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

To cite this version:
Luc Lehéricy. State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models. 2017. <hal-01544390v1>
State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models

Luc Lehéricy  
Laboratoire de Mathématiques d’Orsay  
Univ. Paris-Sud, CNRS, Université Paris-Saclay  
91405 Orsay, France

Abstract

This paper considers the problem of estimating the emission densities of a nonparametric finite state space hidden Markov model in a way that is state-by-state adaptive and leads to minimax rates for each emission density—as opposed to globally minimax estimators, which adapt to the worst regularity among the emission densities. We propose a model selection procedure based on the Goldenschluger-Lepski method. Our method 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. Finally, numerical experiments assess the performance of the method and illustrate how to calibrate it in practice. Our method is not specific to hidden Markov models and can be applied to nonparametric multiview mixture models.

Keywords: hidden Markov models; model selection; nonparametric density estimation; 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 basic idea is similar to mixture models, where each observation depends only on the population it comes from. In a HMM, each observation depends only on a hidden variable, its hidden state, with markovian regime.

Formally, a hidden Markov model is a process \((X_j, Y_j)\) where \((X_j)\) is a Markov chain on \(\mathcal{X}\) and the \(Y_j\)'s are independent conditionally on \((X_j)\) with a distribution depending only on \(X_j\). The parameters of the HMM are the initial distribution and transition matrix of the Markov chain and the emission distributions, that is the distributions \((\nu^*_k)_{k \in \mathcal{X}}\) where \(\nu^*_k\) is the distribution of \(Y_j\) conditionally to \(X_j = k\). The learning problem is to infer the parameters from the observations \((Y_j)_j\) only.

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 \(\mu\), 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 the best possible con-
vergence rate when the emission densities are not restricted to belong to a set of densities described by finitely many parameters.

**Nonparametric state-by-state adaptivity.** Theoretical results in the nonparametric setting have only been developed recently. de Castro et al. (to appear) and Bonhomme et al. (2016b) introduce spectral methods, and the latter is proved to be minimax but not adaptive. 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 do not handle them separately. This translates into their error criterion, which is the supremum of the errors on all densities. Thus, 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. Such results are not entirely satisfactory, since the regularity of each emission density could be different, leading to different convergence rates. In particular, this means that having just one emission density that is very hard to estimate is enough to deteriorate the convergence rate of all emission densities. Solving this problem requires a different approach.

This is the main point of our paper. We show that it is possible to reach state-by-state adaptive minimax convergence rate. To do so, we propose a new method which is able to estimate each emission density separately.

**Plug-in procedure.** The method we propose is based on the Goldenschluger-Lepski method developed in the seminal papers by Goldenshluger and Lepski (2011, 2014) for density estimation, extended by Goldenshluger and Lepski (2013) to the white noise and regression models. It consists in using an auxiliary family of estimators and choosing the estimator that performs a good bias-variance tradeoff separately for each hidden state.

The method and assumptions are detailed in Section 2. Let us give a quick overview of the method. Given estimators $\hat{f}_k(M)$ of the emission densities for each hidden state $k$ and each model $M$ in a given family of models, one computes a substitute for the bias of the estimators by

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

for some penalty $\sigma$. 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 $\sigma$ 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.

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 convergence rate that is different for each state. These convergence rates will even be adaptive minimax up to a logarithmic factor when the method is applied to our two preliminary estimators: a spectral estimator and a least squares estimator. 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 $\sigma$ is essential in order to obtain minimax convergence rates. 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 auxiliary 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 a 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 convergence rate 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 4.5 for a more detailed discussion about the algorithmic complexity of the algorithms.

**Auxiliary estimators.** We use two methods to construct auxiliary estimators and apply the selection estimator. The motivation and key result of this part of the paper is to control the variances of the estimators by the right penalty $\sigma$. This part is crucial if one wants to get adaptive minimax rates, and has not been addressed in previous papers. For both methods, we develop new theoretical results that allow to obtain a penalty $\sigma$ that leads to adaptive minimax convergence rates 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. (to appear) 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. (to appear). However, their estimator cannot reach the minimax convergence rate: 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 Lehéry (2016) 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 Lehéry (2016) for the stability. However, since they rely on the minimization of a non-convex 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 $\theta$ and $\theta'$ respectively (these parameters actually correspond to the transition matrix and the emission densities of the HMM). Then one would like to get

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

where $d$ is the natural $L^2$ distance on the parameters and $C(\theta)$ is a positive constant which does not depend on $\theta'$. 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 $\theta^*_M$ 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

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

Such a result is crucial to control the variance of the estimators by a penalty term $\sigma$, 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. Lehéry (2016) 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 $\theta$) is unknown, which makes it impossible to use in our setting: one needs the constants $C(\theta^*_M)$ to be lower bounded by the same positive constant, which requires some sort of regularity on the function $\theta \mapsto C(\theta)$ 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 $C$ to be lower semicontinuous and positive without any additional assumption. In addition, we
give an explicit formula for the constant when \( \theta' \) and \( \theta \) are close, which gives an explicit bound for the asymptotical convergence rate. This result allows us to control the variance of the least squares estimators by a penalty \( \sigma \) which ensures that the state-by-state method reaches the adaptive minimax rate up to a logarithmic term.

**Numerical validation.** Section 4 shows how to apply the state-by-state selection method in practice and shows its performance on simulated data.

Note that the theoretical results give a penalty term \( \sigma \) 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. Nonetheless, 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 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. Finally, Section 4.5 discusses the complexities of the auxiliary estimation methods and of the selection procedures.

Section 5 contains a conclusion and perspectives for this work. Finally, Section 6 is dedicated to the proofs.

**Notations** We will use the following notations throughout the paper.

- \([K] = \{1, \ldots, K\}\) is the set of integers between 1 and \( K \).
- \(\mathcal{S}(K)\) is the set of permutations of \([K]\).
- \(\|\cdot\|_F\) is the Frobenius norm.
- \(\text{Span}(A)\) 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 \( A \in \mathbb{R}^{n \times p} \).
- For \( f = (f_1, \ldots, f_K) \in L^2(\mathcal{Y}, \mu)^K \), \( G(f) \) is the Gram matrix of \( f \), defined by \( G(f)_{i,j} = \langle f_i, f_j \rangle \) 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 \( Q^* \) be its transition matrix and \( \pi^* \) be its initial distribution. Let \((Y_j)_{j \geq 1}\) be random variables on a measured space \((\mathcal{Y}, \mu)\) with \( \mu \) \( \sigma \)-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 $\nu_k^*$ be the distribution of $Y_j$ conditionally to \( \{X_j = k\} \). Assume that $\nu_k^*$ has density $f_k^*$ with respect to $\mu$. We call $(\nu_k^*)_{k \in \mathcal{X}}$ the emission distributions and $\mathbf{f}^*(M) = (f_k^*(M))_{k \in \mathcal{X}}$ the emission densities. Then $(X_j, Y_j)_{j \geq 1}$ is a hidden Markov model with parameters $(\pi^*, Q^*, \mathbf{f}^*)$. 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 $\mathcal{P}_M \subseteq \mathcal{K}$ be a nested family of finite-dimensional subspaces such that their union is dense in $L^2(\mathcal{Y}, \mu)$. 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 $(\mathcal{P}_M)^K$.

In order to estimate the emission densities, we will only use some of the $\mathcal{P}_M$. Let $\mathcal{M}_n \subseteq \mathbb{N}$ 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 $\hat{f}_n^{(M)} = (\hat{f}_{n,k}^{(M)})_{k \in \mathcal{X}}$. We will need to assume that for all models, the distance between $\hat{f}_{n,k}^{(M)}$ and $f_k^{*,(M)}$ is small with high probability. In the following, we will drop the dependency in $n$ and simply write $\mathcal{M}$ and $\hat{f}^{(M)}$.

