



HAL
open science

MINIMAX ADAPTIVE ESTIMATION OF NONPARAMETRIC HIDDEN MARKOV MODELS

Yohann de Castro, Élisabeth Gassiat, Claire Lacour

► **To cite this version:**

Yohann de Castro, Élisabeth Gassiat, Claire Lacour. MINIMAX ADAPTIVE ESTIMATION OF NONPARAMETRIC HIDDEN MARKOV MODELS. 2015. hal-01105424v2

HAL Id: hal-01105424

<https://hal.science/hal-01105424v2>

Preprint submitted on 24 Jul 2015 (v2), last revised 27 Dec 2015 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MINIMAX ADAPTIVE ESTIMATION OF NONPARAMETRIC HIDDEN MARKOV MODELS

Y. DE CASTRO, É. GASSIAT, AND C. LACOUR

ABSTRACT. We consider stationary hidden Markov models with finite state space and nonparametric modeling of the emission distributions. It has remained unknown until very recently that such models are identifiable. In this paper, we propose a new penalized least-squares estimator for the emission distributions which is statistically optimal and practically tractable. We prove a non asymptotic oracle inequality for our nonparametric estimator of the emission distributions. A consequence is that this new estimator is rate minimax adaptive up to a logarithmic term. Our methodology is based on projections of the emission distributions onto nested subspaces of increasing complexity. The popular spectral estimators are unable to achieve the optimal rate but may be used as initial points in our procedure. Simulations are given that show the improvement obtained when applying the least-squares minimization consecutively to the spectral estimation.

1. INTRODUCTION

1.1. Context and motivations. Finite state space hidden Markov models (HMMs for short) are widely used to model data evolving in time and coming from heterogeneous populations. They seem to be reliable models to depict practical situations in a variety of applications such as economics, genomics, signal processing and image analysis, ecology, environment, speech recognition, to name but a few. From a statistical view point, finite state space HMMs are stochastic processes $(X_j, Y_j)_{j \geq 1}$ where $(X_j)_{j \geq 1}$ is a Markov chain living in a finite state space and conditionally on $(X_j)_{j \geq 1}$ the Y_j 's are independent with a distribution depending only on X_j and living in \mathcal{Y} . The observations are $Y_{1:N} = (Y_1, \dots, Y_N)$ and the associated states $X_{1:N} = (X_1, \dots, X_N)$ are unobserved. The parameters of the model are the initial distribution, the transition matrix of the hidden chain, and the emission distributions of the observations, that is the probability distributions of the Y_j 's conditionally to $X_j = x$ for all possible x 's. In this paper we shall consider stationary ergodic HMMs so that the initial distribution is the stationary distribution of the (ergodic) hidden Markov chain.

Until very recently, asymptotic performances of estimators were proved theoretically only in the parametric frame (that is, with finitely many unknown parameters). Though, nonparametric methods for HMMs have been considered in applied papers, but with no theoretical guarantees, see for instance [CC00] for voice activity detection, [LWM03] for climate state identification, [Lef03] for automatic speech recognition, [SC09] for facial expression recognition, [VBMMR13] for methylation comparison of proteins, [YPRH11] for copy number variants identification in DNA analysis.

The preliminary obstacle to obtain theoretical results on general finite state space nonparametric HMMs was to understand when such models are indeed identifiable.

Date: July 24, 2015.

Key words and phrases. Nonparametric estimation; Hidden markov models; Spectral Method; Concentration inequality; Model selection; Adaptive estimation.

Marginal distributions of finitely many observations are finite mixtures of products of the emission distributions. It is clear that identifiability can not be obtained based on the marginal distribution of only one observation. It is needed, and it is enough, to consider the marginal distribution of at least three consecutive observations to get identifiability, see [GCR15], following [AMR09] and [HKZ12].

1.2. Contribution. The aim of our paper is to propose a new approach to estimate nonparametric HMMs with a statistically optimal and practically tractable method. We obtain this way nonparametric estimators of the emission distributions that achieve the minimax rate of estimation in an adaptive setting.

Our perspective is based on estimating the projections of the emission laws onto nested subspaces of increasing complexity. Our analysis encompasses any family of nested subspaces of Hilbert spaces and works with a large variety of models. In this framework one could think to use the spectral estimators as proposed by [HKZ12, AHK12] in the parametric frame, by extending them to the nonparametric frame. But a careful analysis of the tradeoff between sampling size and approximation complexity shows that they do not lead to rate optimal estimators of the emission densities, see [DCGLC15] for a formal statement and proof. This can be easily understood. Indeed, the spectral estimators of the emission densities are computed as functions of the empirical estimator of the marginal distribution of three consecutive observations on \mathcal{Y}^3 , for which, roughly speaking, when \mathcal{Y} is a subset of \mathbb{R} , the optimal rate is $N^{-s/(2s+3)}$, N being the number of observations and s the smoothness of the emission densities. Thus the rate obtained this way for the emission densities is also $N^{-s/(2s+3)}$. But since those emission densities describe one dimensional random variables on \mathcal{Y} , one could hope to be able to obtain the sharper rate $N^{-s/(2s+1)}$. This is the rate we obtain, up to a $\log N$ term, with our new method. Let us explain how it works.

Using the HMM modeling, and using sieves for the emission densities on \mathcal{Y} , we propose a penalized least squares estimator in the model selection frame. We prove an oracle inequality for the L_2 -risk of the estimator of the density of (Y_1, Y_2, Y_3) , see Theorem 2. Since the complexity of the model is that given by the sieves for the emission densities, this leads, up to a $\log N$ term, to the adaptive minimax rate computed as for the density of only one observation Y_1 though we estimate the density of (Y_1, Y_2, Y_3) . Roughly speaking, when the observations are one dimensional, that is when \mathcal{Y} is a subset of \mathbb{R} , the obtained rate for the density of (Y_1, Y_2, Y_3) is of order $N^{-s/(2s+1)}$ up to a $\log N$ term, N being the number of observations and s the smoothness of the emission densities.

The key point is then to be able to go back to the emission densities. This is the cornerstone of our main result. We prove in Theorem 3 that, under the assumption [HD] defined in Section 4.2, the quadratic risk for the density of (Y_1, Y_2, Y_3) is lower bounded by some positive constant multiplied by the quadratic risk for the emission densities. This appropriate assumption is generically satisfied in the sense that it holds for all possible emission densities for which the L_2 -norms and Hilbert dot products do not lie on a particular algebraic surface with coefficients depending on the transition matrix of the hidden chain. Moreover, we prove that, when the number of hidden states equals two, this appropriate assumption is always verified when the two emission densities are distinct, see Lemma 2.

Our methodology requires that we have a preliminary estimator of the transition matrix. To get such an estimator, it is possible to use spectral methods. Thus our approach is the following. First, get a preliminary estimator of the initial distribution and the transition matrix of the hidden chain. Second, apply penalized least squares estimation on the density of three consecutive observations, using HMM modeling, model selection on the emission densities, and initial distribution

and stationary matrix of the hidden chain set at the estimated value. This gives emission density estimators which have minimax adaptive rate, as our main result states, see Theorem 4. A simplified version of this theorem can be given as follows.

Theorem 1 — *Assume $(Y_j)_{j \geq 1}$ is a Hidden Markov chain on \mathbb{R} , with latent Markov chain $(X_j)_{j \geq 1}$ with K possible values and true transition matrix \mathbf{Q}^* . Denote f_k^* the density of Y_n given $X_n = k$, for $k = 1, \dots, K$. Assume the true transition matrix \mathbf{Q}^* is full rank and the true emission densities f_k^* , $k = 1, \dots, K$ are linearly independent, with smoothness s . Assume that [HD] holds true. Then, up to label switching, for N the number of observations large enough, the estimators $\hat{\mathbf{Q}}, \hat{f}_k$, $k = 1, \dots, K$ built in Section 3 and 5 enjoy*

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{Q}^* - \hat{\mathbf{Q}}\|^2 \right] &= O\left(\frac{\log N}{N}\right), \\ \mathbb{E} \left[\|f_k^* - \hat{f}_k\|_2^2 \right] &= O\left(\left[\frac{\log N}{N}\right]^{\frac{s}{2s+1}}\right), \quad k = 1, \dots, K. \end{aligned}$$

Moreover, since the family of sieves we consider is that given by finite dimensional spaces described by an orthonormal basis, we are able to use the spectral estimators of the coefficients of the densities as initial points in the least squares minimization. This is important since, in the HMM framework, least squares minimization does not have an explicit solution and may lead to several local minima. However, since the spectral estimates are proved to be consistent, we may be confident that their use as initial point is enough. Simulations indeed confirm this point.

To conclude we claim that our results support a powerful new approach to estimate, for the first time, nonparametric HMMs with a statistically optimal and practically tractable method.

1.3. Related works. The papers [AMR09], [HKZ12] and [AHK12] paved the way to obtain identifiability under reasonable assumptions. In [AHK12] the authors point out a structural link between multivariate mixtures with conditionally independent observations and finite state space HMMs. In [HKZ12] the authors propose a spectral method to estimate all parameters for finite state space HMMs (with finitely many observations), under the assumption that the transition matrix of the hidden chain is non singular, and that the (finitely valued) emission distributions are linearly independent. Extension to emission distributions on any space, under the linear independence assumptions (and keeping the assumption of non singularity of the transition matrix), allowed to prove the general identifiability result for finite state space HMMs, see [GCR15], where also model selection likelihood methods and nonparametric kernel methods are proposed to get nonparametric estimators. Let us notice also [Ver15] that proves theoretical consistency of the posterior in nonparametric Bayesian methods for finite state space HMMs with adequate assumptions. Later, [AH14] obtained identifiability when the emission distributions are all distinct (not necessarily linearly independent) and still when the transition matrix of the hidden chain is full rank. In the nonparametric multivariate mixture model, [SADX14] prove that any linear functional of the emission distributions may be estimated with parametric rate of convergence in the context of reproducing kernel Hilbert spaces. The latter uses spectral methods, not the same but similar to the ones proposed in [HKZ12] and [AHK12].

Recent papers that contain theoretical results on different kinds of nonparametric HMMs are [GR13], where the emitted distributions are translated of each other, and [DL12] in which the authors consider regression models with hidden regressor variables that can be Markovian on a continuous state space.

1.4. Outline of the paper. In Section 2, we set the notations, the model we shall study, and the assumptions we shall consider. We then state an identifiability

Lemma that will be useful for our estimation method. In Sections 3 and 4 we give our main results. We explain the penalized least-squares estimation method in Section 3, and we prove in Section 4 that, when the transition matrix is irreducible and aperiodic, when the emission distributions are linearly independent and the penalty is adequately chosen, then, under a generic assumption, the penalized least squares estimator is asymptotically minimax adaptive up to a $\log N$ term, see Theorem 4 and Corollary 2. For this, we first prove an oracle inequality for the estimation of the density of (Y_1, Y_2, Y_3) , see Theorem 2, then we prove the key result relating the risk of the density of (Y_1, Y_2, Y_3) to that of the emission densities, see Theorem 3. The latter holds under a generic assumption which we prove to be always verified in case $K = 2$, see Lemma 2. Finally, we need the performances of the spectral estimator of the transition matrix and of the stationary distribution which are given in Section 5, see Theorem 5, proved in [DCGLC15]. We finally present simulations in Section 6 to illustrate our theoretical results. Those simulations show in particular the improvement obtained when applying the least-squares minimization consecutively to the spectral estimation. Detailed proofs are given in Section 7.

2. NOTATIONS AND ASSUMPTIONS

2.1. Nonparametric Hidden Markov Model. Let K, D be positive integers and let \mathcal{L}^D be the Lebesgue measure on \mathbb{R}^D . Denote by \mathcal{X} the set $\{1, \dots, K\}$ of hidden states, $\mathcal{Y} \subset \mathbb{R}^D$ the observation space, and Δ_K the space of probability measures on \mathcal{X} identified to the $(K - 1)$ -dimensional simplex. Let $(X_n)_{n \geq 1}$ be a Markov chain on \mathcal{X} with $K \times K$ transition matrix \mathbf{Q}^* and initial distribution $\pi^* \in \Delta_K$. Let $(Y_n)_{n \geq 1}$ be a sequence of observed random variables on \mathcal{Y} . Assume that, conditional on $(X_n)_{n \geq 1}$, the observations $(Y_n)_{n \geq 1}$ are independent and, for all $n \in \mathbb{N}$, the distribution of Y_n depends only on X_n . Denote by μ_k^* the conditional law of Y_n conditional on $\{X_n = k\}$, and assume that μ_k^* has density f_k^* with respect to the measure \mathcal{L}^D on \mathcal{Y} :

$$\forall k \in \mathcal{X}, \quad d\mu_k^* = f_k^* d\mathcal{L}^D.$$

Denote by $\mathfrak{F}^* := \{f_1^*, \dots, f_K^*\}$ the set of emission densities with respect to the Lebesgue measure. Then, for any integer n , the distribution of (Y_1, \dots, Y_n) has density with respect to $(\mathcal{L}^D)^{\otimes n}$

$$\sum_{k_1, \dots, k_n=1}^K \pi^*(k_1) \mathbf{Q}^*(k_1, k_2) \dots \mathbf{Q}^*(k_{n-1}, k_n) f_{k_1}^*(y_1) \dots f_{k_n}^*(y_n).$$

We shall denote g^* the density of (Y_1, Y_2, Y_3) .

In this paper we shall address two observations schemes. We shall consider N i.i.d. samples $(Y_1^{(s)}, Y_2^{(s)}, Y_3^{(s)})_{s=1}^N$ of three consecutive observations (**Scenario A**) or consecutive observations of the same chain (**Scenario B**):

$$\forall s \in \{1, \dots, N\}, \quad (Y_1^{(s)}, Y_2^{(s)}, Y_3^{(s)}) := (Y_s, Y_{s+1}, Y_{s+2}).$$

2.2. Projections of the population joint laws. Denote by $(\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D), \|\cdot\|_2)$ the Hilbert space of square integrable functions on \mathcal{Y} with respect to the Lebesgue measure \mathcal{L}^D equipped with the usual inner product $\langle \cdot, \cdot \rangle$ on $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Assume $\mathfrak{F}^* \subset \mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$.

Let $(M_r)_{r \geq 1}$ be an increasing sequence of integers, and let $(\mathfrak{P}_{M_r})_{r \geq 1}$ be a sequence of nested subspaces such that their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Let $\Phi_{M_r} := \{\varphi_1, \dots, \varphi_{M_r}\}$ be an orthonormal basis of \mathfrak{P}_{M_r} . Recall that for all $f \in \mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$,

$$(1) \quad \lim_{r \rightarrow \infty} \sum_{m=1}^{M_r} \langle f, \varphi_m \rangle \varphi_m = f,$$

in $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Note that changing M_r may change all functions φ_m , $1 \leq m \leq M_r$ in the basis Φ_{M_r} , which we shall not indicate in the notation for sake of readability. Also, we drop the dependence on r and write M instead of M_r . Define the projection of the emission laws onto \mathfrak{F}_M by

$$\forall k \in \mathcal{X}, \quad f_{M,k}^* := \sum_{m=1}^M \langle f_k^*, \varphi_m \rangle \varphi_m.$$

We shall write $\mathbf{f}_M^* := (f_{M,1}^*, \dots, f_{M,K}^*)$ and $\mathbf{f}^* := (f_1^*, \dots, f_K^*)$ throughout this paper.

Remark — *One can consider the following standard examples:*

- (**Spline**) *The space of piecewise polynomials of degree bounded by d_r based on the regular partition with p_r^D regular pieces on $\mathcal{Y} = [0, 1]^D$. It holds that $M_r = (d_r + 1)^D p_r^D$.*
- (**Trig.**) *The space of real trigonometric polynomials on $\mathcal{Y} = [0, 1]^D$ with degree less than r . It holds that $M_r = (2r + 1)^D$.*
- (**Wav.**) *A wavelet basis Φ_{M_r} of scale r on $\mathcal{Y} = [0, 1]^D$, see [Mey92]. It holds that $M_r = 2^{(r+1)D}$.*

2.3. Assumptions. We shall use the following assumptions on the hidden chain.

- [**H1**] *The transition matrix \mathbf{Q}^* has full rank,*
- [**H2**] *The Markov chain $(X_n)_{n \geq 1}$ is irreducible and aperiodic,*
- [**H3**] *The initial distribution $\pi^* = (\pi_1^*, \dots, \pi_K^*)$ is the stationary distribution*

Notice that under [**H1**], [**H2**] and [**H3**], one has

$$\forall k \in \mathcal{X}, \quad \pi_k^* \geq \pi_{\min}^* > 0.$$

We shall use the following assumption on the emission densities.

- [**H4**] *The family of emission densities $\mathfrak{F}^* := \{f_1^*, \dots, f_K^*\}$ is linearly independent.*
- Those assumptions appear in spectral methods, see for instance [HKZ12, AHK12], and in identifiability issues, see for instance [AMR09, GCR15, AH14].

2.4. Identifiability lemma. For any $\mathbf{f} = (f_1, \dots, f_K) \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$ and any irreducible transition matrix \mathbf{Q} , denote $g^{\mathbf{Q}, \mathbf{f}} : \mathcal{Y}^3 \rightarrow \mathbb{R}$ the function given by

$$(2) \quad g^{\mathbf{Q}, \mathbf{f}}(y_1, y_2, y_3) = \sum_{k_1, k_2, k_3=1}^K \pi(k_1) \mathbf{Q}(k_1, k_2) \mathbf{Q}(k_2, k_3) f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3),$$

where π is the stationary distribution of \mathbf{Q} . When $\mathbf{Q} = \mathbf{Q}^*$ and $\mathbf{f} = \mathbf{f}^*$, we get $g^{\mathbf{Q}^*, \mathbf{f}^*} = g^*$. When f_1, \dots, f_K are probability densities on \mathcal{Y} , $g^{\mathbf{Q}, \mathbf{f}}$ is the probability distribution of three consecutive observations of a stationary HMM. We now state a lemma that gathers all what we need about identifiability.

For any transition matrix \mathbf{Q} , let $T_{\mathbf{Q}}$ be the set of permutations τ such that for all i and j , $\mathbf{Q}(\tau(i), \tau(j)) = \mathbf{Q}(i, j)$. The permutations in $T_{\mathbf{Q}}$ describe how the states of the Markov chain may be permuted without changing the distribution of the whole chain: for any τ in $T_{\mathbf{Q}}$, $(\tau(X_n))_{n \geq 1}$ has the same distribution as $(X_n)_{n \geq 1}$. Since the hidden chain is not observed, if the emission distributions are permuted using τ , we get the same HMM. In other words, if $\mathbf{f}^\tau = (f_{\tau(1)}, \dots, f_{\tau(K)})$, then $g^{\mathbf{Q}, \mathbf{f}^\tau} = g^{\mathbf{Q}, \mathbf{f}}$. Since identifiability up to permutation of the hidden states is obtained from the marginal distribution of three consecutive observations, we get the following lemma whose detailed proof is given in Section 7.1.

