State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models

Luc Lehéricy

State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models

Luc Lehéricy LUC.LEHERICY@MATH.U-PSUD.FR
Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France

Abstract

In this paper, we introduce a new estimator for the emission densities of a nonparametric hidden Markov model. It is adaptive and minimax with respect to each state's regularity– as opposed to globally minimax estimators, which adapt to the worst regularity among the emission densities. Our method is based on Goldenshluger and Lepski's methodology. It is computationally efficient and only requires a family of preliminary estimators, without any restriction on the type of estimators considered. We present two such estimators that allow to reach minimax rates up to a logarithmic term: a spectral estimator and a least squares estimator. We show how to calibrate it in practice and assess its performance on simulations and on real data.

Keywords: hidden Markov model; model selection; nonparametric density estimation; oracle inequality; adaptive minimax estimation; spectral method; least squares method.

1. Introduction

Finite state space hidden Markov models, or HMMs in short, are powerful tools for studying discrete time series and have been used in a variety of applications such as economics, signal processing and image analysis, genomics, ecology, speech recognition and ecology among others. The core idea is that the behaviour of the observations depends on a hidden variable that evolves like a Markov chain.

Formally, a hidden Markov model is a process $(X_j, Y_j)_{j\geq 1}$ in which $(X_j)_j$ is a Markov chain on $\mathcal{X}$, the Y_i's are independent conditionally on $(X_j)_j$ and the conditional distribution of Y_i given $(X_j)_j$ depends only on X_i. The parameters of the HMM are the parameters of the Markov chain, that is its initial distribution and transition matrix, and the parameters of the observations, that is the *emission distributions* $(\nu_k^*)_{k\in\mathcal{X}}$ where ν_k^* is the distribution of Y_j conditionally to $X_j = k$. Only the observations $(Y_j)_j$ are available.

In this article, we focus on estimating the emission distributions in a nonparametric setting. More specifically, assume that the emission distributions have a density with respect to some known dominating measure μ, and write f_k^* their densities–which we call the *emission densities*. The goal of this paper is to estimate all f_k^*'s with their minimax rate of convergence when the emission densities are not restricted to belong to a set of densities described by finitely many parameters.

1.1 Nonparametric state-by-state adaptivity

Theoretical results in the nonparametric setting have only been developed recently. De Castro et al. (2017) and Bonhomme et al. (2016b) introduce spectral methods, and the latter is proved to be minimax but not adaptive–which means one needs to know the regularity of the densities beforehand to reach the minimax rate of convergence. De Castro et al. (2016) introduce a least squares estimator which is shown to be minimax adaptive up to a logarithmic term. However, all these papers have a common drawback: they study the emission densities as a whole and can not handle them separately. This comes from their error criterion, which is the supremum of the errors on all densities: what they actually prove is that $\max_{k\in\mathcal{X}} \|\hat{f}_k - f_k^*\|_2$ converges with minimax rate when $(\hat{f}_k)_k$ are their density estimators. In general, the regularity of each emission density could be different, leading to different rates of convergence. This means that having just one emission density that is very hard to estimate is enough to deteriorate the rate of convergence of all emission densities.

In this paper, we construct an estimator that is adaptive and estimates each emission density with its own minimax rate of convergence. We call this property state-by-state adaptivity. Our method does so by handling each emission density individually in a way that is theoretically justified–reaching minimax and adaptive rates of convergence with respect to the regularity of the emission densities–and computationally efficient thanks to its low computational and sample complexity.

Our approach for estimating the densities nonparametrically is model selection. The core idea is to approximate the target density using a family of parametric models that is dense within the nonparametric class of densities. For a square integrable density f^*, we consider its projection f_M^* on a finite-dimensional space $\mathfrak{P}_M$ (the parametric model), where M is a model index. This projection introduces an error, the *bias*, which is the distance $\|f^* - f_M^*\|_2$ between the target quantity and the model. The larger the model, the smaller the bias. On the other hand, larger models will make the estimation harder, resulting in a larger *variance* $\|\hat{f}_M - f_M^*\|_2^2$. The key step of model selection is to select a model with a small total error–or alternatively, a good *bias-variance tradeoff*.

In many situations, it is possible to reach the minimax rate of convergence with a good bias-variance tradeoff. Previous estimators of the emission densities of a HMM perform such a tradeoff based on an error that takes the transition matrix and all emission densities into account. Such an error leads to a rate of convergence that corresponds to the slowest minimax rate amongst the different parameters. In contrast, our method performs a bias-variance tradeoff for each emission density using an error term that depends only on the density in question, which makes it possible to reach the minimax rates for each density.

1.2 Plug-in procedure

The method we propose is based on the method developed in the seminal papers of Goldenshluger and Lepski (2011, 2014) for density estimation, extended by Goldenshluger and Lepski (2013) to the white noise and regression models. It takes a family of estimators as input and chooses the estimator that performs a good bias-variance tradeoff separately for each hidden state. We recommend the article of Lacour et al. (2016) for an insightful presentation of this method in the case of conditional density estimation.

Our method and assumptions are detailed in Section 2. Let us give a quick overview of the method. Assume the densities belong to a Hilbert space $\mathcal{H}$. Given a family of subsets of finite-dimensional subspaces of $\mathcal{H}$ (the models) indexed by M and estimators $\hat{f}_k^{(M)}$ of the emission densities for each hidden state k and each model M, one computes a substitute for the bias of the estimators by

$$A_k(M) = \sup_{M'} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M \wedge M')} \right\|_2 - \sigma(M') \right\}.$$

for some penalty σ. Then, for each state k, one selects the estimator $\hat{M}_k$ from the model M minimizing the quantity $A_k(M) + 2\sigma(M)$. The penalty σ can also be interpreted as a variance bound, so that this penalization procedure can be seen as performing a bias-variance tradeoff. The novelty of this method is that it selects a different $\hat{M}_k$, that is a different model, for each hidden state: this is where the state-by-state adaptivity comes from. Also note that contrary to Goldenshluger and Lepski (2013), we do not make any assumption on how the estimators are computed, provided a variance bound holds.

The main theoretical result is an oracle inequality on the selected estimators $\hat{f}_k^{(\hat{M}_k)}$, see Theorem 2. As a consequence, we are able to get a rate of convergence that is different for each state. These rates of convergence will even be adaptive minimax up to a logarithmic factor when the method is applied to our two families of estimators: spectral estimators and least squares estimators. To the best of our knowledge, this is the first state-by-state adaptive algorithm for hidden Markov models.

Note that finding the right penalty term σ is essential in order to obtain minimax rates of convergence. This requires a fine theoretical control of the variance of the auxiliary estimators, in the form of assumption $[\mathbf{H}(\epsilon)]$ (see Section 2.1). To the best of our knowledge, there is no suitable result in the literature. This is the second theoretical contribution of this paper: we control two families of estimators in a way that makes it possible to reach adaptive minimax rate with our state-by-state selection method, up to a logarithmic term.

On the practical side, we run this method and several variants on data simulated from a HMM with three hidden states and one irregular density, as illustrated in Section 4. The simulations confirm that it converges with a different rate for each emission density, and that the irregular density does not alter the rate of convergence of the other ones, which is exactly what we wanted to achieve.

Better still, the added computation time is negligible compared to the computation time of the estimators: even for the spectral estimators of Section 3.2 (which can be computed much faster than the least squares estimators and the maximum likelihood estimators using EM), computing the estimators on 200 models for 50,000 observations (the lower bound of our sample sizes) of a 3-states HMM requires a few minutes, compared to a couple of seconds for the state-by-state selection step. The difference becomes even larger for more observations, since the complexity of the state-by-state selection step is independent of the sample size: for instance, computing the spectral estimators on 300 models for 2,200,000 observations requires a bit less than two hours, and a bit more than ten hours for 10,000,000 observations, compared to less than ten seconds for the selection step in both cases. We refer to Section 4.6 for a more detailled discussion about the algorithmic complexity of the algorithms.

1.3 Families of estimators

We use two methods to construct families of estimators and apply the selection algorithm. The motivation and key result of this part of the paper is to control the variances of the estimators by the right penalty σ. This part is crucial if one wants to get adaptive minimax rates, and has not been adressed in previous papers. For both methods, we develop new theoretical results that allow to obtain a penalty σ that leads to adaptive minimax rates of convergence up to a logarithmic term. We present the algorithms and theoretical guarantees in Section 3.

The first method is a spectral method and is detailed in Section 3.2. Several spectral algorithms were developed, see for instance Anandkumar et al. (2012) and Hsu et al. (2012) in the parametric setting, and Bonhomme et al. (2016b) and De Castro et al. (2017) in a nonparametric framework. The main advantages of spectral methods are their computational efficiency and the fact that they do not resort to optimization procedure such as the EM and more generally nonconvex optimization algorithm, thus avoiding the well-documented issue of getting stuck into local sub-optimal minima.

Our spectral algorithm is based on the one studied in De Castro et al. (2017). However, their estimator cannot reach the minimax rate of convergence: the variance bound $\sigma(M)$ deduced from their results is proportional to M^3, while reaching the minimax rate requires $\sigma(M)$ to be proportional to M. To solve this issue, we introduce a modified version of their algorithm and show that it has the right variance bound, so that it is able to reach the adaptive minimax rate after our state-by-state selection procedure, up to a logarithmic term. Our algorithm also has an improved complexity: it is at most quasi-linear in the number of observations and in the model dimension, instead of cubic in the model dimension for the original algorithm.

The second method is a least squares method and is detailed in Section 3.3. Nonparametric least squares methods were first introduced by De Castro et al. (2016) to estimate the emission densities and extended by Lehéricy (to appear) to estimate all parameters at once. They rely on estimating the density of three consecutive observations of the HMM using a least squares criterion. Since the model is identifiable from the distribution of three consecutive observations when the emission distributions are linearly independent, it is possible to recover the parameters from this density. In practice, these methods are more accurate than the spectral methods and are more stable when the models are close to not satisfying the identifiability condition, see for instance De Castro et al. (2016) for the accuracy and Lehéricy (to appear) for the stability. However, since they rely on the minimization of a nonconvex criterion, the computation times of the corresponding algorithms are often longer than the ones from spectral methods.

A key step in proving theoretical guarantees for least squares methods is to relate the error on the density of three consecutive observations to the error on the HMM parameters in order to obtain an oracle inequality on the parameters from the oracle inequality on the density of three observations. More precisely, the difficult part is to lower bound the error on the density by the error on the parameters. Let us write g and g' the densities of the first three observations of a HMM with parameters θ and θ' respectively (these parameters actually correspond to the transition matrix and the emission densities of the HMM). Then

one would like to get

$$\|g - g'\|_2 \geq \mathcal{C}(\theta)\, d(\theta, \theta')$$

where d is the natural $\mathbf{L}^2$ distance on the parameters and $\mathcal{C}(\theta)$ is a positive constant which does not depend on θ'. Such inequalities are then used to lower bound the variance of the estimator of the density of three observations g^* by the variance of the parameter estimators: let g be the projection of g^* and g' be the estimator of g^* on the current approximation space (with index M). Denote θ^*_M and $\hat{\theta}_M$ the corresponding parameters and assume that the error $\|g - g'\|_2$ is bounded by some constant $\sigma'(M)$, then the result will be that

$$d(\hat{\theta}_M, \theta^*_M) \leq \frac{\sigma'(M)}{\mathcal{C}(\theta^*_M)}.$$

Such a result is crucial to control the variance of the estimators by a penalty term σ, which is the result we need for the state-by-state selection method. In the case where only the emission densities vary, De Castro et al. (2016) proved that such an inequality always holds for HMMs with 2 hidden states using brute-force computations, but it is still unknown whether it is always true for larger number of states. When the number of states is larger than 2, they show that this inequality holds under a generic assumption. Lehéricy (to appear) extended this result to the case where all parameters may vary. However, the constants deduced from both articles are not explicit, and their regularity (when seen as a function of θ) is unknown, which makes it impossible to use in our setting: one needs the constants $\mathcal{C}(\theta^*_M)$ to be lower bounded by the same positive constant, which requires some sort of regularity on the function $\theta \longmapsto \mathcal{C}(\theta)$ in the neighborhood of the true parameters.

To solve this problem, we develop a finer control of the behaviour of the difference $\|g - g'\|_2$, which is summarized in Theorem 10. We show that it is possible to assume $\mathcal{C}$ to be lower semicontinuous and positive without any additional assumption. In addition, we give an explicit formula for the constant when θ' and θ are close, which gives an explicit bound for the asymptotical rate of convergence. This result allows us to control the variance of the least squares estimators by a penalty σ which ensures that the state-by-state method reaches the adaptive minimax rate up to a logarithmic term.

1.4 Numerical validation and application to real data sets

Section 4 shows how to apply the state-by-state selection method in practice and shows its performance on simulated data and a comparison with a method based on cross validation that does note estimate state by state.

Note that the theoretical results give a penalty term σ known only up to a multiplicative constant which is unknown in practice. This problem, the *penalty calibration* issue, is usual in model selection methods. It can be solved using algorithms such as the dimension jump heuristics, see for instance Birgé and Massart (2007), who introduce this heuristics and prove that it leads to an optimal penalization in the special case of Gaussian model selection framework. This method has been shown to behave well in practice in a variety of domains, see for instance Baudry et al. (2012). We describe the method and show how to use this heuristics to calibrate the penalties in Section 4.2.

We propose and compare several variants of our algorithm. Section 4.2 shows some variants in the calibration of the penalties and Section 4.3 shows other ways to select the

final estimator. We discuss the result of the simulations and the convergence of the selected estimators in Section 4.4.

In Section 4.5, we compare our method with a non state-by-state adaptive method based on cross validation. Finally, we discuss the complexities of the auxiliary estimation methods and of our selection procedures in Section 4.6.

In Section 5, we apply our algorithm to two sets of GPS tracks. The first data set contains trajectories of artisanal fishers from Madagascar, recorded using a regular sampling with 30 seconds time steps. The second data set contains GPS positions of Peruvian seabird, recorded with 1 second time steps. We convert these tracks into the average velocity during each time step and apply our method using spectral estimators as input. The observed behaviour confirms the ability of our method to adapt to the different regularities by selecting different dimensions for each emission density.

Section 6 contains a conclusion and perspectives for this work.

Finally, Appendix A contains the details of our spectral algorithm and Appendix B is dedicated to the proofs.

1.5 Notations

We will use the following notations throughout the paper.

- $[K] = \{1, \dots, K\}$ is the set of integers between 1 and K.
- $\mathfrak{S}(K)$ is the set of permutations of $[K]$.
- $\|\cdot\|_F$ is the Frobenius norm. We implicitely extend the definition of the Frobenius norm to tensors with more than 2 dimensions.
- $\mathrm{Span}(A)$ is the linear space spanned by the family A.
- $\sigma_1(A) \geq \cdots \geq \sigma_{p\wedge n}(A)$ are the singular values of the matrix $A \in \mathbb{R}^{n\times p}$.
- $\mathbf{L}^2(\mathcal{Y}, \mu)$ is the set of real square integrable measurable functions on $\mathcal{Y}$ with respect to the measure μ.
- For $\mathbf{f} = (f_1, \dots, f_K) \in \mathbf{L}^2(\mathcal{Y}, \mu)^K$, $G(\mathbf{f})$ is the Gram matrix of $\mathbf{f}$, defined by $G(\mathbf{f})_{i,j} = \langle f_i, f_j \rangle$ for all $i, j \in [K]$.

2. The state-by-state selection procedure

In this section, we introduce the framework and our state-by-state selection method.

In Section 2.1, we introduce the notations and assumptions. In Section 2.2, we present our selection method and prove that it satisfies an oracle inequality.

2.1 Framework and assumptions

Let $(X_j)_{j\geq 1}$ be a Markov chain with finite state space $\mathcal{X}$ of size K. Let $\mathbf{Q}^*$ be its transition matrix and π^* be its initial distribution. Let $(Y_j)_{j\geq 1}$ be random variables on a measured space $(\mathcal{Y}, \mu)$ with μ σ-finite such that conditionally on $(X_j)_{j\geq 1}$ the Y_j's are independent with a distribution depending only on X_j. Let ν_k^* be the distribution of Y_j conditionally to $\{X_j = k\}$. Assume that ν_k^* has density f_k^* with respect to μ. We call $(\nu_k^*)_{k\in\mathcal{X}}$ the *emission distributions* and $\mathbf{f}^* = (f_k^*)_{k\in\mathcal{X}}$ the *emission densities*. Then $(X_j, Y_j)_{j\geq 1}$ is a hidden Markov

model with parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. The hidden chain $(X_j)_{j\geq 1}$ is assumed to be unobserved, so that the estimators are based only on the observations $(Y_j)_{j\geq 1}$.

Let $(\mathfrak{P}_M)_{M\in\mathbb{N}}$ be a nested family of finite-dimensional subspaces such that their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mu)$. The spaces $(\mathfrak{P}_M)_{M\in\mathbb{N}}$ are our models; in the following we abusively call M the model instead of $\mathfrak{P}_M$. For each index $M \in \mathbb{N}$, we write $\mathbf{f}^{*,(M)} = (f_k^{*,(M)})_{k\in\mathcal{X}}$ the projection of $\mathbf{f}^*$ on $(\mathfrak{P}_M)^K$. It is the best approximation of the true densities within the model M.

In order to estimate the emission densities, we do not need to use every models. Typically there is no point in taking models with more dimensions than the sample size, since they will likely be overfitting. Let $\mathcal{M}_n \subset \mathbb{N}$ be the set of indices which will be used for the estimation from n observations. For each $M \in \mathcal{M}_n$, we assume we are given an estimator $\hat{\mathbf{f}}_n^{(M)} = (\hat{f}_{n,k}^{(M)})_{k\in\mathcal{X}} \in (\mathfrak{P}_M)^K$. We will need to assume that for all models, the variance–that is the distance between $\hat{\mathbf{f}}_n^{(M)}$ and $\mathbf{f}^{*,(M)}$–is small with high probability. In the following, we drop the dependency in n and simply write $\mathcal{M}$ and $\hat{\mathbf{f}}^{(M)}$.

The following result is what one usually obtains in model selection. It bounds the distance between the estimators $\hat{\mathbf{f}}^{(M)}$ and the projections $\mathbf{f}^{*,(M)}$ by some penalty function σ. Thus, $\sigma/2$ can be seen as a bound of the variance term.

