Bernoulli **25**(1), 2019, 464–498
https://doi.org/10.3150/17-BEJ993

Consistent order estimation for nonparametric hidden Markov models

LUC LEHÉRICY

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France. E-mail: luc.lehericy@math.u-psud.fr

We consider the problem of estimating the number of hidden states (the *order*) of a nonparametric hidden Markov model (HMM). We propose two different methods and prove their almost sure consistency without any prior assumption, be it on the order or on the emission distributions. This is the first time a consistency result is proved in such a general setting without using restrictive assumptions such as *a priori* upper bounds on the order or parametric restrictions on the emission distributions. Our main method relies on the minimization of a penalized least squares criterion. In addition to the consistency of the order estimation, we also prove that this method yields rate minimax adaptive estimators of the parameters of the HMM – up to a logarithmic factor. Our second method relies on estimating the rank of a matrix obtained from the distribution of two consecutive observations. Finally, numerical experiments are used to compare both methods and study their ability to select the right order in several situations.

Keywords: hidden Markov model; least squares method; model selection; nonparametric density estimation; order estimation; spectral method

1. Introduction

1.1. Context and motivation

Hidden Markov models (HMM in short) are powerful tools to study time-evolving processes on heterogeneous populations. Nonparametric HMMs – that is, hidden Markov models where the parameters are not restricted to a finite-dimensional space – have proved useful in a wide range of applications, see, for instance, Couvreur and Couvreur [9] for voice activity detection, Lambert, Whiting and Metcalfe [22] for climate state identification, Lefèvre [24] for automatic speech recognition, Shang and Chan [31] for facial expression recognition, Volant et al. [34] for methylation comparison of proteins, Yau et al. [35] for copy number variants identification in DNA analysis.

In practice, the hidden states often have an interpretation in the modelling of the phenomenon. It is thus important to be able to infer the right order in addition to the parameters when dealing with hidden Markov models. However, this task is notoriously difficult: Gassiat and Keribin [17] show that the likelihood ratio statistic is unbounded even in the simple case where one wants to test if a HMM has 1 or 2 hidden states. As far as we know, no consistency result has been proved about order selection for nonparametric HMMs. Even for parametric HMMs, no estimator has been proved to be consistent in a general setting without assuming that an *a priori* upper bound on the order is known beforehand.

Not only is the order estimation useful in order to interpret the model, it is also necessary to ensure stability. This is because over estimating the order causes a loss of identifiability: there are several ways to add one state to a HMM without changing its distribution. The spectral estimators (Anandkumar, Hsu and Kakade [3], de Castro, Gassiat and Le Corff [11]) are especially sensitive to this problem, as shown by Lehéricy [25] and Figure 6: as soon as the HMM becomes close to a HMM with fewer hidden states, the estimators give absurd results. Thus, estimating the right order is crucial for such methods to be effective.

Formally, a hidden Markov model is a markovian process $(X_t, Y_t)_{t\geq 1}$ taking value in $\mathcal{X} \times \mathcal{Y}$. $(X_t)_{t\geq 1}$ is a Markov chain and the observations Y_t depend only on the associated X_t (i.e., the $(Y_t)_{t\geq 1}$ are independent conditionally on $(X_t)_{t\geq 1}$). The states $(X_t)_{t\geq 1}$ are assumed to be hidden, so that one has only access to the observations $(Y_t)_{t\geq 1}$. When the number of hidden states $|\mathcal{X}|$ (which we call the *order* of the HMM) is finite, the model is completely defined by its order, the initial distribution and the transition matrix of the hidden Markov chain, and the possible distributions of an observation Y_t conditionally to the values of its hidden state X_t, which we call the *emission distributions*. The goal of the estimation procedures is to recover these parameters by using only the observations $(Y_t)_{t\geq 1}$.

Up to now, most theoretical results on hidden Markov models dealt with the parametric framework, that is, with a finite number of parameters. However, it is not always possible to restrict the model to such a convenient finite-dimensional space. Theoretical results in the nonparametric framework were only developed recently and do not address the order estimation problem. de Castro, Gassiat and Lacour [10] propose a least squares method, that is, minimax adaptive up to a logarithmic factor. de Castro, Gassiat and Le Corff [11] and Robin, Bonhomme and Jochmans [29] study spectral methods. The latter is also proved to reach the minimax convergence rate but is not adaptive: it requires the regularity of the emission distributions to be known. All these methods require the order of the HMM to be known.

Our work is novel on three points. First, it deals with the nonparametric setting: we need no parametric or regularity assumption on the emission densities. Note that all our results also apply to parametric settings or even to finite observation spaces, since these are just special cases of nonparametric estimation. Secondly, we do not require any *a priori* upper bound on the order, an assumption that is often made in earlier works, both frequentist and bayesian. Finally, our least squares method yields estimators of all model parameters at the same time, without requiring any prior information. Oracle inequalities show that these estimators are rate minimax adaptive up to a logarithmic factor.

1.2. Related works

The first step to obtain theoretical results was to understand when hidden Markov models are identifiable. This challenging issue was only solved a few years ago, see Gassiat, Cleynen and Robin [16] (following Allman, Matias and Rhodes [2] and Hsu, Kakade and Zhang [20]) and with weaker assumptions Alexandrovich, Holzmann and Leister [1]. Both proved that under generic assumptions, the parameters of the HMM can be recovered from the distribution of a finite number of consecutive observations, thus paving the way for guarantees on parameter estimation.

HMM inference is generally decomposed in two parts. The first one is the estimation of the order, and the second one is the estimation of the parameters once the order is known.

From a theoretical point of view, the order estimation problem remains widely open in the HMM framework. One can distinguish two kinds of results. The first kind does not need an *a priori* upper bound on the order, but is only applicable to restrictive cases. For instance, using tools from coding theory, Gassiat and Boucheron [15] introduced a penalized maximum likelihood order estimator for which they prove strong consistency without *a priori* upper bound on the order of the HMM. Nevertheless, their result is restricted to a finite observation space and they have to use heavy penalties that grow as a power of the order. For the special case of Gaussian or Poisson emission distributions, Chambaz, Garivier and Gassiat [8] showed that the penalized maximum likelihood estimator is strongly consistent without any *a priori* upper bound on the order. The second kind of results is more general but requires an *a priori* upper bound of the order just to get weak consistency of order estimators, for penalized likelihood criterion (Gassiat [14]) as well as Bayesian approaches (Gassiat and Rousseau [18], van Havre et al. [33]).

On a practical side, several order estimation methods using penalized likelihood criterion have been studied numerically, see, for instance, Volant et al. [34] when emission distributions are a mixture of parametric densities or Celeux and Durand [7] for parametric HMMs. The latter also introduced cross-validation procedures that aimed for circumventing the lack of independance of the observations. In the case of nonparametric HMMs, Langrock et al. [23] studied a method using P-splines with a custom penalization.

Then comes the question of estimating the parameters of the HMM once its order is known. In the parametric setting, the asymptotic behaviour of the maximum likelihood estimator is rather well understood (see, for instance, Bickel et al. [5] or Douc et al. [13] using techniques from Le Gland and Mevel [27]), but so far the question of its nonasymptotic behaviour remains open. Hsu, Kakade and Zhang [20] and Anandkumar, Hsu and Kakade [3] proposed a spectral method for parametric HMMs based on joint diagonalization of a set of matrices and controlled its nonasymptotic error. Robin, Bonhomme and Jochmans [29] and de Castro, Gassiat and Le Corff [11] extended this method to the nonparametric setting, and de Castro, Gassiat and Lacour [10] used the latter to obtain an estimator of the transition matrix of the hidden chain for a least squares estimator of the emission densities, that is, minimax adaptive up to a logarithmic factor. Our least squares estimation method is a generalization of their procedure that is able to deal with all parameters at once and does not require auxiliary estimators.

1.3. Contribution

The aim of our paper is twofold. First, we introduce two estimators of the order for nonparametric HMMs and show that both converge almost surely to the right order under minimal assumptions. Second, we numerically assess their ability to select the right order and compare their efficiency.

Our first and main method is the penalized least squares estimator. This method is based on estimating the projection of the emission distributions onto a family of nested parametric subspaces. Our results hold for any Hilbert space, including parametric sets of emission densities and finite observation spaces. Then, for each subspace and for each possible value K of the order, we look for the HMM with K hidden states and with emission distributions in the chosen subspace that matches the observations "best" – where "best" means minimizing the empirical equivalent of an $\mathbf{L}^2$ distance. This step provides an empirical distance between the observations

and the model, which is then penalized in order to counterbalance the overfitting phenomenon that occurs when considering large models. Our first main result is that for a suitable choice of the penalty, choosing the model (i.e., the order and the subspace) which minimizes this penalized distance leads to a strongly consistent estimator of the order, see Corollary 5.

In addition, this method also provides estimators of the other parameters of the HMM for free, by taking the parameters of the HMM corresponding to the selected model. We prove an oracle inequality on the $\mathbf{L}^2$ risk of these estimators, which shows that they achieve the minimax adaptive rate of convergence, up to a logarithmic factor, see Theorem 10 and Corollary 11.

Our second estimator comes from spectral methods. Just like for our least squares procedure, we consider a nested family of parametric subspaces of a Hilbert space. Let us choose one of them, and denote by $(\varphi_a)_a$ an orthonormal basis of this subspace. Then, consider the matrix $\mathbf{N}$ defined by

$$\mathbf{N}(a,b) := \mathbb{E}\big[\varphi_a(Y_1)\varphi_b(Y_2)\big].$$

This matrix contains the coordinates of the density of (Y_1, Y_2) in the orthonormal basis $(\varphi_a \otimes \varphi_b)_{a,b}$. It is proved in Section 4 that the rank of $\mathbf{N}$ is exactly equal to the order of the HMM as soon as the subspace is large enough. Therefore, finding its rank means finding the number of hidden states. However, in practice, one only has access to an empirical version of this matrix. The difficulty comes from the fact that this noisy version will almost surely have full rank. Thus, the key point is to recover the order of the true matrix given its empirical (full rank) counterpart. We achieve this by thresholding the spectrum of the empirical matrix. Notice that other methods exist to estimate the rank of a matrix based on a noisy observation, see, for instance, Kleibergen and Paap [21] and references therein. Unfortunately, most can not be applied directly to our setting since they require an invertibility condition on the covariance matrix of the matrix entries. The CRT statistics from Robin and Smith [30] is a notable exception, however their test of rank also requires to calibrate a tuning parameter in order to be weakly consistent.

Then, we run an implementation of these two methods and compare their efficiency on simulated data. The difficulty at this stage comes from the fact that both methods involve a tuning parameter: the constant of the penalty for the least squares method and the constant of the threshold for the spectral method. This is a common issue that appears in every model selection method in one form or another, and many heuristics have been proposed to circumvent this difficulty.

For the least squares estimator, we compare two methods which have been both proved to be theoretically valid in simple cases and empirically validated in a large variety of situations: the slope heuristics (see for instance Baudry, Maugis and Michel [4] and references therein) and the dimension jump heuristics (introduced and proved to lead to an optimal penalization in the gaussian model selection framework by Birgé and Massart [6]). Both behave well with our estimator and lead to a satisfying calibration of the penalty.

For the spectral estimator, we introduce a custom heuristics based on the fact that the smallest singular values of the empirical version of the matrix $\mathbf{N}$ decrease in a simple manner. It is thus possible to calibrate an entirely data-driven threshold to distinguish "significant" singular values – that is, the ones corresponding to non-zero singular values of the real $\mathbf{N}$ – from noise.

The numerical validation shows that our least squares method performs well in almost any situation. It is able to select the right order accurately with notably fewer observations than the spectral estimator, and is easier to calibrate. On the other hand, the spectral method is very fast,

which allows to take more observations into account. This allows to obtain satisfying estimators in a short amount of time.

Regarding the inference of the other parameters, our least squares estimator offers several advantages when compared to previous methods. First, it does not need a preliminary estimation of the transition matrix or of the order, unlike de Castro, Gassiat and Lacour [10] who used the transition matrix given by spectral estimators. Nevertheless, our method still reaches the adaptive minimax convergence rate for the estimation of the emission densities, up to a logarithmic factor. This is especially useful to avoid the cases where their auxiliary estimator fails. For instance, the spectral method that de Castro, Gassiat and Lacour [10] used is unreliable when the order is over estimated or where the states are almost linearly dependent, see, for instance, Lehéricy [25] or Figure 6. Then, our least squares method is robust to an overestimation of the order, both theoretically and numerically, thanks to the iterative initialization procedure that we introduce. This initialization method consists in using estimators from smaller models as initial point for the minimization algorithm in order to avoid getting stuck in suboptimal local extrema. We believe it can be of practical interest since it produces robust estimators and can also be used in other settings, for instance as initialization for expectation maximization algorithm for maximum likelihood estimators.

1.4. Outline of the paper

Our paper is organized as follows.

Section 2 is devoted to the notations, the model and the assumptions.

Our main procedure, the penalized least squares method, is introduced in Section 3. We first state an identifiability proposition which we use to prove strong consistency of the estimator of the order. This is done in two steps. First, we control the probability to underestimate the order with Proposition 1. This gives an exponential bound on the probability of error, see Theorem 3. Second, we control the probability to overestimate the order, see Theorem 4. For this, we introduce a general condition on the penalty, which we use to prove a bound with polynomial rate of decrease on the probability of error, and illustrate how to easily satisfy this condition. Finally, we state oracle inequalities on the estimators of the density of L consecutive observations and on the parameters of the hidden Markov model under a generic assumption, see Theorem 10 and Corollary 11, which shows that they reach the minimax convergence rate up to a logarithmic factor.

In Section 4, we introduce the spectral algorithm and propose a strongly consistent estimator of the order. This is done by thresholding the spectrum of the empirical version of the matrix $\mathbf{N}$, which describes the projection of the distribution of two observations onto an orthonormal basis, see Theorem 13.

