{43}$$

by the entropy chain rule. Since for all $y \in \mathbb{R}^d$,

$$\inf_{x\in\Lambda} \gamma_t(y-x) \geqslant \delta_t \inf_{x\in\Lambda} b(y-x)/\|b\|_1 \geqslant \delta_t \sup_{x'\in\Lambda} \gamma^\star(y-x')/\|b\|_1 \,,$$

for all $(y_1,\dots,y_k) \in (\mathbb{R}^d)^k$,

$$p_{K_t,\gamma_t}(y_1,\dots,y_k) \geqslant \left(\frac{\delta_t}{\|b\|_1}\right)^k p_{K^\star,\gamma^\star}(y_1,\dots,y_k)\,,$$

so that Lemma 4 of [Shen et al., 2013] entails

$$\begin{aligned} KL(\mathbb{P}^{(k)}_{K^\star,\gamma^\star} \| \mathbb{P}^{(k)}_{K_t,\gamma_t}) &\leqslant \left(1 + 2k \log \frac{\|b\|_1}{\delta_t}\right) h^2(\mathbb{P}^{(k)}_{K^\star,\gamma^\star}, \mathbb{P}^{(k)}_{K_t,\gamma_t})\,, \\ &\leqslant 2\left(1 + 2k \log \frac{\|b\|_1}{\delta_t}\right) d_{\mathrm{TV}}(P^{(k)}_{K^\star,\gamma^\star}, P^{(k)}_{K_n,\gamma_n})\,, \end{aligned}$$

using that the square of the Hellinger distance is upper bounded by the $\mathbf{L}^1$ distance, that is twice the total variation distance. Together with Equations (42) and (43) and for the sequence $(k_t)_{t\geqslant 1}$ of Equation (42), this implies that for t large enough,

$$\mathbf{K}(\mathbb{P}_{K^\star,\gamma^\star} \| \mathbb{P}_{K_t,\gamma_t}) \leqslant 16 k_t^2 \delta_t \log \frac{\|b\|_1}{\delta_t} + C^2 \left(1 - \frac{1}{C^2}\right)^{k_t - 2}$$

which tends to zero by construction of $(k_t)_{t\geqslant 1}$. The last step is to prove that there exists a sequence $(K_t)_{t\geqslant 1}$ of kernels in $\bigcup_{M,D,n} S_{M,D,n}$ that converge to $K^\star$.

Lemma 2. *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 $(M_t)_{t\geqslant 1} \longrightarrow +\infty$ and a sequence $((A_i^t)_{1\leqslant i\leqslant M_t})_{t\geqslant 1}$ of measurable partitions of the support of λ such that*

$$\begin{cases} D_t := \sup\limits_{1\leqslant i\leqslant M_t} \mathit{diam}(A_i^t) \underset{t\to+\infty}{\longrightarrow} 0\,, \\ \forall t \geqslant 1, \quad \forall 1 \leqslant i \leqslant M_t, \quad \lambda(A_i^t) \in \left[\frac{1}{2M_t}, \frac{2}{M_t}\right]\,. \end{cases} \tag{44}$$

To address the case where $\lambda^\star$ is not absolutely continuous with respect to the Lebesgue measure, we 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 $W_1(\mu * U_\epsilon, \nu * U_\epsilon) = W_1(\mu,\nu)$ 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 $\epsilon \longrightarrow 0$. 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, we assume that $\lambda^\star$ is absolutely continuous with respect to the Lebesgue measure and construct a sequence approximating $K^\star$.

Let $(M_t)_{t\geqslant 1}$ and $((A_i^t)_{1\leqslant i\leqslant M_t})_{t\geqslant 1}$ be the sequences obtained by applying Lemma 2 to $\lambda^\star$. For all $t \geqslant 1$ and $i \in \{1,\dots,M_t\}$, let m_i^t be an element of A_i^t. For all $t \geqslant 1$, the elements of the vector m^t are distinct

because $(A_i^t)_{1\leqslant i\leqslant M_t}$ 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 $\{m_i^t\}_{1\leqslant i\leqslant M_t}$ defined by

$$\forall x \in \Lambda \cap (\eta_t \mathbb{Z}^d), \quad \forall i \in \{1, \dots, M_t\}, \quad \widetilde{K}_t(x, m_i) = K^\star(x, A_i^t). \tag{45}$$

Note that by the Lemma and assumption H4, $\widetilde{K}_t(x, m_i) \in [\frac{1}{CM_t}, \frac{C}{M_t}]$ for all x and i. Moreover, for all $x, x' \in \Lambda \cap (\eta_t \mathbb{Z}^d)$,

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

by taking $\eta_t \geqslant 4D_t / \inf_{u\in(0,\operatorname{diam}(\Lambda)]} \omega(u)/u$, 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 $\{m_i^t\}_{1\leqslant i\leqslant M_t}$ for all $x \in \Lambda$. Let Q_t be the matrix defined by $Q_t(i,j) = K_t(m_i^t, m_j^t)$ and let us show that K_{m^t,Q_t} approximates $K^\star$.