[$\mathbf{H}(\epsilon)$] With probability $1-\epsilon$,

$$\forall M \in \mathcal{M}, \quad \inf_{\tau_{n,M}\in\mathfrak{S}(K)} \max_{k\in\mathcal{X}} \left\| \hat{f}_k^{(M)} - f_{\tau_{n,M}(k)}^{*,(M)} \right\|_2 \leq \frac{\sigma(M,\epsilon,n)}{2}$$

where the upper bound $\sigma : (M, \epsilon, n) \in \mathcal{M} \times [0,1] \times \mathbb{N}^* \longmapsto \sigma(M,\epsilon,n) \in \mathbb{R}_+$ is nondecreasing in M. We show in Sections 3.2 and 3.3 how to obtain such a result for a spectral method and for a least squares method (using an algorithm from Lehéricy (to appear)). In the following, we omit the parameters ϵ and n in the notations and only write $\sigma(M)$.

What is important for the selection step is that the permutation $\tau_{n,M}$ does not depend on the model M: one needs all estimators $(\hat{f}_k^{(M)})_{M\in\mathcal{M}}$ to correspond to the same emission density, namely $f_{\tau_n(k)}^*$ when $\tau_{n,M} = \tau_n$ is the same for all models M. This can be done in the following way: let $M_0 \in \mathcal{M}$ and let

$$\hat{\tau}^{(M)} \in \operatorname*{arg\,min}_{\tau\in\mathfrak{S}(K)} \left\{ \max_{k\in\mathcal{X}} \left\| \hat{f}_{\tau(k)}^{(M)} - \hat{f}_k^{(M_0)} \right\|_2 \right\}$$

for all $M \in \mathcal{M}$. Then, consider the estimators obtained by swapping the hidden states by these permutations. In other words, for all $k \in \mathcal{X}$, consider

$$\hat{f}_{k,\text{new}}^{(M)} = \hat{f}_{\hat{\tau}^{(M)}(k)}^{(M)}.$$

Now, assume that the error on the estimators is small enough. More precisely, write $B_{M,M_0} = \max_{k\in\mathcal{X}} \left\| f_k^{*,(M)} - f_k^{*,(M_0)} \right\|_2$ the distance between the projections of $\mathbf{f}^*$ on the models M and M_0 and assume that $2\left[\sigma(M)/2 + \sigma(M_0)/2 + B_{M,M_0}\right]$ (that is twice the upper bound of the distance between two estimated emission densities corresponding to the same hidden states in models M and M_0) is smaller than $m(\mathbf{f}^*, M_0) := \min_{k'\neq k} \left\| f_k^{*,(M_0)} - f_{k'}^{*,(M_0)} \right\|_2$, which is the smallest distance between two different densities of the vector $\mathbf{f}^{*,(M_0)}$.

Then $[\mathbf{H}(\epsilon)]$ ensures that with probability as least $1-\epsilon$, for all k, there exists a single component of $\hat{\mathbf{f}}^{(M)}$ that is closer than $\sigma(M)/2 + \sigma(M_0)/2$ of $f_k^{*,(M_0)}$, and this component will be $\hat{f}^{(M)}_{\hat{\tau}^{(M)}(k)}$ by definition. This is summarized in the following lemma.

Lemma 1 *Assume $[\boldsymbol{H}(\epsilon)]$ holds. Then with probability $1-\epsilon$, there exists a permutation $\tau_n \in \mathfrak{S}(K)$ such that for all $k \in \mathcal{X}$ and for all $M \in \mathcal{M}$ such that*

$$\sigma(M) + \sigma(M_0) + 2B_{M,M_0} < m(\mathbf{f}^*, M_0),$$

one has

$$\max_{k\in\mathcal{X}} \left\| \hat{f}^{(M)}_{k,new} - f^{*,(M)}_{\tau_n(k)} \right\|_2 \leq \frac{\sigma(M)}{2}. \tag{1}$$

Proof Proof in Section B.1 ■

Thus, this property holds asymptotically as soon as $\inf \mathcal{M}$ tends to infinity and $\sup_{M\in\mathcal{M}} \sigma(M)$ tends to zero.

2.2 Estimator and oracle inequality

Let us now introduce our selection procedure. This method and the following theorem are based on the approach of Goldenshluger and Lepski (2011), but do not require any assumption on the structure of the estimators, provided a variance bound such as Equation (1) holds.

For each $k \in \mathcal{X}$ and $M \in \mathcal{M}$, let

$$A_k(M) = \sup_{M'\in\mathcal{M}} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M\wedge M')} \right\|_2 - \sigma(M') \right\}.$$

$A_k(M)$ serves as a replacement for the bias of the estimator $\hat{f}_k^{(M)}$, as can be seen in Equation (2). This comes from the fact that for large M', the quantity $\|\hat{f}_k^{(M')} - \hat{f}_k^{(M)}\|_2$ is upper bounded by the variances $\|\hat{f}_k^{(M')} - f_k^{*,(M')}\|_2$ and $\|\hat{f}_k^{(M)} - f_k^{*,(M)}\|_2$ (which are bounded by $\sigma(M')/2$) plus the bias $\|f_k^{*,(M)} - f_k^*\|_2$. Thus, only the bias term remains after substracting the variance bound $\sigma(M')$.

Then, for all $k \in \mathcal{X}$, select a model through the bias-variance tradeoff

$$\hat{M}_k \in \arg\min_{M\in\mathcal{M}} \{A_k(M) + 2\sigma(M)\}$$

and finally take

$$\hat{f}_k = \hat{f}_k^{(\hat{M}_k)}.$$

The following theorem shows an oracle inequality on this estimator.

Theorem 2 *Let $\epsilon \geq 0$ and assume equation (1) holds for all $k \in \mathcal{X}$ with probability $1-\epsilon$. Then with probability $1-\epsilon$,*

$$\forall k \in \mathcal{X}, \quad \|\hat{f}_k - f^*_{\tau_n(k)}\|_2 \leq 4 \inf_{M\in\mathcal{M}} \left\{ \|f^{*,(M)}_{\tau_n(k)} - f^*_{\tau_n(k)}\|_2 + \sigma(M,\epsilon) \right\}.$$

Proof We restrict ourselves to the event of probability at least $1-\epsilon$ where equation (1) holds for all $k \in \mathcal{X}$.

The first step consists in decomposing the total error: for all $M \in \mathcal{M}$ and $k \in \mathcal{X}$,

$$\begin{aligned}\left\|\hat{f}_k^{(\hat{M}_k)} - f^*_{\tau_n(k)}\right\|_2 \leq& \left\|\hat{f}_k^{(\hat{M}_k)} - \hat{f}_k^{(\hat{M}_k \wedge M)}\right\|_2 + \left\|\hat{f}_k^{(\hat{M}_k \wedge M)} - \hat{f}_k^{(M)}\right\|_2 \\ &+ \left\|\hat{f}_k^{(M)} - f^{*,(M)}_{\tau_n(k)}\right\|_2 + \left\|f^{*,(M)}_{\tau_n(k)} - f^*_{\tau_n(k)}\right\|_2.\end{aligned}$$

From now on, we will omit the subscripts k and $\tau_n(k)$. Using equation (1) and the definition of $A(M)$ and $\hat{M}$, one gets

$$\begin{aligned}\left\|\hat{f}^{(\hat{M})} - f^*\right\|_2 \leq& (A(M) + \sigma(\hat{M})) + (A(\hat{M}) + \sigma(M)) \\ &+ \sigma(M) + \left\|f^{*,(M)} - f^*\right\|_2 \\ \leq& 2A(M) + 4\sigma(M) + \left\|f^{*,(M)} - f^*\right\|_2.\end{aligned}$$

Then, notice that $A(M)$ can be bounded by

$$\begin{aligned}A(M) \leq& \sup_{M'} \left\{\left\|\hat{f}^{(M')} - f^{*,(M')}\right\|_2 + \left\|\hat{f}^{(M \wedge M')} - f^{*,(M \wedge M')}\right\|_2 - \sigma(M')\right\} \\ &+ \sup_{M'} \left\|f^{*,(M')} - f^{*,(M \wedge M')}\right\|_2.\end{aligned}$$

Since σ is nondecreasing, $\sigma(M \wedge M') \leq \sigma(M')$, so that the first term is upper bounded by zero thanks to equation (1). The second term can be controlled since the orthogonal projection is a contraction. This leads to

$$A(M) \leq \left\|f^* - f^{*,(M)}\right\|_2, \tag{2}$$

which is enough to conclude. ■

Remark 3 *The oracle inequality also holds when taking*

$$A_k(M) = \sup_{M' \geq M} \left\{\left\|\hat{f}_k^{(M')} - \hat{f}_k^{(M)}\right\|_2 - \sigma(M')\right\}_+.$$

Remark 4 *Note that the selected $\hat{M}_k$ implicitely depends on the probability of error ϵ through the penalty σ.*

In the asymptotic setting, we take ϵ as a function of n, so that $\hat{M}_k$ is a function of n only. This will be used to get rid of ϵ when proving that the estimators reach the minimax rates of convergence.

3. Plug-in estimators and theoretical guarantees

In this section, we introduce two methods to construct families of estimators of the emission densities. We show that they satisfy assumption $[\mathbf{H}(\epsilon)]$ for a given variance bound σ.

In Section 3.1, we introduce the assumptions we will need for both methods. Section 3.2 is dedicated to the spectral estimator and Section 3.3 to the least squares estimator.

3.1 Framework and assumptions

Recall that we approximate $\mathbf{L}^2(\mathcal{Y},\mu)$ by a nested family of finite-dimensional subspaces $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ such that their union is dense in $\mathbf{L}^2(\mathcal{Y},\mu)$ and write $f_k^{*,(M)}$ the orthogonal projection of f_k^* on $\mathfrak{P}_M$ for all $k\in\mathcal{X}$ and $M\in\mathcal{M}$. We assume that $\mathcal{M}\subset\mathbb{N}$ and that the space $\mathfrak{P}_M$ has dimension M. A typical way to construct such spaces is to take $\mathfrak{P}_M$ spanned by the first M vectors of an orthonormal basis.

Both methods will construct an estimator of the emission densities for each model of this family. These estimators will then be plugged in the state-by-state selection method of Section 2.2, which will select one model for each state of the HMM.

We will need the following assumptions.

[HX] $(X_j)_{j\geq 1}$ is a stationary ergodic Markov chain with parameters $(\pi^*,\mathbf{Q}^*)$;

[Hid] $\mathbf{Q}^*$ is invertible and the family $\mathbf{f}^*$ is linearly independent.

The ergodicity assumption in **[HX]** is standard in order to obtain convergence results. In this case, the initial distribution is forgotten exponentially fast, so that the HMM will essentially behave like a stationary process after a short period of time. For the sake of simplicity, we assume the Markov chain to be stationary.

[Hid] appears in identifiability results, see for instance Gassiat et al. (2015) and Theorem 8. It is sufficient to ensure identifiability of the HMM from the law of three consecutive observations. Note that it is in general not possible to recover the law of a HMM from two observations (see for instance Appendix G of Anandkumar et al. (2012)), so that three is actually the minimum to obtain general identifiability.

3.2 The spectral method

Algorithm 1 is a variant of the spectral algorithm introduced in De Castro et al. (2017). Unlike the original one, it is able to reach the minimax rate of convergence thanks to two improvements. The first one consists in decomposing the joint density on different models, hence the use of two dimensions m and M. The second one consists in trying several randomized joint diagonalizations instead of just one, and selecting the best one, hence the parameter r. These additional parameters do not actually add much to the complexity of the algorithm: in theory, the choice $m,r\approx\log(n)$ is fine (see Corollary 6), and in practice, any large enough constant works, see Section 4 for more details.

For all $M\in\mathcal{M}$, let $(\varphi_1^M,\dots,\varphi_M^M)$ be an orthonormal basis of $\mathfrak{P}_M$. Let

$$\eta_3(m,M)^2 := \sup_{y,y'\in\mathcal{Y}^3} \sum_{a,c=1}^{m}\sum_{b=1}^{M} (\varphi_a^m(y_1)\varphi_b^M(y_2)\varphi_c^m(y_3) - \varphi_a^m(y_1')\varphi_b^M(y_2')\varphi_c^m(y_3'))^2.$$

The following theorem follows the proof of Theorem 3.1 of De Castro et al. (2017), with modifications that allow to control the error of the spectral estimators in expectation and are essential to obtain the right rates of convergence in Corollary 6.

Theorem 5 *Assume **[HX]** and **[Hid]** hold. Then there exists a constant M_0 depending on $\mathbf{f}^*$ and constants C_σ and n_1 depending on $\mathbf{f}^*$ and $\mathbf{Q}^*$ such that for all $\epsilon\in(0,1)$, for all*

Algorithm 1: Spectral estimation of the emission densities of a HMM (short version)

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, two dimensions $m \leq M$, an orthonormal basis $(\varphi_1, \dots, \varphi_M)$ and number of retries r.

Result: Spectral estimators $(\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$.

[Step 1] Consider the following empirical estimators: for any $a, c \in [m]$ and $b \in [M]$,

- $\hat{\mathbf{M}}_{m,M,m}(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$
- $\hat{\mathbf{P}}_{m,m}(a,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_c(Y_{s+2})$.

[Step 2] Let $\hat{\mathbf{U}}_m$ be the $m \times K$ matrix of orthonormal left singular vectors of $\hat{\mathbf{P}}_{m,m}$ corresponding to its top K singular values. $\hat{\mathbf{U}}_m$ can be seen as a projection. Denote by $\mathbf{P'}$ and $\mathbf{M'}(\cdot, b, \cdot)$ the projected tensors, defined by $\mathbf{P'} = \hat{\mathbf{U}}_m^\top \hat{\mathbf{P}}_{m,m}\hat{\mathbf{U}}_m$ and likewise for $\mathbf{M'}$.

[Step 3] Form the matrices $\mathbf{B}(b) := (\mathbf{P'})^{-1}\mathbf{M'}$ for all $b \in [M]$.

[Step 4] Construct a matrix $\hat{\mathbf{O}}$ by taking the best approximate simultaneous diagonalization of all $\mathbf{B}(b)$ among r attempts: for all $b \in [M]$, $\mathbf{B}(b) \approx \mathbf{R}\mathrm{Diag}[\hat{\mathbf{O}}(b,\cdot)]\mathbf{R}^{-1}$ for some matrix $\mathbf{R}$ (see details in Algorithm 3, in Appendix A).

[Step 5] Define the emission densities estimators $\hat{\mathbf{f}}^{(M,r)} := (\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$ by: for all $k \in \mathcal{X}$, $\hat{f}_k^{(M,r)} := \sum_{b=1}^{M} \hat{\mathbf{O}}(b,k)\varphi_b$.

$m, M \in \mathcal{M}$ such that $M \geq m \geq M_0$ and for all $n \geq n_1\eta_3^2(m,M)(-\log\epsilon)^2$, with probability greater than $1-6\epsilon$,

$$\inf_{\tau\in\mathfrak{S}(K)} \max_{k\in\mathcal{X}} \|\hat{f}_k^{(M,\lceil t\rceil)} - f_{\tau(k)}^{*,(M)}\|_2^2 \leq C_\sigma \eta_3^2(m,M)\frac{(-\log\epsilon)^2}{n}$$

Proof Proof in Section B.2. ■

Note that the constants n_1 and C_σ depend on $\mathbf{Q}^*$ and $\mathbf{f}^*$. This dependency will not affect the rates of convergence of the estimators (with respect to the sample size n), but it can change the constants of the bounds and the minimum sample size needed to reach the asymptotic regime.

Let us now apply the state-by-state selection method to these estimators. The following corollary shows that it is possible to reach the minimax rate of convergence up to a logarithmic term separately for each state under standard assumptions. Note that we need to bound the resulting estimators by some power of n, but this assumption is not very restrictive since α can be arbitrarily large.

Corollary 6 *Assume **[HX]** and **[Hid]** hold. Also assume that $\eta_3^2(m,M) \leq C_\eta m^2 M$ for a constant $C_\eta > 0$ and that for all $k \in \mathcal{X}$, there exists s_k such that $\|f_k^{*,(M)} - f_k^*\|_2 = O(M^{-s_k})$. Then there exists a constant C_σ depending on $\mathbf{f}^*$ and $\mathbf{Q}^*$ such that the following holds.*

Let $\alpha > 0$ and $C \geq 2(1+2\alpha)\sqrt{C_\eta C_\sigma}$. Let $\hat{\mathbf{f}}^{sbs}$ be the estimators selected from the family $(\hat{\mathbf{f}}^{(M,\lceil(1+2\alpha)\log(n)\rceil)})_{M\leq M_{\max}(n)}$ with $M_{\max}(n) = n/\log(n)^5$, $m_M = \log(n)$ and $\sigma(M) = C\sqrt{\frac{M\log(n)^4}{n}}$ for all M. Then there exists a sequence of random permutations $(\tau_n)_{n\geq 1}$ such that

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\|(-n^\alpha)\vee(\hat{f}^{sbs}_{\tau_n(k)}\wedge n^\alpha) - f^*_k\right\|_2^2\right] = O\left(\left(\frac{n}{\log(n)^4}\right)^{\frac{-2s_k}{2s_k+1}}\right).$$

The novelty of this result is that each emission density is estimated with its own rate of convergence: the rate $\frac{-s_k}{2s_k+1}$ is different for each emission density, even though the original spectral estimators did not handle them separately. This is due to our state-by-state selection method.

Moreover, it is able to reach the minimax rate for each density in an adaptive way. For instance, in the case of a β-Hölder density on $\mathcal{Y} = [0,1]^D$ (equipped with a trigonometric basis), one can easily check the control of η_3, and the control $\|f_k^{*,(M)} - f_k^*\|_2 = O(M^{-\beta/D})$ follows from standard approximation results, see for instance DeVore and Lorentz (1993). Thus, our estimators converge with the rate $(n/\log(n)^4)^{-2\beta/(2\beta+D)}$ to this density: this is the minimax rate up to a logarithmic factor.

Remark 7 *By aligning the estimators like in Section 2.1, one can replace the sequence of permutations in Corollary 6 by a single permutation, in other words there exists a random permutation τ which does not depend on n such that*

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\|(-n^\alpha)\vee(\hat{f}^{sbs}_{\tau(k)}\wedge n^\alpha) - f^*_k\right\|_2^2\right] = O\left(\left(\frac{n}{\log(n)^4}\right)^{\frac{-2s_k}{2s_k+1}}\right).$$

This means that the sequence $(\hat{f}^{sbs}_k)_{n\geq 1}$ is an adaptive rate-minimax estimator of f^*_k–or more precisely of one of the emission densities $(f^*_{k'})_{k'\in\mathcal{X}}$, but since the distribution of the HMM is invariant under relabelling of the hidden states, one can assume the limit to be f^*_k without loss of generality–up to a logarithmic term.