Lemma 1 — *Assume that \mathbf{Q} is a transition matrix for which [**H1**] and [**H2**] hold. Assume that [**H4**] holds. Then for any $\mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$,*

$$g^{\mathbf{Q}, \mathbf{f}^* + \mathbf{h}} = g^{\mathbf{Q}, \mathbf{f}^*} \iff \exists \tau \in T_{\mathbf{Q}} \text{ such that } h_j = f_{\tau(j)}^* - f_j^*, \quad j = 1, \dots, K.$$

In particular, if $T_{\mathbf{Q}}$ reduces to the identity permutation,

$$g^{\mathbf{Q}, \mathbf{f}^* + \mathbf{h}} = g^{\mathbf{Q}, \mathbf{f}^*} \iff \mathbf{h} = (0, \dots, 0).$$

3. THE PENALIZED LEAST-SQUARES ESTIMATOR

In this section we shall estimate the emission densities using the so-called penalized least squares method. Here, the least squares adjustment is made on the density g^* of (Y_1, Y_2, Y_3) . Starting from the operator $\Gamma : t \mapsto \|t - g^*\|_2^2 - \|g^*\|_2^2 = \|t\|_2^2 - 2 \int t g^*$ which is minimum for the target g^* , we introduce the corresponding empirical contrast γ_N . Namely, for any $t \in \mathbf{L}^2(\mathcal{Y}^3, \mathcal{L}^{D \otimes 3})$, set

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

with $Z_s := (Y_1^s, Y_2^s, Y_3^s)$ (**Scenario A**) or $Z_s := (Y_s, Y_{s+1}, Y_{s+2})$ (**Scenario B**). As N tends to infinity, $\gamma_N(t) - \gamma_N(g^*)$ converges almost surely to $\|t - g^*\|_2^2$, thus the name least squares contrast function. A natural estimator is then a function t such that $\gamma_N(t)$ is minimum over a judicious approximation space which is a set of functions of form $g^{\mathbf{Q}, \mathbf{f}}$, \mathbf{Q} a transition matrix and $\mathbf{f} \in \mathcal{F}^K$, for \mathcal{F} a subset of $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$. We thus define a whole collection of estimates \hat{g}_M , each M indexing an approximation subspace (also called model). Considering (2) we shall introduce a collection of model of functions by projection of possible \mathbf{f} 's on the subspaces $(\mathfrak{P}_M)_M$. Thus, for any irreducible transition matrix \mathbf{Q} with stationary distribution π , we define $\mathcal{S}(\mathbf{Q}, M)$ as the set of functions $g^{\mathbf{Q}, \mathbf{f}}$ such that $\mathbf{f} \in \mathcal{F}^K$ and, for each $k = 1, \dots, K$, there exists $(a_{m,k})_{1 \leq m \leq M} \in \mathbb{R}^M$ such that

$$f_k = \sum_{m=1}^M a_{m,k} \varphi_m.$$

We now assume that we have in hand an estimator $\hat{\mathbf{Q}}$ of \mathbf{Q}^* . For instance, one can use a spectral estimator, we recall such a construction in Section 5. Then, $(\mathcal{S}(\hat{\mathbf{Q}}, M))_M$ is the collection of models we use for the least squares minimization. For any M , define \hat{g}_M as a minimizer of $\gamma_N(t)$ for $t \in \mathcal{S}(\hat{\mathbf{Q}}, M)$. Then \hat{g}_M can be written as

$$\hat{g}_M = g^{\hat{\mathbf{Q}}, \hat{\mathbf{f}}_M}$$

with $\hat{\mathbf{f}}_M \in \mathcal{F}^K$ and

$$\hat{f}_{M,k} = \sum_{m=1}^M \hat{a}_{m,k} \varphi_m, \quad k = 1, \dots, K$$

for some $(\hat{a}_{m,k})_{1 \leq m \leq M} \in \mathbb{R}^M$, $k = 1, \dots, K$. It then remains to select the best model, that is to choose M which minimizes $\|\hat{g}_M - g^*\|_2^2 - \|g^*\|_2^2$. This quantity is close to $\gamma_N(\hat{g}_M)$, but we need to take into account the deviations of the process $\Gamma - \gamma_N$. Then we rather minimize $\gamma_N(\hat{g}_M) + \text{pen}(N, M)$ where $\text{pen}(N, M)$ is a penalty term to be specified. Our final estimator will be a penalized least squares estimator. For this purpose we set a penalty function $\text{pen}(N, M)$ and choose

$$\hat{M} = \arg \min_{M=1, \dots, N} \{\gamma_N(\hat{g}_M) + \text{pen}(N, M)\}.$$

Notice that, with N observations, we consider N subspaces as candidates for model selection. Then the estimator of g^* is $\hat{g} = \hat{g}_{\hat{M}}$, and the estimator of \mathbf{f}^* is $\hat{\mathbf{f}} := \hat{\mathbf{f}}_{\hat{M}}$ so that $\hat{g} = g^{\hat{\mathbf{Q}}, \hat{\mathbf{f}}}$.

The least squares estimator does not have an explicit form such as in usual non-parametric estimation, so that one has to use numerical minimization algorithms.

As initial point of the minimization algorithm, we shall use the spectral estimator, see Section 6 for more details. Since the spectral estimator is consistent, see [DCGLC15], the algorithm does not suffer from initialization problems.

4. ADAPTIVE ESTIMATION OF THE EMISSION DISTRIBUTIONS

4.1. Oracle inequality for the estimation of g^* . We now fix a bounded subset \mathcal{F} of $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$, and we shall use the following assumption:

[HF] \mathcal{F} is a closed bounded subset of $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$ such that: for any $f \in \mathcal{F}$, $\int f d\mathcal{L}^D = 1$ and $\|f\|_\infty \leq C_{\mathcal{F}, \infty}$ for some fixed $C_{\mathcal{F}, \infty} > 0$.

We denote $C_{\mathcal{F}, 2} = \sup_{f \in \mathcal{F}} \|f\|_2 < \infty$.

Our first main result is an oracle inequality for the estimation of g^* which is stated below and proved in Section 7.2. We denote by \mathfrak{S}_K the set of permutations of $\{1, \dots, K\}$. When a is a vector, $\|a\|_2$ denotes its euclidian norm, when A is a matrix, $\|A\|_F$ denotes its Frobenius norm and $\|A\|$ its operator norm.

Theorem 2 — Assume [H1]-[H4] and [HF]. Assume also $\mathbf{f}^* \in \mathcal{F}^K$, and for all M , $\mathbf{f}_M^* \in \mathcal{F}^K$. Then, there exists positive constants N_0 , ρ^* and A_1^* (depending on $C_{\mathcal{F}, 2}$ and $C_{\mathcal{F}, \infty}$ (**Scenario B**) or on \mathbf{Q}^* , $C_{\mathcal{F}, 2}$ and $C_{\mathcal{F}, \infty}$ (**Scenario A**)) such that, if

$$\text{pen}(N, M) \geq \rho^* \frac{M \log N}{N}$$

then for all $x > 0$, for all $N \geq N_0$, one has with probability $1 - (e - 1)^{-1}e^{-x}$, for any permutation $\tau \in \mathfrak{S}_K$,

$$\begin{aligned} \|\hat{g} - g^*\|_2^2 &\leq 6 \inf_M \left\{ \|g^* - g^{\mathbf{Q}^*, \mathbf{f}_M^*}\|_2^2 + \text{pen}(N, M) \right\} + A_1^* \frac{x}{N} \\ &\quad + 18C_{\mathcal{F}, 2}^6 (2\|\mathbf{Q}^* - \mathbb{P}_\tau \hat{\mathbf{Q}}_N \mathbb{P}_\tau^\top\|_F^2 + \|\pi^* - \mathbb{P}_\tau \hat{\pi}\|_2^2). \end{aligned}$$

Here, \mathbb{P}_τ is the permutation matrix associated to τ .

The important fact in this oracle inequality is that the minimal possible penalty is of order M/N (up to logarithmic terms) and not M^3/N as is usually the case when estimating a joint density of three random variables, so that we get a minimax rate adaptive estimator of g^* .

4.2. Main result. The problem is now to deduce from Theorem 2 a result on $\|f_k^* - \hat{f}_k\|_2^2$, $k = 1, \dots, K$. This is the cornerstone of our work: we prove that, under a generic assumption on the parameters of the unknown HMM, a direct lower bound links, up to some positive constant, $\|\hat{g} - g^*\|_2^2$ to $\sum_{k=1}^K \|f_k^* - \hat{f}_k\|_2^2$. Let us now describe the assumption and comment on its genericity.

For any $\mathbf{f} \in \mathcal{F}^K$, define $G(\mathbf{f})$ the $K \times K$ matrix with coefficients $G(\mathbf{f})_{i,j} = \langle f_i, f_j \rangle$, $i, j = 1, \dots, K$. Notice that under the assumption [H4], $G(\mathbf{f}^*)$ is positive definite. Let \mathbf{Q} be a transition matrix verifying [H1]-[H2] and let A_Q be the diagonal matrix having the stationary distribution of \mathbf{Q} on the diagonal. We shall now define a quadratic form with coefficients depending on \mathbf{Q} and $G(\mathbf{f})$. If U is a $K \times K$

matrix such that $U\mathbf{1}_K = 0$,

$$\begin{aligned} & \sum_{i,j=1}^K \left\{ (\mathbf{Q}^T A_Q U G(\mathbf{f}) U^T A_Q \mathbf{Q})_{i,j} (G(\mathbf{f}))_{i,j} (\mathbf{Q} G(\mathbf{f}) \mathbf{Q}^T)_{i,j} \right. \\ & \quad + (\mathbf{Q}^T A_Q G(\mathbf{f}) A_Q \mathbf{Q})_{i,j} (U G(\mathbf{f}) U^T)_{i,j} (\mathbf{Q} G(\mathbf{f}) \mathbf{Q}^T)_{i,j} \\ & \quad + (\mathbf{Q}^T A_Q G(\mathbf{f}) A_Q \mathbf{Q})_{i,j} (G(\mathbf{f}))_{i,j} (\mathbf{Q} U G(\mathbf{f}) U^T \mathbf{Q}^T)_{i,j} \left. \right\} \\ & \quad + 2 \sum_{i,j} \left\{ (\mathbf{Q}^T A_Q U G(\mathbf{f}) A_Q \mathbf{Q})_{i,j} (U G(\mathbf{f}))_{j,i} (\mathbf{Q} G(\mathbf{f}) \mathbf{Q}^T)_{i,j} \right. \\ & \quad + (\mathbf{Q}^T A_Q U G(\mathbf{f}) A_Q \mathbf{Q})_{i,j} (\mathbf{Q} U G(\mathbf{f}) \mathbf{Q}^T)_{j,i} (G(\mathbf{f}))_{i,j} \\ & \quad \left. + (U G(\mathbf{f}))_{i,j} (\mathbf{Q} U G(\mathbf{f}) \mathbf{Q}^T)_{j,i} (\mathbf{Q}^T A_Q G(\mathbf{f}) A_Q \mathbf{Q})_{i,j} \right\} \end{aligned}$$

defines a semidefinite positive quadratic form \mathcal{D} in the coefficients $U_{i,j}$, $i = 1, \dots, K$, $j = 1, \dots, K - 1$. The determinant of this quadratic form is a polynomial in the coefficients of the matrices \mathbf{Q} , A_Q and $G(\mathbf{f})$. Since the coefficients of A_Q are rational functions of the coefficients of the matrix \mathbf{Q} , this determinant is also a rational function of the coefficients of the matrices \mathbf{Q} and $G(\mathbf{f})$. Define $H(\mathbf{Q}, G(\mathbf{f}))$ the numerator of the determinant. Then $H(\mathbf{Q}, G(\mathbf{f}))$ is a polynomial in the coefficients of the matrices \mathbf{Q} and $G(\mathbf{f})$. Our assumption will be:

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

Since H is a polynomial function of $Q_{i,j}^*$, $i = 1, \dots, K$, $j = 1, \dots, K - 1$, and $\langle f_i^*, f_j^* \rangle$, $i, j = 1, \dots, K$, the assumption $[\mathbf{HD}]$ is generically satisfied. We have been able to prove that Assumption $[\mathbf{HD}]$ always holds in the case $K = 2$. We were only able to prove this result by brute force, it is given in Section 7.4 and Appendix B.

Lemma 2 — Assume $K = 2$. Then for all \mathbf{Q}^* and \mathbf{f}^* such that $[\mathbf{H1}]-[\mathbf{H4}]$ hold, $H(\mathbf{Q}^*, G(\mathbf{f}^*)) > 0$.

Notice now that, when $[\mathbf{HD}]$ and $[\mathbf{H1}]-[\mathbf{H3}]$ hold, it is possible to define a compact neighborhood \mathcal{V} of \mathbf{Q}^* such that, for all $\mathbf{Q} \in \mathcal{V}$, $H(\mathbf{Q}, G(\mathbf{f}^*)) \neq 0$ and $[\mathbf{H1}]-[\mathbf{H3}]$ hold for \mathbf{Q} .

For any $\mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$, define $\|\mathbf{h}\|_{\mathbf{Q}}^2 := \min_{\tau \in \mathcal{T}_{\mathbf{Q}}} \{ \sum_{k=1}^K \|h_k + f_k^* - f_{\tau(k)}^*\|_2^2 \}$. We may now state the theorem which is the cornerstone of our main result.

Theorem 3 — Assume $[\mathbf{H1}]-[\mathbf{H4}]$ and $[\mathbf{HD}]$. Let \mathcal{K} be a closed bounded subset of $(\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$ such that if $\mathbf{h} \in \mathcal{K}$, then $\int h_i d\mathcal{L}^D = 0$, $i = 1, \dots, K$. Let \mathcal{V} be a compact neighborhood of \mathbf{Q}^* such that, for all $\mathbf{Q} \in \mathcal{V}$, $H(\mathbf{Q}, G(\mathbf{f}^*)) \neq 0$ and $[\mathbf{H1}]-[\mathbf{H3}]$ holds for \mathbf{Q} . Then there exists a positive constant $c(\mathcal{K}, \mathcal{V}, \mathfrak{F}^*)$ such that

$$\forall \mathbf{h} \in \mathcal{K}, \forall \mathbf{Q} \in \mathcal{V}, \|g^{\mathbf{Q}, \mathbf{f}^* + \mathbf{h}} - g^{\mathbf{Q}, \mathbf{f}^*}\|_2 \geq c(\mathcal{K}, \mathcal{V}, \mathfrak{F}^*) \|\mathbf{h}\|_{\mathbf{Q}}.$$

This theorem is proved in Section 7.3.

We are now ready to prove our main result on the penalized least squares estimator of the emission densities. The following theorem gives an oracle inequality for the estimators of the emission distributions provided the penalty is adequately chosen.

Theorem 4 (Adaptive estimation) — Assume $[\mathbf{H1}]-[\mathbf{H4}]$, $[\mathbf{HF}]$ and $[\mathbf{HD}]$. Assume also that for all M , $\mathbf{f}_M^* \in \mathcal{F}^K$. Let \mathcal{V} be a compact neighborhood of \mathbf{Q}^* such that, for all $\mathbf{Q} \in \mathcal{V}$, $H(\mathbf{Q}, G(\mathbf{f}^*)) \neq 0$ and $[\mathbf{H1}]-[\mathbf{H3}]$ holds for \mathbf{Q} . Then, there exists a positive constant A^* (depending on \mathcal{V} , \mathbf{f}^* , $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$) and positive

constants N_0 and ρ^* (depending on $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ (**Scenario A**) or on \mathbf{Q}^* , $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ (**Scenario B**)) such that, if

$$\text{pen}(N, M) \geq \rho^* \frac{M \log N}{N}$$

then for all $x > 0$, for all $N \geq N_0$, for any permutation $\tau_N \in \mathfrak{S}_K$, with probability larger than $1 - (e-1)^{-1}e^{-x} - \mathbb{P}\left(\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top \notin \mathcal{V}\right)$,

$$\begin{aligned} \sum_{k=1}^K \|f_k^* - \hat{f}_{\tau_N(k)}\|_2^2 &\leq A^* \left[\inf_M \left\{ \sum_{k=1}^K \|f_k^* - f_{M,k}^*\|_2^2 + \text{pen}(N, M) \right\} \right. \\ &\quad \left. + \|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\|_F^2 + \|\pi^* - \mathbb{P}_{\tau_N} \hat{\pi}\|_2^2 + \frac{x}{N} \right]. \end{aligned}$$

Remark — As usual in HMM or mixture model estimation, it is only possible to estimate the model up to label switching of the hidden states, this is the meaning of the permutation τ_N .

Remark — An important consequence of the theorem is that a right choice of the penalty leads to a rate minimax adaptive estimator up to a $\log N$ term, see Corollary 1 below. For this purpose, one has to choose an estimator $\hat{\mathbf{Q}}$ of \mathbf{Q}^* which is, up to label switching, consistent with controlled rate. One possible choice is a spectral estimator.

Proof. Let $\mathcal{K} = \{\mathbf{h} = \mathbf{f} - \mathbf{f}^*, \mathbf{f} \in \mathcal{F}^K\}$. Using Theorem 2 we get that for all $x > 0$, for all $N \geq N_0$, with probability $1 - (e-1)^{-1}e^{-x}$, one has for any permutation $\tilde{\tau}_N$,

$$(3) \quad \begin{aligned} \|\hat{g} - g^*\|_2^2 &\leq 6 \inf_M \left\{ \|g^* - g^{\mathbf{Q}^*, \mathbf{f}^*}\|_2^2 + \text{pen}(N, M) \right\} + A_1^* \frac{x}{N} \\ &\quad + 18C_{\mathcal{F},2}^6 \left(\|\mathbf{Q}^* - \mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top\|_F^2 + \|\pi^* - \mathbb{P}_{\tilde{\tau}_N} \hat{\pi}\|_2^2 \right). \end{aligned}$$

Notice that writing

$$\begin{aligned} \hat{g}(y_1, y_2, y_3) &= \sum_{k_1, k_2, k_3=1}^K (\mathbb{P}_{\tilde{\tau}_N} \hat{\pi}^*)(k_1) (\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top)(k_1, k_2) (\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top)(k_2, k_3) \\ &\quad \times \hat{f}_{\tilde{\tau}_N(k_1)}(y_1) \hat{f}_{\tilde{\tau}_N(k_2)}(y_2) \hat{f}_{\tilde{\tau}_N(k_3)}(y_3), \end{aligned}$$

and applying Theorem 3 we get, on the event $\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top \in \mathcal{V}$, that there exists $\tau \in T_{\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top}$ such that

$$\sum_{k=1}^K \|f_{\tau(k)}^* - \hat{f}_{\tilde{\tau}_N(k)}\|_2^2 \leq \frac{1}{c(\mathcal{K}, \mathcal{V}, \mathfrak{F}^*)^2} \|\hat{g} - g^{\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top, \mathbf{f}^*}\|_2^2.$$

Define $\tau_N = \tilde{\tau}_N \circ \tau^{-1}$. Since $\tau \in T_{\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top}$, we get that $\mathbb{P}_{\tilde{\tau}_N} \hat{\mathbf{Q}} \mathbb{P}_{\tilde{\tau}_N}^\top = \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top$ and $\mathbb{P}_{\tilde{\tau}_N} \hat{\pi} = \mathbb{P}_{\tau_N} \hat{\pi}$, and the previous equation rewrites

$$(4) \quad \sum_{k=1}^K \|f_k^* - \hat{f}_{\tau_N(k)}\|_2^2 \leq \frac{1}{c(\mathcal{K}, \mathcal{V}, \mathfrak{F}^*)^2} \|\hat{g} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*}\|_2^2.$$

Now by the triangular inequality

$$(5) \quad \|\hat{g} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*}\|_2 \leq \|\hat{g} - g^*\|_2 + \|g^{\mathbf{Q}^*, \mathbf{f}^*} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*}\|_2$$

We have

$$\begin{aligned} & \left(g^{\mathbf{Q}^*, \mathbf{f}^*} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*} \right) (y_1, y_2, y_3) = \\ & \sum_{k_1, k_2, k_3=1}^K \left(\pi^*(k_1) \mathbf{Q}^*(k_1, k_2) \mathbf{Q}^*(k_2, k_3) - (\mathbb{P}_{\tau_N} \hat{\pi}^*)(k_1) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_1, k_2) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_2, k_3) \right) \\ & \quad \times f_{k_1}^*(y_1) f_{k_2}^*(y_2) f_{k_3}^*(y_3). \end{aligned}$$

