Finite state space non parametric Hidden Markov Models are in general identifiable

E. Gassiat, A. Cleynen and S. Robin

July 4, 2013

Abstract

1 Introduction

Finite mixtures are widely used in applications to model data coming from different populations. Let X be the latent random variable whose value is the label of the population the observation comes from, and let Y be the observed random variable. With finitely many populations, X takes values in $\{1, \dots, k\}$ for some fixed integer k, and conditionally to $X = j$, Y has distribution μ_j. Here, $\mu_1, \dots, \mu_k$ are probability distributions on the observation space $\mathcal{Y}$ endowed with its Borel sigma-field and are called emission distributions. Assume that we are given n observations $Y_1, \dots, Y_n$ with the same distribution as Y, that is with distribution

$$\sum_{j=1}^{k} \pi_j \mu_j \tag{1}$$

where $\pi_j = \mathbb{P}(X = j)$, $j = 1, \dots, k$. If the latent variables $X_1, \dots, X_n$ are i.i.d., then so are the observed variables $Y_1, \dots, Y_n$. If the latent variables are not independent, then the observed variables are not either. The dependency structure of the observed variables is given by that of the latent variables.

To be able to infer about the population structures, that is about the emission distributions, one usually states parametric models, saying that the emission distributions belong to some set parametrized by finitely many parameters (for instance Poisson distributions, or Gaussian distributions). Indeed, one may not recover the individual emission distributions from a convex combination of them without further information. Since independence is often a crude approximation of the joint behavior of observed variables, mixture models were generalized to hidden Markov models to be used for clustering purposes. In hidden Markov models (shortened as HMMs in the paper), the latent variables

form a Markov chain. Efficient algorithms allow to compute the likelihood and to build practical inference methods, see [5] for a recent state of the art in HMMs.

But parametric modeling of emission distributions may lead to poor results, in particular for clustering purposes. Recent interest in using nonparametric HMMs appeared in applications, see for instance [6] for voice activity detection, [21] for climate state identification, [15] for automatic speech recognition, [23] for facial expression recognition, [26] for methylation comparison of proteins. These papers propose algorithms to get non parametric estimators and perform classification, but none of them gives theoretical results to support the methods. As noted before, theoretical results in the independent context may only be obtained under further assumptions on the emission distributions. But it has been proved recently by one of the authors ([12]) that, for translations mixtures, that is when the emission distributions are all translated from an unknown one, identifiability holds without any assumption on the translated distribution provided that the latent variables are indeed not independent.

The aim of this paper is to prove that for HMMs, this result may be generalized to any mixtures with Markov regime. As a consequence, consistent estimators of the distribution of the latent variables and of the emission distributions may be built, leading to non parametric classification procedures. In some sense, the message of our paper may be summarized as: *non independence of the observed variables helps if one wants to identify the population structure of the data and to cluster the observations.*

In section 2 we state and prove our main result: non parametric HMMs may be fully identified provided the transition matrix of the hidden Markov chain has full rank, and the emission distributions are linearly independent, see Theorem 1. We then propose various estimation procedures, that should be consistent, thanks to identifiability. In section 3 we show how our identifiability result applies to models used in applications. Finally, in section 4 we present a simulation study mimicking RNA-Seq data and an application to transcriptomic tiling array data.

2 The identifiability result and consequences

2.1 Main theorem

Let $(X_i)_{i\geq 1}$ be a stationary Markov chain on $\{1,\dots,k\}$. Let $(Y_i)_{i\geq 1}$ be a (possibly multidimensional) real valued HMM, that is, a sequence of random variables taking values in $\mathbb{R}^d$ such that, conditionally to $(X_i)_{i\geq 1}$, the Y_i's are independent, and their distribution depends only on the current X_i. If Q is the transition matrix of the Markov chain, and $M = (\mu_1,\dots,\mu_k)$ are k probability distributions on $\mathbb{R}^d$, we denote by $\mathbb{P}_{Q,M}$ the distribution of $(Y_i)_{i\geq 1}$, where $(X_i)_{i\geq 1}$ has transition Q and μ_i is the distribution of Y_1 conditionally to $X_1 = i$, $i = 1,\dots,k$. We call $\mu_1,\dots,\mu_k$ the emission distributions. Notice that in case the Markov chain is irreducible, there exists a unique stationary distribution and $\mathbb{P}_{Q,M}$

is well defined, while in the case where the Markov chain is not irreducible, there might exist several stationary distributions so that the distribution of X_1 has to be specified.