At this point, it is important to note that the choice of the constant $C \geq 2(1+2\alpha)\sqrt{C_\eta C_\sigma}$ depends on the hidden parameters of the HMM and as such is unknown. This penalty calibration problem is very common in the model selection framework and can be solved in practice using methods such as the slope heuristics or the dimension jump method which have been proved to be theoretically valid in several cases, see for instance Baudry et al. (2012) and references therein. We use the dimension jump method and explain its principle and implementation in Section 4.2.

Proof Using Theorem 5, one gets that for all n and for all $M \in \mathcal{M}$ such that $n \geq n_1\eta_3^2(m_M, M)(1+2\alpha)^2\log(n)^2$, with probability $1 - 6n^{-1-2\alpha}$,

$$\begin{aligned}\inf_{\tau\in\mathfrak{S}(K)}\max_{k\in\mathcal{X}}\|\hat{f}_k^{(M,\lceil t\rceil)} - f^{*,(M)}_{\tau(k)}\|_2^2 \leq& C_\sigma\eta_3^2(m_M,M)\frac{(1+2\alpha)^2\log(n)^2}{n}\\ \leq&(1+\alpha)^2C_\sigma C_\eta M\frac{\log(n)^4}{n}\\ \leq&\frac{\sigma(M)^2}{4}\end{aligned}$$

where $\sigma(M) = C\sqrt{\frac{M\log(n)^4}{n}}$ with C such that $C^2 \geq 4(1+2\alpha)^2 C_\sigma C_\eta$.

The condition on M becomes

$$n \geq n_1 \log(n)^4 M(1+2\alpha)^2$$

and is asymptotically true for all $M \leq M_{\max}(n)$ as soon as $M_{\max}(n) = o(n/\log(n)^4)$.

Thus, $[\mathbf{H}(6n^{-(1+2\alpha)})]$ is true for the family $(\hat{\mathbf{f}}^{(M,\lceil(1+2\alpha)\log(n)\rceil)})_{M\leq M_{\max}(n)}$. Note that the assumption $M_{\max}(n) = o(n/\log(n)^4)$ also implies that there exists M_1 such that for n large enough, Lemma 1 holds for all $M \geq M_1$, so that Theorem 2 implies that for n large enough, there exists a permutation τ_n such that with probability $1 - 6n^{-(1+2\alpha)}$, for all $k \in \mathcal{X}$,

$$\begin{aligned}
\|\hat{f}^{\text{sbs}}_{\tau_n(k)} - f^*_k\|_2 &\leq 4 \inf_{M_1\leq M\leq M_{\max}} \{\|f^{*,(M)}_k - f^*_k\|_2 + \sigma(M)\} \\
&= O\left(\inf_{M_1\leq M\leq M_{\max}} \left\{M^{-s_k} + \sqrt{\frac{M\log(n)^4}{n}}\right\}\right) \\
&= O\left(\left(\frac{n}{\log(n)^4}\right)^{-s_k/(1+2s_k)}\right),
\end{aligned}$$

where the tradeoff is reached for $M = (\frac{n}{\log(n)^4})^{1/(1+2s_k)}$, which is in $[M_1, M_{\max}(n)]$ for n large enough.

Finally, write A the event of probability smaller than $6n^{-(1+2\alpha)}$ where $[\mathbf{H}(6n^{-(1+\alpha)})]$ doesn't hold, then for n large enough and for all $k \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}\left[\left\|(-n^\alpha) \vee (\hat{f}^{\text{sbs}}_{\tau_n(k)} \wedge n^\alpha) - f^*_k\right\|_2^2\right] &\leq \mathbb{E}\left[\mathbf{1}_A \left\|\hat{f}^{\text{sbs}}_{\tau_n(k)} - f^*_k\right\|_2^2\right] + \mathbb{E}\left[\mathbf{1}_{A^c}(n^{2\alpha} + \|f^*_k\|_2^2)\right] \\
&= O\left(\left(\frac{n}{\log(n)^4}\right)^{-2s_k/(1+2s_k)}\right) + O\left(\frac{n^{2\alpha} + \|f^*_k\|_2^2}{n^{1+2\alpha}}\right) \\
&= O\left(\left(\frac{n}{\log(n)^4}\right)^{-2s_k/(1+2s_k)}\right).
\end{aligned}$$

■

3.3 The penalized least squares method

Let $\mathcal{F}$ be a subset of $\mathbf{L}^2(\mathcal{Y}, \mu)$. We will need the following assumption on $\mathcal{F}$ in order to control the deviations of the estimators:

[HF] $\mathbf{f}^* \in \mathcal{F}^{K^*}$, $\mathcal{F}$ is closed under projection on $\mathfrak{P}_M$ for all $M \in \mathcal{M}$ and

$$\forall f \in \mathcal{F}, \quad \begin{cases} \|f\|_\infty \leq C_{\mathcal{F},\infty} \\ \|f\|_2 \leq C_{\mathcal{F},2} \end{cases}$$

with $C_{\mathcal{F},\infty}$ and $C_{\mathcal{F},2}$ larger than 1.

A simple way to construct such a set $\mathcal{F}$ when μ is a finite measure is to take the sets $(\mathfrak{P}_M)_M$ spanned by the first M vectors of an orthonormal basis $(\varphi_i)_{i\geq 0}$ whose first vector φ_0 is proportional to $\mathbf{1}$. Then any set $\mathcal{F}$ of densities such that $\int f d\mu = 1$, $\sum_i \langle f, \varphi_i\rangle^2 \leq C_{\mathcal{F},2}$ and $\sum_i |\langle f, \varphi_i\rangle| \|\varphi_i\|_\infty \leq C_{\mathcal{F},\infty}$ for given constants $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ and for all $f \in \mathcal{F}$ satisfies **[HF]**.

When $\mathbf{Q} \in \mathbb{R}^{K\times K}$, $\pi \in \mathbb{R}^K$ and $\mathbf{f} \in (\mathbf{L}^2(\mathcal{Y},\mu))^K$, let

$$g^{\pi,\mathbf{Q},\mathbf{f}}(y_1,y_2,y_3) = \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3) f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3).$$

When π is a probability distribution, $\mathbf{Q}$ a transition matrix and $\mathbf{f}$ a K-uple of probability densities, then $g^{\pi,\mathbf{Q},\mathbf{f}}$ is the density of the first three observations of a HMM with parameters $(\pi,\mathbf{Q},\mathbf{f})$. The motivation behind estimating $g^{\pi,\mathbf{Q},\mathbf{f}}$ is that it allows to recover the true parameters under the identifiability assumption **[Hid]**, as shown in the following theorem.

Let $\mathcal{Q}$ be the set of transition matrices on $\mathcal{X}$ and Δ the set of probability distributions on $\mathcal{X}$. For a permutation $\tau \in \mathfrak{S}(K)$, write $\mathbb{P}_\tau$ its matrix (that is the matrix defined by $\mathbb{P}_\tau(i,j) = \mathbf{1}_{\{j=\tau(i)\}}$). Finally, define the distance on the HMM parameters

$$\begin{aligned} &d_{\mathrm{perm}}((\pi_1,\mathbf{Q}_1,\mathbf{f}_1),(\pi_2,\mathbf{Q}_2,\mathbf{f}_2))^2 \\ &\qquad = \inf_{\tau\in\mathfrak{S}(K)} \left\{ \|\pi_1 - \mathbb{P}_\tau \pi_2\|_2^2 + \|\mathbf{Q}_1 - \mathbb{P}_\tau \mathbf{Q}_2 \mathbb{P}_\tau^\top\|_F^2 + \sum_{k\in\mathcal{X}} \|f_{1,k} - f_{2,\tau(k)}\|_2^2 \right\}. \end{aligned}$$

This distance is invariant under permutation of the hidden states. This corresponds to the fact that a HMM is only identifiable up to relabelling of its hidden states.

Theorem 8 (Identifiability) *Let $(\pi^*,\mathbf{Q}^*,\mathbf{f}^*) \in \Delta\times\mathcal{Q}\times(\mathbf{L}^2(\mathcal{Y},\mu))^K$ such that $\pi^*_x > 0$ for all $x\in\mathcal{X}$ and **[Hid]** holds. Then for all $(\pi,\mathbf{Q},\mathbf{f}) \in \Delta\times\mathcal{Q}\times(\mathbf{L}^2(\mathcal{Y},\mu))^K$,*

$$(g^{\pi,\mathbf{Q},\mathbf{f}} = g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}) \;\Rightarrow\; d_{perm}((\pi,\mathbf{Q},\mathbf{f}),(\pi^*,\mathbf{Q}^*,\mathbf{f}^*)) = 0.$$

Proof The spectral algorithm of De Castro et al. (2017) applied on the finite dimensional space spanned by the components of $\mathbf{f}$ and $\mathbf{f}^*$ allows to recover all the parameters even when the emission densities are not probability densities and when the Markov chain is not stationary. ∎

Define the empirical contrast

$$\gamma_n(t) = \|t\|_2^2 - \frac{2}{n}\sum_{j=1}^{n} t(Z_j)$$

where $Z_j := (Y_j, Y_{j+1}, Y_{j+2})$ and $(Y_j)_{1\leq j\leq n+2}$ are the observations. It is a biased estimator of the $\mathbf{L}^2$ loss: for all $t \in (\mathbf{L}^2(\mathcal{Y},\mu))^3$,

$$\mathbb{E}[\gamma_n(t)] = \|t - g^*\|_2^2 - \|g^*\|_2^2$$

Algorithm 2: Least squares estimation of the emission densities of a HMM

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, a dimension M and an orthonormal basis $\Phi = (\varphi_1, \dots, \varphi_M)$.

Result: Least squares estimators $\hat{\pi}^{(M)}$, $\hat{\mathbf{Q}}^{(M)}$ and $(\hat{f}_k^{(M)})_{k\in\mathcal{X}}$.

[Step 1] Compute the tensor $\hat{\mathbf{M}}_M$ defined by $\hat{\mathbf{M}}_M(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$ for all $a, b, c \in [M]$.

[Step 2] Find a minimizer $(\hat{\pi}^{(M)}, \hat{\mathbf{Q}}^{(M)}, \hat{\mathbf{O}})$ of $(\pi, \mathbf{Q}, \mathbf{O}) \longmapsto \|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} - \hat{\mathbf{M}}_M\|_F^2$ where

- $\pi \in \mathbb{R}^K$ is a probability distribution on $\mathcal{X}$, i.e. $\sum_{k\in\mathcal{X}} \pi_k = 1$;
- $\mathbf{Q} \in \mathbb{R}^{K\times K}$ is a transition matrix on $\mathcal{X}$, i.e. $\sum_{k'\in\mathcal{X}} Q(k,k') = 1$ for all $k \in \mathcal{X}$;
- $\mathbf{O}$ is a $M \times K$ matrix such that for all $k \in \mathcal{X}$, $\sum_{b=1}^{M} \mathbf{O}(b,k)\varphi_b \in \mathcal{F}$;
- $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} \in \mathbb{R}^{M\times M\times M}$ is defined by $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})}(\cdot, b, \cdot) = \mathbf{O}\mathrm{Diag}[\pi]\mathbf{Q}\mathrm{Diag}[\mathbf{O}(b,\cdot)]\mathbf{Q}\mathbf{O}^\top$ for all $b \in [M]$.

[Step 3] Consider the emission densities estimators $\hat{\mathbf{f}}^{(M)} := (\hat{f}_k^{(M)})_{k\in\mathcal{X}}$ defined by for all $k \in \mathcal{X}$, $\hat{f}_k^{(M)} := \sum_{b=1}^{M} \hat{\mathbf{O}}(b,k)\varphi_b$.

where $g^* = g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}$. Since the bias does not depend on the function t, one can hope that the minimizers of γ_n are close to minimizers of $\|t - g^*\|_2$. We will show that this is indeed the case.

The least squares estimators of all HMM parameters are defined for each model $\mathfrak{P}_M$ by

$$(\hat{\pi}^{(M)}, \hat{\mathbf{Q}}^{(M)}, \hat{\mathbf{f}}^{(M)}) \in \underset{\pi\in\Delta,\, \mathbf{Q}\in\mathcal{Q},\, \mathbf{f}\in(\mathfrak{P}_M\cap\mathcal{F})^K}{\arg\min} \gamma_n(g^{\pi,\mathbf{Q},\mathbf{f}}).$$

The procedure is summarized in Algorithm 2. Note that with the notations of the algorithm,

$$\gamma_n(g^{\pi,\mathbf{Q},\mathbf{O}^\top\Phi}) = \|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} - \hat{\mathbf{M}}_M\|_F^2 - \|\hat{\mathbf{M}}_M\|_F^2.$$

Then, the proof of the oracle inequality of Lehéricy (to appear) allows to get the following result.

Theorem 9 *Assume **[HF]**, **[HX]** and **[Hid]** hold.*

Then there exists constants C and n_0 depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$ and $\mathbf{Q}^*$ such that for all $n \geq n_0$, for all $t > 0$, with probability greater than $1 - e^{-t}$, one has for all $M \in \mathcal{M}$ such that $M \leq n$:

$$\|\hat{g}^{\hat{\pi}^{(M)},\hat{\mathbf{Q}}^{(M)},\hat{\mathbf{f}}^{(M)}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^{*,(M)}}\|_2^2 \leq C\Big(\frac{t}{n} + M\frac{\log(n)}{n}\Big).$$

In order to deduce a control of the error on the parameters–and in particular on the emission densities–from the previous result, we will need to assume that the quadratic form derived from the second-order expansion of $(\pi, \mathbf{Q}, \mathbf{f}) \in \Delta \times \mathcal{Q} \times \mathcal{F}^K \longmapsto \|g^{\pi,\mathbf{Q},\mathbf{f}} - g^*\|_2^2$ around $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ is nondegenerate.

It is still unknown whether this nondegeneracy property is true for all parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ such that **[Hid]** and **[HX]** hold. De Castro et al. (2016) prove it for $K = 2$ hidden states when only the emission densities are allowed to vary by using brute-force computations. To do so, they introduce an (explicit) polynomial in the coefficients of π^*, $\mathbf{Q}^*$ and of the Gram matrix of $\mathbf{f}^*$ and prove that its value is nonzero if and only if the quadratic form is nondegenerate for the corresponding parameters. The difficult part of the proof is to show that this polynomial is always nonzero.

For the expression of this polynomial–which we will write H–in our setting, we refer to Section B.3. Note that Lehéricy (to appear) proves that this polynomial H is non identically zero: it is shown that there exists parameters $(\pi, \mathbf{Q}, \mathbf{f})$ satisfying **[HX]** and **[Hid]** such that $H(\pi, \mathbf{Q}, \mathbf{f}) \neq 0$, which means that the following assumption is generically satisfied:

[Hdet] $H(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \neq 0$.

The following result allows to lower bound the $\mathbf{L}^2$ error on the density of three consecutive observations by the error on the parameters of the HMM using this condition. It is an improvement of Theorem 6 of De Castro et al. (2016) and Theorem 9 of Lehéricy (to appear). The main difference is that the constant $c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F})$ does not depend on the $\mathbf{f}$ around which the parameters are taken. This is crucial to obtain Corollary 11, from which we will deduce **[H0]**. Note that we do not need $\mathbf{f}$ to be in a compact neighborhood of $\mathbf{f}^*$. Another improvement is that the constant in the minoration only depends on the true parameters and on the set $\mathcal{F}$.

Theorem 10 *1. Assume that **[HF]** holds and that for all $f \in \mathcal{F}$, $\int f d\mu = 1$.*

*Then there exist a lower semicontinuous function $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \longmapsto c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F})$ that is positive when **[Hid]** and **[Hdet]** hold and a neighborhood $\mathcal{V}$ of $\mathbf{f}^*$ in $\mathcal{F}^K$ depending only on π^*, $\mathbf{Q}^*$, $\mathbf{f}^*$ and $\mathcal{F}$ such that for all $\mathbf{f} \in \mathcal{V}$ and for all $\pi \in \Delta$, $\mathbf{Q} \in \mathcal{Q}$ and $\mathbf{h} \in \mathcal{F}^K$,*

$$\|g^{\pi,\mathbf{Q},\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2 \geq c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F}) d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}))^2.$$

*2. There exists a continuous function $\epsilon : (\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \mapsto \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ that is positive when **[Hid]** and **[Hdet]** hold and such that for all $\pi \in \Delta$, $\mathbf{Q} \in \mathcal{Q}$ and $\mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$ a K-uple of probability densities such that $d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)) \leq \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$, one has*

$$\|g^{\pi,\mathbf{Q},\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \geq c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*))^2.$$

where

$$c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) = \frac{(\inf_{k \in \mathcal{X}} \pi^*(k))\, \sigma_K(\mathbf{Q}^*)^4 \sigma_K(G(\mathbf{f}^*))^2}{4} \wedge \frac{H(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)}{2(1 \wedge K\|G(\mathbf{f}^*)\|_\infty)(3K^3(1 \vee \|G(\mathbf{f}^*)\|_\infty^4))^{K^2 - K/2}}.$$

Proof Proof in Section B.4. ∎

Corollary 11 *Assume **[HX]**, **[HF]**, **[Hid]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$.*

Then there exists a constant n_0 depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$ and $\mathbf{Q}^*$ and constants M_0 and C' depending on $\mathcal{F}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ such that for all $n \geq n_0$ and $t > 0$, with probability greater than $1 - e^{-t}$, one has for all $M \in \mathcal{M}$ such that $M_0 \leq M \leq n$:

$$\inf_{\tau \in \mathfrak{S}(K)} \max_{k \in \mathcal{X}} \|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\|_2^2 \leq C' \left(M \frac{\log(n)}{n} + \frac{t}{n} \right).$$

Remark 12 *Using the second point of Theorem 10, one can alternatively take n_0 and M_0 depending on $\mathcal{F}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$, and C' depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ only. For instance, one can take $C' = C/c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ with the notations of Theorems 9 and 10.*

In particular, this means that the asymptotic variance bound of the least squares estimators (and therefore the rate of convergence of the estimators selected by our state-by-state selection method) does not depend on the set $\mathcal{F}$, but only on the HMM parameters and on the bounds $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ on the square and supremum norms of the emission densities. Note that this universality result is essentially an asymptotic one since it requires n_0 to depend on $\mathcal{F}$ in a non-explicit way.

Proof Let $\mathcal{V}$ be the neighborhood given by Theorem 10, then there exists M_0 such that for all $M \geq M_0$, $\mathbf{f}^{*,(M)} \in \mathcal{V}$. Then Theorem 9 and Theorem 10 applied to $\pi = \hat{\pi}^{(M)}$, $\mathbf{Q} = \hat{\mathbf{Q}}^{(M)}$, $\mathbf{h} = \hat{\mathbf{f}}^{(M)}$ and $\mathbf{f} = \mathbf{f}^{*,(M)}$ for all M allow to conclude. ■

We may now state the following result which shows that the state-by-state selection method applied to these estimators reaches the minimax rate of convergence (up to a logarithmic factor) in an adaptive manner under generic assumptions. Its proof is the same as the one of Corollary 6.