In Section 5, we propose practical algorithms to apply both methods and compare them. First, we set the parameters on which we will test both procedures. Second, we compare their results and discuss their performance. Last, we introduce and discuss the heuristics we used to practically implement both methods.

Our main technical result, Lemma 16, can be found at the beginning of Section 6. It is used extensively for both the consistency of the estimator of the order and the oracle inequalities on the HMM parameters. The rest of this section is dedicated to the proofs of the results.

Supplement A (Lehéricy [26]) is organized as follows. Appendix A contains the spectral algorithm from de Castro, Gassiat and Le Corff [11] and de Castro, Gassiat and Lacour [10] that we use in our simulations. Appendix B gathers the proofs of Section 3.4, which deals with the oracle inequalities for the least squares method. Finally, Appendix C contains the proof of Lemma 16, and Appendix D contains miscellaneous lemmas and proofs.

2. Definitions and assumptions

We will use the following notations throughout the paper.

- $\mathbb{N}^* = \{1, 2, \dots\}$ is the set of positive integers.
- For $k \in \mathbb{N}^*$, $[k]$ is the set $\{1, \dots, k\}$.
- If f_1 and f_2 are two functions, we denote by $f_1 \otimes f_2$ their tensor product, defined by $f_1 \otimes f_2(x_1, x_2) = f_1(x_1) f_2(x_2)$.
- $\mathrm{Span}(\Phi)$ is the linear space spanned by the family Φ.
- Proj_E is the orthogonal projection operator onto a linear space E.
- If E_1 and E_2 are two linear spaces, we denote by $E_1 \otimes E_2$ their tensor product, that is, the linear space spanned by the tensor products of their elements: $E_1 \otimes E_2 = \mathrm{Span}(f_1 \otimes f_2 | f_1 \in E_1, f_2 \in E_2)$.
- $\Delta_K = \{\pi \in [0,1]^K | \sum_{k=1}^K \pi_k = 1\}$ is the simplex in dimension K. It will be seen as the set of probability measures on a finite set of size K.
- $\mathcal{Q}_K \subset \mathbb{R}^{K \times K}$ is the set of irreducible transition matrices of size K.
- Id_K is the identity matrix of size K.
- $\mathbf{L}^2(A, \nu)$ is the Hilbert space of square integrable functions on A with respect to the measure ν.
- The notation $C \equiv C(a, b, \dots)$ for a constant C will mean that the value of C depends on the specified parameters $a, b, \dots$ For several constants depending on the same parameters, we will write $(C, D) \equiv (C, D)(a, b, \dots)$.

In the following, L is a positive integer which will denote the number of consecutive observations used for the estimation procedure.

2.1. Hidden Markov models

Let $(X_j)_{j \geq 1}$ be a Markov chain with finite state space $\mathcal{X}$ of size K^* with transition matrix $\mathbf{Q}^*$ and initial distribution π^*. Without loss of generality, we can set $\mathcal{X} = [K^*]$.

Let $(Y_j)_{j \geq 1}$ be random variables on a measured space $(\mathcal{Y}, \mu)$ with μ σ-finite such that conditionally on $(X_j)_{j \geq 1}$ the Y_j's are independent with a distribution depending only on X_j. Let ν_k^* be the distribution of Y_j conditionally to $\{X_j = k\}$. Assume that ν_k^* has density $f_k^* \in \mathbf{L}^2(\mathcal{Y}, \mu)$ with respect to μ. We call $(\nu_k^*)_{k \in \mathcal{X}}$ the *emission distributions* and $\mathbf{f}^* = (f_1^*, \dots, f_{K^*}^*)$ the *emission densities*.

Then $(X_j, Y_j)_{j \geq 1}$ is a hidden Markov model with parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, K^*)$. The hidden chain $(X_j)_{j \geq 1}$ is assumed to be unknown, so that the estimator only has access to the observations $(Y_j)_{j \geq 1}$.

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

$$g^{\pi,\mathbf{Q},\mathbf{f},K} = \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1}, k_i) \bigotimes_{i=1}^{L} f_{k_i}.$$

When π is a probability distribution on $[K]$, $\mathbf{Q}$ a $K \times K$ transition matrix and $\mathbf{f}$ a K-uple of probability densities, $g^{\pi,\mathbf{Q},\mathbf{f},K}$ is the density of the first L observations of a HMM with parameters $(\pi, \mathbf{Q}, \mathbf{f}, K)$.

For the sake of readability, we will drop the dependence in K in the following and write $g^{\pi,\mathbf{Q},\mathbf{f}}$ instead of $g^{\pi,\mathbf{Q},\mathbf{f},K}$. Moreover, if $\mathbf{Q}$ is irreducible with stationary distribution π, we simply write $g^{\mathbf{Q},\mathbf{f}}$, and we write the true density $g^* := g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}$.

2.2. Assumptions

Let $\mathcal{F}$ be a subset of $\mathbf{L}^2(\mathcal{Y}, \mu)$ and $(\mathfrak{P}_M)_{M \in \mathcal{M} \subset \mathbb{N}}$ be a sequence of nested subspaces of $\mathbf{L}^2(\mathcal{Y}, \mu)$ such that $\mathfrak{P}_M$ has dimension M for all $M \in \mathcal{M}$ and their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mu)$. $(\mathfrak{P}_M)_{M \in \mathcal{M}}$ will be the subspaces onto which the projections of the emission densities will be estimated.

We will need the following assumptions.

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

[HidA] $\mathbf{Q}^*$ is invertible, $L \geq 3$ and the family $\mathbf{f}^*$ is linearly independent;

[HidB] $\mathbf{Q}^*$ is invertible, $L \geq (2K^* + 1)((K^*)^2 - 2K^* + 2) + 1$ and the emission densities $(f_k^*)_{k \in \mathcal{X}}$ are all distinct;

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

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

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

The ergodicity assumption in **[HX]** is completely 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. In order to simplify the proofs, we assume the Markov chain to be stationary. One can check that our results are essentially the same when the initial distribution is not the stationary one.

[HidA] appears in spectral methods, with the hypothesis that $\pi^* > 0$ elementwise, see, for instance, Hsu, Kakade and Zhang [20]. **[HidA]** and **[HidB]** also appear in identifiability issues, possibly combined with the stationarity hypothesis, see Alexandrovich, Holzmann and Leister [1] and Gassiat, Cleynen and Robin [16]. Note that the condition on L in **[HidB]** only involves the real order K^*.

Even though **[HidB]** appears less restrictive than **[HidA]** about the emission densities, it is delicate to use here. The problem lies in the condition on the number of consecutive observations L. For **[HidB]**, one has to take L larger than an increasing function of the order, so it requires to have an *a priori* upper bound on the order to choose L. This is less interesting than **[HidA]**, which can work without prior bound since it only requires $L = 3$ for any value of the order.

3. Least squares estimation

In this section, we introduce our penalized least squares estimator and study its asymptotic properties.

3.1. Approximation spaces and estimators

We want to estimate the density of L consecutive observations g^* by minimizing the quadratic loss $t \mapsto \|t - g^*\|_2^2 - \|g^*\|_2^2$. We thus take the corresponding empirical loss

$$\gamma_n(t) = \|t\|_2^2 - \frac{2}{n}\sum_{s=1}^{n} t(Z_s),$$

where $Z_s = (Y_s, \dots, Y_{s+L-1})$ for an observation sequence $(Y_t)_{1\le t\le n+L-1}$ of length $n+L-1$ coming from a single HMM $(X_t, Y_t)_{t\ge 1}$.

Define for all $K \in \mathbb{N}^*$, $M \in \mathcal{M}$:

$$S_{K,M} := \left\{ g^{\mathbf{Q},\mathbf{f}}, \mathbf{Q} \in \mathcal{Q}_K, \mathbf{f} \in (\mathcal{F} \cap \mathfrak{P}_M)^K \right\},$$
$$S_K := \left\{ g^{\mathbf{Q},\mathbf{f}}, \mathbf{Q} \in \mathcal{Q}_K, \mathbf{f} \in \mathcal{F}^K \right\},$$

where $\mathcal{F}$ and $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ are defined in Section 2.2. In the following, we will always implicitly consider $M \in \mathcal{M}$.

Comment. For all $M \in \mathcal{M}$, $(S_{K,M})_{K\ge 1}$ is a sequence of nested subspaces. Indeed, it is possible to add one state to any hidden Markov model without changing anything to the distribution of L consecutive observations while keeping the same emission densities and ensuring that the transition matrix remains irreducible (see function `DUPL` of Algorithm 2 in Appendix A for an example). The same holds for $(S_K)_{K\ge 1}$.

Likewise, for all $K \in \mathbb{N}^*$, $(S_{K,M})_{M\in\mathcal{M}}$ is a sequence of nested subspaces.

For all K and M, we define the corresponding estimators

$$\hat{g}_{K,M} = g^{\hat{\mathbf{Q}}_{K,M}, \hat{\mathbf{f}}_{K,M}} \in \operatorname*{arg\,min}_{t\in S_{K,M}} \gamma_n(t),$$

where we dropped the dependency in n for ease of notation. Then, we select the parameters using the penalized empirical loss:

$$(\hat{K}_{\text{l.s.}}, \hat{M}) \in \operatorname*{arg\,min}_{K\le n, M\le n} \left\{ \gamma_n(\hat{g}_{K,M}) + \text{pen}(n, M, K) \right\}$$

which leads to the estimators

$$\hat{g} := \hat{g}_{\hat{K}_{\text{l.s.}}, \hat{M}},$$

$$\hat{\mathbf{Q}} := \hat{\mathbf{Q}}_{\hat{K}_{\text{l.s.}},\hat{M}},$$
$$\hat{\mathbf{f}} := \hat{\mathbf{f}}_{\hat{K}_{\text{l.s.}},\hat{M}}.$$

3.2. Underestimation of the order

Note that the distribution of the HMM remains unchanged under permutation of the hidden states. We will therefore use a pseudo-distance d_{perm} that is invariant by permutation on the set of parameters.

We define it as follows. Let $K \geq 1$, $\pi_1, \pi_2 \in \Delta_K$, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ transition matrices of size K, $\mathbf{f}_1, \mathbf{f}_2 \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$. Let $\mathfrak{S}(\mathcal{X})$ be the set of permutations of $\mathcal{X}$. For all $\tau \in \mathfrak{S}(\mathcal{X})$, define the swapped parameters $\tau\pi_1$, $\tau\mathbf{Q}_1$ and $\tau\mathbf{f}_1$ by

$$(\tau\pi_1)(k) := \pi_1\big(\tau(k)\big),$$
$$(\tau\mathbf{Q}_1)(k,l) := \mathbf{Q}_1\big(\tau(k), \tau(l)\big),$$
$$(\tau\mathbf{f}_1)_k := f_{1,\tau(k)}$$

and finally

$$d_{\text{perm}}\big((\pi_1, \mathbf{Q}_1, \mathbf{f}_1), (\pi_2, \mathbf{Q}_2, \mathbf{f}_2)\big) := \inf_{\tau \in \mathfrak{S}(\mathcal{X})} \Bigg(\|\tau\pi_1 - \pi_2\|_2^2 + \|\tau\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2 + \sum_{k=1}^{K} \big\|(\tau\mathbf{f}_1)_k - f_{2,k}\big\|_2^2 \Bigg)^{1/2}.$$

The following properties will be of use to prove the consistency of the order estimator, but we think it can also be of independent interest to better understand the identifiability of the model. The first one is a generalization of previous identifiability results from Alexandrovich, Holzmann and Leister [1], Gassiat, Cleynen and Robin [16], de Castro, Gassiat and Lacour [10].

Proposition 1. *Let $K \geq 1$, $\pi \in \Delta_K$ such that $\pi_k > 0$ for all $k \in \mathcal{X}$, $\mathbf{Q}$ transition matrix of size K and $\mathbf{f} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$ such that* **[HidA]** *or* **[HidB]** *hold for the order K. Then, for all $K' \geq 1$, for all $\pi' \in \Delta_{K'}$, for all transition matrix $\mathbf{Q}'$ of size K' and all $\mathbf{f}' \in (\mathbf{L}^2(\mathcal{Y}, \mu))^{K'}$, the following holds*:

$$\big(g^{\pi,\mathbf{Q},\mathbf{f}} = g^{\pi',\mathbf{Q}',\mathbf{f}'} \text{ and } K' \leq K\big)$$
$$\Rightarrow \quad \big(K = K' \text{ and } d_{\text{perm}}\big((\pi, \mathbf{Q}, \mathbf{f}), (\pi', \mathbf{Q}', \mathbf{f}')\big) = 0\big).$$

Comment. This property does not require two assumptions that appear in Alexandrovich, Holzmann and Leister [1] and Gassiat, Cleynen and Robin [16]: that $\mathbf{f}$ is a family of probability densities and that the Markov chain is stationary.

In particular, the fact that $\mathbf{f}$ may not be a family of probability densities is crucial in the proof of Corollary 2, which is necessary to prove the strong consistency of the estimator of the order.

Proof. Assume **[HidA]**. The spectral algorithm from de Castro, Gassiat and Le Corff [11] applied on the linear space spanned by both sets of densities allows to retrieve the order from two consecutive observations and the parameters from three consecutive observations. Their proof works when the emission densities are not probability densities and when the chain is not stationary.

Assume **[HidB]**. A careful reading of the proofs of Alexandrovich, Holzmann and Leister [1] shows that their result can be extended to general observation spaces and do not require the measures to be probabilities. □

The second property is the following corollary, which states that the $\mathbf{L}^2$ distance between the actual model and the models where the order is underestimated is positive. It is worth noting that we do not need $\mathcal{F}$ to be compact.