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

\[ \mathbb{H}(\epsilon) \quad \text{With probability } 1 - \epsilon, \]
\[ \forall M \in \mathcal{M}, \quad \inf_{\tau_n, M \in \mathcal{S}(K)} \max_{k \in \mathcal{X}} \left\| \hat{f}_{k}^{(M)} - f_{\tau_n, M(k)}^{*,(M)} \right\|_2 \leq \frac{\sigma(M, \epsilon, n)}{2} \]

where the upper bound $\sigma : (M, \epsilon, n) \in \mathcal{M} \times [0, 1] \times \mathbb{N}^* \mapsto \sigma(M, \epsilon, n) \in \mathbb{R}_+$ 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 Lehéry (2016)). In the following, we omit the parameters $\epsilon$ 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 $f_{\tau_n, M(k)}^{*,(k)}$ 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

\[ \hat{\tau}^{(M)} = \arg \min_{\tau \in \mathcal{S}(K)} \left\{ \max_{k \in \mathcal{X}} \left\| \hat{f}_{\tau(k)}^{(M)} - \hat{f}_{k}^{(M_{0})} \right\|_2 \right\} \]

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

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

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

one has

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

**Proof** Proof in Section 6.1

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

### 2.2 Estimator and oracle inequality

For each $k \in X$ and $M \in M$, let

$$A_k(M) = \sup_{M' \in M} \left\{ \left\| \tilde{f}^{(M')}_{k} - \tilde{f}^*(M \wedge M') \right\|_2 - \sigma(M') \right\}.$$

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

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

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

and finally take

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

The following theorem shows an oracle inequality on this estimator.

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

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

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

---

7
The first step consists in decomposing the total error: for all $M \in \mathcal{M}$ and $k \in \mathcal{X}$,
\[
\| \hat{f}(\hat{M}_k) - f^\star_{\tau_n(k)} \|_2 \leq \| \hat{f}(\hat{M}_k) - \hat{f}(\hat{M}_k \wedge M) \|_2 + \| \hat{f}(\hat{M}_k \wedge M) - \hat{f}(M) \|_2 \\
+ \| \hat{f}(M) - f^\star_{\tau_n(k)} \|_2 + \| f^\star_{\tau_n(k)} - f^\star \|_2.
\]

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
\[
\| \hat{f}(\hat{M}) - f^\star \|_2 \leq (A(M) + \sigma(\hat{M})) + (A(\hat{M}) + \sigma(M)) \\
+ \sigma(M) + \| f^\star(M) - f^\star \|_2 \\
\leq 2A(M) + 4\sigma(M) + \| f^\star(M) - f^\star \|_2.
\]

Then, notice that $A(M)$ can be bounded by
\[
A(M) \leq \sup_{M'} \left\{ \| \hat{f}(M') - f^\star(M') \|_2 + \| \hat{f}(M \wedge M') - f^\star(M \wedge M') \|_2 - \sigma(M') \right\} \\
+ \sup_{M'} \| f^\star(M') - f^\star(M \wedge M') \|_2.
\]

Since $\sigma$ 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
\[
A(M) \leq \| f^\star - f^\star(M) \|_2,
\]
which is enough to conclude.

\begin{remark}
The oracle inequality also holds when taking
\[
A_k(M) = \sup_{M' \geq M} \left\{ \| \hat{f}(M') - \hat{f}(M) \|_2 - \sigma(M') \right\} + .
\]
This positivity condition appears in the original Goldenschluger-Lepski method.
\end{remark}

\begin{remark}
Note that the selected $\hat{M}_k$ implicitly depends on the probability of error $\epsilon$ through the penalty $\sigma$.

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

3. Plug-in estimators and theoretical guarantees

In this section, we introduce two methods to construct preliminary estimators of the emis-

sion densities. We show that they satisfy assumption $[H(\epsilon)]$ for a given variance bound $\sigma$.

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 $L^2(Y,\mu)$ by a nested family of finite-dimensional subspaces $(\Psi_M)_{M \in \mathcal{M}}$ such that their union is dense in $L^2(Y,\mu)$ and write $f^*_k(M)$ the orthogonal projection of $f^*_k$ on $\Psi_M$ for all $k \in \mathcal{X}$ and $M \in \mathcal{M}$. We assume that $\mathcal{M} \subset \mathbb{N}$ and that the space $\Psi_M$ has dimension $M$. A typical way to construct such spaces is to take $\Psi_M$ 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 $(\pi^*, Q^*)$;

**[Hid]** $Q^*$ is invertible and the family $f^*$ 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 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. (to appear). Unlike the original one, it is able to reach the minimax convergence rate 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 codiagonalizing the matrices for several random rotations instead of just one, and selecting the rotation that separates the singular values of the $\Lambda_i$ “best”, 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 $(\varphi^M_1, \ldots, \varphi^M_M)$ be an orthonormal basis of $\Psi_M$. Let

$$\eta_3(m, M)^2 := \sup_{y,y' \in Y^3} \sum_{a,c=1}^m \sum_{b=1}^M \left( \varphi^m_a(y_1) \varphi^M_b(y_2) \varphi^m_c(y_3) - \varphi^m_a(y'_1) \varphi^M_b(y'_2) \varphi^m_c(y'_3) \right)^2.$$ 

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

**Theorem 5** Assume **[HX]** and **[Hid]** hold. Then there exists a constant $M_0$ depending on $f^*$ and constants $C_\sigma$ and $n_1$ depending on $f^*$ and $Q^*$ such that for all $\epsilon \in (0, 1)$, for all...
Algorithm 1: Spectral estimation of the emission densities of a HMM

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

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

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

\[
\begin{align*}
\hat{L}_m(a) &:= \frac{1}{n} \sum_{s=1}^{n} \varphi_a(Y_s) \\
\hat{M}_{m,M,m}(a, b, c) &:= \frac{1}{n} \sum_{s=1}^{n} \varphi_a(Y_s) \varphi_b(Y_{s+1}) \varphi_c(Y_{s+2}) \\
\hat{N}_{m,M}(a, b) &:= \frac{1}{n} \sum_{s=1}^{n} \varphi_a(Y_s) \varphi_b(Y_{s+1}) \\
\hat{P}_{m,m}(a, c) &:= \frac{1}{n} \sum_{s=1}^{n} \varphi_a(Y_s) \varphi_c(Y_{s+2}).
\end{align*}
\]

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

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

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

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

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

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

\(m, M \in \mathcal{M}\) 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 \(1 - 6\epsilon\),

\[
\inf_{\tau \in \Theta(K)} \max_{k \in X} \left| \|\hat{f}_k^{(M,r)} - f_\tau^*(M)\|_2^2 - C_\sigma \eta_3^2(m, M)(-\log \epsilon)^2 \right| / n.
\]

Proof Proof in Section 6.2.

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 convergence rate 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 \(a\) can be arbitrarily large.
Corollary 6 Assume [HX] and [Hid] hold. Also assume that \( \eta_2^2(m, M) \leq C_\eta m^2 M \) 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 \( C_\sigma \) depending on \( f^* \) and \( Q^* \) such that the following holds.