Corollary 13 *Assume **[HX]**, **[HF]**, **[Hid]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$ and that for all k, there exists s_k such that $\|f_k^{*,(M)} - f_k^*\|_2 = O(M^{-s_k})$. Then there exists a constant C_σ depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ such that the following holds.*

Let $C \geq C_\sigma$ and let $\hat{\mathbf{f}}^{sbs}$ be the estimators selected from the family $(\hat{\mathbf{f}}^{(M)})_{M \leq n}$ with $\sigma(M) = C\sqrt{\frac{M \log(n)}{n}}$ for all M, aligned like in Remark 7. Then there exists a random permutation τ which does not depend on n such that

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\| \hat{f}_{\tau(k)}^{sbs} - f_k^* \right\|_2 \right] = O\left(\left(\frac{n}{\log(n)} \right)^{\frac{-s_k}{2s_k+1}} \right).$$

4. Numerical experiments

This section is dedicated to the discussion of the practical implementation of our method. We run the spectral estimators on simulated data for different number of observations and study the rate of convergence of the selected estimators for several variants of our method.

Finally, we discuss the algorithmic complexity of the different estimators and selection methods.

In Section 4.1, we introduce the parameters with which we generate the observations. In Section 4.2, we discuss how to calibrate the constant of the penalty in practice. In Section 4.3, we introduce two other ways to select the final estimators, the POS and MAX variants. Section 4.4 contains the results of the simulations for each variant and calibration method. In Section 4.5, we present a cross validation procedure and compare its results with the one obtained using our method. Finally, we discuss the algorithmic complexity of the different algorithms and estimators in Section 4.6.

4.1 Setting and parameters

We take $\mathcal{Y} = [0,1]$ equipped with the Lebesgue measure. We choose the approximation spaces spanned by a trigonometric basis: $\mathfrak{P}_M := \mathrm{Span}(\varphi_1, \dots, \varphi_M)$ with

$$\begin{cases} \varphi_1(x) & = 1 \\ \varphi_{2m}(x) & = \sqrt{2}\cos(2\pi m x) \\ \varphi_{2m+1}(x) & = \sqrt{2}\sin(2\pi m x) \end{cases}$$

for all $x \in [0,1]$ and $m \in \mathbb{N}^*$. We will consider a hidden Markov model with $K = 3$ hidden states and the following parameters:

- Transition matrix

$$\mathbf{Q}^* = \begin{pmatrix} 0.7 & 0.1 & 0.2 \\ 0.08 & 0.8 & 0.12 \\ 0.15 & 0.15 & 0.7 \end{pmatrix};$$

- Emission densities (see Figure 1)
  - Uniform distribution on $[0;1]$;
  - Symmetrized Beta distribution, that is a mixture with the same weight of $\frac{2}{3}X$ and $1 - \frac{1}{3}X'$ with X, X' i.i.d. following a Beta distribution with parameters $(3, 1.6)$;
  - Beta distribution with parameters $(3, 7)$.

We generate n observations and run the spectral algorithm in order to obtain estimators for the models $\mathfrak{P}_M$ with $\texttt{M\_min} \leq M \leq \texttt{M\_max}$, $m = 20$ and $r = \lceil 2\log(n) + 2\log(M) \rceil$, where $\texttt{M\_min} = 3$ and $\texttt{M\_max} = 300$. Finally, we use the state-by-state selection method to choose the final estimator for each emission density. The main reason for using spectral estimators instead of maximum likelihood estimation or least squares estimation is its computational speed: it is much faster for large n than the least squares algorithm or the EM algorithm, which makes studying asymptotic behaviours possible.

We made 300 simulations, 20 per value of n, with n taking values in $\{5\times10^4, 7\times10^4, 1\times10^5, 1.5\times10^5, 2.2\times10^5, 3.5\times10^5, 5\times10^5, 7\times10^5, 1\times10^6, 1.5\times10^6, 2.2\times10^6, 3.5\times10^6, 5\times10^6, 7\times10^6, 1\times10^7\}$.

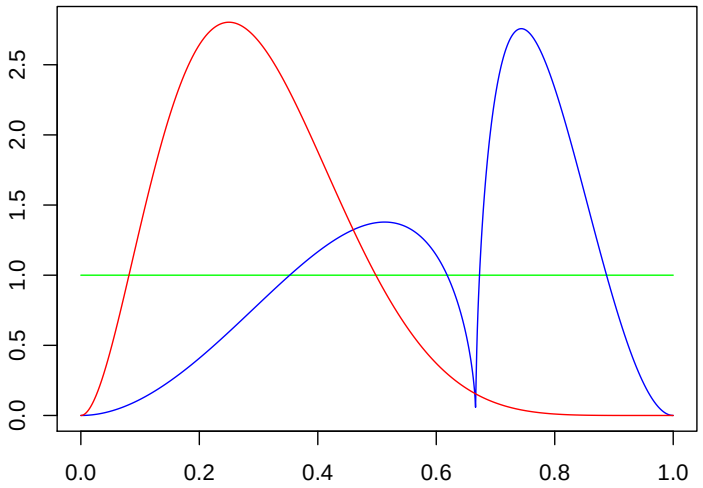


Figure 1: Emission densities. In all following figures, the uniform distribution corresponds to the green lines, the Beta distribution to the red lines and the symmetrized Beta distribution to the blue lines.

4.2 Penalty calibration

It is important to note that when considering spectral and least squares methods, the penalty σ in the state-by-state selection procedure depends on the hidden parameters of the HMM and as such is unknown in practice. This penalty calibration problem is well known and several procedures exist that allow to solve it, for instance the slope heuristics and the dimension jump method (see for instance Baudry et al. (2012) and references therein). In the following, we will use the dimension jump method to calibrate the penalty in the state-by-state selection procedure.

Consider a penalty shape $\mathrm{pen}_{\mathrm{shape}}$ and define $\hat{M}_k(\rho)$ the model selected for the hidden state k by the state-by-state selection estimator using the penalty $\rho\,\mathrm{pen}_{\mathrm{shape}}$:

$$\hat{M}_k(\rho) \in \operatorname*{arg\,min}_{M\in\mathcal{M}}\{A_k(M) + 2\rho\,\mathrm{pen}_{\mathrm{shape}}(M)\}.$$

where

$$A_k(M) = \sup_{M'\in\mathcal{M}}\left\{\left\|\hat{f}_k^{(M')} - \hat{f}_k^{(M\wedge M')}\right\|_2 - \rho\,\mathrm{pen}_{\mathrm{shape}}(M')\right\}.$$

The dimension jump method relies on the heuristics that there exists a constant C such that $C\,\mathrm{pen}_{\mathrm{shape}}$ is a *minimal penalty*. This means that for all $\rho < C$, the selected models $\hat{M}_k(\rho)$ will be very large, while for $\rho > C$, the models will remain small. This translates into a sharp jump located around a value $\rho_{\mathrm{jump},k} = C$ in the plot of $\rho \longmapsto \hat{M}_k(\rho)$. The final step consists in taking twice this value to calibrate the constant of the penalty, thus

selecting the model $\hat{M}(2\rho_{\text{jump},k})$. In practice, we take $\rho_{\text{jump},k}$ as the position of the largest jump of the function $\rho \longmapsto \hat{M}_k(\rho)$.

Figure 2 shows the resulting dimension jumps for $n = 220,000$ observations. Each curve corresponds to one of the $\hat{M}_k(\rho)$ and has a clear dimension jump, which confirms the relevance of the heuristics. Several methods may be used to calibrate the constant of the penalty:

eachjump. Calibrate the constant independently for each state. This method has the advantage of being easy to calibrate since there is usually a single sharp jump in each state's complexity. However, our theoretical results do not suggest that the penalty constant is different for each state;

jumpmax. Calibrate the constant for all states together using only the latest jump. This consists in taking the maximum of the $\rho_{\text{jump},k}$ to select the final models. Since the penalty is known up to a multiplicative constant and taking a constant larger than needed does not affect the rates of convergence–contrary to smaller constants–this is the "safe" option;

jumpmean. Calibrate the constant for all states together using the mean of the positions of the different jumps.

We try and compare these calibration methods in Section 4.4.

4.3 Alternative selection procedures

4.3.1 Variant POS.

As mentionned in Section 2.2, it is also possible to select the estimators using the criterion

$$A_k(M) = \sup_{M' \geq M} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M)} \right\|_2 - \sigma(M') \right\}_+$$

followed by

$$\hat{M}_k \in \arg\min_{M \in \mathcal{M}} \{A_k(M) + 2\sigma(M)\}.$$

This positivity condition was in the original Goldenshluger-Lepski method. The theoretical guarantees remain the same as the previous method and both behave almost identically in practice, as shown in Section 4.4.

4.3.2 Variant MAX.

In the context of kernel density estimation, Lacour et al. (2016) show that the Goldenshluger-Lepski method still works when the biais estimate $A_k(M)$ of the model M is replaced by the distance between the estimator of the model M and the estimator with the smallest bandwidth (the analog of the largest model in our setting). They also prove an oracle inequality for this method after adding a corrective term to the penalty.

The following variant is based on the same idea. It consists in selecting the model

$$\hat{M}_k \in \arg\min_{M \in \mathcal{M}} \{\| \hat{f}_k^{(M_{\max})} - \hat{f}_k^{(M)} \|_2 + \sigma(M)\}$$

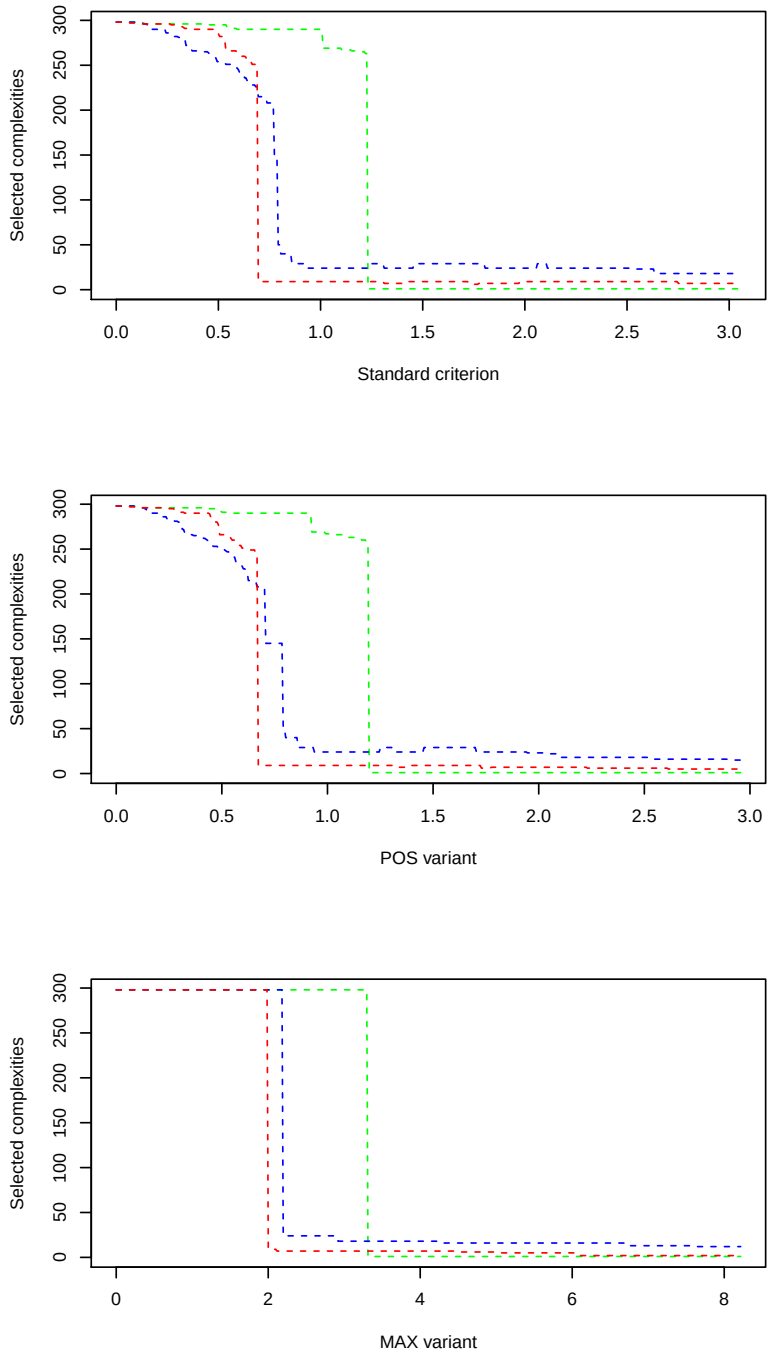


Figure 2: Selected complexities with respect to the penalty constant ρ for the same simulation of $n = 500,000$ observations. The colored dashed lines correspond to the single-state complexities $M_k(\rho)$.

for each $k \in \mathcal{X}$ and takes

$$\hat{f}_k = \hat{f}_k^{(\hat{M}_k)},$$

where σ is the same penalty as the one in the usual state-by-state selection method.

An advantage of this algorithm is its lower complexity, since it requires $O(M_{\max})$ computations of $\mathbf{L}^2$ norms instead of $O(M_{\max}^2)$. We do not study this method theoretically in our setting. However, the simulations (and in particular Figure 4) show that it behaves similarly to the standard state-by-state selection method in the asymptotic regime and even has a smaller error for small number of observations. In addition, the dimension jumps are much sharper for this method than for the usual state-by-state selection method (see Figure 2), which makes the calibration heuristics easier to use.

4.4 Results

Figure 3 shows the evolution of the error $\|\hat{f}_k - f_k^*\|_2$ for each state k with respect to the number of observations n, for all penalty calibration methods and all variants of the model selection procedure. Figure 4 compares the evolution of the median error for the different calibration methods and for the different selection variants, and Figure 5 compares two estimators with the oracle estimators.

When the number of observations n is large enough, the logarithm of the error decreases linearly with respect to $\log(n)$. This corresponds to the asymptotic convergence regime: the error is expected to decrease as a power of the number of observations n when n tends to infinity. The corresponding slopes are listed in Table 1.

For each state, the confidence intervals of the rates of all estimators–including the oracle estimators–have a common intersection (except for the symmetrized Beta distribution in the jumpmax MAX variant, whose estimators seem to converge faster than the others). This tends to confirm that the calibration and selection variants are asymptotically equivalent. This phenomenon is also visible in Figures 3 and 4: in the asymptotic regime, the errors decrease in a similar way for all methods.

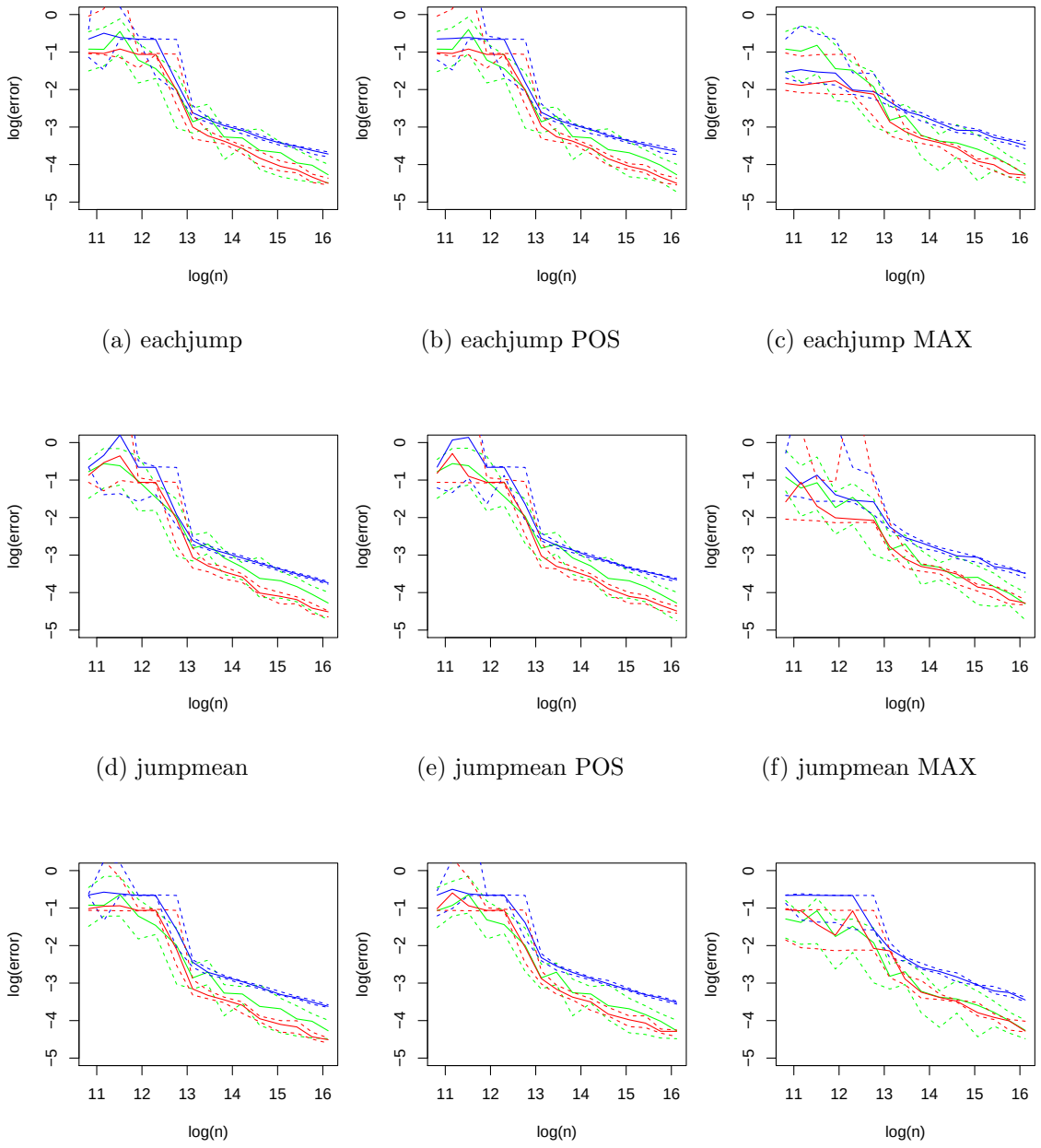
Furthermore, the rates of convergence are clearly distinct. The uniform distribution is estimated with a rate of convergence of approximately $n^{-1/2}$, which is also the best possible rate (it corresponds to a parametric estimation rate). In comparison, the rate of convergence for the symmetrized Beta distribution is much slower (around $n^{-0.36}$). This shows that the algorithm effectively adapts to the regularity of each state and that one irregular emission density does not deteriorate the rates of convergence of the other densities.