Thus by Cauchy-Schwarz

$$\begin{aligned} & \|g^{\mathbf{Q}^*, \mathbf{f}^*} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*}\|_2^2 \leq \\ & \int \sum_{k_1, k_2, k_3=1}^K \left(\pi^*(k_1) \mathbf{Q}^*(k_1, k_2) \mathbf{Q}^*(k_2, k_3) - (\mathbb{P}_{\tau_N} \hat{\pi}^*)(k_1) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_1, k_2) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_2, k_3) \right)^2 \\ & \quad \times \left(\sum_{k_1, k_2, k_3=1}^K f_{k_1}^*(y_1)^2 f_{k_2}^*(y_2)^2 f_{k_3}^*(y_3)^2 \right) d\mathcal{L}^D(y_1) d\mathcal{L}^D(y_2) d\mathcal{L}^D(y_3) \\ & \leq K^3 C_{\mathcal{F}, 2}^6 \sum_{k_1, k_2, k_3=1}^K \left(\pi^*(k_1) \mathbf{Q}^*(k_1, k_2) \mathbf{Q}^*(k_2, k_3) - (\mathbb{P}_{\tau_N} \hat{\pi}^*)(k_1) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_1, k_2) (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_2, k_3) \right)^2 \\ & \leq 3K^3 C_{\mathcal{F}, 2}^6 \sum_{k_1, k_2, k_3=1}^K \left[(\pi^*(k_1) - \mathbb{P}_{\tau_N} \hat{\pi}^*(k_1))^2 \mathbf{Q}^*(k_1, k_2)^2 \mathbf{Q}^*(k_2, k_3)^2 \right. \\ & \quad \left. + (\mathbb{P}_{\tau_N} \hat{\pi}^*)(k_1)^2 \left(\mathbf{Q}^*(k_1, k_2) - (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_1, k_2) \right)^2 \mathbf{Q}^*(k_2, k_3)^2 \right. \\ & \quad \left. + (\mathbb{P}_{\tau_N} \hat{\pi}^*)(k_1)^2 (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_1, k_2)^2 \left(\mathbf{Q}^*(k_2, k_3) - (\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top)(k_2, k_3) \right)^2 \right] \end{aligned}$$

so that

$$(6) \quad \|g^{\mathbf{Q}^*, \mathbf{f}^*} - g^{\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top, \mathbf{f}^*}\|_2 \leq \sqrt{3} K^{3/2} C_{\mathcal{F}, 2}^3 \left[\|\pi^* - \mathbb{P}_{\tau_N} \hat{\pi}^*\|_2 + 2 \|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\|_F \right].$$

In the same way,

$$\begin{aligned} & \left(g^* - g^{\mathbf{Q}^*, \mathbf{f}_M^*} \right) (y_1, y_2, y_3) = \\ & \sum_{k_1, k_2, k_3=1}^K \pi^*(k_1) \mathbf{Q}^*(k_1, k_2) \mathbf{Q}^*(k_2, k_3) \left(f_{k_1}^*(y_1) f_{k_2}^*(y_2) f_{k_3}^*(y_3) - f_{M, k_1}^*(y_1) f_{M, k_2}^*(y_2) f_{M, k_3}^*(y_3) \right) \end{aligned}$$

so that

$$\|g^* - g^{\mathbf{Q}^*, \mathbf{f}_M^*}\|_2 \leq \sqrt{3} K^{3/2} C_{\mathcal{F}, 2}^2 \max\{\|f_k^* - f_{M, k}^*\|_2, k = 1, \dots, K\}.$$

Thus collecting (3), (4), (5), (6) and with an appropriate choice of A^* we get Theorem 4. \square

To apply Theorem 4 one has to choose an estimator $\hat{\mathbf{Q}}$ with controlled behavior, to be able to evaluate the probability of the event $\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top \in \mathcal{V}$ and the rate of convergence of $\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top$ and $\mathbb{P}_{\tau_N} \hat{\pi}$. One possibility is to use the spectral estimator described in Section 5. To get the following result, we propose to use the spectral estimator with, for each N , the dimension M_N chosen such that $\eta_3(\Phi_{M_N}) = O((\log N)^{1/4})$ (see Section 5).

Corollary 1 — *With this choice of $\hat{\mathbf{Q}}$, under the assumptions of Theorem 4, there exists a sequence of permutations $\tau_N \in \mathfrak{S}_K$ such that as N tends to infinity,*

$$\mathbb{E} \left[\sum_{k=1}^K \|f_k^* - \hat{f}_{\tau_N(k)}\|_2^2 \right] = O \left(\inf_{M'} \left\{ \sum_{k=1}^K \|f_k^* - f_{M',k}^*\|_2^2 + \text{pen}(N, M') \right\} + \frac{\log N}{N} \right).$$

Thus, choosing $\text{pen}(N, M) = \rho M \log N / N$ for a large ρ leads to the minimax asymptotic rate of convergence up to a power of $\log N$. Indeed, standard results in approximation theory [DL93] show that one can upper bound the approximation error $\|f_k^* - f_{M,k}^*\|_2$ by $\mathcal{O}(M^{-\frac{s}{D}})$ where $s > 0$ denotes a regularity parameter. Then the trade-off is obtained for $M^{\frac{1}{D}} \sim (N / \log N)^{\frac{1}{2s+D}}$, which leads to the quasi-optimal rate $(N / \log N)^{-\frac{s}{2s+D}}$ for the non parametric estimation when the minimal smoothness of the emission densities is s . Notice that the algorithm automatically selects the best M leading to this rate.

To implement the estimator, it remains to choose a value for ρ in the penalty. The calibration of this parameter is a classical issue and could be the subject of a full paper. In practice one can use the slope heuristic [BMM12].

Proof. We shall apply Theorem 5 where, for each N , we define δ_N such that $(-\log \delta_N) / \delta_N^2 := (\log N)^{1/2}$. Notice first that δ_N goes to 0 and that M_N tends to infinity as N tends to infinity, so that for large enough N , $M_N \geq M_{\mathfrak{F}^*}$. By denoting τ_N the τ_{M_N} given by Theorem 5 we get that for all $x \geq x(\mathbf{Q}^*)$, for all $N \geq \mathbf{N}(\mathbf{Q}^*, \mathfrak{F}^*)x \log N$, with probability $1 - [4 + (e-1)^{-1}]e^{-x} - 2\delta_N$,

$$\|\pi^* - \mathbb{P}_{\tau_N} \hat{\pi}\|_2 \leq \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \sqrt{\frac{\log N}{N}} \sqrt{x}$$

and

$$\|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\| \leq \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \sqrt{\frac{\log N}{N}} \sqrt{x}.$$

We first obtain that

$$\begin{aligned} & \limsup_{N \rightarrow +\infty} \mathbb{E} \left[\frac{N}{\log N} \|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\|^2 \right] \leq \\ & \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*)^2 \int_0^{+\infty} \limsup_{N \rightarrow +\infty} \mathbb{P} \left(\frac{\sqrt{N}}{\mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \sqrt{\log N}} \|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\| \geq \sqrt{x} \right) dx \leq \\ & \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*)^2 x(\mathbf{Q}^*) + \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*)^2 \int_{x(\mathbf{Q}^*)}^{+\infty} [4 + (e-1)^{-1}] e^{-x} dx < +\infty \end{aligned}$$

so that

$$\mathbb{E} \left[\|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\|^2 \right] = O \left(\frac{\log N}{N} \right).$$

Similarly,

$$\mathbb{E} \left[\|\pi^* - \mathbb{P}_{\tau_N} \hat{\pi}\|^2 \right] = O \left(\frac{\log N}{N} \right).$$

We also obtain, by taking $x = N / (\log N)^{1/4}$, that

$$\limsup_{N \rightarrow +\infty} \mathbb{P} \left(\mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top \notin \mathcal{V} \right) = 0,$$

so that, using Theorem 4, we get

$$\begin{aligned} & \limsup_{N \rightarrow +\infty} \mathbb{P} \left(\frac{N}{A^*} \left[\sum_{k=1}^K \|f_k^* - \hat{f}_{\tau_N(k)}\|_2^2 - \inf_M \left\{ \sum_{k=1}^K \|f_k^* - f_{M,k}^*\|_2^2 + \text{pen}(N, M) \right\} \right. \right. \\ & \quad \left. \left. - \|\mathbf{Q}^* - \mathbb{P}_{\tau_N} \hat{\mathbf{Q}} \mathbb{P}_{\tau_N}^\top\|_F^2 - \|\pi^* - \mathbb{P}_{\tau_N} \hat{\pi}\|_2^2 + \frac{x}{N} \right] \geq x \right) \leq (e-1)^{-1} e^{-x}. \end{aligned}$$

Thus, by integration and the previous results, Corollary 1 follows. \square

5. NONPARAMETRIC SPECTRAL METHOD

This section is devoted to a short description of the nonparametric spectral method for sake of completeness: we describe the algorithm, and give the results we need to support the use of spectral estimators to initialize our algorithm. A detailed study of the nonparametric spectral method is given in [DCGLC15].

The following procedure (see Algorithm 1) describes a tractable approach to estimate the transition matrix in a way that can be used for the penalized least squares estimator of the emission densities, and also for the estimation of the projections of the emission densities that may be used to initialize the least squares algorithm. The procedure is based on recent developments in parametric estimation of HMMs. For each fixed M , we estimate the projection of the emission distributions on the basis Φ_M using the spectral method proposed in [AHK12]. As the authors of the latter paper explain, this allows further to estimate the transition matrix (we use a modified version of their estimator), and we set the estimator of the stationary distribution as the stationary distribution of the estimator of the transition matrix. The computation of those estimators is particularly simple: it is based on one SVD, some matrix inversions and one diagonalization. One can prove, with overwhelming probability, all matrix inversions and the diagonalization can be done rightfully, see [DCGLC15]. In the following, when A is a $(p \times q)$ matrix with $p \geq q$, A^\top denotes the transpose matrix of A , $A(k, l)$ its (k, l) th entry, $A(\cdot, l)$ its l th column and $A(k, \cdot)$ its k th line. When v is a vector of size p , we denote by $\mathfrak{Diag}[v]$ the diagonal matrix with diagonal entries v_i and, by abuse of notation, $\mathfrak{Diag}[v] = \mathfrak{Diag}[v^\top]$.

We now state a result which allows to derive the asymptotic properties of the spectral estimators. Let us define

$$\eta_3^2(\Phi_M) := \sup_{y, y' \in \mathcal{Y}^3} \sum_{a, b, c=1}^M (\varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3) - \varphi_a(y'_1)\varphi_b(y'_2)\varphi_c(y'_3))^2.$$

Note that in the examples (**Spline**), (**Trig.**) and (**Wav.**) we have:

$$\eta_3(\Phi_M) \leq C_\eta M^{\frac{3}{2}}$$

where $C_\eta > 0$ is a constant. The following Theorem is proved in [DCGLC15].

Theorem 5 (Spectral estimators) — *Assume [H1]-[H4]. Then, there exist positive constant numbers $M_{\mathfrak{F}^*}$, $x(\mathbf{Q}^*)$, $\mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*)$ and $\mathbf{N}(\mathbf{Q}^*, \mathfrak{F}^*)$ such that the following holds. For any $x \geq x(\mathbf{Q}^*)$, for any $\delta \in (0, 1)$, for any $M \geq M_{\mathfrak{F}^*}$, there exists a permutation $\tau_M \in \mathcal{S}_K$ such that the spectral method estimators $\hat{f}_{M,k}$, $\hat{\pi}$ and $\hat{\mathbf{Q}}$ enjoy: for any $N \geq \mathbf{N}(\mathbf{Q}^*, \mathfrak{F}^*)\eta_3(\Phi_M)^2 x(-\log \delta)/\delta^2$, with probability greater than $1 - 2\delta - 4e^{-x}$,*

$$\begin{aligned} \|f_{M,k}^* - \hat{f}_{M,\tau_M(k)}\|_2 &\leq \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \frac{\sqrt{-\log \delta}}{\delta} \frac{\eta_3(\Phi_M)}{\sqrt{N}} \sqrt{x}, \\ \|\pi^* - \mathbb{P}_{\tau_M} \hat{\pi}\|_2 &\leq \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \frac{\sqrt{-\log \delta}}{\delta} \frac{\eta_3(\Phi_M)}{\sqrt{N}} \sqrt{x}, \\ \|\mathbf{Q}^* - \mathbb{P}_{\tau_M} \hat{\mathbf{Q}} \mathbb{P}_{\tau_M}^\top\| &\leq \mathcal{C}(\mathbf{Q}^*, \mathfrak{F}^*) \frac{\sqrt{-\log \delta}}{\delta} \frac{\eta_3(\Phi_M)}{\sqrt{N}} \sqrt{x}. \end{aligned}$$

6. NUMERICAL EXPERIMENTS

In this section we show numerical experiments on the performances of the adaptive estimation method studied in this paper. We recall that the experimenter

Algorithm 1: Nonparametric spectral estimation of the transition matrix and the emission laws

Data: An observed chain (Y_1, \dots, Y_N) and a number of hidden states K .

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

[Step 1] Consider the following empirical estimators: For any a, b, c in $\{1, \dots, M\}$,

$$\begin{aligned}\hat{\mathbf{L}}_M(a) &:= \frac{1}{N} \sum_{s=1}^N \varphi_a(Y_1^{(s)}), \\ \hat{\mathbf{M}}_M(a, b, c) &:= \frac{1}{N} \sum_{s=1}^N \varphi_a(Y_1^{(s)}) \varphi_b(Y_2^{(s)}) \varphi_c(Y_3^{(s)}), \\ \hat{\mathbf{N}}_M(a, b) &:= \frac{1}{N} \sum_{s=1}^N \varphi_a(Y_1^{(s)}) \varphi_b(Y_2^{(s)}), \\ \text{and } \hat{\mathbf{P}}_M(a, c) &:= \frac{1}{N} \sum_{s=1}^N \varphi_a(Y_1^{(s)}) \varphi_c(Y_3^{(s)}).\end{aligned}$$

[Step 2] Let $\hat{\mathbf{U}}$ be the $M \times K$ matrix of orthonormal right singular vectors of $\hat{\mathbf{P}}_M$ corresponding to its top K singular values.

[Step 3] Form the matrices:

$$\forall b \in \{1, \dots, M\}, \quad \hat{\mathbf{B}}(b) := (\hat{\mathbf{U}}^\top \hat{\mathbf{P}}_M \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{M}}_M(\cdot, b, \cdot) \hat{\mathbf{U}}.$$

[Step 4] Set Θ a $(K \times K)$ random unitary matrix uniformly drawn and form the matrices:

$$\forall k \in \{1, \dots, K\}, \quad \hat{\mathbf{C}}(k) := \sum_{b=1}^M (\hat{\mathbf{U}} \Theta)(b, k) \hat{\mathbf{B}}(b).$$

[Step 5] Compute $\hat{\mathbf{R}}$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{\mathbf{C}}(1)$:

$$\hat{\mathbf{R}}^{-1} \hat{\mathbf{C}}(1) \hat{\mathbf{R}} = \mathfrak{D}\text{diag}[(\hat{\Lambda}(1, 1), \dots, \hat{\Lambda}(1, K))].$$

[Step 6] Set:

$$\forall k, k' \in \mathcal{X}, \quad \hat{\Lambda}(k, k') := (\hat{\mathbf{R}}^{-1} \hat{\mathbf{C}}(k) \hat{\mathbf{R}})(k', k'),$$

$$\text{and } \hat{\mathbf{O}}_M := \hat{\mathbf{U}} \Theta \hat{\Lambda}.$$

[Step 7] Consider the emission laws estimator $\hat{\mathbf{f}} := (\hat{f}_{M,k})_{k \in \mathcal{X}}$ defined by:

$$\forall k \in \mathcal{X}, \quad \hat{f}_{M,k} := \sum_{m=1}^M \hat{\mathbf{O}}_M(m, k) \varphi_m,$$

[Step 8] Set

$$\hat{\pi} := (\hat{\mathbf{U}}^\top \hat{\mathbf{O}}_M)^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{L}}_M.$$

[Step 9] Consider the transition matrix estimator:

$$\hat{\mathbf{Q}} := \Pi_{\text{TM}} \left((\hat{\mathbf{U}}^\top \hat{\mathbf{O}}_M \mathfrak{D}\text{diag}[\hat{\pi}])^{-1} \hat{\mathbf{U}}^\top \hat{\mathbf{N}}_M \hat{\mathbf{U}} (\hat{\mathbf{O}}_M^\top \hat{\mathbf{U}})^{-1} \right),$$

where Π_{TM} denotes the projection (with respect to the scalar product given by the Frobenius norm) onto the convex set of transition matrices, and define $\hat{\pi}$ as the stationary distribution of $\hat{\mathbf{Q}}$.

knows nothing about the underlying hidden Markov model but the number of hidden states K . In this set of experiments, we consider the regular histogram basis

for estimating $K = 2$ emission laws given by beta laws of parameters $(2, 5)$ and $(4, 3)$ from a single chain of size $N = 30,000$.

Our method is based on the computation of least squares estimators \hat{g}_M defined as minimizers of the empirical contrast γ_N . Since the function γ_N is non-convex, we use an iterative algorithm to search for the minimum of γ_N (see below) with the starting point given by the spectral estimator.

Then, we compute the adaptive choice of the size of the model, namely

$$\hat{M} = \arg \min_{M=1, \dots, N} \{\gamma_N(\hat{g}_M) + \text{pen}(N, M)\} .$$

Interestingly, we are able to apply the slope heuristic procedure in the nonparametric HMM's frame. This compelling data-driven procedure allows us to tune the penalty appearing in our estimator. More precisely, typical behaviors of the function $M \mapsto \gamma(\hat{g}_M)$ and the function $\rho \mapsto \arg \min\{\gamma(\hat{g}_M) + \rho \text{pen}(N, M)\}$ usher the experimenter to the right tuning parameter, an interested reader may consult [BMM12].

Hence we compute \hat{g}_M for each $M = 1, \dots, N$ as follows

- First compute the spectral estimator. This is straightforward using the procedure described by [Step1-9], Section 5. In particular, the spectral estimator gives an estimation $\hat{\mathbf{Q}}, \hat{\pi}$ of the transition matrix and its stationary distribution which is used to compute the least squares contrast function.
- Use the spectral estimator of the emission densities as a starting point for ‘‘Covariance Matrix Adaptation Evolution Strategy’’ (CMA-ES), see [Han06]. This iterative algorithm may ultimately find a local/global minimum of the contrast function.

Note that the size M of the projection space for the spectral estimator has been set as the one chosen by the slope heuristic for the least squares estimators.

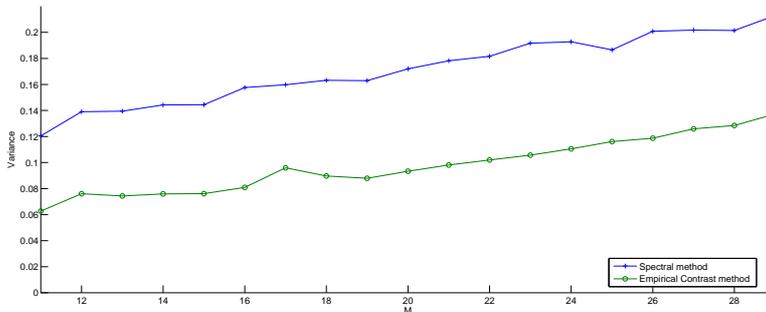


FIGURE 1. Comparison of the variances of the spectral and the least squares estimators.

A first numerical experiment, depicted in Figure 1, compares, for each M , the variances (i.e. the ℓ_2 -distance between the estimator and the orthogonal projection onto the subspace generated by the basis Φ_M) obtained by the spectral method and the empirical least squares method over 100 iterations on chains of length 40,000. It consolidates the idea that the least square method significantly improves the ℓ_2 -distance to the best approximation of the emission laws. Indeed, even for small values of M , one may see that the variance is divided by two in Figure 1.

One can see on Figure 2 that our method also qualitatively improve upon the spectral method.

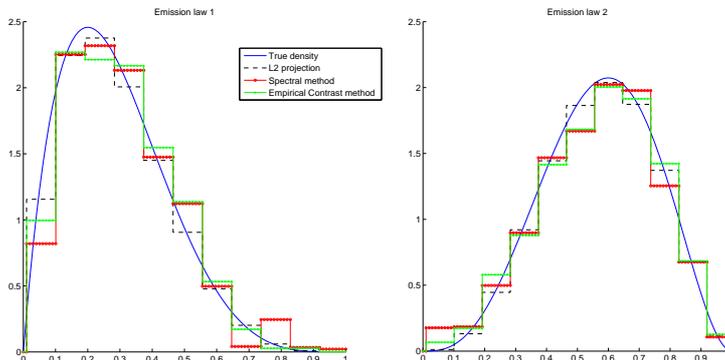


FIGURE 2. Estimators of the emissions densities.