Let \( \alpha > 0 \) and \( C \geq 2(1 + 2\alpha) \sqrt{C_\eta C_\sigma} \). Let \( \hat{f}^{sbs} \) be the estimators selected from the family \( \{ \hat{f}^{,(M),\left[(1+2\alpha)\log(n)\right]} \}_{M \leq M_{\text{max}}(n)} \) with \( M_{\text{max}}(n) = n / \log(n) \), \( m_M = \log(n) \) 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

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

The novelty of this result is that each emission density is estimated with its own convergence rate: the rate \( \alpha^{-\frac{2\beta}{2\beta+1}} \) 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 \( \beta \)-Hölder density on \( [0,1]^D \) (equipped with a trigonometric basis), one can easily check the control of \( \eta_3 \), 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 \( (n / \log(n)^4)^{\frac{2\beta}{2\beta+D}} \) 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 \( \tau \) which does not depend on \( n \) such that

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

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 \( 1 - 6n^{-1-2\alpha} \),

\[
\inf_{\tau \in \mathcal{S}(K)} \max_{k \in \mathcal{X}} \| f_k^{(M, \tau)} - f_k^{*(M)} \|^2 \leq C_\sigma \eta_3^2(m_M, M) \frac{(1 + 2\alpha)^2 \log(n)^2}{n} \\
\leq (1 + \alpha)^2 C_\sigma C_\eta M \frac{\log(n)^4}{n} \\
\leq \frac{\sigma(M)^2}{4}
\]

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 \( M \leq M_{\text{max}}(n) \) as soon as \( M_{\text{max}}(n) = o(n/\log(n)^4) \).

Thus, \( [H(6n^{-(1+2\alpha)})] \) is true for the family \( (f^{(M, \lceil (1+2\alpha) \log(n) \rceil)})_{M \leq M_{\text{max}}(n)} \). Note that the assumption \( M_{\text{max}}(n) = o(n/\log(n)^4) \) also implies that there exists \( M_1 < M_{\text{max}}(n) \) 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 \( \tau_n \) such that with probability \( 1 - 6n^{-(1+2\alpha)} \), for all \( k \in \mathcal{X} \),

\[
\| \hat{f}_{\tau_n(k)} - f_k^* \|_2 \leq 4 \inf_{M_1 \leq M \leq M_{\text{max}}} \{ \| f_k^{*(M)} - f_k^* \|_2 + \sigma(M) \}
\]

\[
= O \left( \inf_{M_1 \leq M \leq M_{\text{max}}} \left( M^{-s_k} + \sqrt{\frac{M \log(n)^4}{n}} \right) \right)
\]

\[
= O \left( \left( \frac{n}{\log(n)^4} \right)^{-s_k/(1+2s_k)} \right),
\]

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

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

\[
\mathbb{E} \left[ \left\| (-n^\alpha) \vee (\hat{f}_{\tau_n(k)}^{\text{obs}} \wedge n^\alpha) - f_k^* \right\|_2^2 \right] \leq \mathbb{E} \left[ 1_A \left\| \hat{f}_{\tau_n(k)}^{\text{obs}} - f_k^* \right\|_2^2 \right] + \mathbb{E} \left[ 1_{A^c} (n^{2\alpha} + \| f_k^* \|_2^2) \right]
\]

\[
= O \left( \left( \frac{n}{\log(n)^4} \right)^{-2s_k/(1+2s_k)} \right) + O \left( \frac{n^{2\alpha} + \| f_k^* \|_2^2}{n^{1+2\alpha}} \right)
\]

\[
= O \left( \left( \frac{n}{\log(n)^4} \right)^{-2s_k/(1+2s_k)} \right).
\]
3.3 The penalized least squares method

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

\[ \text{[HF]} \quad f^* \in F^{K^*}, \quad F \text{ is closed under projection on } \Psi_M \text{ for all } M \in M \quad \text{and} \]

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

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

A simple way to construct such a set $F$ when $\mu$ is a finite measure is to take the sets $(\Psi_M)_M$ spanned by the first $M$ vectors of an orthonormal basis $(\varphi_i)_{i \geq 0}$ whose first vector $\varphi_0$ is proportional to 1. Then any set $F$ of densities such that $\int f d\mu = 1$, $\sum_i \langle f, \varphi_i \rangle^2 \leq C_{F,2}$ and $\sum_i \|\langle f, \varphi_i \rangle\|_\infty \leq C_{F,\infty}$ for given constants $C_{F,2}$ and $C_{F,\infty}$ and for all $f \in F$ satisfies [HF].

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

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

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

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

\[
d_{\text{perm}}((\pi_1, Q_1, f_1), (\pi_2, Q_2, f_2))^2 = \inf_{\tau \in \mathfrak{S}(K)} \left\{ \|\pi_1 - P_{\tau} \pi_2\|_2^2 + \|Q_1 - P_{\tau} Q_2 P_{\tau}^\top\|_F^2 + \sum_{k \in X} \|f_{1,k} - f_{2,\tau(k)}\|_2^2 \right\}.\]

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 $(\pi^*, Q^*, f^*) \in \Delta \times Q \times (L^2(Y, \mu))^K$ such that $\pi_x^* > 0$ for all $x \in X$ and [Hid] holds. Then for all $((\pi, Q, f) \in \Delta \times Q \times (L^2(Y, \mu))^K$, $g^{\pi,Q,f} = g^{\pi^*,Q^*,f^*} \Rightarrow d_{\text{perm}}((\pi, Q, f), (\pi^*, Q^*, f^*)) = 0$.

**Proof** The spectral algorithm of de Castro et al. (to appear) applied on the finite dimensional space spanned by the components of $f$ and $f^*$ allows to recover all the parameters even when the emission densities are not probability densities and when the Markov chain is not stationary. $\blacksquare$
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 \(L^2\) loss: for all \(t \in (L^2(\mathcal{Y}, \mu))^3\),
\[
\mathbb{E}[\gamma_n(t)] = \|t - g^*\|^2_2 - \|g^*\|^2_2
\]
where \(g^* = g^{\pi^*, Q^*, f^*}\). Since the bias does not depend on the function \(t\), one can hope that the minimizers of \(\gamma_n\) 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
\[
(\hat{\pi}^{(M)}, \hat{Q}^{(M)}, \hat{f}^{(M)}) \in \arg\min_{\pi \in \Delta, Q \in \mathfrak{Q}, f \in (\mathfrak{Q} \cap \mathfrak{F})^K} \gamma_n(g^{\pi, Q, f}).
\]
The procedure is summarized in Algorithm 2. Note that with the notations of the algorithm,
\[
\gamma_n(g^{\pi^*, Q^*, f^*} \Phi) = \|M_{(\pi, Q, O)} - \hat{M}_M\|^2_F + \|\hat{M}_M\|^2_F.
\]

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

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

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

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

**Step 2** Find a minimizer \((\hat{\pi}^{(M)}, \hat{Q}^{(M)}, \hat{O})\) of \((\pi, Q, O) \mapsto \|M_{(\pi, Q, O)} - \hat{M}_M\|^2_F\) where
- \(\pi \in \mathbb{R}^K\) is a probability distribution on \(\mathcal{X}\), i.e. \(\sum_{k \in \mathcal{X}} \pi_k = 1\);
- \(Q \in \mathbb{R}^{K \times K}\) is a transition matrix on \(\mathcal{X}\), i.e. \(\sum_{k' \in \mathcal{X}} Q(k, k') = 1\) for all \(k \in \mathcal{X}\);
- \(O\) is a \(M \times K\) matrix such that for all \(k \in \mathcal{X}\), \(\sum_{b=1}^M O(b, k) \varphi_b \in \mathcal{F}\);
- \(M_{(\pi, Q, O)} \in \mathbb{R}^{M \times M \times M}\) is defined by \(M_{(\pi, Q, O)}(b, \cdot, \cdot) = \text{ODiag}[\pi]Q\text{Diag}[O(b, \cdot)]QO^\top\) for all \(b \in [M]\).