Note that the above rates are in accordance with the minimax rates as far as the Hölder regularity is concerned. The minimax Hölder rate for the symmetrized Beta (which is 0.6-Hölder) is $n^{-3/11}$, or approximately $n^{-0.27}$, which means our estimator converges faster than the minimax rate would suggest. The minimax Hölder rate for the Beta distribution (which is 3-Hölder) is $n^{-3/7}$, or approximately $n^{-0.43}$, which is around the observed value.

4.5 Comparison with cross validation

In this section, we use a cross validation procedure based on our spectral estimators to check whether our method actually improves estimation accuracy.œ

When estimating a density by taking an estimator within some class (the model), two sources of error appear: the bias, that is the (deterministic) distance between the true

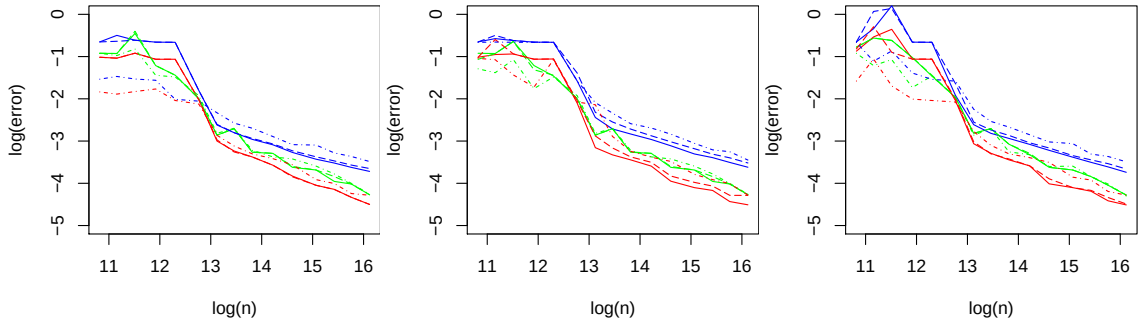


(a) eachjump (b) eachjump POS (c) eachjump MAX

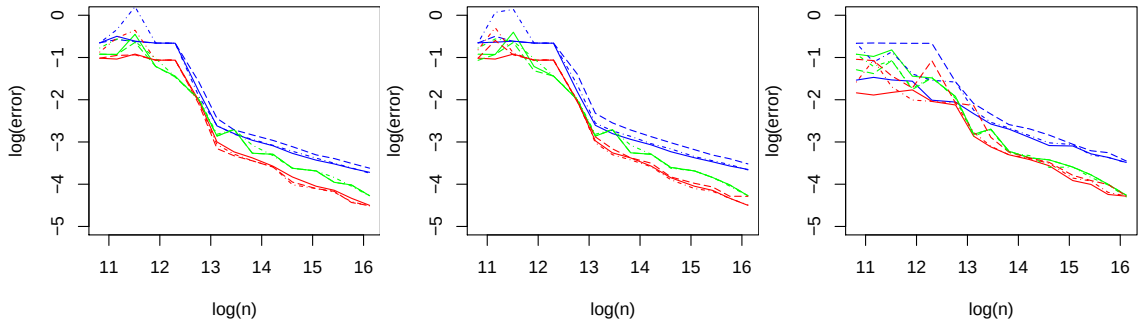
(d) jumpmean (e) jumpmean POS (f) jumpmean MAX

(g) jumpmax (h) jumpmax POS (i) jumpmax MAX

Figure 3: Logarithm of the $\mathbf{L}^2$ error on each emission densities depending on the logarithm of the number of observations for each of the selection and calibration methods. Each color corresponds to one emission density. The full lines are the medians of the 20 observations and the dashed ones are the 25 and 75 percentiles.



(a) eachjump summary (b) jumpmax summary (c) jumpmean summary



(d) none summary (e) POS summary (f) MAX summary

Figure 4: Superposition of the median lines of Figure 3 by selection method and by calibration variant. Each color corresponds to one emission density. In Subfigures (a)-(c), the full lines correspond to the basic selection method, the dashed ones to the POS method and the dotted ones to the MAX method. In Subfigures (d)-(f), the full lines correspond to the eachjump method, the dashed ones to the jumpmax method and the dotted ones to the jumpmean method.

density and the model, and the variance, that is the (random) error of the estimation within the model. Small models will have a large bias but a small variance, while large models will have a small bias and a large variance. The core issue of model selection is to select a model that minimizes the total error, that is large enough to accurately describe the true densities and small enough to prevent overfitting: in other words, perform a bias-variance tradeoff.

Cross validation seeks to achieve such a tradeoff by computing an estimate of the total error. This is done by splitting the sample into two sets, the training sample being used for the calibration of the estimator and the validation sample for measuring the error.

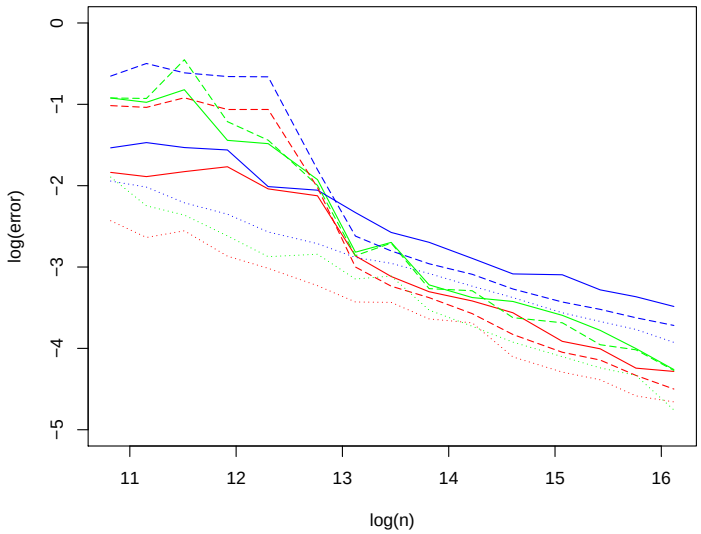


Figure 5: Comparison of the errors for the eachjump MAX method (full lines), for the eachjump method (dashed lines) and for the oracle estimators (dotted lines). For each k, the oracle estimator $\hat{f}_k^{\text{oracle}}$ is defined as the $\hat{f}_k^{(M)}$ which minimizes $\|\hat{f}_k^{(M)} - f_k^*\|_2$.

Taking the mean of these errors for different splits between training and validation samples provides an estimator of the total error. This method has become popular for its simplicity of use. We refer to the survey of Arlot et al. (2010) for an overview on this method and its guarantees.

4.5.1 Risk

We use the least squares criterion of Algorithm 2 to quantify the error of the estimators. Since the guarantees on spectral estimators rely on the $\mathbf{L}^2$ norm, a least squares criterion is more natural than the likelihood. In addition, the spectral estimators might take negative values depending on the orthonormal basis, which is not a problem as far as $\mathbf{L}^2$ error is concerned but can be an issue for the likelihood.

Let us first recall this criterion. Given an orthonormal basis $(\varphi_i)_{i\in\mathbb{N}}$ of $\mathbf{L}^2(\mathcal{Y},\mu)$, define the coordinate tensor of the empirical distribution of the triplet (Y_1,Y_2,Y_3) on this basis by

$$\hat{\mathbf{M}}(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}) \quad \text{for all } a,b,c \in \mathbb{N}.$$

Given a transition matrix $\mathbf{Q}$ of size K, a stationary distribution π of $\mathbf{Q}$ and a vector of densities $\mathbf{f} = (f_1,\dots,f_K)$, define the coordinate matrix $\mathbf{O}$ of $\mathbf{f}$ by $\mathbf{O}(b,k) = \langle \varphi_b, f_k\rangle$. Let $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})}$ be the coordinate tensor of the distribution of (Y_1,Y_2,Y_3) under the parameters $(\pi,\mathbf{Q},\mathbf{f})$, that is

$$\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})}(\cdot,b,\cdot) = \mathbf{O}\text{Diag}[\pi]\mathbf{Q}\text{Diag}[\mathbf{O}(b,\cdot)]\mathbf{Q}\mathbf{O}^\top \quad \text{for all } b \in \mathbb{N}.$$

Estimator	Convergence rate exponents		
	Uniform	Sym. Beta	Beta
Eachjump	-0.500 ± 0.046	-0.347 ± 0.007	-0.470 ± 0.015
Eachjump POS	-0.503 ± 0.047	-0.327 ± 0.008	-0.469 ± 0.015
Eachjump MAX	-0.480 ± 0.052	-0.335 ± 0.009	-0.449 ± 0.015
Jumpmean	-0.532 ± 0.048	-0.349 ± 0.006	-0.471 ± 0.017
Jumpmean POS	-0.540 ± 0.048	-0.350 ± 0.006	-0.456 ± 0.016
Jumpmean MAX	-0.493 ± 0.049	-0.374 ± 0.009	-0.437 ± 0.015
Jumpmax	-0.500 ± 0.046	-0.349 ± 0.006	-0.464 ± 0.016
Jumpmax POS	-0.492 ± 0.046	-0.358 ± 0.006	-0.442 ± 0.015
Jumpmax MAX	-0.480 ± 0.052	-0.404 ± 0.009	-0.466 ± 0.015
Cross Validation	-0.434 ± 0.007	-0.263 ± 0.011	-0.377 ± 0.008
Oracle	-0.517 ± 0.048	-0.360 ± 0.006	-0.459 ± 0.017
Minimax (Hölder)	-0.5	$-3/11 \approx -0.273$	$-3/7 \approx -0.429$

Table 1: Exponents of the rates of convergence for the different algorithms. The rates are obtained from a linear regression with the relation $\log(\|\hat{f}_k - f_k^*\|_2) \sim \log(n)$ for the estimators $\hat{f}_k$ computed with $n \geq 700,000$ observations ($n \geq 1,000,000$ for the cross validation estimators from Section 4.5). The smaller the exponent, the faster the estimators converge.

The empirical least squares criterion is $\|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})} - \hat{\mathbf{M}}\|_F^2$. It corresponds to the $\mathbf{L}^2$ error between the empirical distribution of three consecutive observations and the theoretical distribution under the parameters $(\pi, \mathbf{Q}, \mathbf{f})$.

4.5.2 Implementation

We use 10-fold cross validation, that is we split the sequence into 10 segments of same size $I_1, \dots, I_{10}$. In order to avoid interferences between samples, we prune the ends of each segment, so that the observations in each segment can be considered independent. In practice, we take a gap of 30 observations between two segments.

We ran 150 simulations, 10 per value of n, with the same parameters as in Section 4.1. Each simulation is as follows.

For each segment I_j, we run the spectral algorithm on all models $\mathfrak{P}_M$ for `M_min` $\leq M \leq$ `M_max` using only the observations from the other segments. The transition matrix is estimated using an additional step of the spectral method which is adapted from Steps 8 and 9 of Algorithm 1 of De Castro et al. (2017). Then, we compute the least squares criterion for the estimated parameters using the segment I_j as observed sample. Finally,

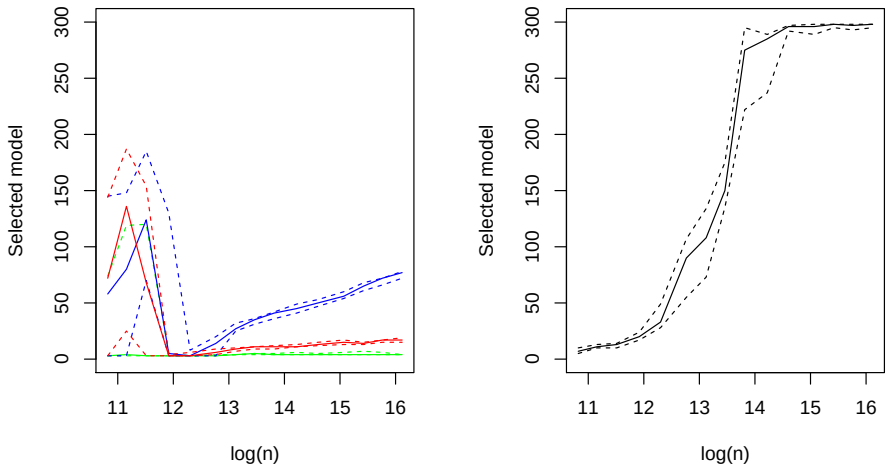


Figure 6: Selected model dimensions for each n using our state-by-state selection method (left) and 10-fold cross validation (right). The full lines are the median model dimensions and the dashed lines are the 25 and 75 percentiles.

for each M, we average this error on all segments I_j, which gives the least squares cross validation error $E_{\mathrm{VC}}(M)$.

This cross validation criterion is used to select one model $\hat{M}_{\mathrm{VC}} \in \arg\min_M E_{\mathrm{VC}}(M)$, from which we construct the final estimators of the emission densities $\hat{f}_k = \hat{f}_k^{(\hat{M}_{\mathrm{VC}})}$ for all k. Note that the selected model is the same for all emission densities.

4.5.3 Results

Figure 6 compares the selected model dimensions for each n using our state-by-state selection method and using the cross validation method. When the number of observations n becomes larger than 10^6, the cross validation tends to always pick the largest model, which means that it does not prevent overfitting as well as our method.

The $\mathbf{L}^2$ errors on the emission densities are shown in Figure 7. It appears that the cross validation has a lower error for small n ($n \leq 350,000$) than our method. However, for larger values of n, the errors becomes larger than the ones of our method (see Figure 5) by up to one order of magnitude, and only start decreasing once the selected model is set to the maximum dimension.

Finally, the estimated rates of convergence are shown in Table 1. Our state-by-state method outperforms the cross validation method for all emission densities. The cross validation estimators only reach the minimax rate of convergence for the less regular density: the symmetrized Beta, and even then they converge slower than the state-by-state estimator. All other emission densities are estimated slower than their minimax rate.

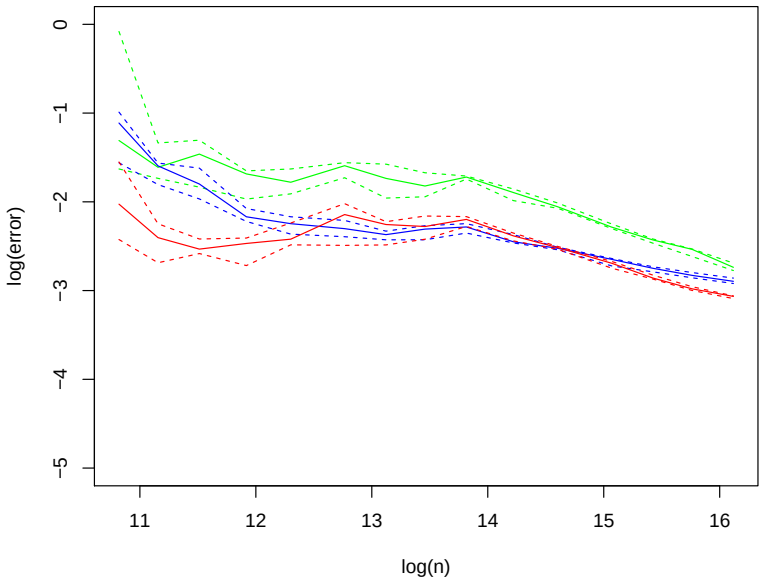


Figure 7: Error of the cross validation estimators for each n using 10-fold cross validation. The full lines are the median errors for each density and the dashed lines are the 25 and 75 percentiles.

4.6 Algorithmic complexity

In the following, we treat K as a constant as far as the algorithmic complexity is concerned. The different complexities are summarized in Table 2.

4.6.1 Spectral algorithm (see Section 3.2)

We consider the algorithmic complexity of estimating the emission densities for all models M such that $M_{\min} \leq M \leq M_{\max}$ with n observations and auxiliary parameters r and m depending on n and M (upper bounded by $m_{\max}$ and $r_{\max}$).

Step 1 can be computed for all models with $O(nM_{\max}m_{\max}^2)$ operations. It is the only step whose complexity depends on n. Steps 2 and 3 require $O(m^3M)$ operations for each model and Steps 4 to 7 require $O(Mr)$ operations for each model, for a total of $O(nM_{\max}m_{\max}^2 + M_{\max}^2m_{\max}^3 + M_{\max}^2r_{\max})$ operations.

In practice, one takes $m \propto \log(n)$, $r \propto \log(n) + \log(M)$ and $M_{\max} \leq n$, so that the total complexity of the spectral algorithm is $O(n\log(n)^2M_{\max})$.

In comparison, the complexity of the spectral algorithm of De Castro et al. (2017) is $O(nM_{\max}^3)$ because of Step 1. This becomes much larger than our complexity when $M_{\max}$ grows as a power of n (which is necessary in order to reach minimax rates).

4.6.2 Least squares algorithm (see Section 3.3)

We consider the algorithmic complexity of estimating the emission densities for all models M such that $M_{\min} \leq M \leq M_{\max}$ with n observations.

Step 1 is similar to the one of the spectral algorithm, but with $O(nM_{\max}^3)$ operations. The complexity of Step 2 is more difficult to evaluate. Since the criterion is nonconvex, finding the minimizer requires to run an approximate minimization algorithm whose complexity C_n will depend on the desired precision–which will in turn depend on the number of observations n–and on the initial points. As discussed in Lehéricy (to appear), this is usually the longest step when computing least squares estimators. Thus, the total complexity of the least squares algorithm is $O(nM_{\max}^3 + C_n)$.

Note that despite the worse sample complexity, the least squares algorithm is tractable and can greatly improve the estimation for small sample size. As shown in Section 4.4, the spectral algorithm is unstable for small samples, which makes the state-by-state selection procedure return abnormal results. This can be explained by the matrix inversions of the spectral method, which sometimes lead to nearly singular matrices when the noise is too large. On the other hand, the least squares method does not involve any matrix inversion, and often gives better results than the spectral estimators, as shown in De Castro et al. (2016), thus making it a relevant choice for small to medium data sets.

4.6.3 Selection method and POS variant (see Sections 2.2 and 4.3)

We consider the algorithmic complexity of selecting estimators from a family $(\hat{\mathbf{f}}^{(M)})_{M_{\min}\leq M\leq M_{\max}}$ of estimators. The selection algorithms can be decomposed in two parts.