DISCUSSION

We have proposed a penalized least squares method to estimate the emission densities of the hidden chain when the transition matrix of the hidden chain is full rank and the emission probability distributions are linearly independent. The algorithm may be initialized using spectral estimators. The obtained estimators are adaptive rate optimal up to a log factor, where adaptivity is upon the family of emission densities. The results hold under a generic assumption of the parameter. We have proved that this generic assumption is always verified when there are two hidden states. A natural question is to ask if, when the number of hidden states is $K > 2$, this assumption is also always verified. The proof for $K = 2$ appeared to be surprisingly hard.

Another question arising from this work is whether it is possible to adapt to different smoothnesses of the emission densities.

7. PROOFS

7.1. Proof of lemma 1. In [HKZ12] it is proved that when [H1], [H2], [H3] hold and when the rank of the matrix $\mathbf{O}_M := (\langle \varphi_m, f_k^* \rangle)_{1 \leq m \leq M, 1 \leq k \leq K}$ is K , the knowledge of the tensor \mathbf{M}_M given by $\mathbf{M}_M(a, b, c) = \mathbb{E}(\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3))$ for all a, b, c in $\{1, \dots, M\}$ allows to recover \mathbf{O}_M and \mathbf{Q} up to relabelling of the hidden states. Thus, when [H1], [H2], [H3] and [H4] hold, the knowledge of $g^{\mathbf{Q}, \mathbf{f}^*}$ is equivalent to the knowledge of the sequence $(\mathbf{M}_M)_M$, which allows to recover \mathbf{Q} and the sequence $(\mathbf{O}_M)_M$, up to relabelling of the hidden states, which allows to recover $\mathbf{f}^* = (f_1^*, \dots, f_K^*)$ up to relabelling of the hidden states, thanks to (1). See also [GCR15].

7.2. Proof of Theorem 2. Throughout the proof N is fixed, and we write γ (instead of γ_N) for the contrast function.

7.2.1. Beginning of the proof: algebraic manipulations.

Let us fix some M and some permutation τ . Using the definitions of \hat{g}_M and \hat{M} , we can write

$$\gamma(\hat{g}_{\hat{M}}) + \text{pen}(N, \hat{M}) \leq \gamma(\hat{g}_M) + \text{pen}(N, M) \leq \gamma(g^{\hat{\mathbf{Q}}, \mathbf{f}^*_{M, \tau^{-1}}}) + \text{pen}(N, M),$$

where $\mathbf{f}^*_{M, \tau^{-1}} = (f^*_{M, \tau^{-1}(1)}, \dots, f^*_{M, \tau^{-1}(K)})$ (here we use that $\mathbf{f}^*_{M, \tau^{-1}} \in \mathcal{F}^K$). But we can compute for all functions t_1, t_2 ,

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

where ν is the centered empirical process

$$\nu(t) = \frac{1}{N} \sum_{s=1}^N t(Y_1^{(s)}, Y_2^{(s)}, Y_3^{(s)}) - \int t g^*.$$

This gives

$$(7) \quad \|\hat{g}_{\hat{M}} - g^*\|_2^2 \leq \|g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} - g^*\|_2^2 + 2\nu(\hat{g}_{\hat{M}} - g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*}) + \text{pen}(N, M) - \text{pen}(N, \hat{M})$$

Now, we denote $R_{\hat{M}} = \|\hat{g}_{\hat{M}} - g^*\|_2^2$ the squared risk and $B_M = \|g^{\mathbf{Q}^*, \mathbf{f}_M^*} - g^*\|_2^2$ a bias term. We also set $S_M = \cup_{\mathbf{Q}} \mathcal{S}(\mathbf{Q}, M)$ and

$$Z_M = \sup_{t \in S_M} \left[\frac{|\nu(t - g^*)|}{\|t - g^*\|_2^2 + x_M^2} \right]$$

for x_M to be determined later. Notice that $g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} = g^{\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top, \mathbf{f}_M^*}$. Then

$$\begin{aligned} \|g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} - g^*\|_2^2 &\leq 2\|g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} - g^{\mathbf{Q}^*, \mathbf{f}_M^*}\|_2^2 + 2\|g^{\mathbf{Q}^*, \mathbf{f}_M^*} - g^*\|_2^2 \\ &\leq 2\|g^{\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top, \mathbf{f}_M^*} - g^{\mathbf{Q}^*, \mathbf{f}_M^*}\|_2^2 + 2B_M. \end{aligned}$$

But, using Schwarz inequality, $\|g^{\mathbf{Q}_1, \mathbf{f}_M^*} - g^{\mathbf{Q}_2, \mathbf{f}_M^*}\|_2^2$ can be bounded by

$$\begin{aligned} &\sum_{m_1, m_2, m_3=1}^M \left| \sum_{k_1, k_2, k_3=1}^K (\pi_1(k_1) \mathbf{Q}_1(k_1, k_2) \mathbf{Q}_1(k_2, k_3) - \pi_2(k_1) \mathbf{Q}_2(k_1, k_2) \mathbf{Q}_2(k_2, k_3)) \right. \\ &\quad \left. \langle f_{k_1}^*, \varphi_{m_1} \rangle \langle f_{k_2}^*, \varphi_{m_2} \rangle \langle f_{k_3}^*, \varphi_{m_3} \rangle \right|^2 \\ &\leq \left(\sum_{k_1, k_2, k_3=1}^K (\pi_1(k_1) \mathbf{Q}_1(k_1, k_2) \mathbf{Q}_1(k_2, k_3) - \pi_2(k_1) \mathbf{Q}_2(k_1, k_2) \mathbf{Q}_2(k_2, k_3))^2 \right) \\ &\quad \sum_{m_1, m_2, m_3=1}^M \left| \sum_{k_1, k_2, k_3=1}^K \langle f_{k_1}^*, \varphi_{m_1} \rangle \langle f_{k_2}^*, \varphi_{m_2} \rangle \langle f_{k_3}^*, \varphi_{m_3} \rangle \right|^2 \\ &\leq 3K^3 C_{\mathcal{F}, 2}^6 (\|\pi_1 - \pi_2\|_2^2 + 2\|\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2) \end{aligned}$$

so that

$$\|g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} - g^*\|_2^2 \leq 6K^3 C_{\mathcal{F}, 2}^6 \left(\|\mathbb{P}_\tau \hat{\pi} - \pi^*\|_2^2 + 2\|\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top - \mathbf{Q}^*\|_F^2 \right) + 2B_M.$$

Next

$$\begin{aligned} \nu(\hat{g}_{\hat{M}} - g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*}) &= \nu(\hat{g}_{\hat{M}} - g^*) + \nu(g^* - g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*}) \\ &\leq Z_{\hat{M}} (\|\hat{g}_{\hat{M}} - g^*\|_2^2 + x_{\hat{M}}^2) + Z_M (\|g^{\hat{\mathbf{Q}}, \mathbf{f}_{M, \tau}^*} - g^*\|_2^2 + x_M^2) \end{aligned}$$

so that (7) becomes

$$\begin{aligned} R_{\hat{M}} &\leq 6K^3 C_{\mathcal{F}, 2}^6 \left(\|\mathbb{P}_\tau \hat{\pi} - \pi^*\|_2^2 + 2\|\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top - \mathbf{Q}^*\|_F^2 \right) + 2B_M + 2Z_{\hat{M}} (R_{\hat{M}} + x_{\hat{M}}^2) \\ &\quad + 2Z_M (6K^3 C_{\mathcal{F}, 2}^6 \left(\|\mathbb{P}_\tau \hat{\pi} - \pi^*\|_2^2 + 2\|\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top - \mathbf{Q}^*\|_F^2 \right) + 2B_M + x_M^2) \\ &\quad + 2\text{pen}(N, M) - \text{pen}(N, \hat{M}) - \text{pen}(N, M), \\ R_{\hat{M}} (1 - 2Z_{\hat{M}}) &\leq (2 + 4Z_M) B_M + 2\text{pen}(N, M) \\ &\quad + (1 + 2Z_M) 6K^3 C_{\mathcal{F}, 2}^6 \left(\|\mathbb{P}_\tau \hat{\pi} - \pi^*\|_2^2 + 2\|\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top - \mathbf{Q}^*\|_F^2 \right) \\ &\quad + 2 \sup_{M'} (2Z_{M'} x_{M'}^2 - \text{pen}(N, M')). \end{aligned}$$

To conclude it is then sufficient to establish that, with probability larger than $1 - (e - 1)^{-1}e^{-x}$, it holds

$$\sup_{M'} Z_{M'} \leq \frac{1}{4} \quad \text{and} \quad \sup_{M'} (2Z_{M'} x_{M'}^2 - \text{pen}(N, M')) \leq A \frac{x}{N},$$

with A a constant depending only on \mathbf{Q}^* and \mathbf{f}^* and not on N, M, x . Thus we will have, for any M , with probability larger than $1 - (e - 1)^{-1}e^{-x}$,

$$\begin{aligned} \frac{1}{2} R_{\hat{M}} &\leq 3B_M + 2\text{pen}(N, M) + 2A \frac{x}{N} \\ &\quad + 9C_{\mathcal{F},2}^6 \left(\|\mathbb{P}_\tau \hat{\pi} - \pi^*\|_2^2 + 2\|\mathbb{P}_\tau \hat{\mathbf{Q}} \mathbb{P}_\tau^\top - \mathbf{Q}^*\|_F^2 \right) \end{aligned}$$

which is the announced result.

The heart of the proof is then the study of Z_M . We introduce u_M a projection of g^* on S_M and we split Z_M in two terms: $Z_M \leq 4Z_{M,1} + Z_{M,2}$ with

$$\begin{cases} Z_{M,1} = \sup_{t \in S_M} \left[\frac{|\nu(t - u_M)|}{\|t - u_M\|_2^2 + 4x_M^2} \right] \\ Z_{M,2} = \frac{|\nu(u_M - g^*)|}{\|u_M - g^*\|_2^2 + x_M^2} \end{cases}$$

Indeed u_M verifies: for all $t \in S_M$,

$$\|u_M - g^*\|_2 \leq \|t - g^*\|_2 \quad \text{and} \quad \|u_M - t\|_2 \leq 2\|t - g^*\|_2.$$

7.2.2. Deviation inequality for $Z_{M,2}$.

Bernstein's inequality (16) for HMMs (see Appendix A) gives, with probability larger than $1 - e^{-z}$:

$$|\nu(u_M - g^*)| \leq 2\sqrt{2c^* \|u_M - g^*\|_2^2 \|g^*\|_\infty} \frac{z}{N} + 2\sqrt{2}c^* \|u_M - g^*\|_\infty \frac{z}{N}.$$

Then, using $a^2 + b^2 \geq 2ab$, with probability larger than $1 - e^{-z}$:

$$\frac{|\nu(u_M - g^*)|}{\|u_M - g^*\|_2^2 + x_M^2} \leq 2\sqrt{2c^* \|g^*\|_\infty} \frac{1}{2x_M} \sqrt{\frac{z}{N}} + 2\sqrt{2}c^* \frac{\|u_M\|_\infty + \|g^*\|_\infty}{x_M^2} \frac{z}{N}.$$

But any function t in S_M can be written

$$t = \sum_{k_1, k_2, k_3=1}^K \pi(k_1) \mathbf{Q}(k_1, k_2) \mathbf{Q}(k_2, k_3) f_{k_1} \otimes f_{k_2} \otimes f_{k_3},$$

with $f_k \in \mathcal{F}$ for $k = 1, \dots, K$, so that $\sup_{t \in S_M} \|t\|_\infty \leq C_{\mathcal{F},\infty}^3$. Then, with probability larger than $1 - e^{-z_M - z}$

$$Z_{M,2} \leq \sqrt{2c^* \|g^*\|_\infty} \sqrt{\frac{z_M + z}{x_M^2 N}} + 4\sqrt{2}c^* C_{\mathcal{F},\infty}^3 \frac{z_M + z}{x_M^2 N}.$$

7.2.3. Deviation inequality for $Z_{M,1}$.

We shall first study the term $\sup_{t \in B_\sigma} |\nu(t - u_M)|$ where

$$B_\sigma = \{t \in S_M, \|t - u_M\|_2 \leq \sigma\}.$$

Remark that, for all $t \in \mathcal{S}(\mathbf{Q}, M)$,

$$\|t\|_2^2 \leq \sum_{k_1, k_2, k_3=1}^K \pi^2(k_1) \mathbf{Q}^2(k_1, k_2) \mathbf{Q}^2(k_2, k_3) \sum_{k_1, k_2, k_3=1}^K C_{\mathcal{F},2}^2 C_{\mathcal{F},2}^2 C_{\mathcal{F},2}^2 \leq K^3 C_{\mathcal{F},2}^6$$

Then, if $t \in B_\sigma$, $\|t - u_M\|_2 \leq \sigma \wedge 2K^{3/2}C_{\mathcal{F},2}^3$. Notice also that for all $t \in S_M$, $\|t - u_M\|_\infty \leq 2C_{\mathcal{F},\infty}^3$. Now Proposition 7 in Appendix A (applied to a countable dense set in B_σ) gives that for any measurable set A such that $\mathbb{P}(A) > 0$,

$$\mathbb{E}^A\left(\sup_{t \in B_\sigma} |\nu(t - u_M)|\right) \leq C^* \left[\frac{E}{N} + \sigma \sqrt{\frac{1}{N} \log\left(\frac{1}{\mathbb{P}(A)}\right)} + \frac{2C_{\mathcal{F},\infty}^3}{N} \log\left(\frac{1}{\mathbb{P}(A)}\right) \right],$$

and

$$E = \sqrt{N} \int_0^\sigma \sqrt{H(u) \wedge N} du + (2C_{\mathcal{F},\infty}^3 + 2K^{3/2}C_{\mathcal{F},2}^3)H(\sigma).$$

Here, for any integrable random variable Z , $E^A[Z]$ denotes $E[Z\mathbf{1}_A]/\mathbb{P}(A)$.

We shall compute E later and find σ_M and φ such that

$$(8) \quad \forall \sigma \geq \sigma_M \quad E \leq (1 + 2C_{\mathcal{F},\infty}^3 + 2K^{3/2}C_{\mathcal{F},2}^3)\varphi(\sigma)\sqrt{N}.$$

(see Section 7.2.4). We then use Lemma 4.23 in [Mas07] to write (for $x_M \geq \sigma_M$)

$$\mathbb{E}^A\left(\sup_{t \in S_M} \left[\frac{|\nu(t - u_M)|}{\|t - u_M\|_2^2 + 4x_M^2} \right] \right) \leq \frac{C^*}{x_M^2} \left[C \frac{\varphi(2x_M)}{\sqrt{N}} + 2x_M \sqrt{\frac{1}{N} \log\left(\frac{1}{\mathbb{P}(A)}\right)} + \frac{2C_{\mathcal{F},\infty}^3}{N} \log\left(\frac{1}{\mathbb{P}(A)}\right) \right]$$

Finally, Lemma 2.4 in [Mas07] ensures that, with probability $1 - e^{-z_M - z}$:

$$(9) \quad Z_{M,1} = \sup_{t \in S_M} \left[\frac{|\nu(t - u_M)|}{\|t - u_M\|_2^2 + 4x_M^2} \right] \leq C^* \left[C \frac{\varphi(2x_M)}{x_M^2 \sqrt{N}} + 2 \sqrt{\frac{z_M + z}{x_M^2 N}} + 2C_{\mathcal{F},\infty}^3 \frac{z_M + z}{x_M^2 N} \right].$$

7.2.4. Computation of the entropy and function φ .

The definition of H given in Proposition 7 shows that $H(\delta)$ is bounded by the classical bracketing entropy for \mathbf{L}^2 distance at point $\delta/C_{\mathcal{F},\infty}^3$ (where $C_{\mathcal{F},\infty}^3$ bounds the sup norm of g^*): $H(\delta) \leq H(\delta/C_{\mathcal{F},\infty}^3, S_M, \mathbf{L}^2)$. We denote by $N(u, S, \mathbf{L}^2) = e^{H(u, S, \mathbf{L}^2)}$ the minimal number of brackets of radius u to cover S . Recall that when t_1 and t_2 are real valued functions, the bracket $[t_1, t_2]$ is the set of real valued functions t such that $t_1(\cdot) \leq t(\cdot) \leq t_2(\cdot)$, and the radius of the bracket is $\|t_2 - t_1\|_2$. Now, observe that $S_M = \cup_{\mathbf{Q}} \mathcal{S}(\mathbf{Q}, M)$ is a set of mixtures of parametric functions. Denoting $\mathbf{k} = (k_1, k_2, k_3)$, S_M is included in

$$\left\{ \begin{array}{l} \sum_{\mathbf{k} \in \{1, \dots, K\}^3} \mu(\mathbf{k}) f_{k_1} \otimes f_{k_2} \otimes f_{k_3}, \mu \geq 0, \quad \sum_{\mathbf{k} \in \{1, \dots, K\}^3} \mu(\mathbf{k}) = 1, \\ f_{k_i} \in \mathcal{F} \cap \text{Span}(\varphi_1, \dots, \varphi_M), i = 1, 2, 3 \end{array} \right\}.$$

Set

$$\mathcal{A} = \{f_1 \otimes f_2 \otimes f_3, f_i \in \mathcal{F} \cap \text{Span}(\varphi_1, \dots, \varphi_M), i = 1, 2, 3\}.$$

Then following the proof in Appendix A of [BT13], we can prove

$$(10) \quad N(\varepsilon, S_M, \mathbf{L}^2) \leq \left(\frac{C_1}{\varepsilon}\right)^{K^3-1} \left[N\left(\frac{\varepsilon}{3}, \mathcal{A}, \mathbf{L}^2\right) \right]^{K^3}.$$

where C_1 depends on K and $C_{\mathcal{F},2}$. Denote $\mathcal{B} = \mathcal{F} \cap \text{Span}(\varphi_1, \dots, \varphi_M)$. Let $a = (a_m)_{1 \leq m \leq M} \in \mathbb{R}^M$ and $b = (b_m)_{1 \leq m \leq M} \in \mathbb{R}^M$ such that $a_m < b_m$, $m = 1, \dots, M$. For each $m = 1, \dots, M$ and $y \in \mathcal{Y}$, let

$$u_m(y) = \begin{cases} a_m & \text{if } \varphi_m(y) \geq 0 \\ b_m & \text{otherwise} \end{cases}$$

$$v_m(y) = a_m + b_m - u_m(y).$$

Then, if $(c_m)_{1 \leq m \leq M} \in \mathbb{R}^M$ is such that for all $m = 1, \dots, M$, $a_m \leq c_m \leq b_m$, then

$$U_{a,b}^1(y) := \sum_{m=1}^M u_m(y) \varphi_m(y) \leq \sum_{m=1}^M c_m \varphi_m(y) \leq \sum_{m=1}^M v_m(y) \varphi_m(y) = U_{a,b}^2(y).$$

Moreover,

$$\begin{aligned} \|U_{a,b}^2 - U_{a,b}^1\|_2^2 &= \left\| \sum_{m=1}^M |b_m - a_m| \cdot |\varphi_m| \right\|_2^2 \\ &\leq M \|b - a\|_2^2 \end{aligned}$$

using Cauchy-Schwarz inequality. Thus, one may cover \mathcal{B} with brackets of form $[U_{a,b}^1, U_{a,b}^2]$. Also, for $i = 1, 2$,

$$\begin{aligned} \|U_{a,b}^i\|_2^2 &\leq \left\| \sum_{m=1}^M |b_m + a_m| \cdot |\varphi_m| \right\|_2^2 \\ &\leq 2M (\|a\|_2^2 + \|b\|_2^2). \end{aligned}$$