**Step 3** Consider the emission densities estimators \(\hat{f}^{(M)} := (\hat{f}_k^{(M)})_{k \in \mathcal{X}}\) defined by for all \(k \in \mathcal{X}\), \(\hat{f}_k^{(M)} := \sum_{b=1}^M \hat{O}(b, k) \varphi_b\).
Theorem 9 Assume \([HF]\), \([HX]\) and \([Hid]\) hold.

Then there exists constants \(C\) and \(n_0\) depending on \(C_{F,2}, C_{F,\infty}\) and \(Q^*\) 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 \(M \leq n:\)

\[
\|g^{\hat{\pi}(M)}f^{\hat{M}}, \hat{f}^{(M)} - g^{\pi*,Q^*,f^*}f^{(M)}\|_2^2 \leq C\left(\frac{t}{n} + M\frac{\log(n)}{n}\right).
\]

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 \((\pi, Q, f) \in \Delta \times Q \times F^K \mapsto \|g^{\pi,Q,f} - g^{\pi*}\|_2^2\) around \((\pi^*, Q^*, f^*)\) is nondegenerate.

It is still unknown whether this nondegeneracy property is true for all parameters \((\pi^*, Q^*, f^*)\) 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 \(\pi^*, Q^*\) and of the Gram matrix of \(f^*\) 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 6.3. Note that Lehéricy (2016) proves that this polynomial \(H\) is non identically zero: it is shown that there exists parameters \((\pi, Q, f)\) satisfying \([HX]\) and \([Hid]\) such that \(H(\pi, Q, f) \neq 0\), which means that the following assumption is generically satisfied:

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

The following result allows to lower bound the \(L^2\) 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 (2016). The main difference is that the constant \(c^*(\pi^*, Q^*, f^*, F)\) does not depend on the \(f\) around which the parameters are taken. This is crucial to obtain Corollary 11, from which we will deduce \([HO]\). Note that we do not need \(f\) to be in a compact neighborhood of \(f^*\). Another improvement is that the constant in the minoration only depends on the true parameters and on the set \(F\).

Theorem 10 1. Assume that \([HF]\) holds and that for all \(f \in F, \int f \, d\mu = 1\).

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

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

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

\[
\|g^{\pi,Q,h} - g^{\pi^*,Q^*,f^*}\|_2^2 \geq c_0(\pi^*, Q^*, f^*)d_{perm}(\pi, Q, h, (\pi^*, Q^*, f))^2.
\]
\[ c_0(\pi^*, Q^*, f^*) = \frac{(\inf_{k \in \mathcal{X}} \pi^*(k)) \sigma_K(Q^*)^4 \sigma_K(G(f^*))^2}{4} \wedge \frac{H(\pi^*, Q^*, f^*)}{2(1 \wedge K \|G(f^*)\|_\infty) (3K^3(1 \vee G(f^*)\|_{\infty}))^{K^2 - K/2}} \]

**Proof** Proof in Section 6.4.

**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 \( C_{F, 2}, C_{F, \infty} \) and \( Q^* \) and constants \( M_0 \) and \( C' \) depending on \( F, Q^* \) and \( f^* \) 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 \( M_0 \leq M \leq n \):

\[
\inf_{\tau \in \mathcal{S}(K)} \max_{k \in \mathcal{X}} \| f_{\tau(k)}^{(M)} - f_{\tau(k)}^* \|_2 \leq C' \left( M \log(n) + \frac{t}{n} \right).
\]

**Remark 12** Using the second point of Theorem 10, one can alternatively take \( n_0 \) and \( M_0 \) depending on \( F, Q^* \) and \( f^* \), and \( C' \) depending on \( C_{F, 2}, C_{F, \infty} \) and \( f^* \) only. For instance, one can take \( C' = C / c_0(\pi^*, Q^*, f^*) \) 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 convergence rate of the estimators selected by our state-by-state selection method) does not depend on the set \( F \), but only on the HMM parameters and on the bounds \( C_{F, 2} \) and \( C_{F, \infty} \) 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 \( F \) 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 \( M \geq M_0 \), \( f_{\tau(k)}^{(M)} \in \mathcal{V} \). Then Theorem 9 and Theorem 10 applied to \( \pi = \hat{\pi}^{(M)}, Q = Q^{(M)}, h = \hat{f}^{(M)} \) and \( f = f^*, (M) \) 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 convergence rate (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 \( C_\sigma \) depending on \( C_{F, 2}, C_{F, \infty} \) and \( f^* \) such that the following holds.

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

\[
\forall k \in \mathcal{X}, \quad \mathbb{E} \left[ \| f^{\text{obs}}_{\tau(k)} - f_k^* \|_2 \right] = O \left( \left( \frac{n}{\log(n)} \right)^{\frac{s_k}{2s_k + 1}} \right).
\]
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. Finally, we discuss the algorithmic complexity of the different algorithms and estimators in Section 4.5.

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: \( \mathcal{P}_M := \text{Span}(\varphi_1, \ldots, \varphi_M) \) with

\[
\begin{align*}
\varphi_1(x) &= 1 \\
\varphi_{2m}(x) &= \sqrt{2} \cos(2\pi mx) \\
\varphi_{2m+1}(x) &= \sqrt{2} \sin(2\pi mx)
\end{align*}
\]

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

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

- 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 \( \mathcal{P}_M \) with \( M_{\text{min}} \leq M \leq M_{\text{max}} \), \( m = 20 \) and \( r = \lceil 2 \log(n) + 2 \log(M) \rceil \), where \( M_{\text{min}} = 3 \) and \( M_{\text{max}} = 300 \). Finally, we use the state-by-state selection method to choose the final estimator for each emission density.

We made 300 simulations, 20 per value of \( n \), with \( n \) taking values in \( \{ 5 \times 10^4, 7 \times 10^4, 1 \times 10^5, 1.5 \times 10^5, 2.2 \times 10^5, 3.5 \times 10^5, 5 \times 10^5, 7 \times 10^5, 1 \times 10^6, 1.5 \times 10^6, 2.2 \times 10^6, 3.5 \times 10^6, 5 \times 10^6, 7 \times 10^6, 1 \times 10^7 \} \).
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 $\sigma$ 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 $\text{pen}_{\text{shape}}$ and define $\hat{M}_k(\rho)$ the model selected for the hidden state $k$ by the state-by-state selection estimator using the penalty $\rho \text{pen}_{\text{shape}}$:

$$\hat{M}_k(\rho) \in \arg\min_{M \in \mathcal{M}} \{ A_k(M) + 2\rho \text{pen}_{\text{shape}}(M) \}.$$ 

where

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

The dimension jump method relies on the heuristics that there exists a constant $C$ such that $C \text{pen}_{\text{shape}}$ 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 $\rho_{\text{jump},k} = C$ in the plot of $\rho \mapsto \hat{M}_k(\rho)$. 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 $\rho \mapsto \hat{M}_k(\rho)$.

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 convergence rates—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

**Variant POS.** As mentioned 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 Goldenschluger-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.