- Compute the distances $\|\hat{f}_k^{(M)} - \hat{f}_k^{(M')}\|_2$ for all M, M' and k. This has complexity $O(M_{\max}^3)$: it requires to compute the $\mathbf{L}^2$ distance of at most $M_{\max}^2$ couples of functions in a Hilbert space of dimension $M_{\max}$.

- Compute $\hat{\rho}_k$ defined as the abscissa of the largest jump of the function $\rho \longmapsto \hat{M}_k(\rho)$ for all k, where $\hat{M}_k$ is defined as in Section 4.2. Note that computing $\hat{M}_k(\rho)$ requires $O(M_{\max}^2)$ operations. An approximate value of $\hat{\rho}_k$ can be computed in $O(\log(\hat{\rho}_k)M_{\max}^2)$ operations, which is usually $O(M_{\max}^2)$.

Once the $\hat{\rho}_k$ are known, it is possible to calibrate the penalty in constant time for the three calibrations methods (eachjump, jumpmax and jumpmean) and to select the final models in $O(M_{\max}^2)$ operations.

Thus, the total complexity of the selection algorithm and of its POS variant is $O(M_{\max}^3)$.

4.6.4 Selection method, MAX variant (see Section 4.3)

In the MAX variant, the first step of the standard selection procedure is replaced by computing the distances $\|\hat{f}_k^{(M_{\max})} - \hat{f}_k^{(M)}\|_2$ for all M. This has complexity $O(M_{\max}^2)$. The complexity of the other steps remains unchanged.

Thus, the total complexity of the MAX variant of the selection algorithm is $O(M_{\max}^2)$.

5. Application to real data

In this section, we present the results of our method on two sets of trajectories. Trajectories are a typical example of dependent data that shows several behaviours depending on the activity of the entity being tracked, which makes hidden Markov models a popular modelling

	Algorithm	Complexity
Preliminary estimators	Spectral method	$O(n\log(n)^2 M_{\max})$
	Spectral method (De Castro et al. (2017))	$O(nM_{\max}^3)$
	Least squares method	$O(nM_{\max}^3 + C_n)$
Selection step	Standard and POS variant	$O(M_{\max}^3)$
	MAX variant	$O(M_{\max}^2)$

Table 2: Complexities of the different algorithms. n is the number of observations, $M_{\max}$ is the largest model dimension considered.

choice. For instance, the movement of a fisher is not the same depending on whether he's travelling to the next fishing zone or actually fishing.

The first data set follows artisanal fishers in Madagascar. The second one contains seabird movements. Studying the movements of fishers and seabirds has many applications, for instance understanding the fishing habits of the tracked entity, controlling the fishing pressure on local ecosystems and monitoring the dynamics of coastal ecosystems, see for instance Boyd et al. (2014); Vermard et al. (2010) and references therein.

5.1 Artisanal fishery

We use GPS tracks of artisanal fishers with a regular sampling period of 30 seconds. These tracks were produced by Faustinato Behivoke (Institut Halieutiques et des Sciences Marines, Université de Toliara, Madagascar) and Marc Léopold (IRD), who recorded artisanal fishers from Ankilibe, in Madagascar. Their fishing method is a seine netting.

From this data, we compute the velocity of the fisher during each time step. In order to estimate densities on $[0,1]$, we divide this velocity by an upper bound of the maximum observed velocity. We consider the observation space $\mathcal{Y} = [0,1]$ endowed with the dominating measure $\delta_0 + \mathrm{Leb}$, where δ_0 is the dirac measure in zero and Leb is the Lebesgue measure on $[0,1]$. As a proof of concept, we use the orthonormal basis consisting of the trigonometric basis on $[0,1]$ and the indicator function of $\{0\}$, that is the family $(\varphi_m)_{m\in\mathbb{N}}$ defined on $[0,1]$ by

$$\text{if } x = 0, \begin{cases} \varphi_0(x) = 1 \\ \varphi_m(x) = 0 \quad \text{for all } m \in \mathbb{N}^* \end{cases}$$

$$\text{if } x \neq 0, \begin{cases} \varphi_0(x) = 0 \\ \varphi_1(x) = 1 \\ \varphi_{2m}(x) = \sqrt{2}\cos(2\pi m x) \quad \text{for all } m \in \mathbb{N}^* \\ \varphi_{2m+1}(x) = \sqrt{2}\sin(2\pi m x) \quad \text{for all } m \in \mathbb{N}^* \end{cases}$$

The results using `M_max` $= 1000$ are shown in Figures 8 and 9. We took the normalizing velocity large enough that all observed normalized velocities belong to $[0, 0.8]$, hence the plot betweeen 0 and 0.8 for the densities.

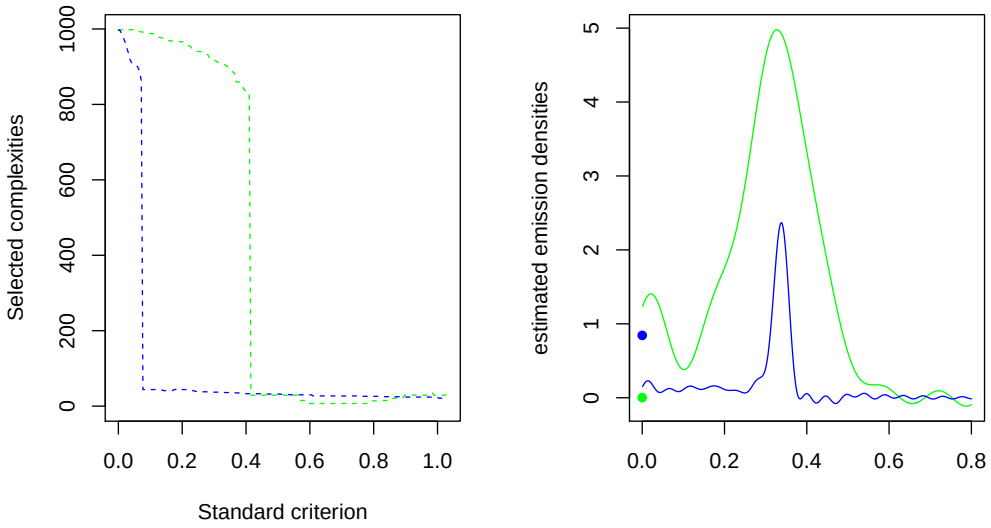


Figure 8: Selected complexities and estimated densities on artisanal fishery data (fisher 1, $n = 17,300$). Green = state 1, blue = state 2. The dirac component is shown as a dot at $y = 0$. The selected dimensions are $(14, 41)$.

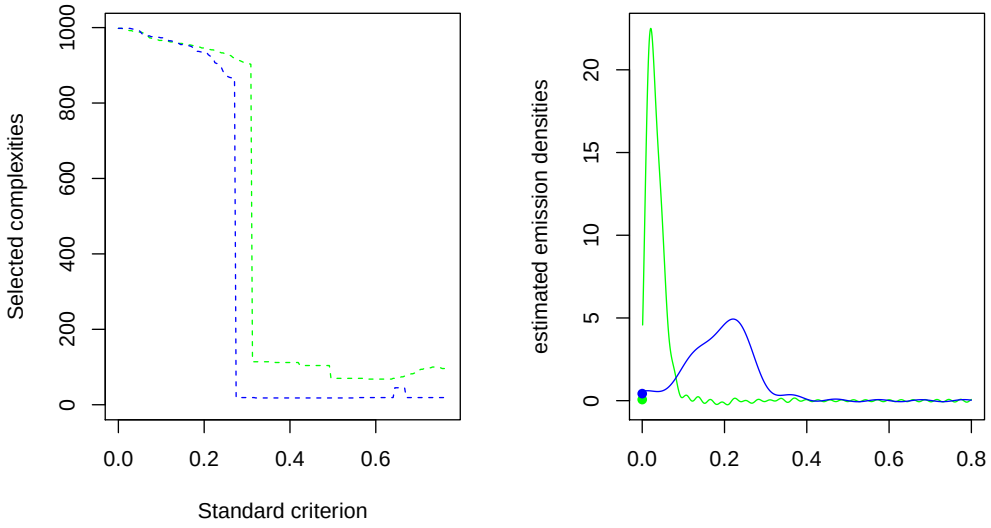


Figure 9: Selected complexities and estimated densities on artisanal fishery data (fisher 2, $n = 11,600$). Green = state 1, blue = state 2. The dirac component is shown as a dot at $y = 0$. The selected dimensions are $(68, 18)$.

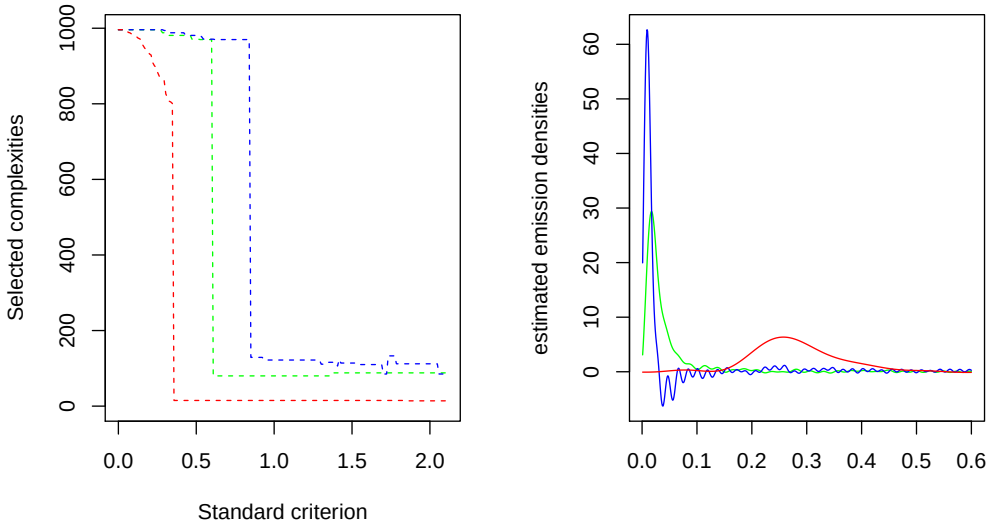


Figure 10: Selected complexities and estimated densities for Cormorant d's trajectory ($n = 2,891$). Green = state 1, blue = state 2, red = state 3. The selected dimensions are $(80, 110, 15)$.

In both cases, the selected model complexities differ greatly depending on the state. This comes from the fact that in both cases, one of the density is spiked, thus requiring more vectors of the orthonormal basis to be approximated. This illustrates that our method is able to estimate the smoother densities with fewer vectors of the basis, thus preventing overfitting.

As a side note, we needed considerably less observations than in the simulations: around 10,000, compared to 500,000 in the simulations. This can be explained by the fact that each state is very stable, with an estimated probability of leaving the states below 0.02–compared to 0.3 in the simulations. This is encouraging, as hidden states in real data are expected to be rather stable, especially when the sampling frequency is high, as long as the conditional independance of the observations can be assumed to hold.

5.2 Seabird foraging

In this Section, we consider the seabird data from Bertrand et al. (2015) and we focus on the tracks named cormorant d in this paper.

We apply the same transformation as in the previous section to obtain normalized velocities in $[0, 0.8]$ (after removal of anomalous velocities exceeding 150 m/s) and run the spectral algorithm with the trigonometric basis on $[0, 1]$ plus the indicator of $\{0\}$. The results are shown in Figure 10.

Note that the use of the trigonometric basis allows the estimated densities to take negative values. This is not a problem as far as minimax rates of convergence (in $\mathbf{L}^2$ norm) are concerned, however this can become an issue if one wants to use these densities in a

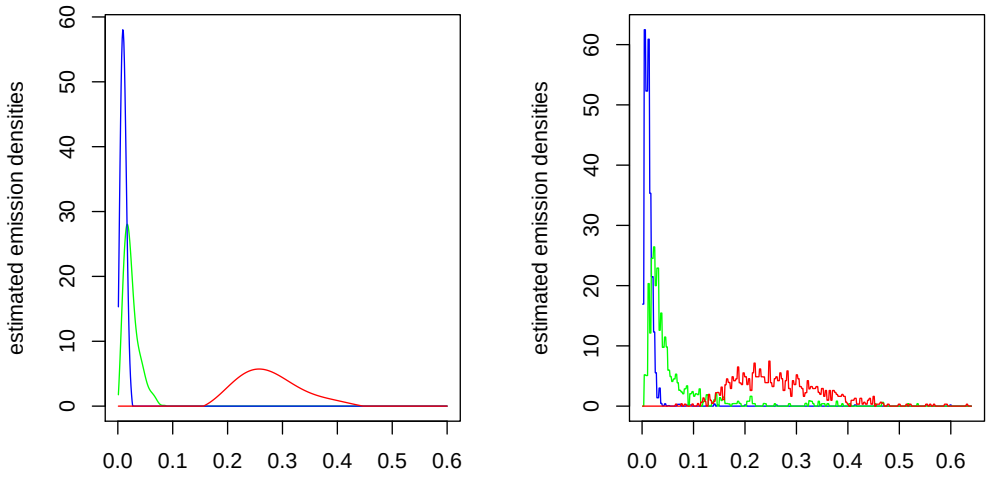


Figure 11: Projection of the estimated densities of Figure 10 for Cormorant d on the set of probability densities (left) and comparison with an estimation with histogram densities on a regular partition of size 300 using the EM algorithm (right).

forward-backward algorithm in order to get an estimator of the hidden states. One way to circumvent this problem is to use simplex projection to compute an approximation of the projection of these estimated density on the simplex of all probability densities. Note that since this is an $\mathbf{L}^2$ projection on a convex set which contains the true densities, the projected densities have an even smaller error, thus keeping the minimax rate of convergence of the original estimators. The resulting densities are shown in Figure 11

The number of observations in this setting is even smaller than for the fishery's data set: our algorithm was able to recover three emission densities from less than 3,000 observations, despite the states being less stable than in the fishery data set: the diagonal terms of the estimated transition matrix using the EM algorithm are $(0.83, 0.93, 0.98)$. In addition, the result of our method is consistent with other estimation methods, as shown in Figure 11: estimating the parameters with the EM algorithm using piecewise constant densities leads to a very similar result.

6. Conclusion and perspectives

We propose a state-by-state selection method to infer the emission densities of a HMM. Using a family of estimators, our method selects one estimator for each hidden state in a way that is adaptive with respect to this state's regularity. This method does not depend on the type of preliminary estimator, as long as a suitable variance bound is available. As such, it may be seen as a plug-in that takes a family of estimators and the corresponding variance bound and outputs the selected estimator. Note that its complexity does not depend on

the number of observations used to compute the estimators, which makes it applicable to arbitrarily large data sets.

To apply this method, we present two families of estimators: a least squares estimator and a spectral estimator. For both, we prove a bound on their variance and show that this bound allows to recover the minimax rate of convergence separately on each hidden state, up to a logarithmic factor. The variance bounds are similar to a BIC penalty, with an additional logarithmic factor for the spectral estimators.

We carry out a numerical study of the method and some variants on simulated data. We use the spectral estimators, which are both fast and don't suffer from initialization issues, unlike the least squares and maximum likelihood estimators. The simulations show that our selection method is very fast compared to the computation of the estimators and that indeed, the final estimators reach the minimax rate of convergence on each state.

Then, we compare our method with a cross validation estimator based on a least square risk. This estimator only reaches the minimax rate corresponding to the worst regularity among the emission densities and fails to select models with small dimensions. It is still noteworthy that the cross validation returns relevant results for small sample sizes, whereas our method requires the sample size to be large enough to work properly. An interesting problem would be to investigate whether cross validation or other methods can be combined with our state-by-state selection method to give an algorithm that is both fast, stable for small sample sizes and optimal in the asymptotic setting.

Finally, we apply our algorithm to real trajectory data sets. On this data, our method proves that it is able to match the regularity of the underlying emission densities. In addition, it is able to produce sensible results with far fewer observations than in our simulation study.

Our state-by-state selection method can be easily applied to multiview mixture models (also named mixture models with repeated measurement, see for instance Bonhomme et al. (2016a) and Gassiat et al. (2016)). Let us first describe the model. A multiview mixture model consists of two random variables, a hidden state U and an observation vector $\mathbf{Y} := (Y_i)_{i\in[m]}$ such that conditionally to U, the components Y_i of $\mathbf{Y}$ are independent with a distribution depending only on U and i. Let us assume that U takes its values in a finite set $\mathcal{X}$ of size K and that the Y_i have some density $f^*_{u,i}$ conditionally to $U = u$ with respect to a dominating measure. A question of interest is to estimate the densities $f^*_{u,i}$ from a sequence of observed $(\mathbf{Y}_n)_{n\geq 1}$.

Our state-by-state selection method can be applied directly to such a model as long as estimators with a proper variance bound are available (see assumption $[\mathbf{H}(\epsilon)]$ in Section 2.1). Indeed, we never use the dependency structure of the model. Regarding the development of preliminary estimators, multiview mixture models appear closely related to hidden Markov models: Anandkumar et al. (2012) and Bonhomme et al. (2016b) develop spectral methods that work for both multiview mixtures and HMMs at the same time using the same theoretical arguments. Thus, it seems clear that variance bounds such as the ones we developed can also be written for multiview mixture models.

Acknowledgments

I am grateful to Elisabeth Gassiat and Claire Lacour for their precious advice. I thank Augustin Touron for providing me with a R implementation of the spectral algorithm. I would also like to thank Marie-Pierre Etienne and of course Faustinato Behivoke (Institut Halieutiques et des Sciences Marines, Université de Toliara, Madagascar), Marc Léopold (IRD) and Sophie Bertrand (IRD) for letting me work on their data sets.

Appendix A. Spectral algorithm, full version

Algorithm 3: Spectral estimation of the emission densities of a HMM (full version)

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, two dimensions $m \leq M$, an orthonormal basis $(\varphi_1, \dots, \varphi_M)$ and number of retries r.

Result: Spectral estimators $(\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$, $\hat{\mathbf{Q}}$ and $\hat{\pi}$.

[Step 1] Consider the following empirical estimators: for any $a, c \in [m]$ and $b \in [M]$,

- $\hat{\mathbf{L}}_m(a) := \frac{1}{n}\sum_{s=1}^n \varphi_a(Y_s)$
- $\hat{\mathbf{M}}_{m,M,m}(a,b,c) := \frac{1}{n}\sum_{s=1}^n \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$
- $\hat{\mathbf{N}}_{m,M}(a,b) := \frac{1}{n}\sum_{s=1}^n \varphi_a(Y_s)\varphi_b(Y_{s+1})$
- $\hat{\mathbf{P}}_{m,m}(a,c) := \frac{1}{n}\sum_{s=1}^n \varphi_a(Y_s)\varphi_c(Y_{s+2})$.

