Stat Comput (2016) 26:61–71
DOI 10.1007/s11222-014-9523-8

Inference in finite state space non parametric Hidden Markov Models and applications

E. Gassiat · A. Cleynen · S. Robin

Received: 28 January 2014 / Accepted: 4 October 2014 / Published online: 6 January 2015

Abstract Hidden Markov models (HMMs) are intensively used in various fields to model and classify data observed along a line (e.g. time). The fit of such models strongly relies on the choice of emission distributions that are most often chosen among some parametric family. In this paper, we prove that finite state space non parametric HMMs are identifiable as soon as the transition matrix of the latent Markov chain has full rank and the emission probability distributions are linearly independent. This general result allows the use of semi- or non-parametric emission distributions. Based on this result we present a series of classification problems that can be tackled out of the strict parametric framework. We derive the corresponding inference algorithms. We also illustrate their use on few biological examples, showing that they may improve the classification performances.

Keywords Identifiability · Hidden Markov Models · Non-parametric

1 Introduction

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 $\{1, \ldots, k\}$ for some fixed integer k, and conditionally on $X = j$, Y has distribution μ_j. Here, $\mu_1, \ldots, \mu_k$ are probability distributions on the observation space $\mathcal{Y}$ and are called emission distributions. Assume that we are given n observations $Y_1, \ldots, Y_n$ with the same distribution as Y, that is with distribution

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

where $\pi_j = \mathbb{P}(X = j)$, $j = 1, \ldots, k$. In mixture models the latent variables $X_1, \ldots, X_n$ are i.i.d., and so are the observed variables $Y_1, \ldots, Y_n$. When the observed data are organized along a line (e.g. along time), independence is often a crude approximation of their joint behavior and hidden Markov models (shortened as HMMs in the paper) are often preferred to independent mixtures for clustering purposes. In HMMs, the latent variables form a Markov chain. As the latent variables are not independent, the observed variables are not either.

In both models, the dependency structure of the observed variables is given by that of the latent variables. Efficient algorithms allow to compute the likelihood and to build practical inference methods for both mixture and HMMs, see e.g. Cappé et al. (2005) for a recent state of the art in HMMs.

To be able to infer about the population structures, one usually states parametric models, saying that the emission distributions belong to some set parametrized by finitely many parameters (for instance Poisson or Gaussian distributions). But parametric modeling of emission distributions may lead to poor results, in particular for clustering purposes.

E. Gassiat
Laboratoire de Mathématique, Université Paris-Sud, Orsay, France
e-mail: Elisabeth.Gassiat@math.u-psud.fr

E. Gassiat
Laboratoire de Mathématique, CNRS, Orsay, France

A. Cleynen · S. Robin (✉)
AgroParisTech, MIA 518, Paris, France
e-mail: Stephane.Robin@agroparistech.fr; robin@agroparistech.fr

A. Cleynen · S. Robin
INRA, MIA 518, Paris, France

However, one may not recover the individual emission distributions from a convex combination of them without further information (see e.g. Bordes et al. 2006a, b or Butucea and Vandekerkhove 2014). Hence, independent non parametric mixture models are not identifiable in general.

Recent interest in non parametric HMMs appeared in applications, see for instance Couvreur and Couvreur (2000) for voice activity detection, Lambert et al. (2003) for climate state identification, Lefèvre (2003) for automatic speech recognition, Shang and Chan (2009) for facial expression recognition, or Volant et al. (2013) for the classification of methylation regions in genomics. These papers propose algorithms to get non parametric estimators and perform classification, but none of them gives theoretical results to support the methods. We emphasize that HMMs with finite space for the emission distributions are fully parametric as the emission distribution are multinomial. Nevertheless, this setting has also recently received some attention and identifiability results have been proved (see e.g. Hsu et al. 2012 or An et al. 2013). Interestingly, it can be noted that these results rely on the assumption that the transition matrix has full rank. We make the same assumption in our main result. See also Tune et al. (2013).

It has been proved recently by Gassiat and Rousseau (2014) 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. In this paper, we prove that this result may be generalized to any non parametric HMMs with finite state space. The underlying idea is again that non independence of the observed variables helps if one wants to identify the population structure of the data and to cluster the observations. See also Dumont and Le Corff (2012). The proof of our result mainly relies on the fact that the joint distribution of three consecutive variables may be written as a mixture of distributions of independent variables, where the distributions are linearly independent signed measures. Identifiability of non parametric mixtures of two multivariate distributions of vectors of independent random variables has been investigated in a seminal paper by Hall and Zhou (2003) and further extended by Allman et al. (2009). Though non independence of the hidden sequence is obviously necessary to obtain identifiability results in the nonparametric finite mixture context, it is not completely understood what is precisely needed. Here and in Gassiat and Rousseau (2014) it is assumed that the matrix of the distribution of two consecutive hidden variables has full rank, but the proofs are quite different.

An important consequence of the identifiability result is that consistent estimators of the distribution of the latent variables and of the emission distributions may be built, leading to non parametric classification procedures.