**Variant MAX.** In the context of kernel density estimation, Lacour et al. (2016) show that the Goldenschluger-Lepski method still works when the bias 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

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

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

$$\hat{f}_k = \hat{f}_k^{(\hat{M}_k)}.$$
Figure 2: Selected complexities with respect to the penalty constant $\rho$ for the same simulation of $n = 500,000$ observations. The colored dashed lines correspond to the single-state complexities $M_k(\rho)$. 
where $\sigma$ 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 $O(M_{\text{max}})$ computations of $L^2$ norms instead of $O(M_{\text{max}}^2)$. 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.
Figure 3: Logarithm of the $L^2$ 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.
In Figures 3 and 4, a phase transition is clearly visible between two regimes. The first regime occurs for \( \log(n) < 13 \) (which corresponds to \( n \leq 220,000 \)). Note that this threshold may vary for other choices of parameters. In this regime, the error of the estimator is roughly constant in \( n \). This phenomenon could be a limitation of spectral algorithms: they are unable to estimate the emission densities correctly when the number of observations is too small.

This issue was brought to light by Lehéry (2015), who shows that the spectral estimators sometimes return aberrant results when the number of observations is small. This problem is closely related to the order estimation issue: when the spectral algorithm is not able to “see” the correct number of states, the estimators it returns are dominated by noise. Lehéry (2016) introduces a consistent order estimator for nonparametric HMMs based on...
thresholding the spectrum of an empirical matrix. The corresponding criterion makes it possible to quantify when the spectral estimators give a sensible result. The main tool of this criterion is the matrix $\hat{N}_{m,M}$ (defined in Algorithm 1). This matrix is a noisy version of the matrix $N = \mathbb{E}[\hat{N}_{m,M}]$, which has a rank equal to the order of the HMM. The phase transition occurs when the noise level is of the order of the smallest (positive) eigenvalue of the matrix $N$: for smaller values of $n$, some singular values of $N$ are drowned by the noise and the corresponding eigenvectors cannot be reliably recovered from the matrix $\hat{N}_{m,M}$, which makes it impossible to recover the model parameters. Conversely, for larger values of $n$, all eigenvectors can be satisfactorily recovered, and the spectral estimators converge to the true parameters.

The second regime occurs for $\log(n) > 13$, that is $n \geq 500,000$. In this situation, the spectral algorithm is able to identify and estimate all states properly. This can be seen through a sharp decrease of the error around the transition, followed by a linear decrease of the logarithm of the error with respect to the logarithm of the number of observations.

**Asymptotic behaviour.** For larger values of the number of observations $n$, 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
Table 1: Exponents 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. The smaller the exponent, the faster the estimators converge.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Uniform</th>
<th>Sym. Beta</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eachjump</td>
<td>-0.500 ± 0.046</td>
<td>-0.347 ± 0.007</td>
<td>-0.470 ± 0.015</td>
</tr>
<tr>
<td>Eachjump POS</td>
<td>-0.503 ± 0.047</td>
<td>-0.327 ± 0.008</td>
<td>-0.469 ± 0.015</td>
</tr>
<tr>
<td>Eachjump MAX</td>
<td>-0.480 ± 0.052</td>
<td>-0.335 ± 0.009</td>
<td>-0.449 ± 0.015</td>
</tr>
<tr>
<td>Jumpmean</td>
<td>-0.532 ± 0.048</td>
<td>-0.349 ± 0.006</td>
<td>-0.471 ± 0.017</td>
</tr>
<tr>
<td>Jumpmean POS</td>
<td>-0.540 ± 0.048</td>
<td>-0.350 ± 0.006</td>
<td>-0.456 ± 0.016</td>
</tr>
<tr>
<td>Jumpmean MAX</td>
<td>-0.493 ± 0.049</td>
<td>-0.374 ± 0.009</td>
<td>-0.437 ± 0.015</td>
</tr>
<tr>
<td>Jumpmax</td>
<td>-0.500 ± 0.046</td>
<td>-0.349 ± 0.006</td>
<td>-0.464 ± 0.016</td>
</tr>
<tr>
<td>Jumpmax POS</td>
<td>-0.492 ± 0.046</td>
<td>-0.358 ± 0.006</td>
<td>-0.442 ± 0.015</td>
</tr>
<tr>
<td>Jumpmax MAX</td>
<td>-0.480 ± 0.052</td>
<td>-0.404 ± 0.009</td>
<td>-0.466 ± 0.015</td>
</tr>
<tr>
<td>Oracle</td>
<td>-0.517 ± 0.048</td>
<td>-0.360 ± 0.006</td>
<td>-0.459 ± 0.017</td>
</tr>
<tr>
<td>Minimax (Hölder)</td>
<td>-0.5</td>
<td>-3/11 ≈ -0.273</td>
<td>-3/7 ≈ -0.429</td>
</tr>
</tbody>
</table>

The 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 convergence rates are clearly distinct. The uniform distribution is estimated with a convergence rate 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 convergence rates 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 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.

**Spectral algorithm** (see Section 3.2). We consider the algorithmic complexity of estimating the emission densities for all models \( M \) such that \( M_{\text{min}} \leq M \leq M_{\text{max}} \) with \( n \)
observations and auxiliary parameters $r$ and $m$ depending on $n$ and $M$ (upper bounded by $m_{\text{max}}$ and $r_{\text{max}}$).

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

In practice, one takes $m \propto \log(n)$, $r \propto \log(n) + \log(M)$ and $M_{\text{max}} \leq n$, so that the total complexity of the spectral algorithm is $O(n \log(n) M_{\text{max}}^3)$.

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

**Least squares algorithm** (see Section 3.3). We consider the algorithmic complexity of estimating the emission densities for all models $M$ such that $M_{\text{min}} \leq M \leq M_{\text{max}}$ with $n$ observations.

Step 1 is similar to the one of the spectral algorithm, but with $O(n M_{\text{max}}^3)$ 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 Lehéricaly (2016), this is usually the longest step when computing least squares estimators.

Thus, the total complexity of the least squares algorithm is $O(n M_{\text{max}}^3 + C_n)$.

**Selection method and POS variant** (see Sections 2.2 and 4.3). We consider the algorithmic complexity of selecting estimators from a family $(\hat{f}(M))_{M_{\text{min}} \leq M \leq M_{\text{max}}}$ 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 $O(M_{\text{max}}^3)$: it requires to compute the $L^2$ distance of at most $M_{\text{max}}^2$ couples of functions in a Hilbert space of dimension $M_{\text{max}}$.
- Compute $\hat{\rho}_k$ defined as the abscissa of the largest jump of the function $\rho \mapsto \hat{M}_k(\rho)$ for all $k$, where $\hat{M}_k$ is defined as in Section 4.2. Note that computing $\hat{M}_k(\rho)$ requires $O(M_{\text{max}}^2)$ operations. An approximate value of $\hat{\rho}_k$ can be computed in $O(\log(\hat{\rho}_k) M_{\text{max}}^2)$ operations, which is usually $O(M_{\text{max}}^2)$.

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 $O(M_{\text{max}}^3)$ operations.

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

**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 $\|\hat{f}_k(M_{\text{max}}) - \hat{f}_k(M)\|_2$ for all $M$. This has complexity $O(M_{\text{max}}^2)$. The complexity of the other steps remains unchanged.

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

26
Table 2: Complexities of the different algorithms. \( n \) is the number of observations, \( M_{\text{max}} \) is the largest model dimension considered.

## 5. Conclusion and perspectives

We have proposed a state-by-state selection method to infer the emission densities of a HMM. Using a family of preliminary estimators, our method selects for each hidden state one of the estimators 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 large datasets.

To apply this method, we have presented two family of estimators: a least squares estimator and a spectral estimator. For both, we proved a bound on their variance and showed that this bound allows to recover the minimax convergence rate 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 carried out a numerical study of the method and some variants on simulated data. We used 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 convergence rate on each state.