Corollary 2. *Assume* **[*HX*]**, (**[*HidA*]** *or* **[*HidB*]**) *and* **[*HF*]** *hold. Then, for all* $K < K^*$:

$$d_K := \inf_{t \in S_K} \left\| t - g^* \right\|_2 > 0$$

Proof. Proof in Section 6.2. □

Our first theorem shows that the probability to underestimate the order decreases exponentially with the number of observations. This comes from Corollary 2: since the empirical criterion converges to the $\mathbf{L}^2$ distance (plus some constant that does not depend on the model), the penalized error will eventually become larger for orders under K^* than for orders over K^*, which means that we won't underestimate the real order. The exponential decrease rate brings to mind the one studied in Gassiat and Boucheron [15]: in both cases, the exponents involve the distance between the actual model and models with underestimated orders, as can be seen in our proof.

Theorem 3. *Assume* **[*HX*]**, (**[*HidA*]** *or* **[*HidB*]**) *and* **[*HF*]** *hold. There exists positive constants* $\rho \equiv \rho(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *and* $\beta \equiv \beta(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, (d_K)_{K<K^*}, L)$ *such that the following holds.*

Assume that

$$\forall n, \forall M, \forall K, \qquad \operatorname{pen}(n, M, K) \geq \rho\big(MK + K^2 - 1\big)\frac{\log(n)}{n},$$

and

$$\forall M, \forall K, \qquad \operatorname{pen}(n, M, K) \underset{n\to\infty}{\longrightarrow} 0$$

then there exists n_0 *such that for all* $n \geq n_0$,

$$\mathbb{P}\big(\hat{K}_{\text{l.s.}} < K^*\big) \leq e^{-\beta n}.$$

Proof. Proof in Section 6.3. □

3.3. Overestimation of the order and consistency

Our second theorem controls the probability to overestimate the order. It consists in overpenalizing large models so that the estimated order remains small.

We will need the following technical condition on the penalty:

Condition ([Hpen](α, ρ)). The penalty function pen satisfies

$$\exists n_1, \forall n \geq n_1, \forall M \leq n, \forall K \leq n \text{ s.t. } K > K^*,$$
$$\operatorname{pen}(n, M, K) - \operatorname{pen}\big(n, M, K^*\big) \geq \rho\big(MK + K^2 - 1\big)\frac{\log(n)}{n} + \alpha\frac{\log(n)}{n}.$$

We can now state the theorem and its corollary proving the strong consistency of our estimator of the order. Note that it does not require any identifiability assumption. In particular, this means that in a non-identified situation, the estimator will select a representation of the HMM with minimal possible number of hidden states.

Theorem 4. *Assume* **[HX]** *and* **[HF]** *hold. There exists positive constants* $(\rho, \beta) \equiv (\rho, \beta)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that the following holds.*

Assume **[Hpen]**(α, ρ) *holds for some* $\alpha \geq 0$, *then there exists* n_0 *such that for all* $n \geq n_0$,

$$\mathbb{P}\big(\hat{K}_{\text{l.s.}} > K^*\big) \leq n^{-\beta\alpha}.$$

Proof. Proof in Section 6.3. □

Corollary 5. *Assume* **[HX]**, **[HF]** *and* (**[HidA]** *or* **[HidB]**) *hold. There exists positive constants* $(\rho, \beta) \equiv (\rho, \beta)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that the following holds.*

Assume that the penalty function satisfies

$$\begin{cases} \forall n, \forall M \leq n, \forall K \leq n, & \operatorname{pen}(n, M, K) \geq \rho\big(MK + K^2 - 1\big)\dfrac{\log(n)}{n}, \\ \forall M, \forall K, & \operatorname{pen}(n, M, K) \underset{n \to +\infty}{\longrightarrow} 0 \end{cases}$$

and [**Hpen**]$(\alpha/\beta, \rho)$ *holds for some* $\alpha > 1$, *then*

$$\mathbb{P}\big(\hat{K}_{\text{l.s.}} \neq K^*\big) = O\big(n^{-\alpha}\big).$$

In particular, $\hat{K}_{\text{l.s.}} \longrightarrow K^*$ *almost surely.*

Let us comment on the condition [**Hpen**] when using a penalty of the form $\operatorname{pen}(n, M, K) = C(MK + K^2 - 1)\log(n)/n$ where C may depend on n.

- If one has an *a priori* bound on the order, that is, if $K^* \leq K_0$ for some known K_0, then direct computations show that for all α, ρ, there exists $C \geq 0$ depending on K_0 (for instance, $C = 2\rho(1 + K_0^2 \vee \frac{\alpha}{\rho})$ works) such that [**Hpen**](α, ρ) holds for all $K^* \leq K_0$ (instead of $K \leq n$).

This means that if one has an *a priori* bound K_0 on the order, then by taking a constant C large enough and $\hat{K}_{\text{l.s.}} \leq K_0$, the estimator $\hat{K}_{\text{l.s.}} > K^*$ is almost surely consistent.

- If one does not have an *a priori* bound on K^*, taking a constant C does not allow to get [**Hpen**](α, ρ) for all possible K^*, which means we can't apply Corollary 5. However, by taking C as a sequence indexed by n that tends to infinity, we get that for all K^* and α, ρ, [**Hpen**](α, ρ) holds. This implies consistency with polynomial decrease of the probability of error, at the cost of overpenalizing.

  Overpenalizing is actually necessary if one wants to satisfy [**Hpen**] for all K^*. This is stated in the following proposition.

Proposition 6. *Let* $\rho > 0$ *and* pen *be a positive penalty such that for all* K^*, [**Hpen**]$(0, \rho)$ *holds, then there exists a sequence* $(u_n)_{n\geq 1} \longrightarrow \infty$ *such that for all* $n \geq 1$, $M \leq n$ *and* $K \leq n$, $\text{pen}(n, M, K) \geq u_n(MK + K^2 - 1)\log(n)/n$.

Proof. Proof in Appendix D.1 of Lehéricy [26]. □

3.4. Oracle inequalities

The results and proofs of this section use similar techniques to the ones in de Castro, Gassiat and Lacour [10]. While they focus on the estimation of the emission densities when the order is known and when one has a preliminary estimator of the transition matrix, our results hold when both the order and the transition matrix are estimated along with the emission densities. A first consequence is our penalty, which now depends on the order K. We also show that not knowing the order does not change the convergence rates for the density of several observations and only adds a logarithmic factor when estimating the emission densities.

Our first result is an oracle inequality on the density of L consecutive observations for the least squares estimator.

Theorem 7. *Assume* [***HX***] *and* [***HF***] *hold. Then there exists positive constants* $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that if the penalty satisfies*

$$\forall n, \forall M \leq n, \forall K \leq n, \qquad \text{pen}(n, M, K) \geq \rho\big(MK + K^2 - 1\big)\frac{\log(n)}{n}$$

then for all $n \geq n_0$, *for all* $x > 0$, *it holds with probability larger than* $1 - e^{-x}$ *that*

$$\big\|\hat{g} - g^*\big\|_2^2 \leq 4 \inf_{K\leq n, M\leq n} \big\{\big\|g^*_{K,M} - g^*\big\|_2^2 + \text{pen}(n, M, K)\big\} + 4A\frac{x}{n},$$

where $g^*_{K,M}$ *is the orthogonal projection of* g^* *onto* $\text{Span}(S_{K,M})$.

Proof. Proof in Section 6.4. □

Comment. The constant 4 before the infimum can be replaced by any constant $\kappa > 1$, at the cost of changing the constants n_0, ρ and A.

We would like to deduce an oracle inequality on the parameters of the HMM from this result. Using Cauchy-Schwarz inequality, it is easy to upper bound the error on the density g^* by the error on the parameters: for all probability distributions π_1 and π_2 on $[K]$, for all transition matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ of size K and for all $\mathbf{f}_1, \mathbf{f}_2 \in \mathcal{F}^K$,

$$\left\| g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2} \right\|_2 \leq C^L_{\mathcal{F},2} \sqrt{LK} d_{\text{perm}}\big((\pi_1, \mathbf{Q}_1, \mathbf{f}_1), (\pi_2, \mathbf{Q}_2, \mathbf{f}_2)\big) \tag{1}$$

as soon as **[HF]** holds. The proof of this equation is detailed in Appendix B.1 of Lehéricy [26].

Thus, all we need to deduce an oracle inequality on the parameters is to lower bound the error on g^* by the error on the parameters. Let $\mathfrak{C} \subset \mathbb{R}^{K^*} \times \mathbb{R}^{K^* \times K^*} \times \mathbb{R}^{K^* \times K^*}$ be the set of parameters (p, q, A) such that

$$\begin{cases} \forall i \in \mathcal{X}, & \displaystyle\sum_{j \in \mathcal{X}} q(i, j) = 0, \\ \forall j \in \mathcal{X}, & \displaystyle\sum_{i \in \mathcal{X}} A(i, j) = 0. \end{cases} \tag{2}$$

Note that $\mathfrak{C}$ can be identified with the set

$$\begin{aligned} \mathfrak{C}_{\text{red}} &:= \left\{ \left((p_i)_{i \geq 2}, \big(q(i, j)\big)_{i,j \geq 2}, \big(A(i, j)\big)_{i \geq 2, j} \right) | (p, q, A) \in \mathfrak{C} \right\} \\ &= \mathbb{R}^{K^*-1} \times \mathbb{R}^{K^* \times (K^*-1)} \times \mathbb{R}^{K^* \times (K^*-1)}. \end{aligned}$$

These assumptions are natural since they are necessary (but not sufficient) to ensure that if $(p, q, A) \in \mathfrak{C}$ and π is a probability distribution, $\mathbf{Q}$ a transition matrix and $\mathbf{f}$ a vector of probability densities, then $\pi + p$ is also a probability distribution, $\mathbf{Q} + q$ a transition matrix and $\mathbf{f} + A\mathbf{f}$ a vector of probability densities.

The first step in order to get a lower bound along the same lines as equation (1) is to control the behaviour of the difference near the true parameters, which comes down to proving that the quadratic form M derived from the second-order expansion of

$$\mathfrak{N} : (p, q, A) \in \mathbb{R}^{K^*} \times \mathbb{R}^{K^* \times K^*} \times \mathbb{R}^{K^* \times K^*} \longmapsto \left\| g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}} \right\|_2^2$$

is positive definite on $\mathfrak{C}$ for $(\pi, \mathbf{Q}, \mathbf{f}) = (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. One can write the coefficients of the matrix of this quadratic form as polynomials in the coefficients of π, $\mathbf{Q}$ and of the Gram matrix $G(\mathbf{f}) := (\langle f_i, f_j \rangle)_{i,j \in \mathcal{X}}$. However, this matrix may not be invertible: one has to consider its restriction to the space $\mathfrak{C}$, which is equivalent to considering the quadratic form $M_{\mathfrak{C}}$ defined on $\mathfrak{C}_{\text{red}}$ by the second-order expansion of $x \in \mathfrak{C}_{\text{red}} \longmapsto \mathfrak{N}(I_{\mathfrak{C}}(x))$ where $I_{\mathfrak{C}}$ is the natural linear injection from $\mathfrak{C}_{\text{red}}$ to $\mathfrak{C}$ (note that $I_{\mathfrak{C}}$ is bijective and bicontinous under **[HF]**). Since the quadratic form $M_{\mathfrak{C}}$ is always nonnegative, we only need its determinant to be non zero in order for the quadratic form M to be positive definite on $\mathfrak{C}$.

Thus, let H be determinant of the matrix of this quadratic form. H is also a polynomial in the coefficients of π, $\mathbf{Q}$ and $G(\mathbf{f})$. The following lemma shows that there exists some parameters π, $\mathbf{Q}$ and $\mathbf{f}$ satisfying the conditions for which H is not zero.

Lemma 8. *There exists some parameters* $(\pi, \mathbf{Q}, \mathbf{f})$ *satisfying the conditions* **[*HX*]** *and* **[*HidA*]** *such that* $H(\pi, \mathbf{Q}, G(\mathbf{f})) \neq 0$.

Proof. Proof in Section 6.5. □

What should be retained from this lemma is that H is a polynomial which is not identically zero on the set of parameters satisfying the identifiability conditions. This means that one can generically assume it to be different from zero, which corresponds to the assumption

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

Since we assumed π^* to be the stationary distribution of $\mathbf{Q}^*$, its coefficients – and by extension H – can be expressed as a rational function of the coefficients of $\mathbf{Q}^*$. Taking H_1 as the numerator of the rational function deduced from H, one gets another polynomial in the coefficients of $\mathbf{Q}^*$ and $G(\mathbf{f}^*)$ which is also non-zero. Thus, the following assumption – which we will need to lower bound the error on the density g^* by the error on the parameters – is generically satisfied.

[HdetStat] $H_1(\mathbf{Q}^*, G(\mathbf{f}^*)) \neq 0$.

Note that **[Hdet]** and **[HdetStat]** are equivalent under the assumption **[HX]**.

Theorem 9. *Assume* ***[HidA]*** *and* ***[Hdet]*** *hold. Then there exists a positive constant* $c(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ *such that for all* $\pi \in \Delta_{K^*}$, *for all transition matrix* $\mathbf{Q}$ *of size* K^* *and for all* $\mathbf{h} \in \mathcal{F}^{K^*}$ *such that* $\int h_i \, d\mu = 1$ *for all* $i \in [K^*]$,

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

Proof. Proof in Appendix B.2 of Lehéricy [26]. □

The following theorem is a direct consequence of the above results. It provides an oracle inequality on the parameters conditionally to the fact that the order has been correctly estimated.

Theorem 10. *Assume* ***[HX]***, ***[HidA]***, ***[HF]*** *and* ***[Hdet]*** *hold. Also assume that for all* $f \in \mathcal{F}$, $\int f \, d\mu = 1$.