[Step 2] Let $\hat{\mathbf{U}}_m$ be the $m \times K$ matrix of orthonormal left singular vectors and $\hat{\mathbf{V}}_M$ be the $M \times K$ matrix of orthonormal right singular vectors of $\hat{\mathbf{N}}_{m,M}$ corresponding to its top K singular values.

[Step 3] Form the matrices for all $b \in [M]$, $\hat{\mathbf{B}}(b) := (\hat{\mathbf{U}}_m^\top \hat{\mathbf{P}}_{m,m}\hat{\mathbf{U}}_m)^{-1}\hat{\mathbf{U}}_m^\top \hat{\mathbf{M}}_{m,M,m}(.,b,.)\hat{\mathbf{U}}_m$.

[Step 4] Set $(\Theta_i)_{1\leq i\leq r}$ r i.i.d. $(K \times K)$ unitary matrix uniformly drawn. Form the matrices for all $k \in \mathcal{X}$ and $i \in [r]$, $\hat{\mathbf{C}}_i(k) := \sum_{b=1}^M (\hat{\mathbf{V}}_M\Theta_i)(b,k)\hat{\mathbf{B}}(b)$.

[Step 5] Compute $\hat{\mathbf{R}}_i$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{\mathbf{C}}_i(1)$: $\hat{\mathbf{R}}_i^{-1}\hat{\mathbf{C}}_i(1)\hat{\mathbf{R}}_i = \mathrm{Diag}(\hat{\Lambda}_i(1,1), \dots, \hat{\Lambda}_i(1,K))$.

[Step 6] Set for all $k, k' \in \mathcal{X}$, $\hat{\Lambda}_i(k,k') := (\hat{\mathbf{R}}_i^{-1}\hat{\mathbf{C}}_i(k)\hat{\mathbf{R}}_i)(k',k')$. Choose i_0 maximizing $\min_k \min_{k_1\neq k_2} |\hat{\Lambda}_i(k,k_1) - \hat{\Lambda}_i(k,k_2)|$ and set $\hat{\mathbf{O}} := \hat{\mathbf{V}}_M\Theta_{i_0}\hat{\Lambda}_{i_0}$.

[Step 7] Consider the emission densities estimators $\hat{\mathbf{f}}^{(M,r)} := (\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$ defined by for all $k \in \mathcal{X}$, $\hat{f}_k^{(M,r)} := \sum_{b=1}^M \hat{\mathbf{O}}(b,k)\varphi_b$.

[Step 8] Let $\hat{\mathbf{O}}_m$ be the $m \times K$ matrix containing the first m rows of $\hat{\mathbf{O}}$. Set $\hat{\pi} = \Pi_\Delta\left((\hat{\mathbf{U}}_m^\top\hat{\mathbf{O}}_m)^{-1}\hat{\mathbf{U}}_m^\top\hat{\mathbf{L}}_m\right)$ where Π_Δ is the $\mathbf{L}^2$ projection onto the probability simplex.

[Step 9] Let $\hat{\mathbf{Q}}$ be the transition matrix defined by $\hat{\mathbf{Q}} = \Pi_{\mathrm{TM}}\left((\hat{\mathbf{U}}_m^\top\hat{\mathbf{O}}_m\mathrm{Diag}[\hat{\pi}])^{-1}\hat{\mathbf{U}}_m^\top\hat{\mathbf{N}}_{m,M}\hat{\mathbf{V}}_M(\hat{\mathbf{O}}^\top\hat{\mathbf{V}}_M)^{-1}\right)$ where Π_{TM} is the projection onto the set of transition matrices. This projection is obtained by projecting each line of the matrix onto the probability simplex.

Appendix B. Proofs

B.1 Proof of Lemma 1

Let $\tau_{n,M}$ be the permutation that minimizes $\tau \longmapsto \max_{k\in\mathcal{X}} \left\|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\right\|_2$. $[\mathbf{H}(\epsilon)]$ means that with probability $1-\epsilon$, one has $\max_{k\in\mathcal{X}} \left\|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\right\|_2 \leq \frac{\sigma(M)}{2}$.

Let $M \in \mathcal{M}$. Let us show that $\left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k')} - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2 > \left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k)} - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2$ for all $k, k' \in \mathcal{X}$ such that $k' \neq k$. If this holds, then the definition of $\hat{\tau}^{(M)}$ implies that $\hat{\tau}^{(M)} = \tau^{-1}_{n,M} \circ \tau_{n,M_0}$. Thus, one has $\max_{k\in\mathcal{X}} \left\|\hat{f}^{(M)}_{k,\text{new}} - f^{*,(M)}_{\tau_{n,M_0}(k)}\right\|_2 \leq \frac{\sigma(M)}{2}$, which is exactly Equation (1) with $\tau_n = \tau_{n,M_0}$.

Applying the triangular inequality leads to

$$\begin{aligned}\left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k)} - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2 \leq& \left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k)} - f^{*,(M)}_k\right\|_2 + \left\|f^{*,(M)}_k - f^{*,(M_0)}_k\right\|_2 + \left\|f^{*,(M_0)}_k - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2 \\ \leq& \frac{\sigma(M)}{2} + B_{M,M_0} + \frac{\sigma(M_0)}{2}\end{aligned}$$

and

$$\begin{aligned}\left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k')} - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2 \geq& \left\|f^{*,(M_0)}_{k'} - f^{*,(M_0)}_k\right\|_2 - \left\|\hat{f}^{(M)}_{\tau^{-1}_{n,M}(k')} - f^{*,(M)}_{k'}\right\|_2 \\ &- \left\|f^{*,(M)}_{k'} - f^{*,(M_0)}_{k'}\right\|_2 - \left\|f^{*,(M_0)}_k - \hat{f}^{(M_0)}_{\tau^{-1}_{n,M_0}(k)}\right\|_2 \\ \geq& m(\mathbf{f}^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2}.\end{aligned}$$

Thus, the result holds as soon as $m(\mathbf{f}^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2} > \frac{\sigma(M)}{2} + B_{M,M_0} + \frac{\sigma(M_0)}{2}$, which is the condition of Lemma 1.

B.2 Proof of Theorem 5

The structure of the proof is the same as the one of Theorem 3.1 of De Castro et al. (2017).

The first difference lies in the fact that we consider different models for each component of the tensors $\hat{\mathbf{N}}_{m,M}$ and $\hat{\mathbf{M}}_{m,M,m}$ in Step 1. As a consequence, we use the left *and* right singular vectors of $\hat{\mathbf{N}}_{m,M}$ instead of just the right singular vectors of $\hat{\mathbf{P}}_{m,m}$. A careful reading shows that their proof can be adapted straightforwardly to this situation.

The second difference consists in generating several independant random unitary matrices in Step 4 and keeping the one that separates the eigenvalues of all $\hat{\mathbf{C}}_i(k)$ best. This allows to replace Lemma F.6 of De Castro et al. (2017) by the following one, based on the independence of the unitary matrices:

Lemma 14 *For all $x > 0$ and $r \in \mathbb{N}^*$,*

$$\mathbb{P}\left[\forall k, k_1 \neq k_2, |\hat{\Lambda}_{i_0}(k, k_1) - \hat{\Lambda}_{i_0}(k, k_2)| \geq \frac{2e^{-x/r}(1 - \epsilon^2_{\mathbf{N}_{m,M}})^{1/2}}{\sqrt{e}K^{5/2}(K-1)}\gamma(\mathbf{O}_M)\right] \geq 1 - e^{-x}$$

and

$$\mathbb{P}\left[\|\hat{\Lambda}_{i_0}\|_\infty \geq \frac{1 + \sqrt{2}\sqrt{x + \log(K^2 r)}}{\sqrt{K}}\|\mathbf{O}_M\|_{2,\infty}\right] \leq e^{-x},$$

The notations $\epsilon_{\mathbf{N}_{m,M}}$ (or $\epsilon_{\mathbf{P}_M}$ in the original proof), $\gamma(\mathbf{O}_M)$ et $\|\mathbf{O}_M\|_{2,\infty}$ are introduced in De Castro et al. (2017).

Using this lemma, their proof leads to our result by taking $r = x = t$.

B.3 Definition of the polynomial H

B.3.1 Definition

We parametrize the application

$$(\pi, \mathbf{Q}, \mathbf{f}) \in \Delta \times \mathcal{Q} \times \mathrm{Span}(\mathbf{f}^*)^K \longmapsto \|g^{\pi,\mathbf{Q},\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \tag{3}$$

in the following way. For $p \in \mathbb{R}^{K-1}$, $q \in \mathbb{R}^{K\times(K-1)}$ and $A \in \mathbb{R}^{K\times(K-1)}$, define the extensions

- $\bar{p} \in \mathbb{R}^K$ defined by $\bar{p}(k) = p(k)$ for all $k \in [K-1]$ and $\bar{p}(K) = -\sum_{k\in[K-1]} p(k)$;
- $\bar{q} \in \mathbb{R}^{K\times K}$ by $\bar{q}(k, K) = -\sum_{k'\in[K-1]} q(k, k')$;
- $\bar{A} \in \mathbb{R}^{K\times K}$ by $\bar{A}(k, K) = -\sum_{k'\in[K-1]} A(k, k')$.

$\bar{p}$ corresponds to $\pi - \pi^*$, $\bar{q}$ to $\mathbf{Q} - \mathbf{Q}^*$ and A to the components of $\mathbf{f} - \mathbf{f}^*$ on $\mathbf{f}^*$ (which is a basis as soon as **[Hid]** holds). The condition on the last component of $\bar{p}$ and of each line of $\bar{q}$ and $\bar{A}$ follows from the fact that $\bar{p}$ corresponds to the difference of two probability vectors, $\bar{q}$ corresponds to the difference of two transition matrices and $\bar{A}$ correspond to the difference of two vectors of probability densities on a basis of probability densities.

Then, consider the quadratic form derived from the Taylor expansion of

$$(p, q, A) \in \mathbb{R}^{K-1} \times \mathbb{R}^{K\times(K-1)} \times \mathbb{R}^{(K-1)\times K} \longmapsto \|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2.$$

Let M be the matrix associated to this quadratic form. We define H as the determinant of M. Direct computations show that H is a polynomial in the coefficients of π^*, $\mathbf{Q}^*$ and $G(\mathbf{f}^*)$.

B.3.2 Link between H and the quadratic form from Equation (3)

The goal of this section is to show how H can be used to lower bound the quadratic form from Equation (3) by a positive constant times the distance between $(\pi, \mathbf{Q}, \mathbf{f})$ and $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. We will not need the assumptions **[Hid]**, **[HF]** or **[Hdet]** unless specified otherwise.

Let us start by the relation between the norms of (p, q, A) and $(\bar{p}, \bar{q}, \bar{A})$.

Lemma 15 *For all $(p, q, A) \in \mathbb{R}^{K-1} \times \mathbb{R}^{K\times(K-1)} \times \mathbb{R}^{(K-1)\times K}$,*

$$\begin{aligned}
\|p\|_2^2 \le \|\bar{p}\|_2^2 \le K\|p\|_2^2,\\
\|q\|_F^2 \le \|\bar{q}\|_F^2 \le K\|q\|_F^2,\\
\|A\|_F^2 \le \|\bar{A}\|_F^2 \le K\|A\|_F^2.
\end{aligned}$$

Proof $\|p\|_2^2 \le \|\bar{p}\|_2^2$ is immediate. Then,

$$\begin{aligned}
\|\bar{p}\|_2^2 =& \|p\|_2^2 + \left(\sum_{k\in[K-1]} p(k)\right)^2\\
\le& \|p\|_2^2 + (K-1)\sum_{k\in[K-1]} p(k)^2\\
=& K\|p\|_2^2.
\end{aligned}$$

The proof is the same for q and A. ■

The next lemma will be used to link the norms of A and $A\mathbf{f}$.

Lemma 16 *For all $\bar{A} \in \mathbb{R}^{K\times K}$ and $\mathbf{f}^* \in (\mathbf{L}^2(\mathcal{Y},\mu))^K$,*

$$\sigma_K(G(\mathbf{f}^*))\|\bar{A}\|_F^2 \leq \sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_2^2 \leq K\|G(\mathbf{f}^*)\|_\infty \|\bar{A}\|_F^2$$

Proof For the first inequality, we use that for all $k \in \mathcal{X}$,

$$\begin{aligned}\|(\bar{A}\mathbf{f}^*)_k\|_2^2 =& \bar{A}(k,\cdot)G(\mathbf{f}^*)\bar{A}(k,\cdot)^\top \\ \geq& \sigma_K(G(\mathbf{f}^*))\|\bar{A}(k,\cdot)\|_2^2\end{aligned}$$

and the inequality follows by summing over k.

For the second inequality,

$$\begin{aligned}\sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_2^2 &= \sum_{k\in[K]} \int (\bar{A}\mathbf{f}^*)_k(x)^2 \mu(dx) \\ &= \sum_{k\in[K]} \int \left(\sum_{j\in[K]} \bar{A}(k,j) f_j^*(x)\right)^2 \mu(dx) \\ &\leq \sum_{k\in[K]} \int K \sum_{j\in[K]} \bar{A}(k,j)^2 (f_j^*)^2(x)\mu(dx) \\ &\leq K\left(\sum_{k,j\in[K]} \bar{A}(k,j)^2\right) \sup_{j\in\mathcal{X}} \int (f_j^*)^2(x)\mu(dx) \\ &= K\|\bar{A}\|_F^2 \|G(\mathbf{f}^*)\|_\infty.\end{aligned}$$

■

Finally, we will use the following result of Lehéricy (to appear) (Section B.2) in order to upper bound the spectrum of the matrix M.

Lemma 17 *For all $\pi_1, \pi_2 \in \Delta$, for all $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{Q}$ and for all $\mathbf{f}_1, \mathbf{f}_2 \in (\mathbf{L}^2(\mathcal{Y},\mu))^K$,*

$$\|g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2}\|_2 \leq \sqrt{3K(\|G(\mathbf{f}_1)\|_\infty^3 \vee \|G(\mathbf{f}_2)\|_\infty^3)}\, d_{perm}((\pi_1,\mathbf{Q}_1,\mathbf{f}_1),(\pi_2,\mathbf{Q}_2,\mathbf{f}_2))$$

Together, these results imply that for all (p,q,A),

$$\begin{aligned}\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2& \\ \leq& 3K(\|G(\mathbf{f}^*+\bar{A}\mathbf{f}^*)\|_\infty^3 \vee \|G(\mathbf{f}^*)\|_\infty^3)(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f})_k\|_F^2) \\ \leq& 3K\|G(\mathbf{f}^*)\|_\infty^3 (1+K^2\|A\|_F^2)^3 (K\|p\|_2^2 + K\|q\|_F^2 + K^2\|G(\mathbf{f}^*)\|_\infty \|A\|_F^2)\end{aligned}$$

so that $\sigma_1(M) \leq \sqrt{3K^3}(1 \vee \|G(\mathbf{f})\|_\infty^2)$. Since $H = \prod_{i=1}^{(K-1)(2K+1)} \sigma_i(M)$, one has

$$\sigma_{(K-1)(2K+1)}(M) \geq \frac{H}{(3K^3(1 \vee \|G(\mathbf{f})\|_\infty^4))^{K^2-K/2}}.$$

Now, assume that **[Hid]** holds, so that $\sigma_K(G(\mathbf{f}^*)) > 0$, then

$$\begin{aligned}
\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \geq& \sigma_{(K-1)(2K+1)}(M)(\|p\|_2^2 + \|q\|_F^2 + \|A\|_F^2) \\
&+ o(\|p\|_2^2 + \|q\|_F^2 + \|A\|_F^2) \\
\geq& \frac{\sigma_{(K-1)(2K+1)}(M)}{1 \wedge K\|G(\mathbf{f}^*)\|_\infty} \left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k \in \mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_F^2 \right) \\
&+ o\left(\frac{1}{1 \wedge \sigma_K(G(\mathbf{f}^*))} \left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k \in \mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_F^2 \right) \right)
\end{aligned}$$

and finally

$$\begin{aligned}
&\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \\
&\qquad \geq c_2(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k \in \mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_F^2 \right) \\
&\qquad\qquad + o\left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k \in \mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_F^2 \right) \qquad (4)
\end{aligned}$$

where

$$c_2(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) = \frac{H}{(1 \wedge K\|G(\mathbf{f}^*)\|_\infty)(3K^3(1 \vee \|G(\mathbf{f}^*)\|_\infty^4))^{K^2-K/2}}$$

is positive as soon as **[Hid]** and **[Hdet]** hold.

B.4 Proof of Theorem 10

Let

$$N_{\mathbf{f}}(p, q, \mathbf{h}) = \|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2$$

and

$$\|(p, q, \mathbf{h})\|_{\mathbf{f}}^2 = d_{\mathrm{perm}}((\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}))^2.$$

We want to show that there exists a constant $c^* > 0$ such that there exists a neighborhood $\mathcal{V}$ of $\mathbf{f}^*$ such that if one writes

$$c_{\mathbf{f}} := \inf_{p \in (\Delta-\Delta),\, q \in (\mathcal{Q}-\mathcal{Q}),\, \mathbf{h} \in (\mathcal{F}-\mathcal{F})^K} \frac{N_{\mathbf{f}}(p, q, \mathbf{h})}{\|(p, q, \mathbf{h})\|_{\mathbf{f}}^2}$$

then $\inf_{\mathbf{f} \in \mathcal{V}} c_{\mathbf{f}} \geq c^*$.

The proof follows the structure of the proof of Theorem 6 of De Castro et al. (2016). It consists of three steps: the first one controls the component of $\mathbf{h}$ that is orthogonal to $\mathbf{f}$. This makes it possible to restrict $\mathbf{h}$ to the finite-dimensional space spanned by $\mathbf{f}$ in the two other parts. The second step controls the case when $\mathbf{h}$ is small, so that the behaviour of $N_{\mathbf{f}}$ is given by its quadratic form, and the last step controls the case where $\mathbf{h}$ is far from zero.

B.4.1 The orthogonal part.