Theorem 1 *Assume k is known, that the probability measures $\mu_1, \ldots, \mu_k$ on $\mathbb{R}^d$ are linearly independent, and that Q has full rank. Then the parameters Q and M are identifiable from $\mathbb{P}^{(3)}_{Q,M}$, that is from the distribution of 3 consecutive observations Y_1, Y_2, Y_3, up to label swapping of the hidden states.*

Let us prove Theorem 1. We have to prove that, if $\tilde{Q}$ is a $k \times k$ transition matrix, and if $\tilde{M} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ are k probability distributions on $\mathbb{R}^d$, if $\mathbb{P}^{(3)}_{\tilde{Q},\tilde{M}} = \mathbb{P}^{(3)}_{Q,M}$, then there exists a permutation σ of the set $\{1, \ldots, k\}$ such that, for all $i, j = 1, \ldots, k$, $\tilde{Q}_{i,j} = Q_{\sigma(i),\sigma(j)}$ and $\tilde{\mu}_i = \mu_{\sigma(i)}$.
Now, since $(X_i)_{i \geq 1}$ is a Markov chain, conditionally to X_2, X_1 and X_3 are independent variables. Then, using this fact, the distribution of (Y_1, Y_2, Y_3) under $\mathbb{P}_{Q,M}$ may be written as

$$\mathbb{P}^{(3)}_{Q,M} = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} \pi_j Q_{j,i} \mu_j \right) \otimes \mu_i \otimes \left(\sum_{j=1}^{k} Q_{i,j} \mu_j \right),$$

where $(\pi_1, \ldots, \pi_k)$ is a stationary distribution of Q which is the distribution of X_1. Similarly,

$$\mathbb{P}^{(3)}_{\tilde{Q},\tilde{M}} = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} \tilde{\pi}_j \tilde{Q}_{j,i} \tilde{\mu}_j \right) \otimes \tilde{\mu}_i \otimes \left(\sum_{j=1}^{k} \tilde{Q}_{i,j} \tilde{\mu}_j \right),$$

where $(\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$ is a stationary distribution of $\tilde{Q}$ which is the distribution of X_1 under $\mathbb{P}_{\tilde{Q},\tilde{M}}$. Since Q has full rank and the probability measures $\mu_1, \ldots, \mu_k$ are linearly independent, the measures $\left(\sum_{j=1}^{k} \pi_j Q_{j,i} \mu_j \right)$, $i = 1, \ldots, k$ are linearly independent, and the probability measures $\left(\sum_{j=1}^{k} Q_{i,j} \mu_j \right)$, $i = 1, \ldots, k$ are also linearly independent. Thus, applying Theorem 9 of [1] we get that there exists a permutation σ of the set $\{1, \ldots, k\}$ such that, for all $i = 1, \ldots, k$:

$$\sum_{j=1}^{k} \tilde{\pi}_j \tilde{Q}_{j,i} \tilde{\mu}_j = \sum_{j=1}^{k} \pi_j Q_{j,\sigma(i)} \mu_j, \ \tilde{\mu}_i = \mu_{\sigma(i)}, \ \sum_{j=1}^{k} \tilde{Q}_{i,j} \tilde{\mu}_j = \sum_{j=1}^{k} Q_{\sigma(i),j} \mu_j.$$

This gives easily, for all $i = 1, \ldots, k$,

$$\sum_{j=1}^{k} \tilde{\pi}_j \tilde{Q}_{j,i} \mu_{\sigma(j)} = \sum_{j=1}^{k} \pi_{\sigma(j)} Q_{\sigma(j),\sigma(i)} \mu_{\sigma(j)}, \ \sum_{j=1}^{k} \tilde{Q}_{i,j} \mu_{\sigma(j)} = \sum_{j=1}^{k} Q_{\sigma(i),\sigma(j)} \mu_{\sigma(j)}.$$

Using now the linear independence of $\mu_1, \ldots, \mu_k$ we get that for all $i, j = 1, \ldots, k$,

$$\tilde{Q}_{j,i} = Q_{\sigma(j),\sigma(i)}, \ \tilde{\pi}_j \tilde{Q}_{j,i} = \pi_{\sigma(j)} Q_{\sigma(j),\sigma(i)}$$

and the theorem is proved.

2.2 Non parametric estimation

Identifiability is the building stone for estimation procedures to lead to consistent estimators. We may now propose several estimation procedures. Let us set the ideas for likelihood based procedures, for which the popular EM algorithm may be used to compute the estimators, as we recall in Section 3.1. Assume that the emission distributions are dominated by a measure ν on $\mathcal{Y}$. Let $\theta = (Q, f_1, \ldots, f_k)$, f_j being the density of μ_j with respect to the dominating measure. Then $(Y_1, \ldots, Y_n)$ has a density $p_{n,\theta}$ with respect to $\nu^{\otimes n}$. Denote $\ell_n(\theta) = \log p_{n,\theta}(Y_1, \ldots, Y_n)$ the log-likelihood, and $\tilde{\ell}_n(\theta) = \sum_{i=1}^{n-2} \log p_{3,\theta}(Y_i, Y_{i+1}, Y_{i+2})$ the pseudo log-likelihood. Likelihood (or pseudo-likelihood) based non parametric estimation usually involves a penalty, which might be chosen as a regularization term (as studied in [25] mainly for independent observations) or as a model selection term (see [19]). More precisely:

- Let $I(f)$ be some functional on the density f. For instance, if $\mathcal{Y}$ is the set of non negative integers, one may take $I(f) = \sum_{j \geq 0} j^\alpha f(j)$ for some $\alpha > 0$; if $\mathcal{Y}$ is the set of real numbers, one may take $I(f) = \int_{-\infty}^{+\infty} [f^{(\alpha)}(u)]^2 du$, where $f^{(\alpha)}$ is the α-th derivative of f. Then the estimator may be chosen as a maximizer of

$$\ell_n(\theta) - \lambda_n[I(f_1) + \ldots + I(f_k)], \tag{2}$$

  or of $\tilde{\ell}_n(\theta) - \lambda_n[I(f_1) + \ldots + I(f_k)]$ for some well chosen positive sequence $(\lambda_n)_{n \geq 1}$. In Section 3.2 we provide an application of this estimator which we further illustrate in Section 4.1.