If now for some a^i, b^i in \mathbb{R}^M , $f_i \in [U_{a^i, b^i}^1, U_{a^i, b^i}^2]$, $i = 1, 2, 3$, then

$$f_1 \otimes f_2 \otimes f_3 \in [V, W]$$

with

$$V = \min\{U_{a^1, b^1}^{i_1} U_{a^2, b^2}^{i_2} U_{a^3, b^3}^{i_3}, i_1, i_2, i_3 \in \{1, 2\}\}$$

and

$$W = \max\{U_{a^1, b^1}^{i_1} U_{a^2, b^2}^{i_2} U_{a^3, b^3}^{i_3}, i_1, i_2, i_3 \in \{1, 2\}\},$$

pointwise. Moreover, one can see that

$$\begin{aligned} |W - V| &\leq \left| U_{a^1, b^1}^2 - U_{a^1, b^1}^1 \right| \max_{j_1, j_2 \in \{1, 2\}} \left| U_{a^2, b^2}^{j_1} \right| \cdot \left| U_{a^3, b^3}^{j_2} \right| \\ &\quad + \left| U_{a^2, b^2}^2 - U_{a^2, b^2}^1 \right| \max_{j_1, j_2 \in \{1, 2\}} \left| U_{a^1, b^1}^{j_1} \right| \cdot \left| U_{a^3, b^3}^{j_2} \right| \\ &\quad + \left| U_{a^3, b^3}^2 - U_{a^3, b^3}^1 \right| \max_{j_1, j_2 \in \{1, 2\}} \left| U_{a^1, b^1}^{j_1} \right| \cdot \left| U_{a^2, b^2}^{j_2} \right| \\ &\leq \sum_{i=1}^3 \left| U_{a^i, b^i}^2 - U_{a^i, b^i}^1 \right| \prod_{j \neq i} \left(\left| U_{a^j, b^j}^1 \right| + \left| U_{a^j, b^j}^2 \right| \right) \end{aligned}$$

so that

$$\begin{aligned} \|W - V\|_2^2 &\leq 12 \sum_{i=1}^3 \left\| U_{a^i, b^i}^2 - U_{a^i, b^i}^1 \right\|_2^2 \prod_{j \neq i} \left(\left\| U_{a^j, b^j}^1 \right\|_2^2 + \left\| U_{a^j, b^j}^2 \right\|_2^2 \right) \\ &\leq 48M^3 \sum_{i=1}^3 \|b^i - a^i\|_2^2 \prod_{j \neq i} (\|a^j\|_2^2 + \|b^j\|_2^2) \\ &\leq 192M^3 C_{\mathcal{F}, 2}^4 \sum_{i=1}^3 \|b^i - a^i\|_2^2. \end{aligned}$$

Thus one may cover \mathcal{A} by covering the ball of radius $C_{\mathcal{F}, 2}$ in \mathbb{R}^M with hypercubes $[a, b]$, for which $\|a\|_2, \|b\|_2$ are less than $C_{\mathcal{F}, 2}$. To get a bracket with radius u , it is enough that $\|b^i - a^i\|_2^2 \leq u^2 / (576M^3 C_{\mathcal{F}, 2}^4)$, $i = 1, 2, 3$. We finally obtain that

$$(11) \quad N(u, \mathcal{A}, \mathbf{L}^2) \leq \left(\frac{48\sqrt{3}M^{3/2}C_{\mathcal{F}, 2}^3}{u} \right)^{3M}.$$

We deduce from (10) and (11) that

$$N(u, S_M, \mathbf{L}^2) \leq \left(\frac{C_1}{u}\right)^{K^3-1} \prod_{\mathbf{k}} \left(\frac{48\sqrt{3}M^{3/2}C_{\mathcal{F},2}^3}{u}\right)^{3M},$$

and then

$$H(u, S_M, \mathbf{L}^2) \leq (K^3 - 1) \log\left(\frac{C_1}{u}\right) + 3MK^3 \log\left(\frac{C_2 M^{3/2}}{u}\right),$$

with C_2 depending on K and $C_{\mathcal{F},2}$. To conclude we use that $\int_0^\sigma \sqrt{\log\left(\frac{1}{x}\right)} dx \leq \sigma(\sqrt{\pi} + \sqrt{\log\left(\frac{1}{\sigma}\right)})$, see [BMM12]. Finally we can write for $\sigma \leq M^{3/2}$:

$$\int_0^\sigma \sqrt{H(u)} du \leq C_3 \sqrt{M} \sigma \left(1 + \sqrt{\log\left(\frac{M^{3/2}}{\sigma}\right)}\right),$$

where C_3 depends on K , $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$. Set

$$\varphi(x) = C_3 \sqrt{M} x \left(1 + \sqrt{\log\left(\frac{M^{3/2}}{x}\right)}\right)$$

The function φ is increasing on $]0, M^{3/2}[$, and $\varphi(x)/x$ is decreasing. Moreover $\varphi(\sigma) \geq \int_0^\sigma \sqrt{H(u)} du$ and $\varphi^2(\sigma) \geq \sigma^2 H(\sigma)$.

7.2.5. End of the proof, choice of parameters.

As soon as $N \geq C_3^2/M^2 := N_0$, we may define σ_M as the solution of equation $\varphi(x) = \sqrt{N}x^2$. Then, for all $\sigma \geq \sigma_M$,

$$H(\sigma) \leq \frac{\varphi(\sigma)^2}{\sigma^2} \leq \frac{\varphi(\sigma)}{\sigma} \sigma \sqrt{N}.$$

This yields, for all $\sigma \geq \sigma_M$,

$$E \leq (1 + 2C_{\mathcal{F},\infty}^3 + 2K^{3/2}C_{\mathcal{F},2}^3)\varphi(\sigma)\sqrt{N},$$

which was required in (8).

Moreover $\frac{\varphi(2x_M)}{x_M\sqrt{N}} \leq 2\sigma_M$ as soon as $x_M \geq \sigma_M$. Combining (9) and (8), we obtain, with probability $1 - e^{-z_M - z}$:

$$Z_M \leq C^{**} \left[\frac{\sigma_M}{x_M} + \sqrt{\frac{z_M + z}{x_M^2 N}} + \frac{z_M + z}{x_M^2 N} \right],$$

where C^{**} depends on $K, C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*$. Now let us choose $x_M = \theta^{-1} \sqrt{\sigma_M^2 + \frac{z_M + z}{N}}$ with θ such that $2\theta + \theta^2 \leq (C^{**})^{-1}/4$. This choice entails: $x_M \geq \theta^{-1}\sigma_M$ and $x_M^2 \geq \theta^{-2} \frac{z_M + z}{N}$. Then with probability $1 - e^{-z_M - z}$:

$$Z_M \leq C^{**}(\theta + \theta + \theta^2).$$

We now choose $z_M = M$ which implies $\sum_{M \geq 1} e^{-z_M} = (e - 1)^{-1}$. Then, with probability $1 - (e - 1)^{-1}e^{-z}$,

$$\forall M \quad Z_M \leq C^{**}(2\theta + \theta^2) \leq \frac{1}{4},$$

and for all M ,

$$\begin{aligned} Z_M x_M^2 &\leq C^{**} \left[\sigma_M x_M + x_M \sqrt{\frac{z_M + z}{N}} + \frac{z_M + z}{N} \right] \\ &\leq C^{**} \theta^{-1} \left(\sigma_M + \sqrt{\frac{z_M + z}{N}} \right)^2 + C^{**} \frac{z_M + z}{N}. \end{aligned}$$

Then, with probability $1 - (e - 1)^{-1} e^{-z}$, for all M ,

$$Z_M x_M^2 - C^{**} \left(2\theta^{-1} \sigma_M^2 + (2\theta^{-1} + 1) \frac{M}{N} \right) \leq C^{**} (2\theta^{-1} + 1) \frac{z}{N}.$$

Then the result is proved as soon as

$$(12) \quad \text{pen}(N, M) \geq 2C^{**} \left(2\theta^{-1} \sigma_M^2 + (2\theta^{-1} + 1) \frac{M}{N} \right).$$

It remains to get an upper bound for σ_M . Recall that σ_M is defined as the solution of equation $C_3 \sqrt{M} x (1 + \sqrt{\log(\frac{M^3}{x})}) = \sqrt{N} x^2$. Then we obtain that for some C_4

$$\sigma_M \leq C_4 \sqrt{\frac{M}{N}} (1 + \sqrt{\log(N)}),$$

and (12) holds as soon as

$$\text{pen}(N, M) \geq \rho^* \frac{M \log(N)}{N}$$

for some constant ρ^* depending on $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ (**Scenario B**) or on \mathbf{Q}^* , $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ (**Scenario A**).

7.3. Proof of Theorem 3. Denote $N(\mathbf{Q}, \mathbf{h}) = \|g^{\mathbf{Q}, \mathbf{f}^* + \mathbf{h}} - g^{\mathbf{Q}, \mathbf{f}^*}\|_2^2$. What we want to prove is that

$$c := c(\mathcal{K}, \mathcal{V}, \mathfrak{F}^*)^2 := \inf_{\tilde{\mathbf{Q}} \in \mathcal{V}, \tilde{\mathbf{h}} \in \mathcal{K}^{\mathcal{K}}, \|\tilde{\mathbf{h}}\|_{\tilde{\mathbf{Q}}} \neq 0} \frac{N(\tilde{\mathbf{Q}}, \tilde{\mathbf{h}})}{\|\tilde{\mathbf{h}}\|_{\tilde{\mathbf{Q}}}^2} > 0.$$

Let $(\mathbf{Q}_n, \mathbf{h}_n)_n$ be a sequence in $\mathcal{V} \times \mathcal{K}$ such that $c = \lim_n \frac{N(\mathbf{Q}_n, \mathbf{h}_n)}{\|\mathbf{h}_n\|_{\mathbf{Q}_n}^2}$. Let (\mathbf{Q}, \mathbf{h}) be a limit point (for the weak topology) of the sequence in the weakly compact set $\mathcal{V} \times \mathcal{K}$ (indeed we can apply Banach-Alaoglu's theorem since \mathcal{K} is bounded and \mathbf{L}^2 can be identified to its dual). Moreover, note that $N(\mathbf{Q}, \mathbf{h})$ and $\|\mathbf{h}\|_{\mathbf{Q}}^2$ are bounded on $\mathcal{V} \times \mathcal{K}$ and thus, considering a sub-sequence, we can assume that $N(\mathbf{Q}_n, \mathbf{h}_n)$, $\|\mathbf{h}_n\|_{\mathbf{Q}_n}^2$, and $\|\mathbf{h}_n\|_{\mathbf{Q}_n}^2$ converge. Then if τ is a permutation such that, for some particular i, j , $Q(\tau(i), \tau(j)) \neq Q(i, j)$, then for large enough n , $Q_n(\tau(i), \tau(j)) \neq Q_n(i, j)$. Thus

$$\limsup_{n \rightarrow +\infty} T_{\mathbf{Q}_n} \subset T_{\mathbf{Q}},$$

so that

$$\lim_{n \rightarrow +\infty} \|\mathbf{h}_n\|_{\mathbf{Q}_n} \geq \lim_{n \rightarrow +\infty} \|\mathbf{h}_n\|_{\mathbf{Q}}.$$

Moreover, since $h \mapsto \|\mathbf{h}\|_{\mathbf{Q}}$ is strongly convex and strongly lower semi-continuous we know that it is weakly lower semi-continuous and it holds

$$\lim_{n \rightarrow +\infty} \|\mathbf{h}_n\|_{\mathbf{Q}} \geq \|\mathbf{h}\|_{\mathbf{Q}}.$$

Thus if $\|\mathbf{h}\|_{\mathbf{Q}} \neq 0$, then using Lemma 1

$$c = \frac{N(\mathbf{Q}, \mathbf{h})}{\lim_{n \rightarrow +\infty} \|\mathbf{h}_n\|_{\mathbf{Q}_n}^2} > 0.$$

Consider now the situation where $\lim_{n \rightarrow +\infty} \|\mathbf{h}_n\|_{\mathbf{Q}_n} = 0$. In this case, note that $\|\mathbf{h}\|_{\mathbf{Q}} = 0$. If $\|\mathbf{h}\|_{\mathbf{Q}}^2 = \sum_{k=1}^K \|h_k + f_k^* - f_{\tau(k)}^*\|_2^2$ for some permutation τ not equal to the identity permutation, then exchange the states in the transition matrix using τ so that we just have to consider the situation where $\|\mathbf{h}\|_{\mathbf{Q}}^2 = \|\mathbf{h}\|_2^2$ with $\|\mathbf{h}\|_2^2 := \sum_{k=1}^K \|h_k\|_2^2$. Then, for large enough n , $\|\mathbf{h}_n\|_{\mathbf{Q}_n} = \|\mathbf{h}_n\|_2^2$.

Direct computation shows that $N(\mathbf{Q}, \mathbf{h})$ is polynomial in the variables $Q_{i,j}$, $\langle f_i^*, f_j^* \rangle$, $\langle h_i, f_j^* \rangle$, $\langle h_i, h_j \rangle$, $i, j = 1, \dots, K$, without linear part. Let $D(\mathbf{Q}, \mathbf{h})$ denote the quadratic part with respect to the variable \mathbf{h} . One gets

$$N(\mathbf{Q}, \mathbf{h}) = D(\mathbf{Q}, \mathbf{h}) + O(\|\mathbf{h}\|_2^3)$$

where the $O(\cdot)$ depends only on \mathbf{f}^* . Let us first notice that $D(\cdot, \cdot)$ is always non negative. Indeed, since for all $\mathbf{Q} \in \mathcal{V}$ and all $\mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$ one has $N(\mathbf{Q}, \mathbf{h}) \geq 0$, it holds

$$\forall \mathbf{Q} \in \mathcal{V}, \forall \mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K, \frac{D(\mathbf{Q}, \mathbf{h})}{\|\mathbf{h}\|_2^2} + O(\|\mathbf{h}\|_2) \geq 0,$$

so that, since for all $\lambda \in \mathbb{R}$, $D(\mathbf{Q}, \lambda \mathbf{h}) = \lambda^2 D(\mathbf{Q}, \mathbf{h})$,

$$(13) \quad \forall \mathbf{Q} \in \mathcal{V}, \forall \mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K, D(\mathbf{Q}, \mathbf{h}) \geq 0.$$

Then we obtain

$$c = \lim_{n \rightarrow +\infty} D(\mathbf{Q}_n, \frac{\mathbf{h}_n}{\|\mathbf{h}_n\|_2}).$$

Indeed in this case for large enough n , $\mathbf{Q}_n(1, 2) \neq \mathbf{Q}_n(2, 1)$ so that $\|\mathbf{h}_n\|_{\mathbf{Q}_n} = \|\mathbf{h}_n\|_2$. We shall now study the function $D(\mathbf{Q}, \mathbf{a})$ for $\mathbf{a} = (a_1, \dots, a_K)$ of form $\frac{\mathbf{h}}{\|\mathbf{h}\|_2}$ with $\mathbf{h} \in \mathcal{K}^K$. Let $\mathbf{u} = (u_1, \dots, u_K)$ be such that u_i , $i = 1, \dots, K$, is the orthogonal projection of a_i on the subspace of $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$ orthogonal to $f_1^* \dots, f_K^*$. Direct computation gives that

$$D(\mathbf{Q}, \mathbf{a}) = T(\mathbf{Q}, \mathbf{u}) + D(\mathbf{Q}, \mathbf{a} - \mathbf{u})$$

where for any \mathbf{Q} and \mathbf{u} ,

$$\begin{aligned} T(\mathbf{Q}, \mathbf{u}) &= \sum_{i,j=1}^K \{ \langle (\mathbf{Q}^T \mathbf{A} \mathbf{u})_i, (\mathbf{Q}^T \mathbf{A} \mathbf{u})_j \rangle \langle f_i^*, f_j^* \rangle \langle (\mathbf{Q} \mathbf{f}^*)_i, (\mathbf{Q} \mathbf{f}^*)_j \rangle \\ &\quad + \langle (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_i, (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_j \rangle \langle u_i, u_j \rangle \langle (\mathbf{Q} \mathbf{f}^*)_i, (\mathbf{Q} \mathbf{f}^*)_j \rangle \\ &\quad + \langle (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_i, (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_j \rangle \langle f_i^*, f_j^* \rangle \langle (\mathbf{Q} \mathbf{u})_i, (\mathbf{Q} \mathbf{u})_j \rangle \}. \end{aligned}$$

Here, $A = \mathfrak{D}\text{diag}[\pi]$ with π the stationary distribution of \mathbf{Q} . But when \mathbf{Q} and A are full rank, and \mathfrak{F}^* is linearly independent, the matrices

$$\langle \langle f_i^*, f_j^* \rangle \rangle_{i,j}, \quad \langle \langle (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_i, (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_j \rangle \rangle_{i,j}, \quad \text{and} \quad \langle \langle (\mathbf{Q} \mathbf{f}^*)_i, (\mathbf{Q} \mathbf{f}^*)_j \rangle \rangle_{i,j},$$

are positive definite.

Remark — We shall now use the fact that if B and C are symmetric positive definite $K \times K$ matrices, then the $K \times K$ matrix E given by $E_{i,j} = B_{i,j} C_{i,j}$, $i, j = 1, \dots, K$ is positive definite. Indeed, let U_1, \dots, U_K be the eigenvectors of B with corresponding eigenvalues $\lambda_1(B) \geq \dots \geq \lambda_K(B) > 0$, so that $B = \sum_{r=1}^K \lambda_r(B) U_r U_r^T$. Let also $\lambda_1(C) \geq \dots \geq \lambda_K(C) > 0$ be the eigenvalues of C . Let

now $x \in \mathbb{R}^K$. Then

$$\begin{aligned} x^T E x &= \sum_{i,j=1}^K x_i B_{i,j} C_{i,j} x_j = \sum_{r=1}^K \lambda_r(B) \sum_{i,j=1}^K x_i (U_r)_i (U_r)_j C_{i,j} x_j \\ &\geq \left(\min_{r=1,\dots,K} \lambda_r(C) \right) \left(\min_{r=1,\dots,K} \lambda_r(B) \right) \sum_{r=1}^K \sum_{i=1}^K (x_i (U_r)_i)^2 \\ &= \left(\min_{r=1,\dots,K} \lambda_r(C) \right) \left(\min_{r=1,\dots,K} \lambda_r(B) \right) \|x\|^2. \end{aligned}$$

Thus, if R, V, W are the matrices given by, for all $i, j = 1, 2$,

$$\begin{aligned} R_{i,j} &= (\langle (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_i, (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_j \rangle)_{i,j} (\langle f_i^*, f_j^* \rangle)_{i,j}, \\ V_{i,j} &= (\langle f_i^*, f_j^* \rangle)_{i,j} (\langle (\mathbf{Q} \mathbf{f}^*)_i, (\mathbf{Q} \mathbf{f}^*)_j \rangle)_{i,j}, \\ W_{i,j} &= (\langle (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_i, (\mathbf{Q}^T \mathbf{A} \mathbf{f}^*)_j \rangle)_{i,j} (\langle (\mathbf{Q} \mathbf{f}^*)_i, (\mathbf{Q} \mathbf{f}^*)_j \rangle)_{i,j}, \end{aligned}$$

and if we denote $\Lambda(\mathbf{Q})$ the minimum of their eigenvalues, then $\Lambda(\mathbf{Q}) > 0$ and we have for any $\mathbf{u} \in (\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D))^K$,

$$\begin{aligned} T(\mathbf{Q}, \mathbf{u}) &= \int ((\mathbf{Q}^T \mathbf{A} \mathbf{u})^T(y) V (\mathbf{Q}^T \mathbf{A} \mathbf{u})(y) + \mathbf{u}^T(y) W \mathbf{u}(y) + (\mathbf{Q} \mathbf{u})^T(y) R (\mathbf{Q} \mathbf{u})(y)) d(\mathcal{L}^D)^{\otimes 3}(y) \\ &\geq \int \Lambda(\mathbf{Q}) (\|\mathbf{Q}^T \mathbf{A} \mathbf{u}(y)\|^2 + \|\mathbf{u}(y)\|^2 + \|\mathbf{Q} \mathbf{u}(y)\|^2) d(\mathcal{L}^D)^{\otimes 3}(y) \\ &= \Lambda(\mathbf{Q}) \sum_{k=1}^K (\|(\mathbf{Q}^T \mathbf{A} \mathbf{u})_k\|_2^2 + \|u_k\|_2^2 + \|(\mathbf{Q} \mathbf{u})_k\|_2^2). \end{aligned}$$

Moreover, we have

$$\lim_{n \rightarrow \infty} \Lambda(\mathbf{Q}_n) = \Lambda(\mathbf{Q}) > 0.$$

Let $\mathbf{a}_n = \frac{\mathbf{h}_n}{\|\mathbf{h}_n\|_2}$, \mathbf{u}_n the orthogonal projection (coordinate by coordinate) of \mathbf{a}_n on the subspace of $\mathbf{L}^2(\mathcal{Y}, \mathcal{L}^D)$ orthogonal to f_1^*, \dots, f_K^* . We get $c \geq \Lambda(\mathbf{Q}) \liminf_{n \rightarrow \infty} \|\mathbf{u}_n\|_2$ so that in case $\liminf_{n \rightarrow \infty} \|\mathbf{u}_n\|_2 > 0$, we get $c > 0$.