Our simulation also seem to show that a phase transition occurs when the number of observations \( n \) used by the spectral estimators increases. For small values of \( n \), the error of the final estimators is large and roughly constant in \( n \)—though the error may be reduced using our MAX variant, which also makes the calibration step easier. Then, around the critical value of \( n \), the error drops sharply. Finally, for larger values of \( n \), the error decreases as a power of \( n \) which corresponds to the minimax convergence rate.

We conjecture that this sharp change is caused by the spectral estimators: since they rely on estimating the eigenvectors of empirical tensors, they may give aberrant result when the noise becomes larger than the spectrum of the matrices they invert. Thus, the phase transition would occur when the noise isn’t able to drown the eigenvalues anymore. An interesting problem would be to investigate this phase transition, and in particular to see whether it is indeed specific to the spectral estimators or if it can be reproduced with other estimators.
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 $Y := (Y_i)_{i \in [m]}$ such that conditionally to $U$, the components $Y_i$ of $Y$ 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 $(Y_n)_{n \geq 1}$.

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 $[H(\epsilon)]$ 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.

6. Proofs

6.1 Proof of Lemma 1

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

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

Applying the triangular inequality leads to

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

and

$$
\left\| \hat{\tau}^{(M)}_{n,M}(k') - \hat{\tau}^{(M)}_{n,M}(k) \right\|_2 \geq \left\| f^*_{\tau(k')} - f^*_{\tau(k)} \right\|_2 - \left\| \hat{\tau}^{(M)}_{n,M}(k') - f^*_{\tau(k')} \right\|_2
\geq m(f^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2}.
$$
Thus, the result holds as soon as
\[ m(f^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2} > \frac{\sigma(M)}{2} + B_{M,M_0} + \frac{\sigma(M_0)}{2}, \]
which is the condition of Lemma 1.

### 6.2 Proof of Theorem 5

The structure of the proof is the same as the one of Theorem 3.1 of de Castro et al. (to appear).

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

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

**Lemma 14** For all \( x > 0 \) and \( r \in \mathbb{N}^* \),
\[
P \left[ \forall k, k_1 \neq k_2, \left| \tilde{A}_{i_0}(k, k_1) - \tilde{A}_{i_0}(k, k_2) \right| \geq \frac{2e^{-x/r}(1 - \epsilon_{\mathbb{N}_{m,M}}^2)^{1/2}}{\sqrt{e}K^{5/2}(K - 1)} \gamma(O_M) \right] \geq 1 - e^{-x} \tag{3}
\]
and
\[
P \left[ \| \tilde{A}_{i_0} \|_\infty \geq \frac{1 + \sqrt{2\sqrt{x + \log(K^2r)}}}{\sqrt{K}} \| O_M \|_{2,\infty} \right] \leq e^{-x}, \tag{4}
\]
The notations \( \epsilon_{\mathbb{N}_{m,M}} \) (or \( \epsilon_{\mathbb{P}_M} \) in the original proof), \( \gamma(O_M) \), and \( \| O_M \|_{2,\infty} \) are introduced in de Castro et al. (to appear).

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

### 6.3 Definition of the polynomial \( H \)

#### 6.3.1 Definition

We parametrize the application
\[(\pi, Q, f) \in \Delta \times Q \times \text{Span}(f^*)^K \mapsto \| g^{\pi,Q,f} - g^{\pi^*,Q^*,f^*} \|^2_2 \tag{5}\]
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 \( Q - Q^* \) and \( \bar{A} \) to the components of \( f - f^* \) on \( f^* \) (which is a basis as soon as \( [\text{Hid}] \) 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

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

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 \( \pi^* \), \( Q^* \) and \( G(f^*) \).

6.3.2 Link between \( H \) and the quadratic form from Equation (5)

The goal of this section is to show how \( H \) can be used to lower bound the quadratic form from Equation (5) by a positive constant times the distance between \( (\pi, Q, f) \) and \( (\pi^*, Q^*, f^*) \).

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 \( (p, q, A) \in \mathbb{R}^{K-1} \times \mathbb{R}^{K \times (K-1)} \times \mathbb{R}^{(K-1) \times K} \),

\[
\|p\|_2^2 \leq \|\bar{p}\|_2^2 \leq K \|p\|_2^2,
\]

\[
\|q\|_F^2 \leq \|\bar{q}\|_F^2 \leq K \|q\|_F^2,
\]

\[
\|A\|_F^2 \leq \|\bar{A}\|_F^2 \leq K \|A\|_F^2.
\]

**Proof** \( \|p\|_2^2 \leq \|\bar{p}\|_2^2 \) is immediate. Then,

\[
\|\bar{p}\|_2^2 = \|p\|_2^2 + \left( \sum_{k \in [K-1]} p(k) \right)^2 \\
\leq \|p\|_2^2 + (K - 1) \sum_{k \in [K-1]} p(k)^2 \\
= K \|p\|_2^2.
\]

The proof is the same for \( q \) and \( A \). \( \square \)

The next lemma will be used to link the norms of \( A \) and \( Af \).

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

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

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

\[
\|(\bar{A}f^*)_k\|_2^2 \leq \bar{A}(k, \cdot)G(f^*)\bar{A}(k, \cdot)^T \\
\geq \sigma_K(G(f^*))\|\bar{A}(k, \cdot)\|_2^2
\]

and the inequality follows by summing over \( k \).
For the second inequality,

\[ \sum_{k \in X} \| (\tilde{A}f^*)_k \|^2_2 = \sum_{k \in [K]} \int (\tilde{A}f^*)_k(x)^2 \mu(dx) \]

\[ = \sum_{k \in [K]} \int \left( \sum_{j \in [K]} \tilde{A}(k,j)f_j^*(x) \right)^2 \mu(dx) \]

\[ \leq \sum_{k \in [K]} \int K \sum_{j \in [K]} \tilde{A}(k,j)^2(f_j^*)^2(x) \mu(dx) \]

\[ \leq K \left( \sum_{k,j \in [K]} \tilde{A}(k,j)^2 \right) \sup_{j \in X} \int (f_j^*)^2(x) \mu(dx) \]

\[ = K \| \tilde{A} \|_F^2 \| G(f^*) \|_\infty. \]

Finally, we will use the following result of Lehéry (2016) (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 \( Q_1, Q_2 \in Q \) and for all \( f_1, f_2 \in (L^2(\mathcal{Y}, \mu))^K \),

\[ \| g^{\pi_1,Q_1,f_1} - g^{\pi_2,Q_2,f_2} \|_2 \leq \sqrt{3K} \| (|G(f_1)|^\frac{3}{\infty} \vee |G(f_2)|^\frac{3}{\infty}) d_{\text{perm}}((\pi_1, Q_1, f_1), (\pi_2, Q_2, f_2)) \]

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

\[ \| g^{\pi^*,\bar{p},Q^*,q,F^*} - g^{\pi^*,Q^*,F^*} \|_2^2 \leq 3K \| (|G(f^*) + \tilde{A}f^*)|^\frac{3}{\infty} \vee |G(f^*)|^\frac{3}{\infty}) \left( \| \bar{p} \|_2^2 + \| \bar{q} \|_F^2 + \sum_{k \in X} \| (\tilde{A}f)_k \|_F^2 \right) \]

\[ \leq 3K \| G(f^*) \|^\frac{3}{\infty} (1 + K^2 \| F \|_F^2)^3 \left( K \| p \|_2^2 + K \| q \|_F^2 + K^2 \| G(f^*) \|_\infty \| A \|_F^2 \right) \]