- If we consider for θ a sequence of models $(\Theta_m)_{m \in \mathcal{M}}$ where Θ_m is the set of possible values for θ for constraint m, one may choose the estimator of m as a maximizer over $\mathcal{M}$ of $\ell_n\left(\widehat{\theta}_m\right) - \text{pen}(n, m)$ (or of $\tilde{\ell}_n\left(\widehat{\theta}_m\right) - \text{pen}(n, m)$), where $\text{pen}(n, m)$ is some penalty term. Here, $\widehat{\theta}_m$ is the maximum likelihood estimator (or the maximum pseudo-likelihood estimator) in model Θ_m for each $m \in \mathcal{M}$. In Section 3.3 we consider for models Θ_m the set of the emission densities which can be modeled as mixture distributions with m components.

We may also consider usual non parametric estimators for emission densities. For instance, in Section 3.4 we consider kernel based estimators computed via maximum likelihood, which we illustrate in Section 4.3.

3 Application to some specific models

In this section we present and discuss a series of hidden Markov models that can be proved to be identifiable thanks to the results above.

3.1 Reminder on the inference of hidden Markov models

A huge variety of techniques have been proposed for the inference of hidden Markov models (see e.g. [5]). The most widely used is probably the E-M algorithm proposed by [7], which can be adapted to several illustrations given below. We recall that this algorithm alternates an expectation (E) step with a maximization (M) step until convergence. At iteration $h+1$, the (M) step retrieves estimates Q^{h+1} and M^{h+1} via the maximization of the conditional expectation

$$\begin{aligned} F^h(Q,M) &= \mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}((Y_i)_{1\le i\le n},(X_i)_{1\le i\le n})|(Y_i)_{1\le i\le n}\right] \\ &= \mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}((X_i)_{1\le i\le n})|(Y_i)_{1\le i\le n}\right] \\ &+ \mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}((Y_i)_{1\le i\le n}|(X_i)_{1\le i\le n})|(Y_i)_{1\le i\le n}\right] \end{aligned} \tag{3}$$

w.r.t. Q and M. This expectation involves the current estimates of the conditional probabilities: $\tau_{ij}^h := \mathbb{P}_{Q^h,M^h}(X_i = j|(Y_i)_{1\le i\le n})$ and $\mathbb{P}_{Q^h,M^h}(X_i = j, X_{i+1} = j'|(Y_i)_{1\le i\le n})$. These conditional probabilities are updated at the next (E) step, using the forward-backward recursion, which takes the current parameter estimates Q^{h+1} and M^{h+1} as inputs. In the sequel, we focus on the estimation of M, the rest of the calculations being standard.

3.2 Non-parametric discrete distributions

We consider a hidden Markov model with discrete observations $(Y_i)_{i\ge1}$ with fully non parametric emission distributions μ_j (denoting $f_j(y) = \mathbb{P}(Y_i = y|X_i = j)$). Theorem 1 ensures that, provided that the distributions μ_j are all linearly independent, the corresponding HMM is identifiable.

Inference. The maximum likelihood inference of this model can be achieved via EM, the M step resulting in

$$f_j^h(y) = S_j^h(y)/N_j^h$$

where $S_j^h(y) = \sum_i \tau_{ij}^h \mathbb{I}(Y_i = y)$ and $N_j^h = \sum_i \tau_{ij}^h$.

Regularization. The EM algorithm can be adapted to the maximization of a penalized likelihood such as (2). Indeed the regularization only affects the (M) step (see [17]). Taking $I(f) = \sum_y m(y)f(y)$ (e.g. $m(y) = y^\alpha$), the estimate of f_j satisfies

$$f_j^h(y) = S_j^h(y) \,/\, \left(\lambda_n m(y) + c_j^h\right)$$

where the constant c_j^h ensures that $\sum_y f_j^h(y) = 1$. Note that this estimate is not explicit but, as $\sum_y f_j^h(y)$ is a monotonous decreasing function of c_j^h, this constant can be efficiently determined using any standard algorithm, such as dichotomy.

RNA-Seq data. In the past few years, next generation sequencing (NGS) technologies have become the state-of-the-art tool for a series of applications in molecular biology such as transcriptome analysis, giving raise to RNA-Seq. Briefly speaking, NGS provide reads that can be aligned along a reference genome, so that a count is associated with each nucleotide. The resulting RNA-Seq count is supposed to reveal the level of transcription of the corresponding nucleotide. HMMs have been proposed ([10, 27]) to determine transcribed regions based on RNA-Seq. The choice of the emission distribution is one of the main issue of such modeling. Poisson distributions display a poor fit to the observed data and the negative binomial has emerged as the the consensus distribution. However, no theoretical justification for such a model exists. Furthermore, the inference of negative binomial models raises several problems, especially for the over-dispersion parameter. The simulation study we perform in Section 4 shows that fully non parametric emission distributions can be used and improve the classification performances.