Else, using the subsequence for which $\liminf_{n \rightarrow \infty} \|\mathbf{u}_n\|_2 = 0$ we have

$$c \geq \liminf_{n \rightarrow \infty} D(\mathbf{Q}_n, \mathbf{a}_n - \mathbf{u}_n).$$

We argue now on this subsequence. The sequence $(\mathbf{a}_n - \mathbf{u}_n)_n$ has coordinates in the finite dimensional space spanned by f_1^*, \dots, f_K^* , so that it has a limit point $\mathbf{a} = (a_1, \dots, a_K)$, and $c \geq D(\mathbf{Q}, \mathbf{a})$. Since on the subsequence $\lim_{n \rightarrow \infty} \|\mathbf{u}_n\|_2 = 0$, we get $\int a_k d\mathcal{L}^D = 0$, $k = 1, \dots, K$. Thus there exists a $K \times K$ matrix U such that $\mathbf{a}^T = U(\mathbf{f}^*)^T$ and $U \mathbf{1}_k = 0$, and computation leads to

$$\begin{aligned} D(\mathbf{Q}, \mathbf{a}) &= \sum_{i,j} \left\{ (\mathbf{Q}^T \mathbf{A} U G^* U^T \mathbf{A} \mathbf{Q})_{i,j} (G^*)_{i,j} (\mathbf{Q} G^* \mathbf{Q}^T)_{i,j} + (\mathbf{Q}^T \mathbf{A} G^* \mathbf{A} \mathbf{Q})_{i,j} (U G^* U^T)_{i,j} (\mathbf{Q} G^* \mathbf{Q}^T)_{i,j} \right. \\ &\quad \left. + (\mathbf{Q}^T \mathbf{A} G^* \mathbf{A} \mathbf{Q})_{i,j} (G^*)_{i,j} (\mathbf{Q} U G^* U^T \mathbf{Q}^T)_{i,j} \right\} + 2 \sum_{i,j} \left\{ (\mathbf{Q}^T \mathbf{A} U G^* \mathbf{A} \mathbf{Q})_{i,j} (U G^*)_{j,i} (\mathbf{Q} G^* \mathbf{Q}^T)_{i,j} \right. \\ &\quad \left. + (\mathbf{Q}^T \mathbf{A} U G^* \mathbf{A} \mathbf{Q})_{i,j} (\mathbf{Q} U G^* \mathbf{Q}^T)_{j,i} (G^*)_{i,j} + (U G^*)_{i,j} (\mathbf{Q} U G^* \mathbf{Q}^T)_{j,i} (\mathbf{Q}^T \mathbf{A} G^* \mathbf{A} \mathbf{Q})_{i,j} \right\}. \end{aligned}$$

with G^* the $K \times K$ matrix such that $(G^*)_{i,j} = \langle f_i^*, f_j^* \rangle$, $i = 1, \dots, K$.

This is the quadratic form \mathcal{D} in $U_{i,j}$, $i = 1, \dots, K$, $j = 1, \dots, K-1$ defined in Section 4.2. This quadratic form is non negative, and as soon as it is positive, we get that $c > 0$. But the quadratic form \mathcal{D} is positive as soon as its determinant is non zero, that is if and only if $H(\mathbf{Q}, G(\mathbf{f}^*)) \neq 0$.

7.4. Proof of Lemma 2. Here we specialize to the situation where $K = 2$. In such a case, $\mathbf{f}^* = (f_1^*, f_2^*)$, and

$$\mathbf{Q}^* = \begin{pmatrix} 1-p^* & p^* \\ q^* & 1-q^* \end{pmatrix}$$

for some p^*, q^* in $[0, 1]$ for which $0 < p^* < 1$, $0 < q^* < 1$, $p^* \neq 1 - q^*$. Now

$$U = \begin{pmatrix} \alpha & -\alpha \\ \beta & -\beta \end{pmatrix}$$

for some real numbers α and β , and brute force computation gives $D(\mathbf{Q}, \mathbf{a}) = D_{1,1}\alpha^2 + 2D_{1,2}\alpha\beta + D_{2,2}\beta^2$ with, denoting $p = \mathbf{Q}(1, 2)$ and $q = \mathbf{Q}(2, 1)$:

$$\begin{aligned} \frac{(p+q)^2 D_{1,1}}{q^2} &= 2(1-p)^2 \|f_1^* - f_2^*\|^2 \|f_1^*\|^2 \|(1-p)f_1^* + pf_2^*\|^2 + \|(1-p)f_1^* + pf_2^*\|^4 \|f_1^* - f_2^*\|^2 \\ &\quad + 4p(1-p) (\langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \langle f_1^*, f_2^* \rangle \|f_1^* - f_2^*\|^2 \\ &\quad + 2p^2 \|f_1^* - f_2^*\|^2 \|f_2^*\|^2 \|qf_1^* + (1-q)f_2^*\|^2 \\ &\quad + 2(1-p)^2 (\langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle)^2 \|f_1^*\|^2 \\ &\quad + 2p^2 (\langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle)^2 \|f_2^*\|^2 \\ &\quad + 4p(1-p) \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_2^* \rangle \\ &\quad + 4(1-p) \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_1^* - f_2^* \rangle \|(1-p)f_1^* + pf_2^*\|^2 \\ &\quad + 4p \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_2^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle, \end{aligned}$$

$$\begin{aligned} \frac{(p+q)^2 D_{2,2}}{p^2} &= 2q^2 \|f_1^* - f_2^*\|^2 \|f_1^*\|^2 \|(1-p)f_1^* + pf_2^*\|^2 + \|qf_1^* + (1-q)f_2^*\|^4 \|f_1^* - f_2^*\|^2 \\ &\quad + 4(1-q)q (\langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \langle f_1^*, f_2^* \rangle \|f_1^* - f_2^*\|^2 \\ &\quad + 2(1-q)^2 \|f_1^* - f_2^*\|^2 \|f_2^*\|^2 \|qf_1^* + (1-q)f_2^*\|^2 \\ &\quad + 2q^2 (\langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle)^2 \|f_1^*\|^2 \\ &\quad + 2(1-q)^2 (\langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle)^2 \|f_2^*\|^2 \\ &\quad + 4q(1-q) \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_2^* \rangle \\ &\quad + 4q \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \\ &\quad + 4(1-q) \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle f_2^*, f_1^* - f_2^* \rangle \|qf_1^* + (1-q)f_2^*\|^2, \end{aligned}$$

and:

$$\begin{aligned}
\frac{(p+q)^2 D_{1,2}}{pq} = & 2(1-p)q \|f_1^* - f_2^*\|^2 \|f_1^*\|^2 \|(1-p)f_1^* + pf_2^*\|^2 \\
& + 2[pq + (1-p)(1-q)] (\langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \langle f_1^*, f_2^* \rangle \|f_1^* - f_2^*\|^2 \\
& + (\langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle)^2 \|f_1^* - f_2^*\|^2 \\
& + 2p(1-q) \|f_1^* - f_2^*\|^2 \|f_2^*\|^2 \|qf_1^* + (1-q)f_2^*\|^2 \\
& + 2q(1-p) (\langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle)^2 \|f_1^*\|^2 \\
& + 2p(1-q) (\langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle)^2 \|f_2^*\|^2 \\
& + 2pq \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, f_1 - f_2^* \rangle \langle f_1^*, f_2^* \rangle \\
& + 2(1-p)(1-q) \langle qf_1^* + (1-q)f_2^*, f_1 - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_2^* \rangle \\
& + q \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_1^* - f_2^* \rangle \|(1-p)f_1^* + pf_2^*\|^2 \\
& + 2(1-p) \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle f_1^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \\
& + 2(1-q) \langle (1-p)f_1^* + pf_2^*, f_1^* - f_2^* \rangle \langle f_2^*, f_1^* - f_2^* \rangle \langle (1-p)f_1^* + pf_2^*, qf_1^* + (1-q)f_2^* \rangle \\
& + 2p \langle qf_1^* + (1-q)f_2^*, f_1^* - f_2^* \rangle \langle f_2^*, f_1^* - f_2^* \rangle \|qf_1^* + (1-q)f_2^*\|^2.
\end{aligned}$$

We have:

$$H(\mathbf{Q}, G(\mathbf{f}^*)) = D_{1,1}D_{2,2} - D_{1,2}^2.$$

We shall now write $H(\mathbf{Q}, G(\mathbf{f}^*))$ using

$$n_1 = \|f_1^*\|_2, \quad n_2 = \|f_2^*\|_2, \quad a = \frac{\langle f_1^*, f_2^* \rangle}{\|f_1^*\|_2 \|f_2^*\|_2},$$

for which the range is $[1, \infty[^2 \times]0, 1[$. Doing so, we obtain a polynomial P_1 in the variables n_1, n_2, a, p and q .

First observe that, by symmetry,

$$P_1(n_1, n_2, a, p, q) = P_1(n_2, n_1, a, q, p).$$

so that it is sufficient to prove that the polynomial P_1 is positive on the domain

$$(14) \quad 1 \leq n_2 \leq n_1,$$

and $0 \leq a < 1$ and $0 < p \neq q < 1$.

Furthermore, consider the change of variable

$$q = 1 - p + d$$

then we have a polynomial P_2 in the variables n_1, n_2, a, p and d which factorizes with

$$\frac{p^2(1-a^2)d^2 n_1^2 n_2^2 (1+d-p)^2}{(1+d)^4}.$$

Dividing by this factor, one gets a polynomial P_3 which is homogeneous of degree 8 in n_1 and n_2 , so that one may set $n_1 = 1$ and keep $b = n_2 \in]0, 1]$ (observe that we have used (14) to reduce the problem to the domain $n_2/n_1 \leq 1$) and obtain a polynomial P_4 in the variables b, a, p and d . It remains to prove that P_4 is positive on $\mathcal{D}_4 = \{b \in]0, 1], a \in [0, 1[, p \in]0, 1[, d \in]p-1, 0[\cup]0, p[\}$.

Consider now the following change of variables

$$b = \frac{1}{1+x^2}, \quad a = \frac{y^2}{1+y^2}, \quad p = \frac{z^2}{1+z^2}, \quad \text{and} \quad d = \frac{(tz)^2 - 1}{(1+t^2)(1+z^2)},$$

mapping $(x, y, z, t) \in \mathbb{R}^4$ onto $(b, a, p, d) \in \mathcal{D}_5 = \{b \in]0, 1], a \in [0, 1[, p \in [0, 1[, d \in]p-1, p[\}$ which contains \mathcal{D}_4 . This change of variables maps P_4 onto a rational fraction with positive denominator, namely

$$(1+t^2)^4(1+y^2)^4(1+z^2)^4(1+x^2)^8$$

So it remains to prove that its numerator P_5 , which is polynomial, is positive on \mathbb{R}^4 . An expression of P_5 can be found in Appendix B. Observe that P_5 is polynomial in x^2, y^2, z^2 and t^2 and there are only three monomials with negative coefficients. These monomials can be expressed as sum of squares using others monomials, namely:

- $-18x^{12}t^2 + 27x^{12} + 1979x^{12}t^4 = 18x^{12} + 9(x^6 - x^6t^2)^2 + 1970x^{12}t^4$,
- $-108x^{10}t^2 + 1970x^{12}t^4 + 495x^8 = 439x^8 + 56(x^4 - x^6t^2)^2 + 1914x^{12}t^4 + 4t^2x^{10}$,
- and $-114x^8t^2 + 972x^4 + 1914x^{12}t^4 = 915x^4 + 57(x^2 - x^6t^2)^2 + 1857x^{12}t^4$.

Thus P_5 is equal to 144 more a sum of squares, hence it is positive. This proves that $H(\mathbf{Q}, G(\mathbf{f}^*))$ is always positive.

APPENDIX A. CONCENTRATION INEQUALITIES

We first recall results that hold both for **(Scenario A)** (where we consider N i.i.d. samples $(Y_1^{(s)}, Y_2^{(s)}, Y_3^{(s)})_{s=1}^N$ of three consecutive observations) and for **(Scenario B)** (where we consider consecutive observations of the same chain).

The following proposition is the classical Bernstein's inequality for **(Scenario A)** and is proved in [Pau13], Theorem 2.4, for **(Scenario B)**.

Proposition 6 — *Let t be a real valued and measurable bounded function on \mathcal{Y}^3 . Let $V = \mathbb{E}[t^2(Z_1)]$. There exists a positive constant c^* depending only on \mathbf{Q}^* such that for all $0 \leq \lambda \leq 1/(2\sqrt{2}c^*\|t\|_\infty)$:*

$$(15) \quad \log \mathbb{E} \exp \left[\lambda \sum_{s=1}^N (t(Z_s) - \mathbb{E}t(Z_s)) \right] \leq \frac{2Nc^*V\lambda^2}{1 - 2\sqrt{2}c^*\|t\|_\infty\lambda}$$

so that for all $x \geq 0$,

$$(16) \quad \mathbb{P} \left(\sum_{s=1}^N (t(Z_s) - \mathbb{E}t(Z_s)) \geq 2\sqrt{2Nc^*Vx} + 2\sqrt{2}c^*\|t\|_\infty x \right) \leq e^{-x}.$$

We now state a deviation inequality, which comes from [Mas07] Theorem 6.8 and Corollary 6.9 for **(Scenario A)**. For **(Scenario B)** the proof of the following proposition follows mutatis mutandis from the proof of Theorem 6.8 (and then Corollary 6.9) in [Mas07] the early first step being equation (15). Recall that when t_1 and t_2 are real valued functions, the bracket $[t_1, t_2]$ is the set of real valued functions t such that $t_1(\cdot) \leq t(\cdot) \leq t_2(\cdot)$. For any measurable set A such that $\mathbb{P}(A) > 0$, and any integrable random variable Z , denote $E^A[Z] = E[Z\mathbb{1}_A]/\mathbb{P}(A)$.

Proposition 7 — *Let \mathcal{T} be some countable class of real valued and measurable functions on \mathcal{Y}^3 . Assume that there exists some positive numbers σ and b such that for all $t \in \mathcal{T}$, $\|t\|_\infty \leq b$ and $\mathbb{E}[t^2(Z_1)] \leq \sigma^2$.*

Assume furthermore that for any positive number δ , there exists some finite set B_δ of brackets covering \mathcal{F} such that for any bracket $[t_1, t_2] \in B_\delta$, $\|t_1 - t_2\|_\infty \leq b$ and $\mathbb{E}[(t_1 - t_2)^2(Z_1)] \leq \delta^2$. Let $e^{H(\delta)}$ denote the minimal cardinality of such a covering. Then, there exists a positive constant C^ depending only on \mathbf{Q}^* such that: for any measurable set A ,*

$$\mathbb{E}^A \left(\sup_{t \in \mathcal{T}} \sum_{s=1}^N (t(Z_s) - \mathbb{E}t(Z_s)) \right) \leq C^* \left[E + \sigma \sqrt{N \log \left(\frac{1}{\mathbb{P}(A)} \right)} + b \log \left(\frac{1}{\mathbb{P}(A)} \right) \right]$$

and for all positive number x

$$\mathbb{P} \left(\sup_{t \in \mathcal{T}} \sum_{s=1}^N (t(Z_s) - \mathbb{E}t(Z_s)) \geq C^* [E + \sigma \sqrt{Nx} + bx] \right) \leq \exp(-x)$$

where

$$E = \sqrt{N} \int_0^\sigma \sqrt{H(u) \wedge N} du + (b + \sigma)H(\sigma)$$

APPENDIX B. EXPRESSION OF POLYNOMIAL P_5

Computer assisted computations (available on Yohann's web page) give that:

$P_5 =$

$$\begin{aligned} & 144 - 114 t^2 x^8 - 108 t^2 x^{10} - 18 t^2 x^{12} + \\ & 192 t^2 + 128 t^4 + 256 t^6 + 176 t^8 + 576 x^2 + 624 t^2 x^2 + \\ & 672 t^4 x^2 + 1776 t^6 x^2 + 1152 t^8 x^2 + 972 x^4 + 720 t^2 x^4 + \\ & 1884 t^4 x^4 + 5496 t^6 x^4 + 3360 t^8 x^4 + 900 x^6 + 264 t^2 x^6 + \\ & 3556 t^4 x^6 + 9920 t^6 x^6 + 5728 t^8 x^6 + 495 x^8 + \\ & 4551 t^4 x^8 + 11424 t^6 x^8 + 6264 t^8 x^8 + 162 x^{10} + \\ & 3810 t^4 x^{10} + 8592 t^6 x^{10} + 4512 t^8 x^{10} + \\ & 27 x^{12} + 1979 t^4 x^{12} + 4120 t^6 x^{12} + \\ & 2096 t^8 x^{12} + 576 t^4 x^{14} + 1152 t^6 x^{14} + 576 t^8 x^{14} + \\ & 72 t^4 x^{16} + 144 t^6 x^{16} + 72 t^8 x^{16} + 144 y^2 + 480 t^2 y^2 + \\ & 784 t^4 y^2 + 704 t^6 y^2 + 256 t^8 y^2 + 576 x^2 y^2 + \\ & 2064 t^2 x^2 y^2 + 4192 t^4 x^2 y^2 + 4496 t^6 x^2 y^2 + \\ & 1792 t^8 x^2 y^2 + 1080 x^4 y^2 + 4104 t^2 x^4 y^2 + \\ & 10760 t^4 x^4 y^2 + 13528 t^6 x^4 y^2 + 5792 t^8 x^4 y^2 + \\ & 1224 x^6 y^2 + 5016 t^2 x^6 y^2 + 17592 t^4 x^6 y^2 + \\ & 25032 t^6 x^6 y^2 + 11232 t^8 x^6 y^2 + 900 x^8 y^2 + \\ & 4224 t^2 x^8 y^2 + 19924 t^4 x^8 y^2 + 30776 t^6 x^8 y^2 + \\ & 14176 t^8 x^8 y^2 + 432 x^{10} y^2 + 2520 t^2 x^{10} y^2 + \\ & 15584 t^4 x^{10} y^2 + 25336 t^6 x^{10} y^2 + 11840 t^8 x^{10} y^2 + \\ & 108 x^{12} y^2 + 936 t^2 x^{12} y^2 + 7916 t^4 x^{12} y^2 + \\ & 13456 t^6 x^{12} y^2 + 6368 t^8 x^{12} y^2 + 144 t^2 x^{14} y^2 + \\ & 2304 t^4 x^{14} y^2 + 4176 t^6 x^{14} y^2 + 2016 t^8 x^{14} y^2 + \\ & 288 t^4 x^{16} y^2 + 576 t^6 x^{16} y^2 + 288 t^8 x^{16} y^2 + 144 y^4 + \\ & 480 t^2 y^4 + 624 t^4 y^4 + 384 t^6 y^4 + 96 t^8 y^4 + 576 x^2 y^4 + \\ & 2208 t^2 x^2 y^4 + 3392 t^4 x^2 y^4 + 2464 t^6 x^2 y^4 + \\ & 704 t^8 x^2 y^4 + 1188 x^4 y^4 + 5256 t^2 x^4 y^4 + \\ & 9636 t^4 x^4 y^4 + 8256 t^6 x^4 y^4 + 2688 t^8 x^4 y^4 + \\ & 1548 x^6 y^4 + 8112 t^2 x^6 y^4 + 18076 t^4 x^6 y^4 + \\ & 18008 t^6 x^6 y^4 + 6496 t^8 x^6 y^4 + 1359 x^8 y^4 + \\ & 8598 t^2 x^8 y^4 + 23375 t^4 x^8 y^4 + 26392 t^6 x^8 y^4 + \\ & 10256 t^8 x^8 y^4 + 810 x^{10} y^4 + 6156 t^2 x^{10} y^4 + \\ & 20442 t^4 x^{10} y^4 + 25656 t^6 x^{10} y^4 + 10560 t^8 x^{10} y^4 + \\ & 243 x^{12} y^4 + 2574 t^2 x^{12} y^4 + 11299 t^4 x^{12} y^4 + \\ & 15848 t^6 x^{12} y^4 + 6880 t^8 x^{12} y^4 + 432 t^2 x^{14} y^4 + \\ & 3456 t^4 x^{14} y^4 + 5616 t^6 x^{14} y^4 + 2592 t^8 x^{14} y^4 + \\ & 432 t^4 x^{16} y^4 + 864 t^6 x^{16} y^4 + 432 t^8 x^{16} y^4 + \\ & 216 x^4 y^6 + 720 t^2 x^4 y^6 + 952 t^4 x^4 y^6 + 608 t^6 x^4 y^6 + \\ & 160 t^8 x^4 y^6 + 648 x^6 y^6 + 2592 t^2 x^6 y^6 + \\ & 4168 t^4 x^6 y^6 + 3152 t^6 x^6 y^6 + 928 t^8 x^6 y^6 + \\ & 918 x^8 y^6 + 4428 t^2 x^8 y^6 + 8502 t^4 x^8 y^6 + \\ & 7392 t^6 x^8 y^6 + 2400 t^8 x^8 y^6 + 756 x^{10} y^6 + \\ & 4392 t^2 x^{10} y^6 + 10036 t^4 x^{10} y^6 + 9920 t^6 x^{10} y^6 + \\ & 3520 t^8 x^{10} y^6 + 270 x^{12} y^6 + 2268 t^2 x^{12} y^6 + \\ & 6766 t^4 x^{12} y^6 + 7808 t^6 x^{12} y^6 + 3040 t^8 x^{12} y^6 + \\ & 432 t^2 x^{14} y^6 + 2304 t^4 x^{14} y^6 + 3312 t^6 x^{14} y^6 + \\ & 1440 t^8 x^{14} y^6 + 288 t^4 x^{16} y^6 + 576 t^6 x^{16} y^6 + \\ & 288 t^8 x^{16} y^6 + 108 x^8 y^8 + 360 t^2 x^8 y^8 + 468 t^4 x^8 y^8 + \\ & 288 t^6 x^8 y^8 + 72 t^8 x^8 y^8 + 216 x^{10} y^8 + 864 t^2 x^{10} y^8 + \\ & 1368 t^4 x^{10} y^8 + 1008 t^6 x^{10} y^8 + 288 t^8 x^{10} y^8 + \\ & 108 x^{12} y^8 + 648 t^2 x^{12} y^8 + 1404 t^4 x^{12} y^8 + \end{aligned}$$