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

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

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

\[ \| g^{\pi^*,\bar{p},Q^*,q,F^*} - g^{\pi^*,Q^*,F^*} \|_2^2 \geq \sigma_{(K-1)(2K+1)}(M) \left( \| \bar{p} \|_2^2 + \| \bar{q} \|_F^2 + \| A \|_F^2 \right) \]

\[ + o \left( \| \bar{p} \|_2^2 + \| \bar{q} \|_F^2 + \| A \|_F^2 \right) \]

\[ \geq \sigma_{(K-1)(2K+1)}(M) \left( \| \bar{p} \|_2^2 + \| \bar{q} \|_F^2 + \sum_{k \in X} \| (\tilde{A}f)_k \|_F^2 \right) \]

\[ + o \left( \frac{1}{1 \wedge \sigma_K(G(f^*))} \left( \| \bar{p} \|_2^2 + \| \bar{q} \|_F^2 + \sum_{k \in X} \| (\tilde{A}f)_k \|_F^2 \right) \right) \]

31
and finally
\[
\|g^{\pi^*+p, Q^*+q, f^*+\tilde{A}^*} - g^{\pi^*, Q^*, f^*}\|^2_2 \\
\geq c_2(\pi^*, Q^*, f^*) \left( \|\tilde{p}\|^2_2 + \|\tilde{q}\|^2_2 + \sum_{k \in X} \|\tilde{A}^*_f\|_F^2 \right) \\
+ o \left( \|\tilde{p}\|^2_2 + \|\tilde{q}\|^2_2 + \sum_{k \in X} \|\tilde{A}^*_f\|_F^2 \right)
\]

where
\[
c_2(\pi^*, Q^*, f^*) = \frac{H}{(1 \wedge \|G(f^*)\|_{\infty})(3K^3(1 \vee \|G(f^*)\|_{\infty})^2)^{K^2-K/2}}
\]
is positive as soon as \[H_{\text{id}}\] and \[H_{\text{det}}\] hold.

\section*{6.4 Proof of Theorem 10}

Let
\[
N_f(p, q, h) = \|g^{\pi^*+p, Q^*+q, f+h} - g^{\pi^*, Q^*, f}\|^2_2
\]
and
\[
\|(p, q, h)\|^2_f = d_{\text{perm}}((\pi^* + p, Q^* + q, f + h), (\pi^*, Q^*, f))^2.
\]
We want to show that there exists a constant \(c^* > 0\) such that there exists a neighborhood \(V\) of \(f^*\) such that if one writes
\[
c_f := \inf_{p \in (\Delta - \Delta), q \in (Q - Q), h \in (F - F)^K} \frac{N_f(p, q, h)}{\|(p, q, h)\|^2_f}
\]
then \(\inf_{f \in V} c_f \geq c^*\).

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 \(h\) that is orthogonal to \(f\). This makes it possible to restrict \(h\) to the finite-dimensional space spanned by \(f\) in the two other parts. The second step controls the case when \(h\) is small, so that the behaviour of \(N_f\) is given by its quadratic form, and the last step controls the case where \(h\) is far from zero.

**The orthogonal part.** Let \(u\) be the orthogonal projection of \(h\) on \(\text{Span}(f)\). Then
\[
N_f(p, q, h) = N_f(p, q, u) + M_f(p, q, u, h - u)
\]
Assume $\sigma$.

Moreover, the Schur product theorem implies that with $U$ an orthogonal matrix and $D$ a diagonal matrix with positive diagonal coefficients, one can lower bound $M_f(p, q, u, a)$ by the second term of the sum, which leads to

$$M_f(p, q, u, a) \geq \sum_{i,j=1}^{K} ((\Pi'Q')^T G(f + u)\Pi'Q')_{i,j} (Q'^TG(f + u)Q')_{i,j} \langle a_i, a_j \rangle$$

Assume $[\text{Hid}]$ holds for the parameters $(\pi^* + p, Q^* + q, f + u)$, then the matrices $(\Pi'Q')^T G(f + u)\Pi'Q'$ and $Q'^TG(f + u)Q'$ are positive symmetric with respective lowest eigenvalue lower bounded by $(\inf_k (\pi_k^* + p_k)\sigma_k(G(Q^* + q))^2\sigma_K(G(f + u)))$ and $\sigma_K(Q^* + q)^2\sigma_K(G(f + u))$. Therefore, their Hadamard product is positive, and one has

$$((\Pi'Q')^T G(f + u)\Pi'Q')_{i,j} (Q'^TG(f + u)Q')_{i,j} = (DU)^T (DU)$$

with $U$ an orthogonal matrix and $D$ a diagonal matrix with positive diagonal coefficients.
is positive when $q$ is close to $(\mathcal{H}_{\text{det}})$, the nondegeneracy of the quadratic form ensured by $f^{*}$ and equation (6) ensure that this form is positive for $f$ around $(0,0,0)$, which leads to a sum of $(63)^{K}$ possibilities, and at least one of them is replaced. There are $63^{K}$ terms, each of which can be bounded by $\|G(f + u)\|_{\infty}$. Let $c_{1}(\pi^{*} + p, Q^{*} + q, f + u) = (\inf_{k}(\pi_{k}^{*} + p_{k}))^{2}\sigma_{K}(Q^{*} + q)^{4}\sigma_{K}(G(f + u))^{2}$. The application $(p, \pi^{*}, q, Q^{*}, u, f) \mapsto c_{1}(\pi^{*} + p, Q^{*} + q, f + u)$ is continuous and nonnegative, it is positive when $[\text{Hid}]$ holds for the parameters $(\pi^{*} + p, Q^{*} + q, f + u)$, and one has

$$M_{f}(p, q, u, a) \geq c_{1}(\pi^{*} + p, Q^{*} + q, f + u)\|a\|_{2}^{2}.$$  

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

**In the neighborhood of $f^{*}$.** The Taylor expansion of

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

around $(0,0,0)$ leads to a nonnegative quadratic form and no linear part. $[\text{Hdet}]$, $[\text{Hid}]$ and equation (6) ensure that this form is positive for $f = f^{*}$. Let $c_{2}(Q^{*}, \pi^{*}, f)$ be as defined in Section 6.3.2, then $f \mapsto c_{2}(Q^{*}, \pi^{*}, f)$ is continuous and it is positive in the neighborhood of $f^{*}$. Moreover, there exists a positive constant $\eta$ depending on $\|G(f)\|_{\infty}$ such that for all $(p, q, u)$ such that $\|(p, q, u)\|_{f} \leq 1$, one has

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

For instance, $\eta = 4000K^{6}\|G(f)\|_{\infty}^{3}$ works: the terms of order 2 or more in the Taylor expansion of $N_{f}$ are the scalar product of sums of terms of the form $\sum_{i,j,k \in \mathcal{X}} \pi^{*}(i)Q^{*}(i,j)Q^{*}(j,k)f_{i}\otimes f_{j}\otimes f_{k}$ where zero to three of the $f$ may be replaced by $u$, zero to two of the $Q^{*}$ by $q$ and $\pi^{*}$ 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 $\|G(f)\|_{\infty}^{3}(\max\{p(i), q(i,j), \|u_{i}\|_{2} \mid i,j \in \mathcal{X}\})^{r}$ where $r$ is the number of replaced terms. By taking the right permutation of states, the max can be bounded by $\|(p, q, u)\|_{f}$, hence the result.
Then, using \( \|(p, q, h)\|_f^2 = \|(p, q, u)\|_f^2 + \|a\|_2^2 \) leads to
\[
\begin{align*}
N_f(p, q, h) \geq & c_1(Q^* + q, \pi^* + p, f + u) \frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2} \\
+ & c_2(Q^*, \pi^*, f) \frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2} - \eta \frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2}^{1/2} \\
\geq & c_1(Q^* + q, \pi^* + p, f + u) \frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2} \\
+ & c_2(Q^*, \pi^*, f) \frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2} - \eta \sqrt{\frac{\|a\|_2^2}{\|(p, q, u)\|_f^2 + \|a\|_2^2}}.
\end{align*}
\]