3.3 Mixtures as emission distributions

Latent variable models with parametric emission distributions often poorly fit the observed data due to the choice of the emission distribution. In the recent years, big efforts have been made to consider more flexible parametric emission distributions (see e.g. [18]). Mixture distribution have recently been proposed to improve flexibility (see [2]). The model is the following: consider a set of m parametric distributions ϕ_ℓ ($\ell = 1 \dots m$) and a $k \times m$ ($m \geq k$) matrix of proportions $\psi = [\psi_{i\ell}]$ such that, for all $j = 1 \dots k$, $\sum_\ell \psi_{j\ell} = 1$. The emission distribution μ_j is defined as

$$\mu_j = \sum_\ell \psi_{j\ell} \phi_\ell. \tag{4}$$

A simple mixture model (i.e. when the hidden variable X_i are iid) with such emission distribution is not identifiable (see [2]). However, its hidden Markov model counterpart is identifiable, under the conditions stated in the following proposition.

Proposition 2 *If the distributions ϕ_ℓ are linearly independent and if the matrix ψ has rank k, then the HMM with emission distribution μ_j defined in* (4) *is identifiable as soon as Q also has full rank.*

Proof. As the distributions ϕ_ℓ are linearly independent, it suffices that the rows of ψ are linearly independent to ensure that so are the distributions μ_j. Identifiability then results from Theorem 1. □

Inference. The maximum likelihood inference of such a model has been studied in [26], although identifiability issues are not theoretically addressed therein. The EM algorithm

can be adapted to this model, considering a second hidden sequence of variables $Z_1, \dots, Z_n$ that are independent conditional on the (X_i) each with multinomial distribution:

$$(Z_i | X_i = j) \sim \mathcal{M}(1; \psi_j)$$

where ψ_j stands for the jth row of ψ. Note that the sequence $Z_1, \dots, Z_n$ is itself a Markov chain, so the conditional probability $\xi^h_{i\ell} := \mathbb{P}(Z_i = \ell | (Y_i)_{i \geq 1})$ can be computed via the forward-backward recursion during the (E) step. See [26].

Mixture of exponential family distributions. In such modeling, the distributions ϕ_ℓ are often chosen within the exponential family, that is

$$\phi_\ell(y) = \exp[\theta_\ell' t(y) - a(y) - b(\theta_\ell)]$$

where $t(y)$ stands for the vector of sufficient statistics, θ_ℓ for the vector of canonical parameters and a and b for the normalizing functions. Standard properties of maximum likelihood estimates in the exponential family yield that the estimates of θ^h_ℓ resulting from the M step must satisfy

$$b'(\theta^h_\ell) = T^h_\ell / N^h_\ell$$

where $T^h_\ell = \sum_i \xi^h_{i\ell} t(Y_i)$ and $N^h_j = \sum_i \xi^h_{i\ell}$. Explicit estimates result from this identity for a series of distribution such as multivariate Gaussian, Poisson, or Binomial. Indeed, Gaussian, Poisson and Binomial $\mathcal{B}(N, p)$ for $N \geq 2m-1$ distributions are linearly independent, as shown in [24].

Convex emission distribution Discrete convex distributions are proved in [11] to be mixtures of triangular discrete distributions. It may be proved, in the same way as in Theorem 8 of [11] that those triangular discrete distributions are in fact linearly independent so that one may use Proposition 2.

Zero-inflated distributions Zero-inflated distributions are mixtures of a Dirac delta distribution δ_0 and a distribution ϕ_j, which is typically chosen from but not limited to the exponential family, so that the emission distribution μ_j can be defined as

$$\mu_j = q_j \delta_0 + (1 - q_j)\phi_j.$$

This model can be expressed as a particular case from that of Equation (4) for which $m = k + 1$ and $\phi_{k+1} = \delta_0$. The matrix ψ is then sparse, with last column $q = (q_1, \dots, q_k)$ and main diagonal $1 - q$. This ensures that provided at most one q_j is equal to one, ψ has full rank. It thus suffices that the ϕ_j are linearly independent to allow the use of Proposition 2, and give support to a vast literature (see [8, 22] for examples of usage of zero-inflated Poisson HMMs to model over-dispersed count datasets).

Non parametric density modeling via mixtures Mixtures, in particular Gaussian mixtures, may be used for a model selection approach for the non parametric estimation of probability densities, see [20]. See also [12] in the HMM context.

3.4 Kernel density estimation

Two major classes of nonparametric density estimators for continuous variables are proposed in the literature in an attempt at capturing the specific shapes of the data where parametric approaches fail: kernel estimates, of which the histogram approach presented in Section 3.2 is a special case, and wavelet-based techniques. We refer to [9] for a complete description of wavelet-estimates properties, or [6] for an example of their use in non-parametric HMMs.

We will focus on kernel-based estimates for the emission densities and for a given bandwith w, we will write $f_j(y)$ of the form

$$f_j(y) = \frac{1}{w} \sum_u \rho_{uj} R\left(\frac{y - y_u}{w}\right)$$

where R is some symmetric kernel function satisfying $\int R = 1$ and where the ρ_{uj} are weights such that, for all u, $\sum_u \rho_{uj} = 1$. A similar estimate was proposed by [13]. We denote $\rho = (\rho_{uj})$ the set of all weights. In this setting, for a given w, the estimation of $(f_1, \ldots, f_k)$ amounts to the estimation of ρ.