Then there exists positive constants $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that if the penalty satisfies*

$$\forall n, \forall M \leq n, \forall K \leq n, \qquad \mathrm{pen}(n, M, K) \geq \rho\left(MK + K^2 - 1\right)\frac{\log(n)}{n}$$

then for all $n \geq n_0$, *for all* $x > 0$, *conditionally to* $\{\hat{K}_{\mathrm{l.s.}} = K^*\}$, *with probability larger than* $1 - e^{-x}$:

$$d_{\mathrm{perm}}\left((\hat{\pi}, \hat{\mathbf{Q}}, \hat{\mathbf{f}}), \left(\pi^*, \mathbf{Q}^*, \mathbf{f}^*\right)\right) \leq \frac{4C_{\mathcal{F},2}^L\sqrt{LK^*}}{c(\mathbf{Q}^*, \mathbf{f}^*)} \times \left[\inf_{M \leq n}\left\{\sum_{k=1}^{K^*}\left\| f_{M,k}^* - f_k^* \right\|_2^2 + \mathrm{pen}\left(n, M, K^*\right)\right\} + A\frac{x}{n}\right],$$

where $f_{M,k}^*$ *is the orthogonal projection of* f_k^* *onto* $\mathfrak{P}_M$.

It is now possible to get the convergence rate of the estimators of the parameters. In order to take the event where $\hat{K}_{\text{l.s.}} \neq K^*$ into account, we agree that the distance between the parameters of two HMMs with different orders is bounded by some constant C_{err}. Note that C_{err} could even be taken as a power of n without changing anything to our result.

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

$$\forall n, \forall M \leq n, \forall K \leq n, \qquad \operatorname{pen}(n, M, K) = \big(MK + K^2 - 1\big)\frac{\log(n)^2}{n}.$$

Then there exists a positive constant $A \equiv A(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that for all* $\beta > 1$, *there exists a positive constant* $n_0 \equiv n_0(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L, \beta)$ *such that for all* $n \geq n_0$ *and for all* $C_{\text{err}} > 0$,

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\hat{K}\neq K^*}C_{\text{err}} + \mathbf{1}_{\hat{K}=K^*}d_{\text{perm}}\big((\hat{\pi},\hat{\mathbf{Q}},\hat{\mathbf{f}}),(\pi^*,\mathbf{Q}^*,\mathbf{f}^*)\big)\big] \\
&\quad\leq \frac{4C_{\mathcal{F},2}^L\sqrt{LK^*}}{c(\mathbf{Q}^*,\mathbf{f}^*)} \\
&\qquad\times \inf_{M\leq n}\left\{\sum_{k=1}^{K^*}\big\|f_{M,k}^* - f_k^*\big\|_2^2 + \operatorname{pen}\big(n,M,K^*\big)\right\} + \frac{A}{c(\mathbf{Q}^*,\mathbf{f}^*)n} + \frac{C_{\text{err}}}{n^\beta},
\end{aligned}$$

and $\mathbb{P}(\hat{K}_{\text{l.s.}} \neq K^*) = O(n^{-\beta})$.

Let us discuss what this corollary implies. The approximation error $\sum_{k=1}^{K^*}\|f_{M,k}^* - f_k^*\|_2^2$ can be bounded in a standard way by $O(M^{-2s/D})$ where $s > 0$ is the regularity of the emission densities, see for instance DeVore and Lorentz [12]. One can obtain a trade-off between approximation error and penalty by choosing $M \approx (n/\log(n)^2)^{D/(2s+D)}$, which leads to the optimal rate of convergence $(n/\log(n)^2)^{-2s/(2s+D)}$, up to a logarithmic factor. This shows that our estimators are minimax adaptive up to a logarithmic factor and converge almost surely to the right number of states, all at the same time.

4. Spectral estimation

In this section, we introduce our spectral order estimator. We will assume **[HX]** and **[HidA]** hold.

The idea of this method is to use the matrix containing the coordinates of the density of two consecutive observations in an orthonormal basis. Take $M \in \mathcal{M}$ and let $\Phi_M = (\varphi_1^{(M)}, \dots, \varphi_M^{(M)})$ be an orthonormal basis of $\mathfrak{P}_M$. For ease of notation, we will drop the dependency in M and write φ_a instead of $\varphi_a^{(M)}$. Let us introduce the matrice $\mathbf{N}_M$ and its empirical estimator, defined by

$$\forall a, b \in [M], \qquad \mathbf{N}_M(a,b) := \mathbb{E}\big[\varphi_a(Y_1)\varphi_b(Y_2)\big],$$

$$\forall a, b \in [M], \qquad \hat{\mathbf{N}}_M(a,b) := \frac{1}{n}\sum_{s=1}^{n}\varphi_a(Y_s)\varphi_b(Y_{s+1}).$$

$\mathbf{N}_M$ contains the coordinates of the density of (Y_1, Y_2) with respect to $\mu^{\otimes 2}$ on the basis Φ_M. It holds that

$$\mathbf{N}_M = \mathbf{O}_M \operatorname{Diag}\big(\pi^*\big)\mathbf{Q}^*\mathbf{O}_M^\top, \tag{3}$$

with $\mathbf{O}_M$ the coordinates of the emission densities on the orthonormal basis:

$$\forall m \in [M], \forall k \in \mathcal{X}, \qquad \mathbf{O}_M(m,k) := \mathbb{E}\big[\varphi_m(Y_1)|X_1 = k\big] = \int \varphi_m f_k^* \, d\mu.$$

When the emission densities are linearly independent, $\mathbf{O}_M$ has full rank for M large enough.

The key remark for our method is that $\mathbf{N}_M$ contains explicit information about the order of the HMM, as stated in the following lemma:

Lemma 12. *There exists $M_0 \equiv M_0(\mathbf{Q}^*, \Phi_M, \mathbf{f}^*)$ such that for all $M \geq M_0$, $\mathbf{N}_M$ has rank K^*.*

In the following, we will assume $M \geq M_0$ for M_0 given by this lemma.

In practice, one only has access to the matrix $\hat{\mathbf{N}}_M$, which can be seen as a noisy version of $\mathbf{N}_M$. In particular, there is no reason for it to have only K^* nonzero singular values. On the contrary, the spectrum becomes noisy, and when some singular values of $\mathbf{N}_M$ are too small, they can be masked by this noise. As seen in equation (3), this can occur when $\mathbf{Q}^*$ or $\mathbf{O}_M$ are close to not having full rank, which means for $\mathbf{O}_M$ that the emission densities are almost linearly dependent.

Denote by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots$ the singular values of the matrix A. We can now state the theorem proving the consistency of the spectral order estimator:

Theorem 13. *Let $\hat{K}_{\text{sp.}}(C) = \#\{i | \sigma_i(\hat{\mathbf{N}}_M) > C\sqrt{\log(n)/n}\}$.*

There exists $C_0 \equiv C_0(\mathbf{Q}^*, \Phi_M)$ and $n_0 \equiv n_0(\mathbf{Q}^*, \Phi_M, \mathbf{O}_M^*)$ such that for all $C \geq C_0$ and $n \geq n_0 C^2(1 + \log(C))$,

$$\mathbb{P}\big(\hat{K}_{\text{sp.}}(C) \neq K^*\big) \leq n^{-2}$$

so that $\hat{K}_{\text{sp.}}(C) \longrightarrow K^*$ almost surely.

Comment. It is possible to take $M \longrightarrow \infty$, n_0 constant and C_0 depending on M in an explicit way as long as M grows slowly enough, that is, $\eta_2(\Phi_M) \leq \text{cst} \cdot \sqrt{n/\log(n)}$ and $C_0 = \text{cst} \cdot \eta_2(\Phi_M)$ where $\eta_2(\Phi_M)$ is defined in Lemma 14.

Proof. We control the difference between the spectra of $\mathbf{N}_M$ and $\hat{\mathbf{N}}_M$ using the following lemma from Appendix E of de Castro, Gassiat and Le Corff [11].

Lemma 14. *There exists some constant $\mathcal{C}_*$ depending only on $\mathbf{Q}^*$ such that for any positive u, M and n,*

$$\mathbb{P}\left[\|\mathbf{N}_M - \hat{\mathbf{N}}_M\|_F \geq \frac{\eta_2(\Phi_M)\mathcal{C}_*}{\sqrt{n}}(1+u)\right] \leq e^{-u^2},$$

where

$$\eta_2^2(\Phi_M) = \sup_{y,y' \in \mathcal{Y}^2} \sum_{a,b=1}^{M} \big(\varphi_a(y_1)\varphi_b(y_2) - \varphi_a\big(y_1'\big)\varphi_b\big(y_2'\big)\big)^2.$$

In particular, taking $u = \sqrt{2\log(n)}$ and assuming $u > 1$ and $n \geq 2$, one has with probability $1 - n^{-2}$ that

$$\sigma_1(\mathbf{N}_M - \hat{\mathbf{N}}_M) \leq C\sqrt{\frac{\log(n)}{n}}$$

for all $C \geq C_0 := 2\sqrt{2}\eta_2(\Phi_M)\mathcal{C}_*$, using that for any matrix A, one has $\sigma_1(A) \leq \|A\|_F$.

Let $C \geq C_0$. We will need Weyl's inequality (a proof may be found in Stewart and Sun [32] for instance):

Lemma 15 (Weyl's inequality). *Let A, B be $p \times q$ matrices with $p \geq q$, then for all $i = 1, \dots, q$,*

$$\big|\sigma_i(A + B) - \sigma_i(A)\big| \leq \sigma_1(B).$$

Using this inequality, one gets that with probability at least $1 - n^{-2}$, for all $1 \leq i \leq K^*$, $\sigma_i(\hat{\mathbf{N}}_M) > \sigma_{K^*}(\mathbf{N}_M) - C\sqrt{\log(n)/n}$ and for all $i > K^*$, $\sigma_i(\hat{\mathbf{N}}_M) < C\sqrt{\log(n)/n}$.

In particular, if $2C\sqrt{\log(n)/n} < \sigma_{K^*}(\mathbf{N}_M)$, then with probability at least $1 - n^{-2}$, the order is exactly the number of singular values of $\hat{\mathbf{N}}_M$ which are larger than $C\sqrt{\log(n)/n}$. Finally, observe that under the condition $n \geq n_0 C^2(1 + \log(C))$,

$$\begin{aligned} C\sqrt{\frac{\log(n)}{n}} &\leq \sqrt{\frac{2\log(C) + \log(1 + \log(C))}{n_0(1 + \log(C))}} \\ &\leq \sqrt{\frac{3}{n_0}}\sqrt{\frac{\log(C)}{1 + \log(C)}}, \end{aligned}$$

since one can assume without loss of generality that $C_0 \geq 1$. By taking $n_0 = 12/\sigma_{K^*}(\mathbf{N}_M)^2$, this concludes the proof. □

5. Numerical experiments

In this section, we show the results of our estimators on simulated data. The simulation parameters are introduced in Section 5.1. We show the numerical results and discuss their ability to select the right order in practice in Section 5.2, and we present the data-driven methods and heuristics we used for the numerical implementation in Section 5.3.

5.1. Simulation parameters

We will consider $\mathcal{Y} = [0, 1]$ with μ being the Lebesgue measure. We will use a trigonometric basis on $\mathbf{L}^2([0, 1])$ to generate the approximation spaces $(\mathfrak{P}_M)_M$. More precisely, define

$$\varphi_0(t) = 1,$$
$$\varphi_a(t) = \sqrt{2}\cos(\pi a t)$$

for all $t \in [0, 1]$ and $a \in \mathbb{N}^*$. We take $\mathfrak{P}_M = \mathrm{Span}(\{\varphi_a | 0 \leq a < M\})$ the spaces induced by the trigonometric basis.

Comment. Taking the same vectors in all bases is not mandatory to ensure theoretical consistency, but in practice it allows us to take an additional initial point for the minimization step and improves the stability of the algorithm (see Step 1 below).

We will assume $\mathbf{f}^*$ to be linearly independent, so that one only needs $L = 3$ observations to recover the parameters of the HMM.

In order to assess the performances of the different procedures, we generate n observations of a HMM of order 3 for several values of n, using the following parameters:

- Emission distributions: Beta distributions with two possible sets of parameters: [(1.5; 5), (7; 2) and (6; 6)] or [(2; 5), (4; 2) and (4; 4)];
- Markov chain parameters:

$$\mathbf{Q}^* = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.07 & 0.13 & 0.8 \end{pmatrix},$$
$$\pi^* = \begin{pmatrix} \frac{47}{120} & \frac{11}{40} & \frac{1}{3} \end{pmatrix}$$
$$\approx (0.3917 \quad 0.2750 \quad 0.3333).$$

Finally, we take $M_{\max} = 50$ the maximum value of M and two possible values for $K_{\max}$, the maximum value of K for which we will compute the estimators : $K_{\max} = 5$ and $K_{\max} = 10$. We took limited values for $K_{\max}$ in order to reduce the computational cost of the estimation, however there is no objection to taking larger values in practice.

The simulation codes are available at https://www.normalesup.org/~lleherice/HMM_order_simfiles/.

5.2. Numerical results

Figure 1 summarizes the results of both procedures. Both select the right order as soon as the number of observations is sufficient.

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$	$\mathbb{P}(\hat{K}_{\text{sp.}} = K^*)$
999	0.2	0
3000	1	0
9999	1	1
19 998	1	1

Beta parameters (1.5; 5), (7; 2) and (6; 6) for $K_{\max} = 5$

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$	$\mathbb{P}(\hat{K}_{\text{sp.}} = K^*)$
7500	0.3	0
19 998	0.9	0
30 000	1	0
49 998	1	0.1

Beta parameters (2; 5), (4; 2) and (4; 4) for $K_{\max} = 5$

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$
999	0.8
3000	0.9

Beta parameters (1.5; 5), (7; 2) and (6; 6) for $K_{\max} = 10$

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$
7500	0.5
19 998	0.9

Beta parameters (2; 5), (4; 2) and (4; 4) for $K_{\max} = 10$

Figure 1. Probability to select the right order for the two methods ($\hat{K}_{\text{l.s.}}$ for the least squares method and $\hat{K}_{\text{sp.}}$ for the spectral method). 10 simulations have been done for each n. Parameters for spectral selection are $M = 40$, $M_{\text{reg}} = 35$ and $\tau = 1.5$ (see Section 5.3.2 for the definition of these parameters).