$$\begin{aligned}
& 1296 t^6 x^{12} y^8 + 432 t^8 x^{12} y^8 + 144 t^2 x^{14} y^8 + \\
& 576 t^4 x^{14} y^8 + 720 t^6 x^{14} y^8 + 288 t^8 x^{14} y^8 + \\
& 72 t^4 x^{16} y^8 + 144 t^6 x^{16} y^8 + 72 t^8 x^{16} y^8 + 192 z^2 + \\
& 416 t^2 z^2 + 288 t^4 z^2 + 320 t^6 z^2 + 256 t^8 z^2 + \\
& 912 x^2 z^2 + 1664 t^2 x^2 z^2 + 1248 t^4 x^2 z^2 + \\
& 2304 t^6 x^2 z^2 + 1808 t^8 x^2 z^2 + 1728 x^4 z^2 + \\
& 2520 t^2 x^4 z^2 + 2776 t^4 x^4 z^2 + 7624 t^6 x^4 z^2 + \\
& 5640 t^8 x^4 z^2 + 1704 x^6 z^2 + 1736 t^2 x^6 z^2 + \\
& 4664 t^4 x^6 z^2 + 14808 t^6 x^6 z^2 + 10176 t^8 x^6 z^2 + \\
& 966 x^8 z^2 + 494 t^2 x^8 z^2 + 6098 t^4 x^8 z^2 + \\
& 18218 t^6 x^8 z^2 + 11648 t^8 x^8 z^2 + 324 x^{10} z^2 + \\
& 36 t^2 x^{10} z^2 + 5468 t^4 x^{10} z^2 + 14444 t^6 x^{10} z^2 + \\
& 8688 t^8 x^{10} z^2 + 54 x^{12} z^2 + 6 t^2 x^{12} z^2 + \\
& 3002 t^4 x^{12} z^2 + 7186 t^6 x^{12} z^2 + 4136 t^8 x^{12} z^2 + \\
& 896 t^4 x^{14} z^2 + 2048 t^6 x^{14} z^2 + 1152 t^8 x^{14} z^2 + \\
& 112 t^4 x^{16} z^2 + 256 t^6 x^{16} z^2 + 144 t^8 x^{16} z^2 + \\
& 480 y^2 z^2 + 1312 t^2 y^2 z^2 + 1888 t^4 y^2 z^2 + \\
& 1760 t^6 y^2 z^2 + 704 t^8 y^2 z^2 + 1776 x^2 y^2 z^2 + \\
& 5248 t^2 x^2 y^2 z^2 + 9504 t^4 x^2 y^2 z^2 + \\
& 10624 t^6 x^2 y^2 z^2 + 4592 t^8 x^2 y^2 z^2 + 3096 x^4 y^2 z^2 + \\
& 9904 t^2 x^4 y^2 z^2 + 23104 t^4 x^4 y^2 z^2 + \\
& 30288 t^6 x^4 y^2 z^2 + 13992 t^8 x^4 y^2 z^2 + 3144 x^6 y^2 z^2 + \\
& 11344 t^2 x^6 y^2 z^2 + 35712 t^4 x^6 y^2 z^2 + \\
& 53424 t^6 x^6 y^2 z^2 + 25912 t^8 x^6 y^2 z^2 + 2064 x^8 y^2 z^2 + \\
& 9016 t^2 x^8 y^2 z^2 + 38552 t^4 x^8 y^2 z^2 + \\
& 63192 t^6 x^8 y^2 z^2 + 31592 t^8 x^8 y^2 z^2 + 936 x^{10} y^2 z^2 + \\
& 5248 t^2 x^{10} y^2 z^2 + 29072 t^4 x^{10} y^2 z^2 + \\
& 50464 t^6 x^{10} y^2 z^2 + 25704 t^8 x^{10} y^2 z^2 + 216 x^{12} y^2 z^2 + \\
& 1872 t^2 x^{12} y^2 z^2 + 14192 t^4 x^{12} y^2 z^2 + \\
& 26056 t^6 x^{12} y^2 z^2 + 13520 t^8 x^{12} y^2 z^2 + \\
& 264 t^2 x^{14} y^2 z^2 + 3896 t^4 x^{14} y^2 z^2 + \\
& 7808 t^6 x^{14} y^2 z^2 + 4176 t^8 x^{14} y^2 z^2 + \\
& 448 t^4 x^{16} y^2 z^2 + 1024 t^6 x^{16} y^2 z^2 + \\
& 576 t^8 x^{16} y^2 z^2 + 480 y^4 z^2 + 1632 t^2 y^4 z^2 + \\
& 2208 t^4 y^4 z^2 + 1440 t^6 y^4 z^2 + 384 t^8 y^4 z^2 + \\
& 1632 x^2 y^4 z^2 + 6528 t^2 x^2 y^4 z^2 + 10688 t^4 x^2 y^4 z^2 + \\
& 8320 t^6 x^2 y^4 z^2 + 2528 t^8 x^2 y^4 z^2 + 3240 x^4 y^4 z^2 + \\
& 14280 t^2 x^4 y^4 z^2 + 27448 t^4 x^4 y^4 z^2 + \\
& 25048 t^6 x^4 y^4 z^2 + 8640 t^8 x^4 y^4 z^2 + 3936 x^6 y^4 z^2 + \\
& 19992 t^2 x^6 y^4 z^2 + 46552 t^4 x^6 y^4 z^2 + \\
& 49352 t^6 x^6 y^4 z^2 + 18856 t^8 x^6 y^4 z^2 + 3198 x^8 y^4 z^2 + \\
& 19518 t^2 x^8 y^4 z^2 + 55218 t^4 x^8 y^4 z^2 + \\
& 66170 t^6 x^8 y^4 z^2 + 27272 t^8 x^8 y^4 z^2 + 1836 x^{10} y^4 z^2 + \\
& 13332 t^2 x^{10} y^4 z^2 + 44988 t^4 x^{10} y^4 z^2 + \\
& 59580 t^6 x^{10} y^4 z^2 + 26088 t^8 x^{10} y^4 z^2 + 486 x^{12} y^4 z^2 + \\
& 5214 t^2 x^{12} y^4 z^2 + 22994 t^4 x^{12} y^4 z^2 + \\
& 34194 t^6 x^{12} y^4 z^2 + 15928 t^8 x^{12} y^4 z^2 + \\
& 792 t^2 x^{14} y^4 z^2 + 6312 t^4 x^{14} y^4 z^2 + \\
& 11136 t^6 x^{14} y^4 z^2 + 5616 t^8 x^{14} y^4 z^2 + \\
& 672 t^4 x^{16} y^4 z^2 + 1536 t^6 x^{16} y^4 z^2 + \\
& 864 t^8 x^{16} y^4 z^2 + 720 x^4 y^6 z^2 + 2480 t^2 x^4 y^6 z^2 + \\
& 3472 t^4 x^4 y^6 z^2 + 2384 t^6 x^4 y^6 z^2 + 672 t^8 x^4 y^6 z^2 + \\
& 1728 x^6 y^6 z^2 + 7440 t^2 x^6 y^6 z^2 + 13072 t^4 x^6 y^6 z^2 + \\
& 10736 t^6 x^6 y^6 z^2 + 3376 t^8 x^6 y^6 z^2 + 2268 x^8 y^6 z^2 + \\
& 11484 t^2 x^8 y^6 z^2 + 23812 t^4 x^8 y^6 z^2 + \\
& 22276 t^6 x^8 y^6 z^2 + 7680 t^8 x^8 y^6 z^2 + 1800 x^{10} y^6 z^2 + \\
& 10568 t^2 x^{10} y^6 z^2 + 25560 t^4 x^{10} y^6 z^2 +
\end{aligned}$$

$$\begin{aligned}
& 26872 t^6 x^{10} y^6 z^2 + 10080 t^8 x^{10} y^6 z^2 + 540 x^{12} y^6 z^2 + \\
& 4836 t^2 x^{12} y^6 z^2 + 15420 t^4 x^{12} y^6 z^2 + \\
& 18964 t^6 x^{12} y^6 z^2 + 7840 t^8 x^{12} y^6 z^2 + \\
& 792 t^2 x^{14} y^6 z^2 + 4520 t^4 x^{14} y^6 z^2 + \\
& 7040 t^6 x^{14} y^6 z^2 + 3312 t^8 x^{14} y^6 z^2 + \\
& 448 t^4 x^{16} y^6 z^2 + 1024 t^6 x^{16} y^6 z^2 + \\
& 576 t^8 x^{16} y^6 z^2 + 360 x^8 y^8 z^2 + 1224 t^2 x^8 y^8 z^2 + \\
& 1656 t^4 x^8 y^8 z^2 + 1080 t^6 x^8 y^8 z^2 + 288 t^8 x^8 y^8 z^2 + \\
& 576 x^{10} y^8 z^2 + 2448 t^2 x^{10} y^8 z^2 + 4176 t^4 x^{10} y^8 z^2 + \\
& 3312 t^6 x^{10} y^8 z^2 + 1008 t^8 x^{10} y^8 z^2 + 216 x^{12} y^8 z^2 + \\
& 1488 t^2 x^{12} y^8 z^2 + 3616 t^4 x^{12} y^8 z^2 + \\
& 3640 t^6 x^{12} y^8 z^2 + 1296 t^8 x^{12} y^8 z^2 + \\
& 264 t^2 x^{14} y^8 z^2 + 1208 t^4 x^{14} y^8 z^2 + \\
& 1664 t^6 x^{14} y^8 z^2 + 720 t^8 x^{14} y^8 z^2 + \\
& 112 t^4 x^{16} y^8 z^2 + 256 t^6 x^{16} y^8 z^2 + 144 t^8 x^{16} y^8 z^2 + \\
& 128 z^4 + 288 t^2 z^4 + 352 t^4 z^4 + 384 t^6 z^4 + 256 t^8 z^4 + \\
& 352 x^2 z^4 + 1056 t^2 x^2 z^4 + 1408 t^4 x^2 z^4 + \\
& 1952 t^6 x^2 z^4 + 1504 t^8 x^2 z^4 + 764 x^4 z^4 + \\
& 2104 t^2 x^4 z^4 + 2616 t^4 x^4 z^4 + 5016 t^6 x^4 z^4 + \\
& 4252 t^8 x^4 z^4 + 804 x^6 z^4 + 1912 t^2 x^6 z^4 + \\
& 2920 t^4 x^6 z^4 + 8536 t^6 x^6 z^4 + 7364 t^8 x^6 z^4 + \\
& 471 x^8 z^4 + 898 t^2 x^8 z^4 + 2694 t^4 x^8 z^4 + \\
& 10058 t^6 x^8 z^4 + 8335 t^8 x^8 z^4 + 162 x^{10} z^4 + \\
& 252 t^2 x^{10} z^4 + 2164 t^4 x^{10} z^4 + 7980 t^6 x^{10} z^4 + \\
& 6226 t^8 x^{10} z^4 + 27 x^{12} z^4 + 42 t^2 x^{12} z^4 + \\
& 1182 t^4 x^{12} z^4 + 4018 t^6 x^{12} z^4 + 2979 t^8 x^{12} z^4 + \\
& 352 t^4 x^{14} z^4 + 1152 t^6 x^{14} z^4 + 832 t^8 x^{14} z^4 + \\
& 44 t^4 x^{16} z^4 + 144 t^6 x^{16} z^4 + 104 t^8 x^{16} z^4 + \\
& 784 y^2 z^4 + 1888 t^2 y^2 z^4 + 2208 t^4 y^2 z^4 + \\
& 1888 t^6 y^2 z^4 + 784 t^8 y^2 z^4 + 2080 x^2 y^2 z^4 + \\
& 5600 t^2 x^2 y^2 z^4 + 8832 t^4 x^2 y^2 z^4 + 9952 t^6 x^2 y^2 z^4 + \\
& 4640 t^8 x^2 y^2 z^4 + 3368 x^4 y^2 z^4 + 9440 t^2 x^4 y^2 z^4 + \\
& 18928 t^4 x^4 y^2 z^4 + 25952 t^6 x^4 y^2 z^4 + \\
& 13224 t^8 x^4 y^2 z^4 + 2840 x^6 y^2 z^4 + 9056 t^2 x^6 y^2 z^4 + \\
& 25872 t^4 x^6 y^2 z^4 + 42464 t^6 x^6 y^2 z^4 + \\
& 23192 t^8 x^6 y^2 z^4 + 1524 x^8 y^2 z^4 + 6072 t^2 x^8 y^2 z^4 + \\
& 25016 t^4 x^8 y^2 z^4 + 46792 t^6 x^8 y^2 z^4 + \\
& 26900 t^8 x^8 y^2 z^4 + 576 x^{10} y^2 z^4 + 3184 t^2 x^{10} y^2 z^4 + \\
& 17216 t^4 x^{10} y^2 z^4 + 35024 t^6 x^{10} y^2 z^4 + \\
& 20928 t^8 x^{10} y^2 z^4 + 108 x^{12} y^2 z^4 + 1008 t^2 x^{12} y^2 z^4 + \\
& 7584 t^4 x^{12} y^2 z^4 + 16968 t^6 x^{12} y^2 z^4 + \\
& 10572 t^8 x^{12} y^2 z^4 + 120 t^2 x^{14} y^2 z^4 + \\
& 1816 t^4 x^{14} y^2 z^4 + 4736 t^6 x^{14} y^2 z^4 + \\
& 3136 t^8 x^{14} y^2 z^4 + 176 t^4 x^{16} y^2 z^4 + \\
& 576 t^6 x^{16} y^2 z^4 + 416 t^8 x^{16} y^2 z^4 + 624 y^4 z^4 + \\
& 2208 t^2 y^4 z^4 + 3168 t^4 y^4 z^4 + 2208 t^6 y^4 z^4 + \\
& 624 t^8 y^4 z^4 + 1600 x^2 y^4 z^4 + 6976 t^2 x^2 y^4 z^4 + \\
& 12672 t^4 x^2 y^4 z^4 + 10816 t^6 x^2 y^4 z^4 + \\
& 3520 t^8 x^2 y^4 z^4 + 3364 x^4 y^4 z^4 + 14456 t^2 x^4 y^4 z^4 + \\
& 29416 t^4 x^4 y^4 z^4 + 29016 t^6 x^4 y^4 z^4 + \\
& 10692 t^8 x^4 y^4 z^4 + 3452 x^6 y^4 z^4 + 17336 t^2 x^6 y^4 z^4 + \\
& 43896 t^4 x^6 y^4 z^4 + 51032 t^6 x^6 y^4 z^4 + \\
& 21020 t^8 x^6 y^4 z^4 + 2495 x^8 y^4 z^4 + 14658 t^2 x^8 y^4 z^4 + \\
& 45814 t^4 x^8 y^4 z^4 + 61162 t^6 x^8 y^4 z^4 + \\
& 27607 t^8 x^8 y^4 z^4 + 1242 x^{10} y^4 z^4 + 8892 t^2 x^{10} y^4 z^4 + \\
& 33252 t^4 x^{10} y^4 z^4 + 49644 t^6 x^{10} y^4 z^4 + \\
& 24234 t^8 x^{10} y^4 z^4 + 243 x^{12} y^4 z^4 + 2914 t^2 x^{12} y^4 z^4 +
\end{aligned}$$