Maximum likelihood. An EM algorithm can be used to get maximum likelihood estimates of Q and ρ. We define

$$G^h(\rho) = \mathbb{E}_{Q^h, M^h}\left[\log p_{n,(Q,M)}((Y_i)_{1\le i\le n}|(X_i)_{1\le i\le n})|(Y_i)_{1\le i\le n}\right],$$

which corresponds to the last term of (3) and is the only term to depend on ρ. As for the estimation of ρ, the (M) step aims at maximizing this function that can be rewritten as

$$\begin{aligned} G^h(\rho) &= \sum_{i,j} \tau_{ij}^h \log\left(\frac{1}{w}\sum_u \rho_{uj} R_{iu}\right) \\ &= \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj} \log\left(\rho_{uj} R_{iu}\right) - \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj} \log \gamma_{iuj} - n\log(w) \end{aligned} \tag{5}$$

where $R_{iu} = R((Y_i - Y_u)/w)$ and $\gamma_{iuj} = \rho_{uj} R_{iu} / \sum_v \rho_{vj} R_{iv}$.

Proposition 3 *The following recursion provides a sequence of increasing value of G^h:*

$$\gamma_{iuj}^\ell = \rho_{uj}^\ell R_{iu} / \sum_v \rho_{vj}^\ell R_{iv}, \qquad \rho_{uj}^{\ell+1} = \sum_i \tau_{ij}^h \gamma_{iuj}^\ell \Big/ \sum_{i,v} \tau_{ij}^h \gamma_{ivj}^\ell$$

satisfies $G^h(P^{\ell+1}) \ge G^h(P^\ell)$.

To proove the proposition, we first remark that $\rho^{\ell+1} = (\rho_{uj}^{\ell+1})$ satisfies

$$\rho^{\ell+1} = \arg\max_{\rho} \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj}^\ell \log\left(\rho_{uj} R_{iu}\right), \qquad \text{s.t. } \forall j : \sum_u \rho_{uj} = 1.$$

It follows that

$$\begin{aligned}
0 \;\leq\; & \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj}^\ell \log\left(\rho_{uj}^{\ell+1} R_{iu}\right) - \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj}^\ell \log\left(\rho_{uj}^{\ell} R_{iu}\right) \\
= \; & \sum_{i,u,j} \tau_{ij}^h \gamma_{iuj}^\ell \log \frac{\rho_{uj}^{\ell+1} R_{iu}}{\rho_{uj}^{\ell} R_{iu}} \quad \leq \quad \sum_{i,j} \tau_{ij}^h \log\left(\sum_u \gamma_{iuj}^\ell \frac{\rho_{uj}^{\ell+1} R_{iu}}{\rho_{uj}^{\ell} R_{iu}}\right) \quad \text{(by Jensen's inequality)} \\
= \; & \sum_{i,j} \tau_{ij}^h \log \frac{\sum_u \rho_{uj}^{\ell+1} R_{iu}}{\sum_v \rho_{vj}^{\ell} R_{iv}} \quad = \quad G^h(\rho^{\ell+1}) - G^h(\rho^{\ell})
\end{aligned}$$

which proves the proposition.

Iterating this recursion therefore improves the objective function $F^h(Q, M)$ – even if convergence is not reached –, which results in a Generalized EM algorithm (GEM: [7]).

Another common approach is to replace the terms ρ_{uj} by the posterior probability that the j^{th} individual belongs to class ℓ. This approximation is encountered in the non-parametric HMM literature both in kernel-based approaches (see for instance [14]) and in wavelet-based approaches (see [6]). However, even if this approximation is very intuitive (and much faster computationally), there is no theoretical guarantee that the EM-like algorithm increases the likelihood. In [3] the authors show through simulation studies that it outperforms other approximation algorithm but fail to obtain descent properties. [16] proposes a very similar algorithm, called Majorization-Minimization, which converges to a local maximum of a smoothed likelihood.

4 Simulation and application

4.1 Simulation study

To study the improvement provided by the use of a non-parametric emission distributions, we designed a simulation study based on a typical application in genomics.

RNA-Seq data. Next generation sequencing (NGS) technologies allow to study gene expression all along the genome. NGS data consist of numbers of reads associated with each nucleotide. These read counts are function of the level of transcription of the considered nucleotide, so NGS allow to detect transcribed regions and to evaluate the level of transcription of each region. The state-of-the-art statistical methods are based on the negative binomial distribution.

Design. Based on the annotation of the yeast genome, we defined regions with four level of expression, from intronic (almost no signal) to highly expressed. We then used RNA-Seq data to define empirical count distributions for each of the four levels (so that $k = 4$), which shall correspond to the hidden states. The data were simulated as follows: 14 regions were defined within a sequence of length $n = 4950$ and associated known states and the count at each position within this regions was sampled in the empirical distribution of the corresponding state. $S = 100$ synthetic datasets were sampled according to this scheme and we denote y_i^s the observation from simulation s ($s = 1 \dots S$) at position i ($i = 1 \dots n$).