The spectral method is easily put in practice and runs extremely fast. It doesn't need a time-consuming contrast minimization step or an initial point. However, the thresholding of the singular values is a delicate issue, and if the order is incorrect, then the theoretical results about the spectral estimators of the parameters don't hold and this method may behave poorly.

The performances of the least squares method are much better and improve for larger values of $K_{\max}$ (see Figure 1 for comparing the order estimators and de Castro, Gassiat and Lacour [10] for comparing the emission densities estimators). In addition, the model selection step is easy to handle and gives an estimator of the order that we proved to be consistent, estimators of the HMM parameters that we proved to be minimax up to a logarithmic factor and a way to check whether the model fits the data well (see Section 5.3.1), all at the same time. However, the minimization of the (non-convex) empirical contrast is a time-consuming step, especially for large samples and large models.

Choosing the right method is thus a question of computational power and amount of available data. For small datasets where one wants to get accurate results, the least squares method is best. Conversely, on large datasets and large models, the spectral method is a good choice in order to obtain many estimators in a reasonable amount of time.

5.3. Practical implementation

5.3.1. *Least squares method*

The first issue that one encounters when trying to minimize the least squares criterion γ_n is that it is not convex. Several algorithms have been proposed to overcome this difficulty. We chose to use CMA-ES (for Covariance Matrix Adaptation Estimation Strategy, see Hansen [19]) in order

to find a minimizer. This estimator is easy to use and works well in many situations, but – like all approximate minimization algorithms – it requires a good initial point since it might otherwise remain stuck in local minima.

One part of our method consists in using previous estimates as initial points for further steps to counter this phenomenon, since it is likely that this way the estimators stay near the real minimizer. More precisely, to obtain $\hat{g}_{K,M}$, we compute several minimizers of γ_n using the previously computed estimators $\hat{g}_{K-1,M}$ where one state has been duplicated and $\hat{g}_{K,M-1}$ extended to the space $\mathfrak{P}_M^{\otimes L}$, and keep the best one. This procedure is detailed in Appendix A, see Algorithm 2.

The underlying heuristcs of this initialization procedure is that when the order is underestimated, then several states are "merged" together. Duplicating a merged state will allow to separate them effectively while still taking advantage of the computations done up to now. It is meant to avoid having to recalculate all states at the same time (which could get us stuck in sub-optimal local minima) when the best solution is likely to be a small modification of the previous estimator. In addition, when the order is overestimated, it allows to make sure the empirical criterion is indeed decreasing with the dimension of the model by giving an estimator that performs at least as well as those from smaller models. This makes our method robust to an overestimation of the order.

The last practical issue is a very common one in the model selection setting: the constant ρ of the penalty is unknown and has to be estimated before one can select the right model. Several data-driven estimators have been proposed to circumvent this difficulty, for instance dimension jump heuristics, slope heuristics, bootstrap or cross validation. We focus on the first two, which have several advantages in our setting. First, they are easy to use, are proved to be theoretically valid in many settings and work well in a wide range of applications (see, for instance, Baudry, Maugis and Michel [4] and references therein). Second, they take advantage of the structure of our problem and both give a qualitative way to check whether the choice of penalty is valid or not, and by extension whether the model is misspecified or not.

Dimension jump heuristics. In this paragraph, we study the selected parameters

$$\rho \longmapsto \big(\hat{M}(\rho), \hat{K}(\rho)\big) \in \arg\min\big\{\gamma_n(\hat{g}_{K,M}) + \rho\,\mathrm{pen}_{\mathrm{shape}}(n, M, K)\big\}$$

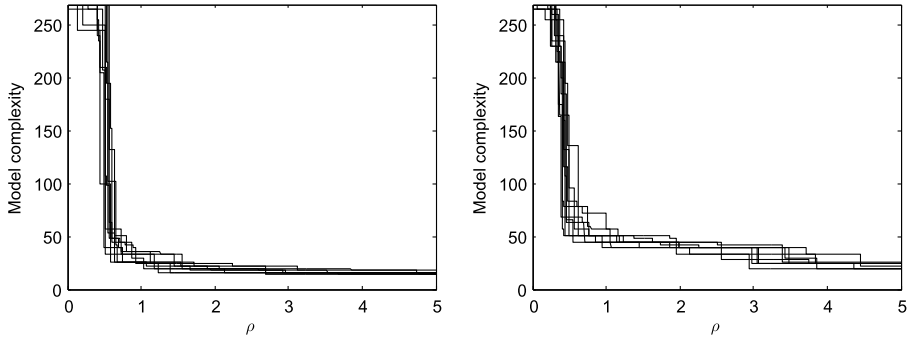
and the selected complexity

$$\rho \longmapsto \mathrm{Comp}(\rho) = \hat{M}(\rho)\hat{K}(\rho) + \hat{K}(\rho))\big[\hat{K}(\rho)) - 1\big]$$

with $\mathrm{pen}_{\mathrm{shape}}(n, M, K) = (MK + K^2 - 1)\log(n)/n$.

Assume that there exists κ such that $\kappa\,\mathrm{pen}_{\mathrm{shape}}$ is a *minimal penalty*, that is, a penalty such that as n tends to infinity, for all $\rho > \kappa$, the size of the model chosen for penalty $\rho\,\mathrm{pen}_{\mathrm{shape}}$ remains small in some sense and for all $\rho < \kappa$, the size of the model becomes huge. Then, for n large enough, this will appear on the graph of the selected model complexity as a "dimension jump": around some constant ρ_{jump}, the complexity will abruptly drop from large models to small models. This is clearly the case in Figure 2. Figure 3 shows the behaviour of $\hat{M}$ and $\hat{K}$ with ρ. A dimension jump also occurs with these functions. It is most visible for $\hat{M}$.

Finally, once the dimension jump location ρ_{jump} has been estimated, we take $\hat{\rho} = 2\rho_{\mathrm{jump}}$ to select the final parameters.



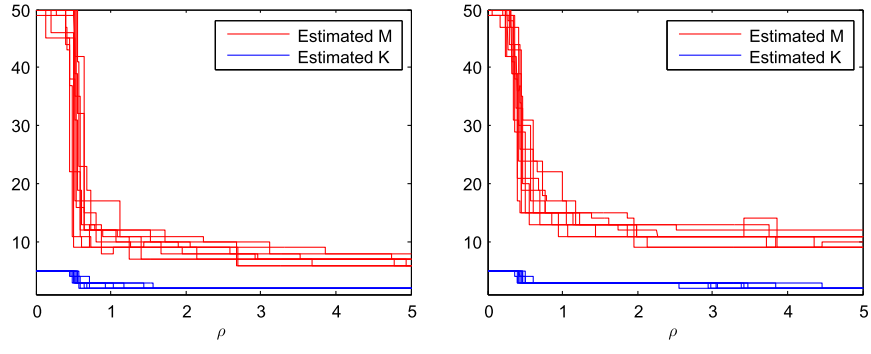
(a) $n = 7500$. (b) $n = 49\,998$.

Figure 2. Graph of $\rho \mapsto \mathrm{Comp}(\rho)$ for 10 sets of n consecutive observations. Here, the parameters of the Beta distribution are $(2; 5)$, $(4; 2)$ and $(4; 4)$, and $K_{\max} = 5$.

It is worth noting that this jump method also gives a qualitative way to check whether the choice of parameters is sensible: if no clear jump can be identified, then either one didn't consider enough models to make the jump clear, or the penalty isn't the right one, or the model cannot approximate the data distribution well.

Slope heuristics. This heuristics relies on the fact that when $\mathrm{pen}_{\mathrm{shape}}$ is a minimal penalty, then the empirical contrast function is expected to behave like $\rho_{\min} \mathrm{pen}_{\mathrm{shape}}$ for large models and for some constant $\rho_{\min}$. This gives both a way to calibrate the constant of the penalty and to check if the chosen penalty has the right shape (see Baudry, Maugis and Michel [4]). The final penalty is then taken as $2\hat{\rho}_{\min} \mathrm{pen}_{\mathrm{shape}}$.

Figure 4 shows the graph of the empirical contrast depending on $\mathrm{pen}_{\mathrm{shape}}$. The slope heuristics works well in this situation, suggesting that our penalty has the right shape.



(a) $n = 7500$. (b) $n = 49\,998$.

Figure 3. Graph of $\rho \mapsto \hat{M}(\rho)$ and $\rho \mapsto \hat{K}(\rho)$ for 10 sets of n consecutive observations. Here, the parameters of the Beta distribution are $(2; 5)$, $(4; 2)$ and $(4; 4)$, and $K_{\max} = 5$.

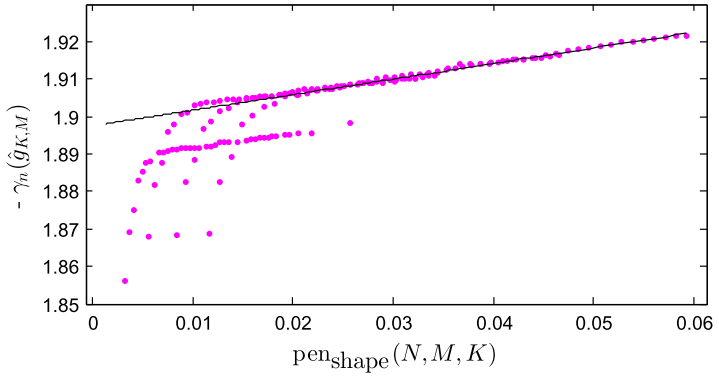


Figure 4. Empirical constrast and calibrated penalty for $n = 49\,998$. Here, the parameters of the Beta distribution are $(2; 5)$, $(4; 2)$ and $(4; 4)$, and $K_{\max} = 5$.

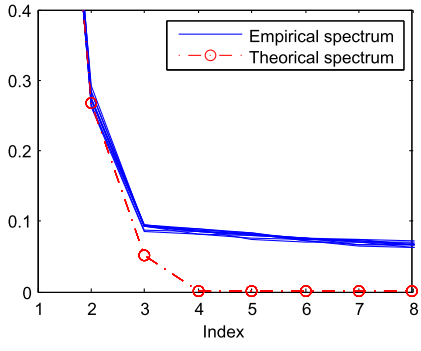
5.3.2. *Spectral method*

The idea of the spectral order estimation is to recover the rank of the matrix $\mathbf{N}_M$. However, this is not always possible: if one singular value of $\mathbf{N}_M$ is smaller than the noise (which is the case when $\mathbf{O}_M$ is close from not being invertible, i.e., when the emission densities are close from being linearly dependent, and when there are only few observations), then this method will not be able to "see" the corresponding hidden state.

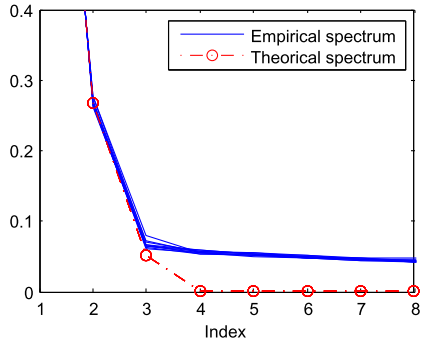
Figure 5 – and in particular Figure 5a – illustrates this problem: the third singular value is smaller than several noisy singular values, which means it won't be possible to recover it. Even if one knows the right order, the fact that the singular value is smaller than the noise can make it impossible for spectral methods to recover the true parameters. Figure 6 shows the result when trying to estimate the densities in the situation of Figure 5a: when the singular value is drowned by the noise, the output of the spectral estimator is aberrant. Notice that it is not a fatality: in the same situation, the least squares method manages to give sensible estimators of the emission densities. This is an intrinsic limitation of the spectral method.

Therefore, what we need is a way to threshold the parameters in order to distinguish noise from significant singular values. The estimator $\hat{K}_{\mathrm{sp.}}(C)$ is one way to achieve this, but the calibration of C is a tricky problem, since the right choice of C depends on the parameters of the HMM. We will use a different method, which relies on the same idea: identifying the noisy singular values which stand out from the others and saying they correspond to nonzero singular values of $\mathbf{N}_M$. Our heuristics relies on the fact that when one sorts the singular values in decreasing order, then the smallest ones approximately follow an affine relation with respect to their index. This tendency is shown in Figure 7.

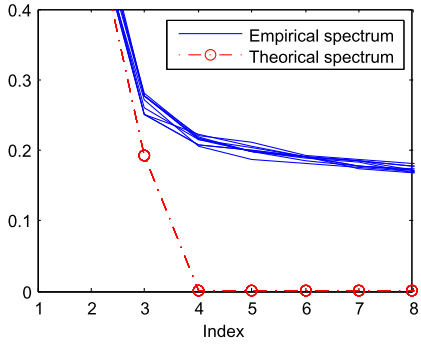
We proceed as follows. Let M and M_{reg} be two positive integers such that $M_{\mathrm{reg}} \leq M \leq M_{\max}$. We estimate the affine dependance of the singular values of $\hat{\mathbf{N}}_M$ with respect to their index with a linear regression using its M_{reg} smallest singular values. Then, we set a thresholding parameter $\tau > 1$. We say a singular value is *significant* if it is above τ times the value that the regression predicts for it. Lastly, we take $\hat{K}_{\mathrm{sp.}}$ as the number of consecutive significant singular values



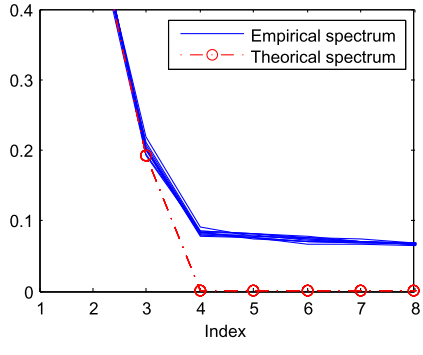
(a) $n = 19\,998$, Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$.



(b) $n = 49\,998$, Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$.



(c) $n = 3000$, Beta parameters $(1.5;5)$, $(7;2)$ and $(6;6)$.