Note that all kernels considered here ($K^\star$, $\widetilde{K}_t$, K_t and K_{m^t,Q_t}) are kernels on the compact set Supp$(\lambda^\star)$. Therefore, we only need to show that $K_{m^t,Q_t} \longrightarrow K$ 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

$$\sup_{x\in\text{Supp}(\lambda^\star)} W_1(K_{m^t,Q_t}(x,\cdot), K^\star(x,\cdot)) \underset{t\to+\infty}{\longrightarrow} 0\,. \tag{46}$$

For all $x \in$ Supp$(\lambda^\star)$, let $X(x)$ (resp. $m(x)$) be one of the elements of $\Lambda \cap (\eta_t\mathbb{Z}^d)$ (resp. $\{m_i^t\}_{1\leqslant i\leqslant M_t}$) closest to x. Then $\sup_{x\in\text{Supp}(\lambda^\star)} |x - m(x)| \leqslant D_t$ 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$(\lambda^\star)$,

$$\begin{aligned} W_1(K_{m^t,Q_t}(x,\cdot), K^\star(x,\cdot)) \leqslant & W_1(K_{m^t,Q_t}(x,\cdot), K_{m^t,Q_t}(m(x),\cdot)) && (47)\\ & + W_1(K_{m^t,Q_t}(m(x),\cdot), K_t(m(x),\cdot)) && (48)\\ & + W_1(K_t(m(x),\cdot), K_t(X(m(x)),\cdot)) \\ & + W_1(K_t(X(m(x)),\cdot), K^\star(X(m(x)),\cdot)) && (49)\\ & + W_1(K^\star(X(m(x)),\cdot), K^\star(x,\cdot))\,. \end{aligned}$$

Note that (48) = (49) = 0 by definition of the kernels. Thus, the regularity assumptions on the kernels ensure that for all $x \in$ Supp$(\lambda^\star)$,

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

which proves Equation (46).

D Further examples

This section highlights some other common models for which the assumptions of Theorem 1 hold.

D.1 Deconvolution with repeated measurements

The model is given by $X_1 = X_2$. Then, $\Phi_R(z_1, z_2) = \Phi_R(z_1 + z_2, 0) = \Phi_R(0, z_1 + z_2)$. Assumption (A2) holds as soon as X_1 is not deterministic and its distribution is in $\mathcal{M}_\rho$ for $\rho < 2$, which holds for instance when it has bounded (unknown) support. Therefore, by Theorem 1, deconvolution with at least two repetitions is identifiable without any assumption on the noise distribution, under the mild assumption that the distribution of the variable of interest has light tails. The model may also contain outliers with unknown probability and still be identifiable.

Identifiability of the model where $X_1 = X_2$ has been proved in this strict submodel by [Li and Vuong, 1998] under the assumption that the characteristic functions of X_1 and of the noise are not vanishing everywhere. Kernel estimators where proved in [Delaigle et al., 2008] equivalent to those for deconvolution with known noise distribution when X_1 has a real characteristic function and for ordinary smooth errors and signal.

D.2 Errors in variable regression models

The model is given by $X_2 = g(X_1)$ where $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$. Note first that if the distribution of (X_1, X_2) is identified, then its support is identified. The support of (X_1, X_2) is the graph of the function g so that g is identified on the support of the distribution of X_1. Assume that the distributions of X_1 and X_2 are in $\mathcal{M}_\rho$ for some $\rho < 2$, which is the case if they are bounded. Assume now that the supports of X_1 and $X_2 = g(X_1)$ have a nonempty interior, which is the case if for instance g is assumed continuous.

If (A1) does not hold, then either there exists $z_0 \in \mathbb{C}^{d_1}$ such that for all $z \in \mathbb{C}^{d_2}$, $\mathbb{E}[\mathrm{e}^{z_0^T X_1 + z^T g(X_1)}] = 0$, or there exists $z_0 \in \mathbb{C}^{d_2}$ such that for all $z \in \mathbb{C}^{d_1}$, $\mathbb{E}[\mathrm{e}^{z^T X_1 + z_0^T g(X_1)}] = 0$. In the last case, since the support of X_1 has a nonempty interior, this is equivalent to $\mathbb{E}[\mathrm{e}^{z_0^T g(X_1)} | X_1] = 0$, which means that $e^{z_0^T g(X_1)} = 0$, which is impossible. Thus, since the support of $g(X_1)$ has a nonempty interior, (A1) does not hold if and only if for some z_0, $\mathbb{E}\left[\mathrm{e}^{z_0^T X_1} \middle| g(X_1)\right] = 0$. Thus, under the assumption that g is continuous and the support of X_1 has a nonempty interior, the error in variables regression model is identifiable without knowing the distribution of the noise as soon as for all z_0,

$$\mathbb{E}\left[\mathrm{e}^{z_0^T X_1} \middle| g(X_1)\right] \neq 0\,. \tag{50}$$

In particular, if g is one-to-one on a subset of the support of X_1 with nonempty interior, then for all z_0, (50) is verified and the model is identifiable. See [Schennach and Hu, 2013].

D.3 Nonparametric hidden regression variables

The model is given by $X_i = g(Z_i)$ where $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ and $(Z_i)_{i \geqslant 1}$ is a sequence of hidden variables, see [Dumont and Le Corff, 2017a, Dumont and Le Corff, 2017b] for nonparametric hidden regression models. In these papers, the counterpart of the nonparametric modeling of the hidden process is the use of partly parametric modeling for the emission densities. 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 (A1) 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.