Evaluation criteria. For each simulation, three HMM models were fitted with, respectively, (a) negative binomial, (b) free non-parametric and (c) regularized emission distributions as defined in Section 3.2, taking

$$I(f) = \sum_y y^2 f(y).$$

For each model, we then inferred the hidden state x_i^s according to both the maximum a posteriori (MAP) rule and the Viterbi most probable path. For each combination of simulation, HMM (a, b or c) and classification rule (MAP or Viterbi), we then computed the rand index between the inferred states ($\hat{x}_i$) and the true one. We recall that the rand index is the proportion of concordant pairs of positions among the $n(n-1)/2$, where the pair (i, i') is said concordant if either $x_i = x_{i'}$ and $\hat{x}_i = \hat{x}_{i'}$, or $x_i \neq x_{i'}$ and $\hat{x}_i \neq \hat{x}_{i'}$.

4.2 Results

MAP and Viterbi classifications achieved very similar performances so we only report the results for Viterbi. Figure 1 displays the rand index for both the parametric (negative binomial) and non-parametric (with no regularization) estimates of the emission distribution for $k = 4$. We observe that, although the mean performances are similar with the two distributions, the parametric negative-binomial sometimes provides poor predictions. The results are very similar for other values of k (not shown).

We then studied the influence on regularization on the performances. We considered a set of values for λ, ranging from 0.25 to 16. Figure 2 shows that regularization can improve the results in a sensible manner. $\lambda = 1$ seems to work best in practice. We do not provide a systematic rule to choose the regularization parameter. Indeed, standard techniques such as cross-validation could be be considered, but would imply an important computational burden.

To illustrate the interest of the non-parametric estimate, we show in Figure 3 the fits obtained with different estimates for a typical simulation. For the regularized version we used $\lambda = 1$ as suggested by the preceeding result. As expected, the unregularized estimate displays the best fit.

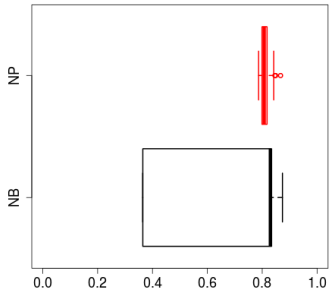

Figure 1: Rand index for the two estimates for $k = 4$: parametric negative binomial (NB: black) and non-parametric (NP: red).

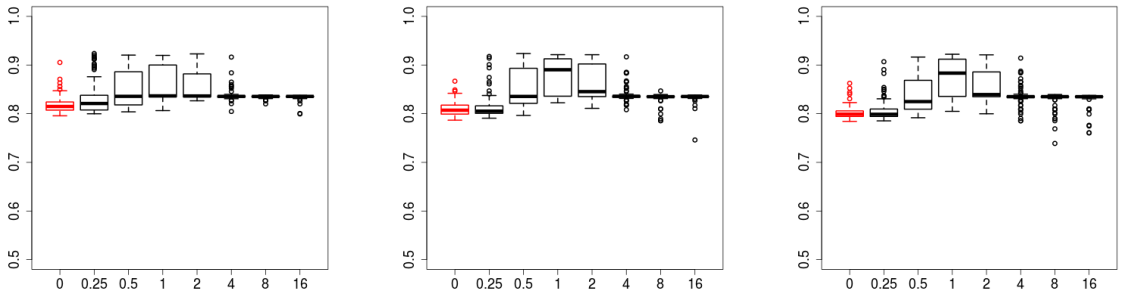

Figure 2: Rand index as a function of the regularization parameter λ. $\lambda = 0$ (in red) corresponds to the non regularized estimate.

4.3 Application to transcriptomic tiling-array

Tiling array. Tiling arrays are a specific microarray technology, where the probes are spread regularly along the genome both in coding and non-coding regions. In transcriptomic applications, tiling arrays capture the intensity of the transcriptional activity at each probe location, thus allowing the detection of transcribed regions. We consider here a comparative experiment were two organs (seed and leaf) of the model plant *A. thaliana* are compared. The data under study corresponds to probes located on chromosome 4. The top left panel of Figure 4 is an idealization of the expected result. Indeed, we expect to find probes being expressed in none of the organs (blue region), probes being expressed in both organs with equal level (black region) and probes being more expressed in one organ than the other (red and green regions). The corresponding region were drawn arbitrarily of the plot to illustrate these four behaviors.