Let \( c_0 = \min(c_1/2, c_2) \), then \( c_0 \) is continuous and there exists a continuous function \( (\pi^*, Q^*, f) \mapsto \epsilon(\pi^*, Q^*, f) \) which is positive as soon as [Hid] and [Hdet] hold for \( (\pi^*, Q^*, f) \) and such that
\[
\|(p, q, u)\|_f \leq \epsilon(\pi^*, Q^*, f) \Rightarrow N_f(p, q, h) \geq c_0(Q^*, \pi^*, f).
\]

Thus, there exists positive constants \( \epsilon_0 \) and \( c_{\text{near}} \) depending on \( Q^*, \pi^* \) and \( f^* \) such that
\[
\forall (p, q, h, f) \in (\Delta - \Delta) \times (Q - Q) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K
\]
\[
\text{s.t. } \|(p, q, u)\|_f \leq \epsilon_0 \text{ and } \sum_{k \in X} \|f_k - f_k^*\|_2^2 \leq \epsilon_0^2, \quad N_f(p, q, h) \geq c_{\text{near}}.
\]

Far from \( f^* \).

**Lemma 18** The application
\[
(p, q, u, f) \in (\Delta - \Delta) \times (Q - Q) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K \mapsto N_f(p, q, u)
\]

restricted to the set of \( (p, q, u, f) \) such that \( u \in \text{Span}(f)^K \) is uniformly continuous for the norm \( \| \cdot \|_{\text{tot}} \) defined by
\[
\|(p, q, u, f)\|_{\text{tot}}^2 := \|p\|_2^2 + \|q\|_2^2 + \sum_{k \in X} (\|u_k\|_2^2 + \|f_k\|_2^2).
\]

Thus, by compacity of \( (\Delta - \Delta) \times (Q - Q) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(f))^K \), the application
\[
c_{\text{far}} : f \mapsto \inf_{(p, q, u) \in (\Delta - \Delta) \times (Q - Q) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(f))^K \text{ s.t. } \|(p, q, u)\|_f > \epsilon_0} N_f(p, q, u)
\]
is continuous. Let us now prove that \( c_{\text{far}}(f^*) > 0 \).

Let \( (p_n, q_n, u_n)_n \in ((\Delta - \Delta) \times (Q - Q) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(f^*))^\mathbb{N} \) be a sequence such that \( \|(p_n, q_n, u_n)\|_f > \epsilon_0 \) for all \( n \) and
\[
c_{\text{far}}(f^*) = \lim_n N_{f^n}(p_n, q_n, u_n).
\]
Then by compacity, this sequences converges towards a limit \((p, q, u)\). Necessarily \(\|(p, q, u)\|_{r} \geq \varepsilon_0\). Since [Hid] holds, Theorem 8 shows that \(N_{F_r}(p, q, u) = 0\), which implies \(c_{\text{far}}(f^*) > 0\) by continuity of \(N_{F_r}\). Note that \(c_{\text{far}}(f^*)\) may depend on \(F\) in addition to the parameters \(\pi^*, Q^*\) and \(f^*\).

Thus, by continuity, there exists \(\varepsilon_1 > 0\) such that for all \(f \in F^K\) such that \(\sum_{k \in X} \|f_k - f_k^*\|_2^2 \leq \varepsilon_1^2\), \(c_{\text{far}}(f) \geq c_{\text{far}}(f^*)/2\).

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

\[
\forall (p, q, h, f) \in (\Delta - \Delta) \times (Q - Q) \times (F - F)^K \times F^K
\]

\[
s.t. \quad \|(p, q, u)\|_{r} \geq \varepsilon_0 \quad \text{and} \quad \sum_{k \in X} \|f_k - f_k^*\|_2^2 \leq \varepsilon_1^2, \quad \frac{N_{F_r}(p, q, h)}{\|(p, q, h)\|_{r}} \geq \frac{N_{F_r}(p, q, u)}{\varepsilon_0^2}
\]

\[
\geq \frac{c_{\text{far}}(f^*)}{2C}.
\]

The theorem follows by taking \(c^*(\pi^*, Q^*, f^*, F) = \min\left(\frac{c_{\text{far}}(f^*)}{2C}, c_{\text{near}}\right)\) and the neighborhood containing all \(f \in F^K\) such that \(\sum_{k \in X} \|f_k - f_k^*\|_2^2 \leq \min(\varepsilon_0, \varepsilon_1)^2\). Moreover, \((\pi, Q, f) \mapsto c^*(\pi, Q, f, F)\) is lower bounded by this value in a neighborhood of \((\pi^*, Q^*, f^*)\), so that it can be assumed to be lower semicontinuous.

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

**Proof of Lemma 18**

\[
|N_{F_r}(p, q, u) - N_{F_r}(p', q', u')| = |\|g^{\pi^* + p, Q^* + q, f + u} - g^{\pi^* + p', Q^* + q', f' + u'} - g^{\pi^* + Q^* f'}\|_2^2|
\]

\[
\leq 2\left|\|g^{\pi^* + p, Q^* + q, f + u} - g^{\pi^* + p', Q^* + q', f' + u'}\|_2^2 + 2\|g^{\pi^* + Q^* f'} - g^{\pi^* + Q^* f'}\|_2^2\right|
\]

\[
+ 2\left|\left\langle g^{\pi^* + p', Q^* + q', f' + u'} - g^{\pi^* + Q^* f'}, g^{\pi^* + p, Q^* + q, f + u} - g^{\pi^* + p', Q^* + q', f' + u'}\right\rangle\right|
\]

\[
+ 2\left|\left\langle g^{\pi^* + p', Q^* + q', f' + u'} - g^{\pi^* + Q^* f' + u'}, g^{\pi^* + Q^* f'} - g^{\pi^* + Q^* f'}\right\rangle\right|
\]

Then, using the fact that \(\|g^{\pi^*, Q^* f'} - g^{\pi^*, Q^* f'}\|_2 \leq \sqrt{3}K C_{F,2}^3 \|\pi - \pi^*, Q - Q^*, f - f', 0\|_{\text{tot}}\) (see Lemma 17), that \(\|g^{\pi, Q, f}\|_2 \leq C_{F,2}^3\) (see for instance Lemma 29 of Lehéricy (2016)) and the Cauchy-Schwarz inequality,

\[
|N_{F_r}(p, q, u) - N_{F_r}(p', q', u')| \leq 6K C_{F,2}^6 \|\|p - p', q - q', f + u - f' - u'\|\|_2^2\]
\]

\[
+ 6K C_{F,2}^6 \|\|0, 0, 0, f - f'\|\|_2^2\]
\]

\[
+ 4\sqrt{3}K C_{F,2}^6 \|\|p - p', q - q', f + u - f' - u'\|\|_2^2\]
\]

\[
+ 4\sqrt{3}K C_{F,2}^6 \|\|0, 0, 0, f - f'\|\|_2^2\]
\]

\[
\leq 24K C_{F,2}^6 \|\|p - p', q - q', u - u', f - f'\|\|_2^2\]
\]

which proves the uniform continuity of the application.
Acknowledgments

We are grateful to Elisabeth Gassiat and Claire Lacour for their precious advice. We would also like to thank Augustin Touron for providing us with a R implementation of the spectral algorithm.

References


Luc Lehéricy. Estimation adaptative non paramétrique pour les modèles à chaîne de Markov cachée. Mémoire de M2, Orsay, 2015.