(d) $n = 19\,998$, Beta parameters $(1.5;5)$, $(7;2)$ and $(6;6)$.

Figure 5. Spectrum of the empirical matrix $\hat{\mathbf{N}}_M$ and the theoretical matrix $\mathbf{N}_M$ for $M = 40$ and 10 simulations. The first singular values are too large to appear here.

starting from the largest one. This heuristics seems to work as soon as τ is large enough, for example, $\tau = 1.5$.

6. Proofs

6.1. Main technical result

The following lemma is the main technical result of this paper. It is the key for both the strong consistency and the oracle inequalities. It allows to control the difference between the empirical criterion γ_n and the theoretical $\mathbf{L}^2$ loss for all models at the same time.

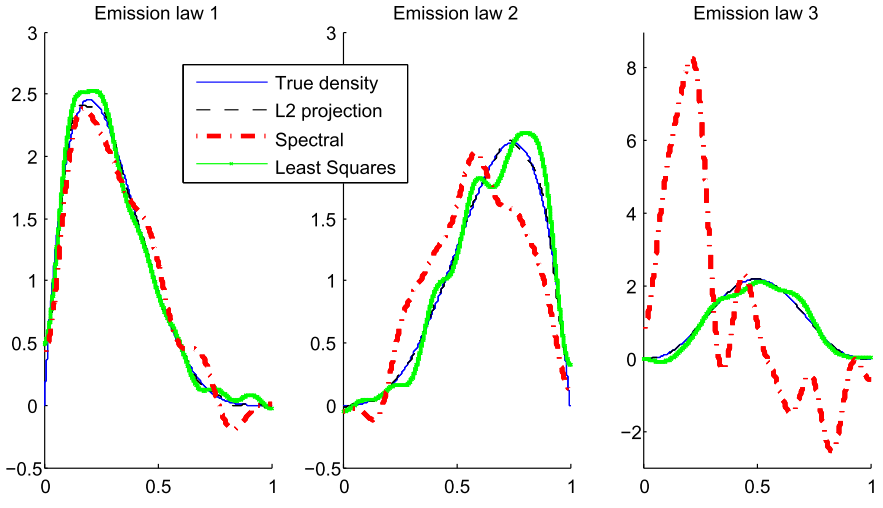


Figure 6. Estimators of the emission densities for $n = 19\,998$ and Beta parameters (2; 5), (4; 2) and (4; 4). We took $K = \hat{K}_{\text{l.s.}} = 3$ and $M = \hat{M} = 13$. The bad behaviour of the spectral algorithm (presented in Appendix A of Lehéricy [26]) when the emission densities are poorly separated is clearly visible on the third emission distribution.

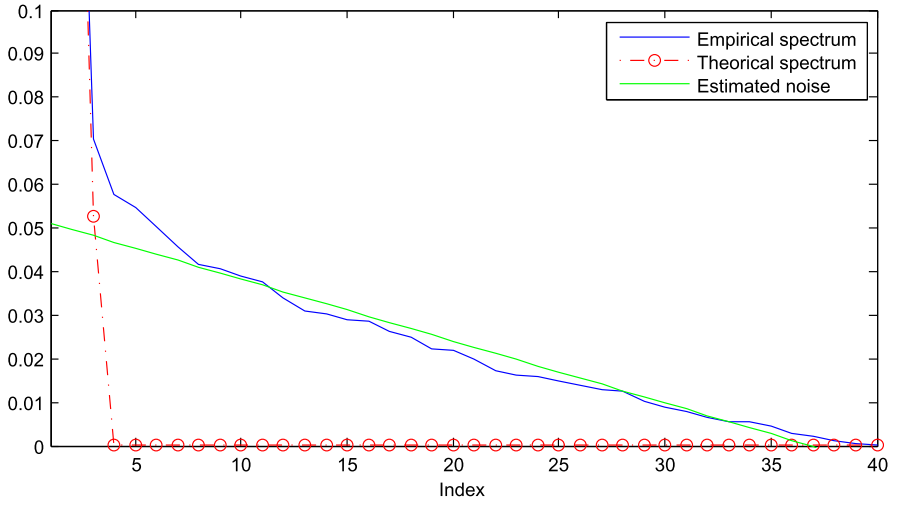


Figure 7. Spectrum of $\mathbf{N}_M$ for $M = 40$ and $n = 49\,998$ for Beta parameters (2; 5), (4; 2) and (4; 4). The regression (green line) has been performed on the 35 smallest singular values. The two largest singular values are too large to appear here.

Define $\nu : t \longmapsto \frac{1}{n}\sum_{s=1}^{n} t(Z_s) - \int t g^*$, so that

$$\forall t \in \mathbf{L}^2\big(\mathcal{Y}^L, \mu^{\otimes L}\big), \qquad \gamma_n(t) + \big\|g^*\big\|_2^2 = \big\|t - g^*\big\|_2^2 - 2\nu(t). \tag{4}$$

Let

$$\begin{aligned} s = (s_{K,M})_{K,M} \in \mathbf{S} &:= \prod_{K\in\mathbb{N}^*, M\in\mathcal{M}} \left(\bigcup_K S_K\right) \\ &\longmapsto \quad \big(Z_{K,M}(s)\big)_{K,M} \\ &:= \left(\sup_{t\in S_{K,M}} \left[\frac{|\nu(t - s_{K,M})|}{\|t - s_{K,M}\|_2^2 + x_{K,M}^2}\right]\right)_{K,M}. \end{aligned} \tag{5}$$

Comment. It is not necessary to assume that $s_{K,M} \in S_{K,M}$. In particular, one can take $s_{K,M} = g^*$ for all K, M. In that case, we will simply write $Z_{K,M}(g^*)$.

Lemma 16. *Assume* **[*HX*]** *and* **[*HF*]** *hold. Then there exists a sequence* $(x_{K,M})_{K,M} \equiv (x_{K,M})_{K,M}(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *and positive constants* $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that if the penalty* $\widetilde{\text{pen}}$ *satisfies*

$$\forall n, \forall M \le n, \forall K \le n, \qquad \widetilde{\text{pen}}(n, M, K) \ge \rho\big(MK + K^2 - 1\big)\frac{\log(n)}{n}$$

then for all $s \in \mathbf{S}$, $n \ge n_0$ *and* $x > 0$, *one has with probability larger than* $1 - e^{-x}$:

$$\begin{cases} \sup\limits_{K'\le n, M'\le n} Z_{K',M'}(s) \le \dfrac{1}{4}, \\ \sup\limits_{K'\le n, M'\le n} \big(2Z_{K',M'}(s)x_{K',M'}^2 - \widetilde{\text{pen}}\big(n, M', K'\big)\big) \le A\dfrac{x}{n}. \end{cases}$$

Comment. One can replace the constant $1/4$ in the first upper bound by any $\epsilon > 0$, at the cost of changing the constants n_0, ρ and A.

The structure of the proof follows the usual method to control empirical processes, see for instance Massart [28], Chapter 6, adapted to the HMM structure by de Castro, Gassiat and Lacour [10]. The novelty and main difficulty of the proof comes from the generalization to both nonparametric densities and an unknown number of states: we had to introduce a much finer control of the constants and of the bracketing entropy of the models in order to take the dependency in the order of the HMM into account.

The details of the proof can be found in Appendix C of Lehéricy [26].

6.2. Proof of Corollary 2

Since the union of $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ is dense in $\mathcal{F}$, we can take M such that **[HidA]** or **[HidB]** holds for $\mathbf{f}_M^* = (f_{M,k}^*)_{k\in\mathcal{X}} := (\text{Proj}_{\mathfrak{P}_M} f_k^*)_{k\in\mathcal{X}}$.

We will need the following lemma.

Lemma 17.

$$\forall \pi \in \mathbb{R}^K, \forall \mathbf{Q} \in \mathbb{R}^{K\times K}, \forall \mathbf{f} \in \mathcal{F}^K, \forall M, \qquad \mathrm{Proj}_{\mathfrak{P}_M^{\otimes L}}\left(g^{\pi,\mathbf{Q},\mathbf{f}}\right) = g^{\pi,\mathbf{Q},\mathrm{Proj}_{\mathfrak{P}_M}(\mathbf{f})}$$

Proof. By linearity of the projection operator, it is enough to prove that for all $(t_1, \ldots, t_L) \in (\mathbf{L}^2(\mathcal{Y}, \mu))^L$,

$$\mathrm{Proj}_{\mathfrak{P}_M^{\otimes L}}(t_1 \otimes \cdots \otimes t_L) = \mathrm{Proj}_{\mathfrak{P}_M}(t_1) \otimes \cdots \otimes \mathrm{Proj}_{\mathfrak{P}_M}(t_L)$$

which is easy to check. □

We will make a proof by contradiction. Assume that $\inf_{t\in S_K} \|t - g^*\|_2 = 0$ for some $K < K^*$. Then there exists a sequence $(g_n)_{n\geq 1} = (g^{\pi_n,\mathbf{Q}_n,\mathbf{f}_n})_{n\geq 1}$ such that $g_n \longrightarrow g^*$ in $\mathbf{L}^2(\mathcal{Y}^L, \mu^{\otimes L})$, with $\pi_n \in \Delta_K$, $\mathbf{Q}_n$ a transition matrix of size K and $\mathbf{f}_n \in \mathcal{F}^K$.

The orthogonal projection onto $\mathfrak{P}_M^{\otimes L}$ is continuous, so by using Lemma 17, one gets that

$$g^{\pi_n,\mathbf{Q}_n,\mathrm{Proj}_{\mathfrak{P}_M}(\mathbf{f}_n)} \longrightarrow g^{\pi^*,\mathbf{Q}^*,\mathbf{f}_M^*}.$$

Then, using the compactness of Δ_K and of the set of transition matrices of size K and the relative compactness of $(\mathcal{F} \cap \mathfrak{P}_M)^K$ (which is a bounded subset of a finite dimension linear space), one gets (up to extraction of a subsequence) that there exists $\pi_\infty \in \Delta_K$, $\mathbf{Q}_\infty$ a transition matrix of size K and $\mathbf{f}_\infty \in (\mathfrak{P}_M)^K$ such that $\pi_n \longrightarrow \pi_\infty$, $\mathbf{Q}_n \longrightarrow \mathbf{Q}_\infty$ and $\mathrm{Proj}_{\mathfrak{P}_M}(\mathbf{f}_n) \longrightarrow \mathbf{f}_\infty$.

Finally, using the continuity of the function $(\pi, \mathbf{Q}, \mathbf{f}) \longmapsto g^{\pi,\mathbf{Q},\mathbf{f}}$ and the uniqueness of the limit, one gets

$$g^{\pi_\infty,\mathbf{Q}_\infty,\mathbf{f}_\infty} = g^{\pi^*,\mathbf{Q}^*,\mathbf{f}_M^*}.$$

Then Proposition 1 contradicts the assumption $K < K^*$, which is enough to conclude.

6.3. Consistency proofs

The definition of $\hat{K}_{\mathrm{l.s.}}$ is equivalent to the following one:

$$\hat{K}_{\mathrm{l.s.}} \in \operatorname*{arg\,min}_{K\leq n}\left\{\gamma_n(\hat{g}_{K,\hat{M}_K}) + \mathrm{pen}(n, \hat{M}_K, K)\right\},$$

where

$$\hat{M}_K \in \operatorname*{arg\,min}_{M\leq n}\left\{\gamma_n(\hat{g}_{K,M}) + \mathrm{pen}(n, M, K)\right\}.$$

Choosing K rather than K^* means that K is better than K^*, that is,

$$\{\hat{K}_{\mathrm{l.s.}} = K\} \subset \Big\{0 \geq \inf_{M\leq n}\left\{\gamma_n(\hat{g}_{K,M}) + \mathrm{pen}(n, M, K)\right\} - \inf_{M\leq n}\left\{\gamma_n(\hat{g}_{K^*,M}) + \mathrm{pen}\big(n, M, K^*\big)\right\}\Big\}.$$

Let

$$\begin{aligned} D_{n,K} &:= \inf_{M\le n}\big\{\gamma_n(\hat g_{K,M}) + \operatorname{pen}(n,M,K)\big\} \\ &\quad - \inf_{M\le n}\big\{\gamma_n(\hat g_{K^*,M}) + \operatorname{pen}\big(n,M,K^*\big)\big\} \\ &= \gamma_n(\hat g_{K,\hat M_K}) + \operatorname{pen}(n,\hat M_K,K) \\ &\quad - \inf_{M\le n}\Big\{\inf_{t\in S_{K^*,M}}\gamma_n(t) + \operatorname{pen}\big(n,M,K^*\big)\Big\}. \end{aligned}$$

Then

$$\{\hat K_{\text{l.s.}} = K\} \subset \{D_{n,K}\le 0\}.$$

We will thus control the probability of the latter event for all $K < K^*$ in the first case and $K > K^*$ in the second case.