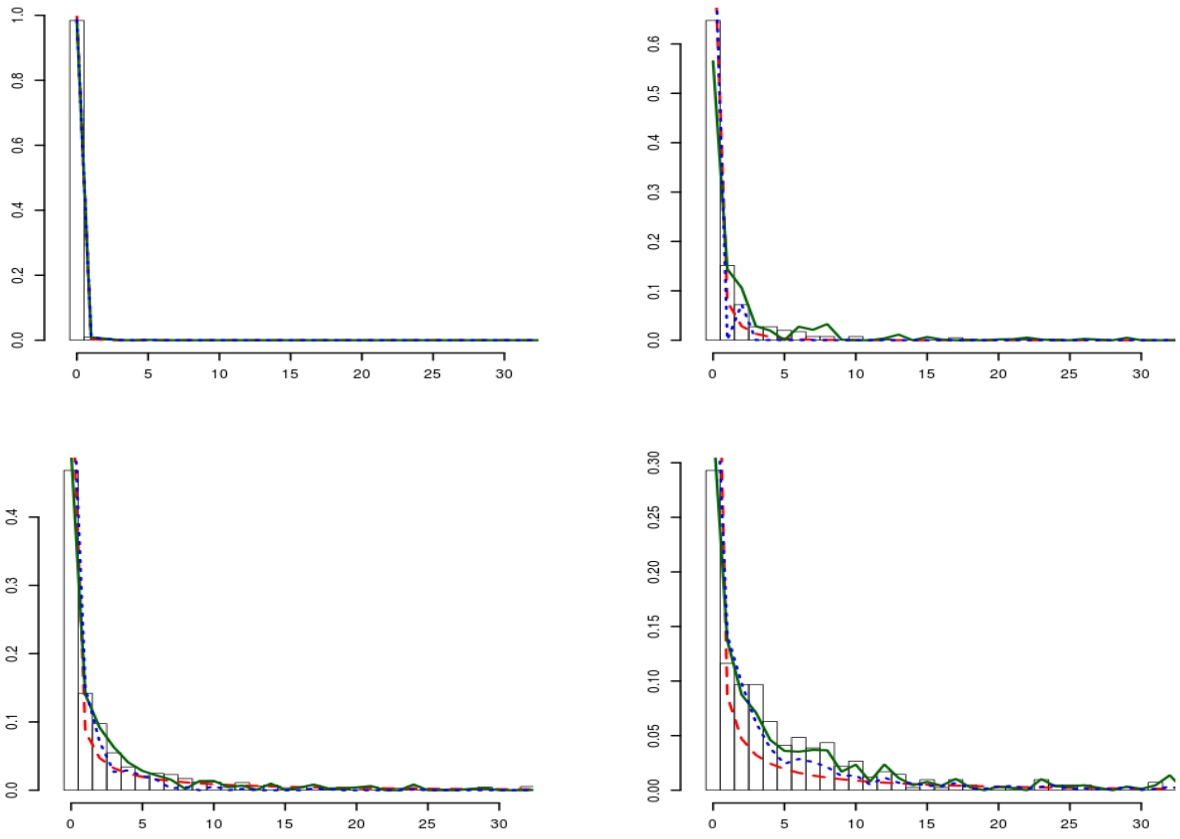

Figure 3: Fit of the estimated distributions with the three estimates: negative binomial (NB: dashed red), non-parametric (NP: solid green) and regularized non-parametric ($\lambda = 1$, rNP: dotted blue).

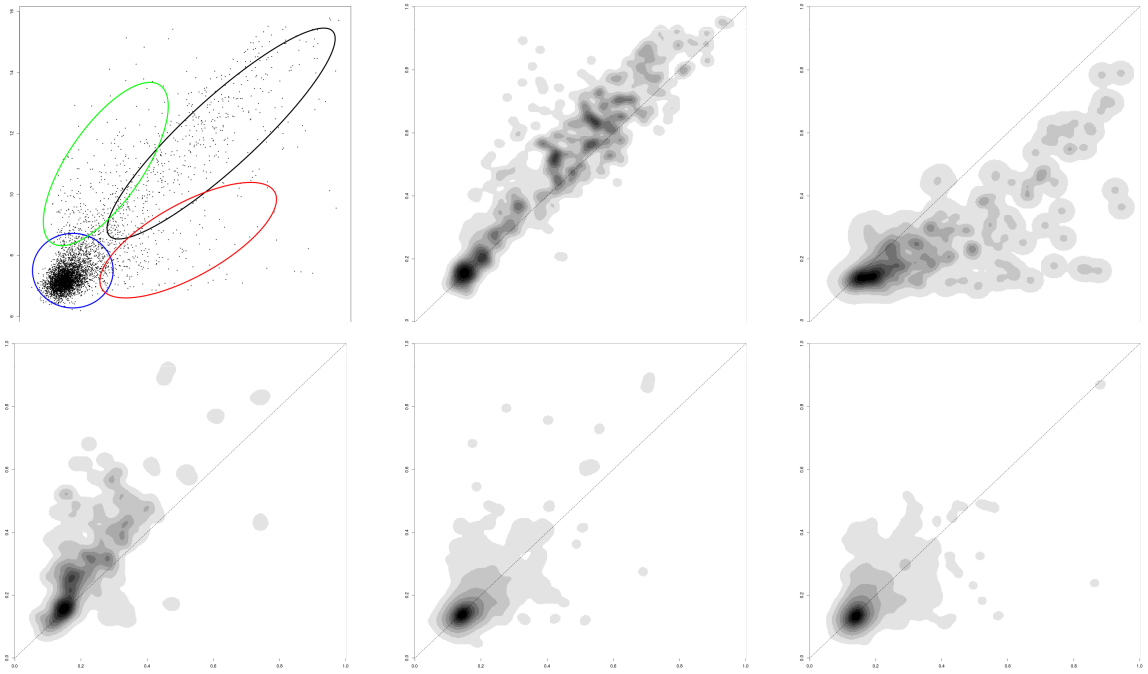

Figure 4: Top left panel: Raw tiling array data from chromosome 4 + idealized groups. Other panels: contour plots of the kernel estimate of each emission distribution for the 5-state non-parametric HMM. The idealized blue group is split into two HMM states (bottom center and bottom right).