Let $\mathbf{u}$ be the orthogonal projection of $\mathbf{h}$ on Span($\mathbf{f}$). Then

$$N_{\mathbf{f}}(p,q,\mathbf{h}) = N_{\mathbf{f}}(p,q,\mathbf{u}) + M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{h}-\mathbf{u})$$

where

$$\begin{aligned}
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) = \sum_{i_1,j_1,k_1} \sum_{i_2,j_2,k_2} & (\pi^*+p)(i_1)(\mathbf{Q}^*+q)(i_1,j_1)(\mathbf{Q}^*+q)(j_1,k_1) \\
& (\pi^*+p)(i_2)(\mathbf{Q}^*+q)(i_2,j_2)(\mathbf{Q}^*+q)(j_2,k_2) \\
& \Big(\langle a_{i_1}, a_{i_2}\rangle \langle (f+u)_{j_1}, (f+u)_{j_2}\rangle \langle (f+u)_{k_1}, (f+u)_{k_2}\rangle \\
& + \langle (f+u)_{i_1}, (f+u)_{i_2}\rangle \langle a_{j_1}, a_{j_2}\rangle \langle (f+u)_{k_1}, (f+u)_{k_2}\rangle \\
& + \langle (f+u)_{i_1}, (f+u)_{i_2}\rangle \langle (f+u)_{j_1}, (f+u)_{j_2}\rangle \langle a_{k_1}, a_{k_2}\rangle \\
& \quad + \langle a_{i_1}, a_{i_2}\rangle \langle a_{j_1}, a_{j_2}\rangle \langle (f+u)_{k_1}, (f+u)_{k_2}\rangle \\
& \quad + \langle a_{i_1}, a_{i_2}\rangle \langle (f+u)_{j_1}, (f+u)_{j_2}\rangle \langle a_{k_1}, a_{k_2}\rangle \\
& \quad + \langle (f+u)_{i_1}, (f+u)_{i_2}\rangle \langle a_{j_1}, a_{j_2}\rangle \langle a_{k_1}, a_{k_2}\rangle \\
& \qquad + \langle a_{i_1}, a_{i_2}\rangle \langle a_{j_1}, a_{j_2}\rangle \langle a_{k_1}, a_{k_2}\rangle \Big).
\end{aligned}$$

Let us write Π' the matrix whose diagonal terms are the elements of π^*+p and $\mathbf{Q}'$ the matrix $\mathbf{Q}^*+q$, then $M_{\mathbf{f}}$ can be written as

$$\begin{aligned}
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) = \sum_{i,j} \Big(& ((\Pi'\mathbf{Q}')^\top G(\mathbf{a})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \qquad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a})\Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \Big).
\end{aligned}$$

By the Schur product theorem, these terms are nonnegative since they correspond to Hadamard products of three Gram matrices which are nonnegative. Thus, one can lower bound $M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a})$ by the second term of the sum, which leads to

$$M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) \geq \sum_{i,j=1}^{K} ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}')_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \langle a_i, a_j\rangle$$

Assume **[Hid]** holds for the parameters $(\pi^*+p, \mathbf{Q}^*+q, \mathbf{f}+\mathbf{u})$, then the matrices $(\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}'$ and $\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}'$ are positive symmetric with respective lowest eigenvalue lower

bounded by $(\inf_k(\pi^*_k+p_k)\sigma_K(\mathbf{Q}^*+q))^2\sigma_K(G(\mathbf{f}+\mathbf{u}))$ and $\sigma_K(\mathbf{Q}^*+q)^2\sigma_K(G(\mathbf{f}+\mathbf{u}))$. Therefore, their Hadamard product is positive, and one has

$$(((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u})\Pi'\mathbf{Q}')_{i,j}(\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j})_{i,j} = (D\mathbf{U})^\top(D\mathbf{U})$$

with $\mathbf{U}$ an orthogonal matrix and D a diagonal matrix with positive diagonal coefficients. Moreover, the Schur product theorem implies that $\sigma_K(D)^2 \geq (\inf_k(\pi^*_k + p_k))^2\sigma_K(\mathbf{Q}^* + q)^4\sigma_K(G(\mathbf{f}+\mathbf{u}))^2$. Then

$$\begin{aligned}
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) \geq& \sum_{i,j=1}^K ((D\mathbf{U})^\top(D\mathbf{U}))_{i,j}\langle a_i, a_j\rangle \\
=& \sum_{j=1}^K \|D\mathbf{U}\mathbf{a}\|_2^2 \\
\geq& \sigma_K(D)^2\|\mathbf{U}\mathbf{a}\|_2^2 \\
\geq& (\inf_k(\pi^*_k + p_k))^2\sigma_K(\mathbf{Q}^* + q)^4\sigma_K(G(\mathbf{f}+\mathbf{u}))^2\|\mathbf{a}\|_2^2.
\end{aligned}$$

Finally, let $c_1(\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{u}) = (\inf_k(\pi^*_k + p_k))^2\sigma_K(\mathbf{Q}^* + q)^4\sigma_K(G(\mathbf{f} + \mathbf{u}))^2$. The application $(p, \pi^*, q, \mathbf{Q}^*, \mathbf{u}, \mathbf{f}) \mapsto c_1(\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{u})$ is continuous and nonnegative, it is positive when **[Hid]** holds for the parameters $(\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{u})$, and one has

$$M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) \geq c_1(\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{u})\|\mathbf{a}\|_2^2.$$

We will now control the term $N_{\mathbf{f}}(p,q,\mathbf{u})$. Two cases appear: when $(\pi^*+p, \mathbf{Q}^*+q, \mathbf{f}+\mathbf{u})$ is close to $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ in some sense and when it is not. The first case will be solved using the nondegeneracy of the quadratic form ensured by **[Hdet]**. The second case will be solved using the identifiability of the HMM.

B.4.2 In the neighborhood of $\mathbf{f}^*$.

The Taylor expansion of

$$(p,q,\mathbf{u}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times ((\mathcal{F} - \mathcal{F}) \cap \mathrm{Span}(\mathbf{f}))^K \mapsto N_{\mathbf{f}}(p,q,\mathbf{u})$$

around $(0,0,0)$ leads to a nonnegative quadratic form and no linear part. **[Hdet]**, **[Hid]** and equation (4) ensure that this form is positive for $\mathbf{f} = \mathbf{f}^*$. Let $c_2(\mathbf{Q}^*, \pi^*, \mathbf{f})$ be as defined in Section B.3.2, then $\mathbf{f} \mapsto c_2(\mathbf{Q}^*, \pi^*, \mathbf{f})$ is continuous and it is positive in the neighborhood of $\mathbf{f}^*$. Moreover, there exists a positive constant η depending on $\|G(\mathbf{f})\|_\infty$ such that for all $(p,q,\mathbf{u})$ such that $\|(p,q,\mathbf{u})\|_{\mathbf{f}} \leq 1$, one has

$$N_{\mathbf{f}}(p,q,\mathbf{u}) \geq c_2(\mathbf{Q}^*, \pi^*, \mathbf{f})\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 - \eta\|(p,q,\mathbf{u})\|_{\mathbf{f}}^3.$$

For instance, $\eta = 4000K^6\|G(\mathbf{f})\|_\infty^3$ works: the terms of order 2 or more in the Taylor expansion of $N_{\mathbf{f}}$ are the scalar product of sums of terms of the form $\sum_{i,j,k\in\mathcal{X}} \pi^*(i)\mathbf{Q}^*(i,j)\mathbf{Q}^*(j,k) f_i \otimes f_j \otimes f_k$ where zero to three of the f may be replaced by u, zero to two of the $\mathbf{Q}^*$ by q and π^* may be replaced by p and at least one of them is replaced. There are 63 possibilities, which leads to a sum of $(63K^3)^2$ terms, each of which can be bounded by

$\|G(\mathbf{f})\|_\infty^3(\max\{p(i), q(i,j), \|u_i\|_2 \mid i,j \in \mathcal{X}\})^r$ where r is the number of replaced terms. By taking the right permutation of states, the max can be bounded by $\|(p,q,\mathbf{u})\|_{\mathbf{f}}$, hence the result.

Then, using $\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2 = \|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2$ leads to

$$\begin{aligned}
\frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2} \geq & c_1(\mathbf{Q}^* + q, \pi^* + p, \mathbf{f} + \mathbf{u}) \frac{\|\mathbf{a}\|_2^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2} \\
& + c_2(\mathbf{Q}^*, \pi^*, \mathbf{f}) \frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2} - \eta \frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{(\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2)^{1/2}} \\
\geq & c_1(\mathbf{Q}^* + q, \pi^* + p, \mathbf{f} + \mathbf{u}) \frac{\|\mathbf{a}\|_2^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2} \\
& + c_2(\mathbf{Q}^*, \pi^*, \mathbf{f}) \frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 + \|\mathbf{a}\|_2^2} - \eta \sqrt{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}
\end{aligned}$$

Let $c_0 = \min(c_1/2, c_2)/2$, then c_0 is continuous and there exists a continuous function $(\pi^*, \mathbf{Q}^*, \mathbf{f}) \mapsto \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f})$ which is positive as soon as **[Hid]** and **[Hdet]** hold for $(\pi^*, \mathbf{Q}^*, \mathbf{f})$ and such that

$$\|(p,q,\mathbf{u})\|_{\mathbf{f}} \leq \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}) \;\Rightarrow\; \frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2} \geq c_0(\mathbf{Q}^*, \pi^*, \mathbf{f}).$$

Thus, there exists positive constants ϵ_0 and c_{near} depending on $\mathbf{Q}^*$, π^* and $\mathbf{f}^*$ such that

$$\forall (p,q,\mathbf{h},\mathbf{f}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K$$
$$\text{s.t. } \|(p,q,\mathbf{u})\|_{\mathbf{f}} \leq \epsilon_0 \text{ and } \sum_{k \in \mathcal{X}} \|f_k - f_k^*\|_2^2 \leq \epsilon_0^2, \qquad \frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2} \geq c_{\text{near}}.$$

B.4.3 Far from $\mathbf{f}^*$.

Lemma 18 *The application*

$$(p,q,\mathbf{u},\mathbf{f}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K \longmapsto N_{\mathbf{f}}(p,q,\mathbf{u})$$

restricted to the set of $(p,q,\mathbf{u},\mathbf{f})$ such that $\mathbf{u} \in Span(\mathbf{f})^K$ is uniformly continuous for the norm $\|\cdot\|_{tot}$ defined by

$$\|(p,q,\mathbf{u},\mathbf{f})\|_{tot}^2 := \|p\|_2^2 + \|q\|_F^2 + \sum_{k \in \mathcal{X}} \left(\|u_k\|_2^2 + \|f_k\|_2^2\right).$$

Thus, by compacity of $(\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(\mathbf{f}))^K$, the application

$$c_{\text{far}} : \mathbf{f} \longmapsto \inf_{(p,q,\mathbf{u}) \in (\Delta-\Delta)\times(\mathcal{Q}-\mathcal{Q})\times((\mathcal{F}-\mathcal{F})\cap\text{Span}(\mathbf{f}))^K \text{ s.t. } \|(p,q,\mathbf{u})\|_{\mathbf{f}} > \epsilon_0} N_{\mathbf{f}}(p,q,\mathbf{u})$$

is continuous. Let us now prove that $c_{\text{far}}(\mathbf{f}^*) > 0$.

Let $(p_n, q_n, \mathbf{u}_n)_n \in ((\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(\mathbf{f}^*))^K)^{\mathbb{N}}$ be a sequence such that $\|(p_n, q_n, \mathbf{u}_n)\|_{\mathbf{f}^*} > \epsilon_0$ for all n and

$$c_{\text{far}}(\mathbf{f}^*) = \lim_n N_{\mathbf{f}^*}(p_n, q_n, \mathbf{u}_n).$$

Then by compacity, this sequences converges towards a limit $(p,q,\mathbf{u})$. Necessarily $\|(p,q,\mathbf{u})\|_{\mathbf{f}^*} \geq \epsilon_0$. Since **[Hid]** holds, Theorem 8 shows that $N_{\mathbf{f}^*}(p,q,\mathbf{u}) > 0$, which implies $c_{\text{far}}(\mathbf{f}^*) > 0$ by continuity of $N_{\mathbf{f}^*}$. Note that $c_{\text{far}}(\mathbf{f}^*)$ may depend on $\mathcal{F}$ in addition to the parameters π^*, $\mathbf{Q}^*$ and $\mathbf{f}^*$.

Thus, by continuity, there exists $\epsilon_1 > 0$ such that for all $\mathbf{f} \in \mathcal{F}^K$ such that $\sum_{k\in\mathcal{X}} \|f_k - f_k^*\|_2^2 \leq \epsilon_1^2$, $c_{\text{far}}(\mathbf{f}) \geq c_{\text{far}}(\mathbf{f}^*)/2$.

Finally, **[HF]** implies that there exists a constant $\mathcal{C}$ depending only on $C_{\mathcal{F},2}$ such that $\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2 \leq \|(p,q,\mathbf{h})\|_{\mathbf{f}}^2 \leq \mathcal{C}$ for all $(p,q,\mathbf{h}) \in (\Delta-\Delta)\times(\mathcal{Q}-\mathcal{Q})\times(\mathcal{F}-\mathcal{F})^K$. Therefore,

$$\begin{aligned}
&\forall (p,q,\mathbf{h},\mathbf{f}) \in (\Delta-\Delta)\times(\mathcal{Q}-\mathcal{Q})\times(\mathcal{F}-\mathcal{F})^K \times \mathcal{F}^K \\
&\text{s.t. } \|(p,q,\mathbf{u})\|_{\mathbf{f}} \geq \epsilon_0 \text{ and } \sum_{k\in\mathcal{X}} \|f_k - f_k^*\|_2^2 \leq \epsilon_1^2, \qquad \frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2} \geq \frac{N_{\mathbf{f}}(p,q,\mathbf{u})}{\mathcal{C}} \\
&\qquad\qquad \geq \frac{c_{\text{far}}(\mathbf{f}^*)}{2\mathcal{C}}.
\end{aligned}$$

The theorem follows by taking $c^*(\pi^*,\mathbf{Q}^*,\mathbf{f}^*,\mathcal{F}) = \min\left(\frac{c_{\text{far}}(\mathbf{f}^*)}{2\mathcal{C}}, c_{\text{near}}\right)$ and the neighborhood containing all $\mathbf{f} \in \mathcal{F}^K$ such that $\sum_{k\in\mathcal{X}} \|f_k - f_k^*\|_2^2 \leq \min(\epsilon_0,\epsilon_1)^2$. Moreover, $(\pi,\mathbf{Q},\mathbf{f}) \longmapsto c^*(\pi,\mathbf{Q},\mathbf{f},\mathcal{F})$ is lower bounded by this value in a neighborhood of $(\pi^*,\mathbf{Q}^*,\mathbf{f}^*)$, so that it can be assumed to be lower semicontinuous.

Note that the dependency of c^* on $\mathcal{F}$ appears during this last step and is made non explicit because of the compacity assumption.

B.4.4 Proof of Lemma 18

$$\begin{aligned}
&\left|N_{\mathbf{f}}(p,q,\mathbf{u}) - N_{\mathbf{f}'}(p',q',\mathbf{u}')\right| \\
&\quad = \left| \|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2 - \|g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}\|_2^2 \right| \\
&\quad \leq 2\|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'}\|_2^2 + 2\|g^{\pi^*,\mathbf{Q}^*,\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}\|_2^2 \\
&\qquad + 2\left|\left\langle g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}, g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} \right\rangle\right| \\
&\qquad + 2\left|\left\langle g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}, g^{\pi^*,\mathbf{Q}^*,\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'} \right\rangle\right|
\end{aligned}$$

Then, using the fact that $\|g^{\pi,\mathbf{Q},\mathbf{f}} - g^{\pi',\mathbf{Q}',\mathbf{f}'}\|_2 \leq \sqrt{3K} C_{\mathcal{F},2}^3 \|(\pi-\pi',\mathbf{Q}-\mathbf{Q}',\mathbf{f}-\mathbf{f}',0)\|_{\text{tot}}$ (see Lemma 17), that $\|g^{\pi,\mathbf{Q},\mathbf{f}}\|_2 \leq C_{\mathcal{F},2}^3$ (see for instance Lemma 29 of Lehéricy (to appear)) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|N_{\mathbf{f}}(p,q,\mathbf{u}) - N_{\mathbf{f}'}(p',q',\mathbf{u}')\right| \leq{}& 6KC_{\mathcal{F},2}^6 \|(p-p',q-q',\mathbf{f}+\mathbf{u}-\mathbf{f}'-\mathbf{u}',0)\|_{\text{tot}}^2 \\
&+ 6KC_{\mathcal{F},2}^6 \|(0,0,0,\mathbf{f}-\mathbf{f}')\|_{\text{tot}}^2 \\
&+ 4\sqrt{3K} C_{\mathcal{F},2}^6 \|(p-p',q-q',\mathbf{f}+\mathbf{u}-\mathbf{f}'-\mathbf{u}',0)\|_{\text{tot}} \\
&+ 4\sqrt{3K} C_{\mathcal{F},2}^6 \|(0,0,0,\mathbf{f}-\mathbf{f}')\|_{\text{tot}}^2 \\
\leq{}& 24KC_{\mathcal{F},2}^6 \Big(\|(p-p',q-q',\mathbf{u}-\mathbf{u}',\mathbf{f}-\mathbf{f}')\|_{\text{tot}}^2 \\
&+ \|(p-p',q-q',\mathbf{u}-\mathbf{u}',\mathbf{f}-\mathbf{f}')\|_{\text{tot}} \Big),
\end{aligned}$$

which proves the uniform continuity of the application.

References

Animashree Anandkumar, Daniel J Hsu, and Sham M Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, volume 1, page 4, 2012.

Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

Sophie Bertrand, Rocío Joo, and Ronan Fablet. Generalized Pareto for pattern-oriented random walk modelling of organisms' movements. *PloS one*, 10(7):e0132231, 2015.

Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):211–229, 2016a.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Estimating multivariate latent-structure models. *The Annals of Statistics*, 44(2):540–563, 2016b.

Charlotte Boyd, André E Punt, Henri Weimerskirch, and Sophie Bertrand. Movement models provide insights into variation in the foraging effort of central place foragers. *Ecological modelling*, 286:13–25, 2014.

Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 17(111): 1–43, 2016.

Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, 2017.

Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric hidden Markov models are in general identifiable. *Stat. Comp.*, pages 1–11, 2015.

Elisabeth Gassiat, Judith Rousseau, and Vernet Elodie. Efficient semiparametric estimation and model selection for multidimensional mixtures. *arXiv preprint arXiv:1607.05430*, 2016.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, pages 1608–1632, 2011.

Alexander Goldenshluger and Oleg Lepski. General selection rule from a family of linear estimators. *Theory of Probability & Its Applications*, 57(2):209–226, 2013.

Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Claire Lacour, Pascal Massart, and Vincent Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *arXiv preprint arXiv:1607.05091*, 2016.

Luc Lehéricy. Order estimation for non-parametric hidden Markov models. *Bernouilli*, to appear.

Youen Vermard, Etienne Rivot, Stéphanie Mahévas, Paul Marchal, and Didier Gascuel. Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian hidden Markov models. *Ecological Modelling*, 221(15):1757–1769, 2010.