Proof of Theorem 3. Let $M_0 \in \mathcal{M}$. We will choose a suitable value for this integer later in the proof. Assume $n \ge M_0$. Then by definition of $D_{n,K}$ and of ν (equation (4)),

$$\begin{aligned} D_{n,K} &\ge \gamma_n(\hat g_{K,\hat M_K}) + \operatorname{pen}(n,\hat M_K,K) - \gamma_n\big(g^*_{K^*,M_0}\big) - \operatorname{pen}\big(n,M_0,K^*\big)) \\ &= \big\|g^* - \hat g_{K,\hat M_K}\big\|_2^2 - \big\|g^* - g^*_{K^*,M_0}\big\|_2^2 - 2\nu\big(\hat g_{K,\hat M_K} - g^*_{K^*,M_0}\big) \\ &\quad + \operatorname{pen}(n,\hat M_K,K) - \operatorname{pen}\big(n,M_0,K^*\big). \end{aligned}$$

Using the definition of $Z_{K,M}$ (equation (5)), one gets that

$$\begin{aligned} \big|\nu\big(\hat g_{K,\hat M_K} - g^*_{K^*,M_0}\big)\big| &\le \big|\nu\big(\hat g_{K,\hat M_K} - g^*\big)\big| + \big|\nu\big(g^* - g^*_{K^*,M_0}\big)\big| \\ &\le Z_{K,\hat M_K}(g^*)\big(\big\|g^* - \hat g_{K,\hat M_K}\big\|_2^2 + x^2_{K,\hat M_K}\big) \\ &\quad + Z_{K^*,M_0}(g^*)\big(\big\|g^* - g^*_{K^*,M_0}\big\|_2^2 + x^2_{K^*,M_0}\big). \end{aligned}$$

Let n_0, ρ and A be as in Lemma 16. We can assume that $n_0 \ge K^*$ so that $K^* \le n$. Let us introduce the function $\widetilde{\operatorname{pen}}(n,M,K) = \rho(MK + K^2 - 1)\frac{\log(n)}{n}$. Let $n \ge n_0$ and $x > 0$ and assume we are in the event of probability $1 - e^{-x}$ of Lemma 16. Then, for all $K \le n$:

$$\begin{aligned} \big|\nu\big(\hat g_{K,\hat M_K} - g^*_{K^*,M_0}\big)\big| &\le \frac14\big\|g^* - \hat g_{K,\hat M_K}\big\|_2^2 + \frac12 A\frac{x}{n} + \frac12\widetilde{\operatorname{pen}}(n,\hat M_K,K) \\ &\quad + \frac14\big\|g^* - g^*_{K^*,M_0}\big\|_2^2 + \frac12 A\frac{x}{n} + \frac12\widetilde{\operatorname{pen}}\big(n,M_0,K^*\big) \end{aligned}$$

and

$$\begin{aligned} D_{n,K} &\ge \frac12\big\|g^* - \hat g_{K,\hat M_K}\big\|_2^2 - \frac32\big\|g^* - g^*_{K^*,M_0}\big\|_2^2 - 2A\frac{x}{n} + \operatorname{pen}(n,\hat M_K,K) \\ &\quad - \operatorname{pen}\big(n,M_0,K^*\big) - \widetilde{\operatorname{pen}}(n,\hat M_K,K) - \widetilde{\operatorname{pen}}\big(n,M_0,K^*\big). \end{aligned}$$

We assumed $\text{pen} \geq \widetilde{\text{pen}}$, so that

$$D_{n,K} \geq \frac{1}{2}\left\|g^* - \hat{g}_{K,\hat{M}_K}\right\|_2^2 - \frac{3}{2}\left\|g^* - g^*_{K^*,M_0}\right\|_2^2 - 2A\frac{x}{n} - 2\,\text{pen}\big(n, M_0, K^*\big).$$

Corollary 2 ensures that

$$d := \inf_{K<K^*} \inf_{t\in S_K} \left\|t - g^*\right\|_2 > 0,$$

so that for all $K < K^*$,

$$D_{n,K} \geq \frac{d^2}{2} - \frac{3}{2}\left\|g^* - g^*_{K^*,M_0}\right\|_2^2 - 2A\frac{x}{n} - 2\,\text{pen}\big(n, M_0, K^*\big).$$

By denseness of $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ in $\mathcal{F}$, one gets that

$$\inf_M \left\|g^*_{K^*,M} - g^*\right\|_2 = 0$$

so that there exists M_0 such that $\|g^* - g^*_{K^*,M_0}\|_2^2 \leq d^2/6$. If we choose this M_0, we get that

$$D_{n,K} \geq \frac{d^2}{4} - 2A\frac{x}{n} - 2\,\text{pen}\big(n, M_0, K^*\big).$$

Which implies that $D_{n,K} > 0$ as soon as $2Ax/n < d^2/4 - 2\,\text{pen}(n, M_0, K^*)$, that is,

$$x < \left(\frac{d^2}{8} - \text{pen}\big(n, M_0, K^*\big)\right)\frac{n}{A}.$$

To conclude, note that there exists $\tilde{n_0} \geq \max(n_0, M_0)$ such that for all $n \geq \tilde{n_0}$, $\text{pen}(n, M_0, K^*) \leq \frac{d^2}{16}$. Then, letting $\beta = \frac{d^2}{16A}$, one has for all $n \geq \tilde{n_0}$, with probability $1 - e^{-\beta n}$, for all $K < K^*$, $D_{n,K} > 0$, which implies that $\hat{K}_{\text{l.s.}} \neq K$. □

Proof of Theorem 4. For all $K \geq K^*$,

$$D_{n,K} \geq \gamma_n(\hat{g}_{K,\hat{M}_K}) + \text{pen}(n, \hat{M}_K, K) - \gamma_n\big(g^*_{K^*,\hat{M}_K}\big) - \text{pen}\big(n, \hat{M}_K, K^*\big)$$

and

$$\gamma_n(\hat{g}_{K,\hat{M}_K}) - \gamma_n\big(g^*_{K^*,\hat{M}_K}\big) = \left\|\hat{g}_{K,\hat{M}_K} - g^*\right\|_2^2 - \left\|g^*_{K^*,\hat{M}_K} - g^*\right\|_2^2 - 2\nu\big(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K}\big).$$

First, note that $g^*_{K^*,\hat{M}_K} = g^*_{K,\hat{M}_K}$. Indeed, $g^*_{K,M}$ was defined as the orthogonal projection of g^* onto $\text{Span}(S_{K,M}) \subset \mathfrak{P}_M^{\otimes L}$. Lemma 17 ensures that $g^{\pi^*,\mathbf{Q}^*,\text{Proj}_{\mathfrak{P}_M}(\mathbf{f}^*)}$ is the orthogonal projection of g^* onto $\mathfrak{P}_M^{\otimes L}$ and since it is in $\text{Span}(S_{K^*,M})$, one has $g^*_{K^*,M} = g^{\pi^*,\mathbf{Q}^*,\text{Proj}_{\mathfrak{P}_M}(\mathbf{f}^*)}$ for all M. Then, we use the fact that $(S_{K,M})_{K\geq 1}$ is a sequence of nested subspaces, so that $g^*_{K^*,M} \in \text{Span}(S_{K,M})$ for all $K \geq K^*$, and since $\text{Span}(S_{K,M}) \subset \mathfrak{P}_M^{\otimes L}$ and $g^*_{K^*,M}$ is the orthogo-

nal projection of g^* onto $\mathfrak{P}_M^{\otimes L}$, it is also the orthogonal projection of g^* onto $\mathrm{Span}(S_{K,M})$. Thus, $g^*_{K^*,M} = g^*_{K,M}$ for all $K \geq K^*$ and M.

Then, one has $\hat{g}_{K,\hat{M}_K} \in S_{K,\hat{M}_K} \subset \mathfrak{P}_{\hat{M}_K}^{\otimes L}$, so that, using the Pythagorean theorem,

$$\|\hat{g}_{K,\hat{M}_K} - g^*\|_2^2 - \|g^*_{K^*,\hat{M}_K} - g^*\|_2^2 = \|\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K}\|_2^2.$$

Let n_0, ρ and A be as in Lemma 16. We can assume that $n_0 \geq K^*$ so that $K^* \leq n$. Let us introduce the function $\widetilde{\mathrm{pen}}(n, M, K) = \rho(MK + K^2 - 1)\frac{\log(n)}{n}$. Let $n \geq n_0$ and $x > 0$ and assume we are in the event of probability $1 - e^{-x}$ of Lemma 16. Then, for all $K \leq n$ such that $K \geq K^*$:

$$\begin{aligned}
\big|\nu\big(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K}\big)\big| &= \big|\nu\big(\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\big)\big| \\
&\leq Z_{K,\hat{M}_K}\big(\big(g^*_{K',M'}\big)_{K',M'}\big)\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 \\
&\quad + Z_{K,\hat{M}_K}\big(\big(g^*_{K',M'}\big)_{K',M'}\big)x^2_{K,\hat{M}_K} \\
&\leq \frac{1}{4}\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 + \frac{1}{2}A\frac{x}{n} + \frac{1}{2}\widetilde{\mathrm{pen}}(n, \hat{M}_K, K),
\end{aligned}$$

which implies

$$\begin{aligned}
\gamma_n(\hat{g}_{K,\hat{M}_K}) - \gamma_n\big(g^*_{K^*,\hat{M}_K}\big) &\geq \frac{1}{2}\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 - A\frac{x}{n} - \widetilde{\mathrm{pen}}(n, \hat{M}_K, K) \\
&\geq -A\frac{x}{n} - \widetilde{\mathrm{pen}}(n, \hat{M}_K, K)
\end{aligned}$$

so that for all $K \leq n$ such that $K \geq K^*$:

$$D_{n,K} \geq \mathrm{pen}(n, \hat{M}_K, K) - \mathrm{pen}\big(n, \hat{M}_K, K^*\big) - \widetilde{\mathrm{pen}}(n, \hat{M}_K, K) - A\frac{x}{n}.$$

Now, assume that [**Hpen**](α, ρ) holds for some $\alpha > 0$ and the above constant ρ. Then there exists n_1 such that for all $n \geq n_1$ and for all $K \leq n$ such that $K \geq K^*$,

$$D_{n,K} \geq \alpha\frac{\log(n)}{n} - A\frac{x}{n},$$

which is strictly positive as soon as $x < \alpha\log(n)/A$. Thus, letting $\beta = 1/(2A)$, one has for all $n \geq \max(n_0, n_1, K^*)$, with probability $1 - n^{-\beta\alpha}$, for all $K \leq n$ such that $K > K^*$, $D_{n,K} > 0$, which implies that $\hat{K}_{\mathrm{l.s.}} \neq K$. This concludes the proof. □

6.4. Proof of the oracle inequality (Theorem 7)

Let $K \leq n$ and $M \leq n$. Then

$$\begin{aligned}
\gamma_n(\hat{g}) + \mathrm{pen}(n, \hat{M}, \hat{K}_{\mathrm{l.s.}}) &\leq \gamma_n(\hat{g}_{K,M}) + \mathrm{pen}(n, M, K) \\
&\leq \gamma_n\big(g^*_{K,M}\big) + \mathrm{pen}(n, M, K),
\end{aligned}$$

where the first inequality comes from the definition of $(\hat{K}_{\text{l.s.}}, \hat{M})$ and the second from the definition of $\hat{g}_{K,M}$. Therefore,

$$\gamma_n(\hat{g}) - \gamma_n\big(g^*_{K,M}\big) \leq \operatorname{pen}(n, M, K) - \operatorname{pen}(n, \hat{M}, \hat{K}_{\text{l.s.}}).$$

By definition of ν (equation (4)),

$$\gamma_n(t_1) - \gamma_n(t_2) = \|t_1 - g^*\|_2^2 - \|t_2 - g^*\|_2^2 - 2\nu(t_1 - t_2)$$

so that

$$\|\hat{g} - g^*\|_2^2 \leq \|g^*_{K,M} - g^*\|_2^2 + \operatorname{pen}(n, M, K) - \operatorname{pen}(n, \hat{M}, \hat{K}_{\text{l.s.}}) + 2\nu\big(\hat{g}_{\hat{M},\hat{K}_{\text{l.s.}}} - g^*_{K,M}\big).$$

Now we want to control the ν term. By linearity,

$$\nu\big(\hat{g}_{\hat{K}_{\text{l.s.}},\hat{M}} - g^*_{K,M}\big) = \nu\big(\hat{g}_{\hat{K}_{\text{l.s.}},\hat{M}} - g^*\big) + \nu\big(g^* - g^*_{K,M}\big).$$

Using the definition of $Z_{K,M}$ (equation (5)), we get that

$$\begin{cases} \big|\nu\big(\hat{g}_{\hat{K}_{\text{l.s.}},\hat{M}} - g^*\big)\big| \leq Z_{\hat{K}_{\text{l.s.}},\hat{M}}(g^*)\big(\|\hat{g}_{\hat{K}_{\text{l.s.}},\hat{M}} - g^*\|_2^2 + x^2_{\hat{K}_{\text{l.s.}},\hat{M}}\big), \\ \big|\nu\big(g^*_{K,M} - g^*\big)\big| \leq Z_{K,M}(g^*)\big(\|g^*_{K,M} - g^*\|_2^2 + x^2_{K,M}\big) \end{cases}$$

so that, using Lemma 16, for all $n \geq n_0$ and $x > 0$, with probability larger than $1 - e^{-x}$, for all $M \leq n$ and $K \leq n$,

$$\begin{aligned} \big|\nu\big(\hat{g}_{\hat{K}_{\text{l.s.}},\hat{M}} - g^*_{K,M}\big)\big| \leq{}& \frac{1}{4}\|\hat{g} - g^*\|_2^2 + \frac{1}{4}\|g^*_{K,M} - g^*\|_2^2 + A\frac{x}{n} \\ &+ \frac{1}{2}\operatorname{pen}(n, \hat{M}, \hat{K}_{\text{l.s.}}) + \frac{1}{2}\operatorname{pen}(n, M, K) \end{aligned}$$

so that

$$\begin{aligned} \|\hat{g} - g^*\|_2^2 \leq{}& \|g^*_{K,M} - g^*\|_2^2 + 2\operatorname{pen}(n, M, K) \\ &+ \frac{1}{2}\|\hat{g} - g^*\|_2^2 + \frac{1}{2}\|g^*_{K,M} - g^*\|_2^2 + 2A\frac{x}{n}, \end{aligned}$$

which means that

$$\frac{1}{2}\|\hat{g} - g^*\|_2^2 \leq \frac{3}{2}\|g^*_{K,M} - g^*\|_2^2 + 2\operatorname{pen}(n, M, K) + 2A\frac{x}{n}$$

and finally

$$\|\hat{g} - g^*\|_2^2 \leq 4 \inf_{K \leq n, M \leq n}\big\{\|g^*_{K,M} - g^*\|_2^2 + \operatorname{pen}(n, M, K)\big\} + 4A\frac{x}{n}$$

which is the expected inequality.