Mixture as emission distributions. The same data has already been analyzed in [26] and [4], using two different kinds of mixture as emission distributions. The former proposed a very problem oriented mixture of elliptic Gaussian distribution, whereas the latter was a generalization of the approach of [2] to hidden Markov models. A consequence of Proposition 2 given above is that both of these models are identifiable.

Non-parametric HMM. Here, we fitted a k-state non-parametric HMM to these data using the kernel method described in Section 3.4. We used a spherical Gaussian kernel for which we first estimated the bandwidth w via cross-validation on (probes expressed in neither organs) into two, resulting in the two bottom left figures in Figure 4. This figure provides the kernel density estimates of the emission distributions under this model. The shape of these distribution turn out to be far from what could be captured by some standard parametric distribution (e.g. 2-dimensional Gaussian). Note that in [26] (see their Fig. 5) $k = 8$ components were needed to recover the expected 4 groups using Gaussian mixture as emission distributions.

5 Conclusion

In this article, we have showed that non-parametric hidden Markov models are identifiable up to state-label switching provided that the transition matrix has full rank and that the emission distributions are linearly independent. This gives support to numerous methods that had previously been proposed for the classification of data using non-parametric HMMs. While they usually proved excellent empirical results, no guarantees on the identifiability of the models had yet been given. We describe multiple examples of procedures for which our result applies, and illustrate the gain provided by the use of a non-parametric emission distribution in two applications. In the first one, we present a simulation study inspired from RNA-Seq experiments. In this context, the addition of a regularization function improves the performances of the non-parametric HMM classification. In the second example, we present the application of kernel based estimation of emission densities to apply on transcriptomic tiling array data. Again, non parametric estimation improves the classification performances. This motivates future work on the choices that are involved in non parametric procedures: selection of the regularizing sequence λ_n in regularized maximum likelihood, proposition of a penalty function for the choice of the number of states, choice of mixture components modeling for the emission distribution, choice of the kernel R in a kernel based maximum likelihood estimation, and choice of the bandwidth w.

Acknowledgments: the authors want to thank Caroline Bérard for providing the transcriptomic tiling array data.

References

[1] Elisabeth S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37:3099–3132, 2009.

[2] Jean-Patrick Baudry, Adrian E Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

[3] Tatiana Benaglia, Didier Chauveau, and David R Hunter. An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[4] C. Bérard, M. L. Martin-Magniette, V. Brunaud, S. Aubourg, and S. Robin. Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome. *Stat Appl Genet Mol Biol*, 10(1), 2011.

[5] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, New York, 2005.

[6] L. Couvreur and C. Couvreur. Wavelet based non-parametric HMMs: theory and methods. In *ICASSP '00 Proceedings*, pages 604–607, 2000.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Society Series B*, 39:1–38, 1977.

[8] Stacia M DeSantis and Dipankar Bandyopadhyay. Hidden Markov models for zero-inflated Poisson counts with an application to substance use. *Statistics in medicine*, 30(14):1678–1694, 2011.

[9] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

[10] Jiang Du, Joel S Rozowsky, Jan O Korbel, Zhengdong D Zhang, Thomas E Royce, Martin H Schultz, Michael Snyder, and Mark Gerstein. A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, 22(24):3016–3024, 2006.

[11] Cécile Durot, Sylvie Huet, François Koladjo, and Stéphane Robin. Least-squares estimation of a convex discrete distribution. *Computational Statistics & Data Analysis*, 2013.

[12] E. Gassiat and J. Rousseau. Non parametric finite translation mixtures with dependent regime. Technical report, 2013.

[13] P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, 31:201–224, 2003.

[14] Ning Jin and Farzin Mokhtarian. A non-parametric HMM learning method for shape dynamics with application to human motion recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 29–32. IEEE, 2006.

[15] F. Lefèvre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speach and Language*, 17:113–136, 2003.

[16] Michael Levine, David R Hunter, and Didier Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416, 2011.

[17] Haifeng Li, Keshu Zhang, and Tao Jiang. The regularized EM algorithm. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 807. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[18] Tsung I. Lin, Jack C. Lee, and Shu Y. Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17(3):909–27, 2007.

[19] Pascal 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.

[20] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68, 2011.

[21] J.P. Whiting M.F. Lambert and A.V. Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences*, 7 (5):652–667, 2003.

[22] Madalina Olteanu, James Ridgway, et al. Hidden Markov models for time series of counts with excess zeros. *Proceedings of ESANN 2012*, pages 133–138, 2012.

[23] 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.

[24] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1985.

[25] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[26] Stevenn Volant, Caroline Bérard, Marie-Laure Martin-Magniette, and Stéphane Robin. Hidden Markov Models with mixtures as emission distributions. *Statistics and Computing*, pages 1–12, 2013.

[27] Zhiyuan Zhai, Shih-Yen Ku, Yihui Luan, Gesine Reinert, Michael S Waterman, and Fengzhu Sun. The power of detecting enriched patterns: An HMM approach. *Journal of Computational Biology*, 17(4):581–592, 2010.