$$\begin{aligned}
& 14758 t^4 x^{12} y^4 z^4 + 25538 t^6 x^{12} y^4 z^4 + \\
& 13643 t^8 x^{12} y^4 z^4 + 360 t^2 x^{14} y^4 z^4 + \\
& 3336 t^4 x^{14} y^4 z^4 + 7296 t^6 x^{14} y^4 z^4 + \\
& 4416 t^8 x^{14} y^4 z^4 + 264 t^4 x^{16} y^4 z^4 + \\
& 864 t^6 x^{16} y^4 z^4 + 624 t^8 x^{16} y^4 z^4 + 952 x^4 y^6 z^4 + \\
& 3472 t^2 x^4 y^6 z^4 + 5232 t^4 x^4 y^6 z^4 + 3856 t^6 x^4 y^6 z^4 + \\
& 1144 t^8 x^4 y^6 z^4 + 1544 x^6 y^6 z^4 + 7760 t^2 x^6 y^6 z^4 + \\
& 15696 t^4 x^6 y^6 z^4 + 14288 t^6 x^6 y^6 z^4 + \\
& 4808 t^8 x^6 y^6 z^4 + 1942 x^8 y^6 z^4 + 10532 t^2 x^8 y^6 z^4 + \\
& 24556 t^4 x^8 y^6 z^4 + 25380 t^6 x^8 y^6 z^4 + \\
& 9414 t^8 x^8 y^6 z^4 + 1332 x^{10} y^6 z^4 + 8408 t^2 x^{10} y^6 z^4 + \\
& 22952 t^4 x^{10} y^6 z^4 + 26776 t^6 x^{10} y^6 z^4 + \\
& 10900 t^8 x^{10} y^6 z^4 + 270 x^{12} y^6 z^4 + 2972 t^2 x^{12} y^6 z^4 + \\
& 11492 t^4 x^{12} y^6 z^4 + 16244 t^6 x^{12} y^6 z^4 + \\
& 7486 t^8 x^{12} y^6 z^4 + 360 t^2 x^{14} y^6 z^4 + \\
& 2632 t^4 x^{14} y^6 z^4 + 4992 t^6 x^{14} y^6 z^4 + \\
& 2752 t^8 x^{14} y^6 z^4 + 176 t^4 x^{16} y^6 z^4 + \\
& 576 t^6 x^{16} y^6 z^4 + 416 t^8 x^{16} y^6 z^4 + 468 x^8 y^8 z^4 + \\
& 1656 t^2 x^8 y^8 z^4 + 2376 t^4 x^8 y^8 z^4 + 1656 t^6 x^8 y^8 z^4 + \\
& 468 t^8 x^8 y^8 z^4 + 504 x^{10} y^8 z^4 + 2448 t^2 x^{10} y^8 z^4 + \\
& 4752 t^4 x^{10} y^8 z^4 + 4176 t^6 x^{10} y^8 z^4 + \\
& 1368 t^8 x^{10} y^8 z^4 + 108 x^{12} y^8 z^4 + 1024 t^2 x^{12} y^8 z^4 + \\
& 3136 t^4 x^{12} y^8 z^4 + 3656 t^6 x^{12} y^8 z^4 + \\
& 1436 t^8 x^{12} y^8 z^4 + 120 t^2 x^{14} y^8 z^4 + \\
& 760 t^4 x^{14} y^8 z^4 + 1280 t^6 x^{14} y^8 z^4 + \\
& 640 t^8 x^{14} y^8 z^4 + 44 t^4 x^{16} y^8 z^4 + 144 t^6 x^{16} y^8 z^4 + \\
& 104 t^8 x^{16} y^8 z^4 + 256 z^6 + 320 t^2 z^6 + 384 t^4 z^6 + \\
& 352 t^6 z^6 + 160 t^8 z^6 + 272 x^2 z^6 + 256 t^2 x^2 z^6 + \\
& 1120 t^4 x^2 z^6 + 1408 t^6 x^2 z^6 + 784 t^8 x^2 z^6 + \\
& 232 x^4 z^6 + 456 t^2 x^4 z^6 + 2104 t^4 x^4 z^6 + \\
& 2712 t^6 x^4 z^6 + 1856 t^8 x^4 z^6 + 96 x^6 z^6 + 472 t^2 x^6 z^6 + \\
& 2072 t^4 x^6 z^6 + 3208 t^6 x^6 z^6 + 2792 t^8 x^6 z^6 + \\
& 24 x^8 z^6 + 298 t^2 x^8 z^6 + 1178 t^4 x^8 z^6 + 2686 t^6 x^8 z^6 + \\
& 2870 t^8 x^8 z^6 + 108 t^2 x^{10} z^6 + 396 t^4 x^{10} z^6 + \\
& 1668 t^6 x^{10} z^6 + 2020 t^8 x^{10} z^6 + 18 t^2 x^{12} z^6 + \\
& 66 t^4 x^{12} z^6 + 726 t^6 x^{12} z^6 + 934 t^8 x^{12} z^6 + \\
& 192 t^6 x^{14} z^6 + 256 t^8 x^{14} z^6 + 24 t^6 x^{16} z^6 + \\
& 32 t^8 x^{16} z^6 + 704 y^2 z^6 + 1760 t^2 y^2 z^6 + \\
& 1888 t^4 y^2 z^6 + 1312 t^6 y^2 z^6 + 480 t^8 y^2 z^6 + \\
& 1136 x^2 y^2 z^6 + 3456 t^2 x^2 y^2 z^6 + 5152 t^4 x^2 y^2 z^6 + \\
& 5248 t^6 x^2 y^2 z^6 + 2416 t^8 x^2 y^2 z^6 + 1768 x^4 y^2 z^6 + \\
& 5200 t^2 x^4 y^2 z^6 + 9152 t^4 x^4 y^2 z^6 + \\
& 11696 t^6 x^4 y^2 z^6 + 6232 t^8 x^4 y^2 z^6 + 1144 x^6 y^2 z^6 + \\
& 3760 t^2 x^6 y^2 z^6 + 9984 t^4 x^6 y^2 z^6 + \\
& 16720 t^6 x^6 y^2 z^6 + 10120 t^8 x^6 y^2 z^6 + 456 x^8 y^2 z^6 + \\
& 1752 t^2 x^8 y^2 z^6 + 7592 t^4 x^8 y^2 z^6 + \\
& 16024 t^6 x^8 y^2 z^6 + 10880 t^8 x^8 y^2 z^6 + 72 x^{10} y^2 z^6 + \\
& 544 t^2 x^{10} y^2 z^6 + 3952 t^4 x^{10} y^2 z^6 + \\
& 10304 t^6 x^{10} y^2 z^6 + 7848 t^8 x^{10} y^2 z^6 + \\
& 72 t^2 x^{12} y^2 z^6 + 1160 t^4 x^{12} y^2 z^6 + \\
& 4192 t^6 x^{12} y^2 z^6 + 3680 t^8 x^{12} y^2 z^6 + \\
& 128 t^4 x^{14} y^2 z^6 + 952 t^6 x^{14} y^2 z^6 + \\
& 1016 t^8 x^{14} y^2 z^6 + 96 t^6 x^{16} y^2 z^6 + 128 t^8 x^{16} y^2 z^6 + \\
& 384 y^4 z^6 + 1440 t^2 y^4 z^6 + 2208 t^4 y^4 z^6 + \\
& 1632 t^6 y^4 z^6 + 480 t^8 y^4 z^6 + 608 x^2 y^4 z^6 + \\
& 3200 t^2 x^2 y^4 z^6 + 6848 t^4 x^2 y^4 z^6 + 6528 t^6 x^2 y^4 z^6 + \\
& 2272 t^8 x^2 y^4 z^6 + 1760 x^4 y^4 z^6 + 7128 t^2 x^4 y^4 z^6 +
\end{aligned}$$

$$\begin{aligned}
& 15128 t^4 x^4 y^4 z^6 + 16008 t^6 x^4 y^4 z^6 + \\
& 6248 t^8 x^4 y^4 z^6 + 1288 x^6 y^4 z^6 + 6856 t^2 x^6 y^4 z^6 + \\
& 19576 t^4 x^6 y^4 z^6 + 25176 t^6 x^6 y^4 z^6 + \\
& 11168 t^8 x^6 y^4 z^6 + 832 x^8 y^4 z^6 + 4730 t^2 x^8 y^4 z^6 + \\
& 17242 t^4 x^8 y^4 z^6 + 26382 t^6 x^8 y^4 z^6 + \\
& 13230 t^8 x^8 y^4 z^6 + 216 x^{10} y^4 z^6 + 1980 t^2 x^{10} y^4 z^6 + \\
& 10092 t^4 x^{10} y^4 z^6 + 18420 t^6 x^{10} y^4 z^6 + \\
& 10476 t^8 x^{10} y^4 z^6 + 274 t^2 x^{12} y^4 z^6 + \\
& 3186 t^4 x^{12} y^4 z^6 + 7806 t^6 x^{12} y^4 z^6 + \\
& 5278 t^8 x^{12} y^4 z^6 + 384 t^4 x^{14} y^4 z^6 + \\
& 1704 t^6 x^{14} y^4 z^6 + 1512 t^8 x^{14} y^4 z^6 + \\
& 144 t^6 x^{16} y^4 z^6 + 192 t^8 x^{16} y^4 z^6 + 608 x^4 y^6 z^6 + \\
& 2384 t^2 x^4 y^6 z^6 + 3856 t^4 x^4 y^6 z^6 + 2992 t^6 x^4 y^6 z^6 + \\
& 912 t^8 x^4 y^6 z^6 + 496 x^6 y^6 z^6 + 3568 t^2 x^6 y^6 z^6 + \\
& 8848 t^4 x^6 y^6 z^6 + 8976 t^6 x^6 y^6 z^6 + 3200 t^8 x^6 y^6 z^6 + \\
& 752 x^8 y^6 z^6 + 4356 t^2 x^8 y^6 z^6 + 11780 t^4 x^8 y^6 z^6 + \\
& 13596 t^6 x^8 y^6 z^6 + 5420 t^8 x^8 y^6 z^6 + 288 x^{10} y^6 z^6 + \\
& 2552 t^2 x^{10} y^6 z^6 + 8984 t^4 x^{10} y^6 z^6 + \\
& 12232 t^6 x^{10} y^6 z^6 + 5512 t^8 x^{10} y^6 z^6 + \\
& 404 t^2 x^{12} y^6 z^6 + 3156 t^4 x^{12} y^6 z^6 + \\
& 5940 t^6 x^{12} y^6 z^6 + 3252 t^8 x^{12} y^6 z^6 + \\
& 384 t^4 x^{14} y^6 z^6 + 1320 t^6 x^{14} y^6 z^6 + \\
& 1000 t^8 x^{14} y^6 z^6 + 96 t^6 x^{16} y^6 z^6 + 128 t^8 x^{16} y^6 z^6 + \\
& 288 x^8 y^8 z^6 + 1080 t^2 x^8 y^8 z^6 + 1656 t^4 x^8 y^8 z^6 + \\
& 1224 t^6 x^8 y^8 z^6 + 360 t^8 x^8 y^8 z^6 + 144 x^{10} y^8 z^6 + \\
& 1008 t^2 x^{10} y^8 z^6 + 2448 t^4 x^{10} y^8 z^6 + \\
& 2448 t^6 x^{10} y^8 z^6 + 864 t^8 x^{10} y^8 z^6 + \\
& 184 t^2 x^{12} y^8 z^6 + 1064 t^4 x^{12} y^8 z^6 + \\
& 1600 t^6 x^{12} y^8 z^6 + 720 t^8 x^{12} y^8 z^6 + \\
& 128 t^4 x^{14} y^8 z^6 + 376 t^6 x^{14} y^8 z^6 + 248 t^8 x^{14} y^8 z^6 + \\
& 24 t^6 x^{16} y^8 z^6 + 32 t^8 x^{16} y^8 z^6 + 176 z^8 + 256 t^2 z^8 + \\
& 256 t^4 z^8 + 160 t^6 z^8 + 48 t^8 z^8 + 256 x^2 z^8 + \\
& 240 t^2 x^2 z^8 + 544 t^4 x^2 z^8 + 496 t^6 x^2 z^8 + \\
& 192 t^8 x^2 z^8 + 224 x^4 z^8 + 152 t^2 x^4 z^8 + 892 t^4 x^4 z^8 + \\
& 848 t^6 x^4 z^8 + 396 t^8 x^4 z^8 + 96 x^6 z^8 + 32 t^2 x^6 z^8 + \\
& 900 t^4 x^6 z^8 + 840 t^6 x^6 z^8 + 516 t^8 x^6 z^8 + 24 x^8 z^8 + \\
& 8 t^2 x^8 z^8 + 575 t^4 x^8 z^8 + 510 t^6 x^8 z^8 + \\
& 463 t^8 x^8 z^8 + 210 t^4 x^{10} z^8 + 180 t^6 x^{10} z^8 + \\
& 290 t^8 x^{10} z^8 + 35 t^4 x^{12} z^8 + 30 t^6 x^{12} z^8 + \\
& 123 t^8 x^{12} z^8 + 32 t^8 x^{14} z^8 + 4 t^8 x^{16} z^8 + 256 y^2 z^8 + \\
& 704 t^2 y^2 z^8 + 784 t^4 y^2 z^8 + 480 t^6 y^2 z^8 + \\
& 144 t^8 y^2 z^8 + 256 x^2 y^2 z^8 + 1040 t^2 x^2 y^2 z^8 + \\
& 1632 t^4 x^2 y^2 z^8 + 1424 t^6 x^2 y^2 z^8 + 576 t^8 x^2 y^2 z^8 + \\
& 416 x^4 y^2 z^8 + 1560 t^2 x^4 y^2 z^8 + 2696 t^4 x^4 y^2 z^8 + \\
& 2760 t^6 x^4 y^2 z^8 + 1336 t^8 x^4 y^2 z^8 + 224 x^6 y^2 z^8 + \\
& 1032 t^2 x^6 y^2 z^8 + 2616 t^4 x^6 y^2 z^8 + 3416 t^6 x^6 y^2 z^8 + \\
& 1992 t^8 x^6 y^2 z^8 + 96 x^8 y^2 z^8 + 472 t^2 x^8 y^2 z^8 + \\
& 1780 t^4 x^8 y^2 z^8 + 2800 t^6 x^8 y^2 z^8 + 1972 t^8 x^8 y^2 z^8 + \\
& 88 t^2 x^{10} y^2 z^8 + 736 t^4 x^{10} y^2 z^8 + 1432 t^6 x^{10} y^2 z^8 + \\
& 1296 t^8 x^{10} y^2 z^8 + 140 t^4 x^{12} y^2 z^8 + \\
& 400 t^6 x^{12} y^2 z^8 + 548 t^8 x^{12} y^2 z^8 + 40 t^6 x^{14} y^2 z^8 + \\
& 136 t^8 x^{14} y^2 z^8 + 16 t^8 x^{16} y^2 z^8 + 96 y^4 z^8 + \\
& 384 t^2 y^4 z^8 + 624 t^4 y^4 z^8 + 480 t^6 y^4 z^8 + \\
& 144 t^8 y^4 z^8 + 64 x^2 y^4 z^8 + 544 t^2 x^2 y^4 z^8 + \\
& 1472 t^4 x^2 y^4 z^8 + 1568 t^6 x^2 y^4 z^8 + 576 t^8 x^2 y^4 z^8 + \\
& 448 x^4 y^4 z^8 + 1696 t^2 x^4 y^4 z^8 + 3524 t^4 x^4 y^4 z^8 + \\
& 3784 t^6 x^4 y^4 z^8 + 1508 t^8 x^4 y^4 z^8 + 224 x^6 y^4 z^8 +
\end{aligned}$$

$$\begin{aligned}
& 1400 t^2 x^6 y^4 z^8 + 4156 t^4 x^6 y^4 z^8 + 5488 t^6 x^6 y^4 z^8 + \\
& 2508 t^8 x^6 y^4 z^8 + 176 x^8 y^4 z^8 + 992 t^2 x^8 y^4 z^8 + \\
& 3367 t^4 x^8 y^4 z^8 + 5190 t^6 x^8 y^4 z^8 + 2735 t^8 x^8 y^4 z^8 + \\
& 264 t^2 x^{10} y^4 z^8 + 1578 t^4 x^{10} y^4 z^8 + \\
& 3084 t^6 x^{10} y^4 z^8 + 1962 t^8 x^{10} y^4 z^8 + \\
& 315 t^4 x^{12} y^4 z^8 + 998 t^6 x^{12} y^4 z^8 + 875 t^8 x^{12} y^4 z^8 + \\
& 120 t^6 x^{14} y^4 z^8 + 216 t^8 x^{14} y^4 z^8 + 24 t^8 x^{16} y^4 z^8 + \\
& 160 x^4 y^6 z^8 + 672 t^2 x^4 y^6 z^8 + 1144 t^4 x^4 y^6 z^8 + \\
& 912 t^6 x^4 y^6 z^8 + 280 t^8 x^4 y^6 z^8 + 32 x^6 y^6 z^8 + \\
& 656 t^2 x^6 y^6 z^8 + 2056 t^4 x^6 y^6 z^8 + 2272 t^6 x^6 y^6 z^8 + \\
& 840 t^8 x^6 y^6 z^8 + 160 x^8 y^6 z^8 + 880 t^2 x^8 y^6 z^8 + \\
& 2534 t^4 x^8 y^6 z^8 + 3100 t^6 x^8 y^6 z^8 + 1286 t^8 x^8 y^6 z^8 + \\
& 320 t^2 x^{10} y^6 z^8 + 1556 t^4 x^{10} y^6 z^8 + \\
& 2408 t^6 x^{10} y^6 z^8 + 1172 t^8 x^{10} y^6 z^8 + \\
& 350 t^4 x^{12} y^6 z^8 + 916 t^6 x^{12} y^6 z^8 + 598 t^8 x^{12} y^6 z^8 + \\
& 120 t^6 x^{14} y^6 z^8 + 152 t^8 x^{14} y^6 z^8 + 16 t^8 x^{16} y^6 z^8 + \\
& 72 x^8 y^8 z^8 + 288 t^2 x^8 y^8 z^8 + 468 t^4 x^8 y^8 z^8 + \\
& 360 t^6 x^8 y^8 z^8 + 108 t^8 x^8 y^8 z^8 + 144 t^2 x^{10} y^8 z^8 + \\
& 504 t^4 x^{10} y^8 z^8 + 576 t^6 x^{10} y^8 z^8 + 216 t^8 x^{10} y^8 z^8 + \\
& 140 t^4 x^{12} y^8 z^8 + 288 t^6 x^{12} y^8 z^8 + 148 t^8 x^{12} y^8 z^8 + \\
& 40 t^6 x^{14} y^8 z^8 + 40 t^8 x^{14} y^8 z^8 + 4 t^8 x^{16} y^8 z^8
\end{aligned}$$

ACKNOWLEDGEMENTS

We would like to thank N. Curien and D. Henrion for their help in the computation of P_5 . We are deeply grateful to V. Magron for pointing us that P_5 can be explicitly expressed as a Sum-Of-Squares. Thanks are due to L. Lehericy for a careful reading of the manuscript.

REFERENCES

- [AH14] G. Alexandrovich and H. Holzmänn. Nonparametric identification of hidden Markov models. *arXiv preprint arXiv:1404.4210*, 2014.
- [AHK12] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. *arXiv preprint arXiv:1203.0683*, 2012.
- [AMR09] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.
- [BMM12] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.
- [BT13] D. Bontemps and W. Toussile. Clustering and variable selection for categorical multivariate data. *Electron. J. Stat.*, 7:2344–2371, 2013.
- [CC00] L. Couvreur and C. Couvreur. Wavelet based non-parametric HMMs: theory and methods. In *ICASSP '00 Proceedings*, pages 604–607, 2000.
- [DCGLC15] Y. De Castro, E. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *arXiv preprint arXiv:1507.06510*, July 2015.
- [DL93] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303. Springer, 1993.
- [DL12] T. Dumont and S. Le Corff. Nonparametric regression on hidden phi-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure. *arXiv preprint arXiv:1209.0633*, 2012.
- [GCR15] É. Gassiat, A. Cleyne, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Stat. Comput.*, pages 1–11, 2015.
- [GR13] É. Gassiat and J. Rousseau. Non parametric finite translation hidden Markov models and extensions. *Bernoulli*, 2013. To appear.
- [Han06] N. Hansen. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer, 2006.
- [HKZ12] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.*, 78(5):1460–1480, 2012.

- [Lef03] F. Lefèvre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speech and Language*, 17:113–136, 2003.
- [LWM03] M. F. Lambert, J. P. Whiting, and A. V. Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences*, 7(5):652–667, 2003.
- [Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- [Mey92] Y. Meyer. *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1992. Translated from the 1990 French original by D. H. Salinger.
- [Pau13] D. Paulin. Concentration inequalities for Markov chains by Marton couplings. *arXiv preprint arXiv:1212.2015v2*, 2013.
- [SADX14] L. Song, A. Anandkumar, B. Dai, and B. Xie. Nonparametric estimation of multi-view latent variable models. In *ICML*, 2014.
- [SC09] L. Shang and K.P. Chan. Nonparametric discriminant HMM and application to facial expression recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 2090–2096, 2009.
- [VBMMR13] S. Volant, C. Bérard, M.-L. Martin-Magniette, and S. Robin. Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, pages 1–12, 2013.
- [Ver15] É. Vernet. Posterior consistency for nonparametric hidden markov models with finite state space. *Electronic Journal of Statistics*, 9:717–752, 2015.
- [YPRH11] C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(1):37–57, 2011.

(LM-Orsay) LABORATOIRE DE MATHÉMATIQUES D’ORSAY (CNRS UMR 8628), UNIVERSITÉ PARIS-SUD, FACULTÉ DES SCIENCES D’ORSAY, 91405 ORSAY, FRANCE.

E-mail address: yohann.decastro,elisabeth.gassiat,claire.lacour@math.u-psud.fr