6.5. Proof of Lemma 8

In the following, we will identify the quadratic form M derived from the second order expansion of $x \mapsto \mathfrak{N}(x)$ and its matrix. Likewise, we will identify the quadratic form $M_{\mathfrak{C}}$ derived from the second order expansion of $x \mapsto \mathfrak{N}(I_{\mathfrak{C}}(x))$ with its matrix. Without loss of generality, one can assume $L = 3$.

Choice of parameters and expression of M

Let $\pi \in \Delta_{K^*}$ be the uniform distribution on $\mathcal{X}$, $\mathbf{Q} = \mathrm{Id}_{K^*}$ and $\mathbf{f}$ such that $\langle f_i, f_j \rangle = F\mathbf{1}_{i=j}$ for some constant $F > 0$. For instance, the f_i's are F times the indicating functions of distinct measurable sets with same measure $\frac{1}{F}$ for μ. In that case, $(f_i/\sqrt{F})_i$ is an orthonormal basis, and the quantity $g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}}$ can be broken down into three order one terms in p, q and A:

- the term in p: $\sum_i p_i f_i \otimes f_i \otimes f_i$;
- the term in q: $\sum_{i,k} q(i,k) f_i \otimes (f_i + f_k) \otimes f_k$;
- the term in A: $\sum_i ((Af)_i \otimes f_i \otimes f_i + f_i \otimes (Af)_i \otimes f_i + f_i \otimes f_i \otimes (Af)_i)$.

Now we can make the list of all second-order terms in the expansion of the quantity $\|g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}}\|_2^2$:

- p and p: $F^3 \sum_i p_i^2$;
- p and q: $2F^3 \sum_i p_i q(i,i)$;
- p and A: $3F^3 \sum_i p_i A_{i,i}$;
- q and q: $2F^3 \sum_{i,k} q(i,k)^2 + 2F^3 \sum_i q(i,i)^2$;
- q and A: $F^3 \sum_{i,k} q(i,k) A_{k,i} + F^3 \sum_{i,k} q(i,k) A_{i,k} + 4F^3 \sum_i q(i,i) A_{i,i}$;
- A and A: $6F^3 \sum_i A_{i,i}^2 + 3F^3 \sum_{i,k} A_{i,k}^2$.

We can now write the matrix M. In order to clarify the structure of this matrix, let us swap the components of the parameters (p, q, A) and consider the new parameters $(A_{\mathrm{diag}}, A_{\mathrm{else}}, p, q_{\mathrm{diag}}, q_{\mathrm{else}})$, where A_{diag} (resp. q_{diag}) is a vector of size K^* containing the diagonal coefficients of A (resp. q) and A_{else} (resp. q_{else}) contains its other coefficients. Then the matrix is:

$$M_{\mathrm{swapped}} = F^3 \left(\begin{array}{cc|c|cc} 9\mathrm{Id}_{K^*} & 0 & 3\mathrm{Id}_{K^*} & 6\mathrm{Id}_{K^*} & 0 \\ 0 & 3\mathrm{Id}_{K^*(K^*-1)} & 0 & 0 & X \\ \hline 3\mathrm{Id}_{K^*} & 0 & \mathrm{Id}_{K^*} & 2\mathrm{Id}_{K^*} & 0 \\ \hline 6\mathrm{Id}_{K^*} & 0 & 2\mathrm{Id}_{K^*} & 4\mathrm{Id}_{K^*} & 0 \\ 0 & X & 0 & 0 & 2\mathrm{Id}_{K^*(K^*-1)} \end{array}\right),$$

where $X[(A_{i,j})_{i\neq j}] = (A_{i,j} + A_{j,i})_{i\neq j}$.

Kernel of M

Subtracting the first block of lines to the third and fourth blocks of lines and then the first block of columns to the third and fourth blocks of columns does not change the rank and leads to the

matrix

$$F^3\left(\begin{array}{cc|c|cc} 9\mathrm{Id}_K & 0 & 0 & 0 & 0 \\ 0 & 3\mathrm{Id}_{K(K-1)} & 0 & 0 & X \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 2\mathrm{Id}_{K(K-1)} \end{array}\right).$$

Thus $\dim(\mathrm{Ker}(M)) \geq 2K$, where $\mathrm{Ker}(M)$ is the kernel of M and dim denotes the dimension. If one takes away the lines and columns corresponding to p and q_{diag}, one gets the matrix

$$F^3\begin{pmatrix} 9\mathrm{Id}_{K^*} & 0 & 0 \\ 0 & 3\mathrm{Id}_{K^*(K^*-1)} & X \\ 0 & X & 2\mathrm{Id}_{K^*(K^*-1)} \end{pmatrix}.$$

This matrix is invertible. Therefore, $\dim(\mathrm{Ker}(M)) = 2K$. Now, for all $i \in [K^*]$, let e_i^1 and e_i^2 be the vectors defined as

$$\begin{cases} \left(e_i^1\right)_{p_k} = 0 & \text{for all } k, \\ \left(e_i^1\right)_{A_{k,l}} = 0 & \text{for all } (k,l) \neq (i,i), \\ \left(e_i^1\right)_{q(k,l)} = 0 & \text{for all } (k,l) \neq (i,i), \\ \left(e_i^1\right)_{A_{i,i}} = 2, \\ \left(e_i^1\right)_{q(i,i)} = -3 \end{cases}$$

and

$$\begin{cases} \left(e_i^2\right)_{p_k} = 0 & \text{for all } k \neq i, \\ \left(e_i^2\right)_{A_{k,l}} = 0 & \text{for all } (k,l) \neq (i,i), \\ \left(e_i^2\right)_{q(k,l)} = 0 & \text{for all } (k,l), \\ \left(e_i^2\right)_{A_{i,i}} = 1, \\ \left(e_i^2\right)_{p_i} = -3. \end{cases}$$

One can easily check that these vectors are linearly independent and are all in $\mathrm{Ker}(M)$. Thus, they are a basis of the kernel of M: $\mathrm{Ker}(M) = \mathrm{Span}(\{e_i^1, e_i^2 | i \in [K]\})$.

Nondegeneracy of M restricted on $\mathfrak{C}$

Since M is symmetric, and thus diagonalisable in an orthonormal basis,

$$M = P_{\mathrm{Ker}(M)^\perp}^\top M_{\mathrm{Ker}(M)^\perp} P_{\mathrm{Ker}(M)^\perp}, \tag{6}$$

where $P_{\mathrm{Ker}(M)^\perp}$ is the orthogonal projection onto the space of vectors orthogonal to $\mathrm{Ker}(M)$ and $M_{\mathrm{Ker}(M)^\perp}$ is a symmetric positive definite matrix, whose smallest eigenvalue will be written c_0 in the following. The last step to conclude will require the two following lemmas:

Lemma 18. $\mathrm{Ker}(M) \cap \mathfrak{C} = \{0\}$.

Proof. Let $x \in \mathrm{Ker}(M) \cap \mathfrak{C}$, then $x = \sum_i (\lambda_i e_i^1 + \mu_i e_i^2)$ because $(e_i^1, e_i^2)_i$ is a basis of $\mathrm{Ker}(M)$. Since $x \in \mathfrak{C}$, one gets $\lambda_i = 0$ for all i because of the conditions on q. Then, the conditions on A imply $\mu_i = 0$ for all i, so that $x = 0$. □

Lemma 19. *There exists a constant $\kappa > 0$ such that for all $x \in \mathfrak{C}$,*

$$\| P_{\mathrm{Ker}(M)^\perp} x \|_F^2 \geq \kappa \| x \|_F^2. \tag{7}$$

Proof. $P_{\mathrm{Ker}(M)^\perp}$ is continuous. By compactness, the quantity

$$\kappa := \inf \left\{ \| P_{\mathrm{Ker}(M)^\perp} x \|_F^2 | x \in \mathfrak{C}, \| x \|_F^2 = 1 \right\}$$

is reached for some $x_0 \in \mathfrak{C} \setminus \{0\}$. If $\kappa = 0$, then $x_0 \in \mathrm{Ker}(M)$, but this is impossible because of Lemma 18. Therefore $\kappa > 0$. □

Finally, for all $x \in \mathfrak{C}$,

$$\begin{aligned} x^\top M x &= x^\top P_{\mathrm{Ker}(M)^\perp}^\top M_{\mathrm{Ker}(M)^\perp} P_{\mathrm{Ker}(M)^\perp} x \\ &= (P_{\mathrm{Ker}(M)^\perp} x)^\top M_{\mathrm{Ker}(M)^\perp} (P_{\mathrm{Ker}(M)^\perp} x) \\ &\geq c_0 \| P_{\mathrm{Ker}(M)^\perp} x \|_F^2 \\ &\geq c_0 \kappa \| x \|_F^2. \end{aligned}$$

Therefore, the quadratic form with matrix M is nondegenerate on $\mathfrak{C}$, which shows that H is non-zero for these $(\pi, \mathbf{Q}, \mathbf{f})$. To conclude, observe that H is continuous and that our choice of parameters can be approximated by parameters satisfying **[HX]** and **[HidA]**.

Acknowledgments

We would like to thank Elisabeth Gassiat for her precious advice. Thanks are also due to Yohann de Castro and Augustin Touron for their code examples which saved us a lot of time during the numerical experiments. We are grateful to the anonymous referee for his careful reading of this work.

Supplementary Material

Supplement A: Additional proofs (DOI: 10.3150/17-BEJ993SUPP; .pdf). We provide the algorithms we used in our simulations as well as the omitted proofs of our results.

References

[1] Alexandrovich, G., Holzmann, H. and Leister, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika* **103** 423–434. MR3509896

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

[3] Anandkumar, A., Hsu, D.J. and Kakade, S.M. (2012). A method of moments for mixture models and hidden Markov models. In *COLT* **1** 4.

[4] Baudry, J.-P., Maugis, C. and Michel, B. (2012). Slope heuristics: Overview and implementation. *Stat. Comput.* **22** 455–470.

[5] Bickel, P.J., Ritov, Y., Ryden, T. et al. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614–1635.

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

[7] Celeux, G. and Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.* **23** 541–564.

[8] Chambaz, A., Garivier, A. and Gassiat, E. (2009). A minimum description length approach to hidden Markov models with Poisson and Gaussian emissions. Application to order identification. *J. Statist. Plann. Inference* **139** 962–977.

[9] Couvreur, L. and Couvreur, C. (2000). Wavelet-based non-parametric HMM's: Theory and applications. In *Acoustics, Speech, and Signal Processing*, 2000. *ICASSP'*00. *Proceedings*. 2000 *IEEE International Conference on* **1** 604–607. New York: IEEE.

[10] de Castro, Y., Gassiat, É. and Lacour, C. (2016). Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.* **17** 1–43.

[11] de Castro, Y., Gassiat, E. and Le Corff, S. (2017). Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Inform. Theory*.

[12] DeVore, R.A. and Lorentz, G.G. (1993). *Constructive Approximation* **303**. Berlin: Springer Science & Business Media.

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

[14] Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. In *Annales de L'IHP Probabilités et Statistiques* **38** 897–906.

[15] Gassiat, E. and Boucheron, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inform. Theory* **49** 964–980.

[16] Gassiat, E., Cleynen, A. and Robin, S. (2015). Finite state space non parametric hidden Markov models are in general identifiable. *Stat. Comput.* 1–11.

[17] Gassiat, E. and Keribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM Probab. Stat.* **4** 25–52.

[18] Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075.

[19] Hansen, N. (2006). The CMA evolution strategy: A comparing review. In *Towards a New Evolutionary Computation* 75–102. Berlin: Springer.

[20] Hsu, D., Kakade, S.M. and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.* **78** 1460–1480.

[21] Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *J. Econometrics* **133** 97–126.

[22] Lambert, M.F., Whiting, J.P. and Metcalfe, A.V. (2003). A non-parametric hidden Markov model for climate state identification. *Hydrol. Earth Syst. Sci. Discuss.* **7** 652–667.

[23] Langrock, R., Kneib, T., Sohn, A. and DeRuiter, S.L. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics* **71** 520–528.
[24] Lefèvre, F. (2003). Non-parametric probability estimation for HMM-based automatic speech recognition. *Comput. Speech Lang.* **17** 113–136.
[25] Lehéricy, L. (2015). Estimation adaptative non paramétrique pour les modèles à chaîne de Markov cachée. Mémoire de M2, Orsay.
[26] Lehéricy, L. (2017). Supplement to "Consistent order estimation for nonparametric hidden Markov models." DOI:10.3150/17-BEJ993SUPP.
[27] Le Gland, F. and Mevel, L. (2000). Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems* **13** 63–93.
[28] Massart, P. (2007). Concentration inequalities and model selection. In *Lecture Notes in Mathematics* **1896**. Berlin: Springer.
[29] Robin, J.-M., Bonhomme, S. and Jochmans, K. (2014). Estimating multivariate latent-structure models.
[30] Robin, J.-M. and Smith, R.J. (2000). Tests of rank. *Econometric Theory* **16** 151–175.
[31] Shang, L. and Chan, K.-P. (2009). Nonparametric discriminant HMM and application to facial expression recognition. In *Computer Vision and Pattern Recognition*, 2009. *CVPR* 2009. *IEEE Conference on* 2090–2096. New York: IEEE.
[32] Stewart, G.W. and Sun, J.-G. (1990). *Matrix Perturbation Theory* (*Computer Science and Scientific Computing*). Boston: Academic Press.
[33] van Havre, Z., Rousseau, J., White, N. and Mengersen, K. (2016). Overfitting hidden Markov models with an unknown number of states. Preprint. Available at arXiv:1602.02466.
[34] Volant, S., Bérard, C., Martin-Magniette, M.-L. and Robin, S. (2014). Hidden Markov models with mixtures as emission distributions. *Stat. Comput.* **24** 493–504.
[35] Yau, C., Papaspiliopoulos, O., Roberts, G.O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 37–57. MR2797735

Received April 2017 and revised September 2017