In Sect. 2 we prove that non parametric HMMs may be fully identified provided that the transition matrix of the hidden Markov chain has full rank, and that the emission distributions are linearly independent (see Theorem 1). We then present and discuss various likelihood-based estimation procedures, and explain briefly how the obtained estimators can be proved to be consistent thanks to the identifiability of HMMs. In Sect. 3 we show how this result applies to models used in applications. Finally, in Sect. 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 $\{1,\ldots,k\}$. 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, if $\pi = (\pi_1,\ldots,\pi_k)$ is a stationary distribution of Q and if $M = (\mu_1,\ldots,\mu_k)$ are k probability distributions on $\mathbb{R}^d$, we denote by $\mathbb{P}_{Q,M,\pi}$ the distribution of $(Y_i)_{i\geq 1}$, where $(X_i)_{i\geq 1}$ has transition Q, X_1 has distribution π and μ_i is the distribution of Y_1 conditionally to $X_1 = i$, $i = 1,\ldots,k$. We call $\mu_1,\ldots,\mu_k$ the emission distributions. Notice that in case the Markov chain is irreducible, there exists a unique stationary distribution π such that for all $i = 1,\ldots,k$, $\pi_i > 0$, 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. In such a case, we assume that for all $i = 1,\ldots,k$, $\pi_i > 0$. if it was not the case, one could reduce the number of hidden states.
For any integer $n \geq 1$, denote by $\mathbb{P}^{(n)}_{Q,M}$ the distribution of the random vector $Y_{1:n} := (Y_1,\ldots,Y_n)$ under $\mathbb{P}_{Q,M}$. We have:

$$\mathbb{P}^{(n)}_{Q,M} = \sum_{i_1,\ldots,i_n=1}^{k} \pi_{i_1} Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n}\mu_{i_1}\otimes\cdots\otimes\mu_{i_n},$$

and if the emission distributions $\mu_1,\ldots,\mu_k$ have densities $f_1,\ldots,f_k$ with respect to some dominating measure ν on $\mathcal{Y}$, then $\mathbb{P}^{(n)}_{Q,M}$ has a density $p_{n,Q,M}$ with respect to $\nu^{\otimes n}$ given by

$$p_{n,Q,M}(y_{1:n}) = \sum_{i_1,\ldots,i_n=1}^{k} \pi_{i_1} Q_{i_1,i_2}\cdots Q_{i_{n-1},i_n} f_{i_1}(y_1)\cdots f_{i_n}(y_n).$$

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 the distribution of* 3 *consecutive observations Y_1,*

Y_2, Y_3, up to label swapping of the hidden states, that is: if $\tilde{Q}$ is a $k \times k$ transition matrix, if $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$ is a stationary distribution of $\tilde{Q}$ such that for all $i = 1, \ldots, k$, $\tilde{\pi}_i > 0$ and if $\tilde{M} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ are k probability distributions on $\mathbb{R}^d$ that verify $\mathbb{P}^{(3)}_{\tilde{Q},\tilde{M}} = \mathbb{P}^{(3)}_{Q,M}$, 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 $\tilde{\mu}_i = \mu_{\sigma(i)}$.

In this statement, measures are supposed linearly independent as elements of the vector space of signed measures. It can be noted that the full rank condition on the transition matrix is classical and already appears in parametric settings such as that of Petrie (1969). In the particular case where $k = 2$, this assumption is equivalent to the non-independence of the variables Y. The linear independence condition on emission distributions is different than that formally stated, in the mixture context, by Yakowitz and Spragins (1968) where the whole set of possible emission distributions has to be linearly independent.

Let us now prove Theorem 1. Let $\tilde{Q}$ be a $k \times k$ transition matrix, $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$ be a stationary distribution of $\tilde{Q}$ such that for all $i = 1, \ldots, k, \tilde{\pi}_i > 0$ and $\tilde{M} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ be k probability distributions on $\mathbb{R}^d$, such that $\mathbb{P}^{(3)}_{\tilde{Q},\tilde{M},\tilde{\pi}} = \mathbb{P}^{(3)}_{Q,M,\pi}$.

The distribution of (Y_1, Y_2, Y_3) under $\mathbb{P}_{Q,M}$ may be written as

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

Here, the fact that $\mathbb{P}^{(3)}_{Q,M}$ factorizes as a sum of k tensorial products of three positive measures will be essential. Notice that this comes from the fact that $(X_i)_{i \geq 1}$ is a Markov chain and that conditionally on X_2, X_1 and X_3 are independent variables. It may be seen from the rewriting

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

where we have used that $\pi_i > 0$ for all $i = 1, \ldots, k$. Similarly,

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

Since Q has full rank and the probability measures $\mu_1, \ldots, \mu_k$ are linearly independent, the probability measures $\left(\sum_{j=1}^{k} \pi_j Q_{j,i} \mu_j / \pi_i \right)$, $i = 1, \ldots, k$ 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 Allman et al. (2009) we get that there exists a permutation σ of the set $\{1, \ldots, k\}$ such that, for all $i = 1, \ldots, k$:

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

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

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

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

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

and the theorem is proved.

An alternative proof of Theorem 1 could be given using the parametric identification result in Hsu et al. (2012) combined with an adequate discretization argument.

2.2 Non parametric estimation

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 Sect. 3.1. Assume that the set of possible emission distributions is 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 van de Geer 2000 mainly for independent observations) or as a model selection term (see Massart 2007). 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

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

  or of $\tilde{\ell}_n(\theta) - \lambda_n[I(f_1) + \cdots + I(f_k)]$ for some well chosen positive sequence $(\lambda_n)_{n \geq 1}$. In Sect. 3.2 we provide an application of this estimator which we further illustrate in Sect. 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 Sect. 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 Sect. 3.4 we consider kernel based estimators computed via maximum likelihood, which we illustrate in Sect. 4.2.

Identifiability is the building stone to prove the consistency of estimation procedures. For pseudo-likelihood based non-parametric estimators, it is likely provable that the Hellinger distance between $p_{3,\theta^\star}$ and $p_{3,\widehat{\theta}}$ converges in probability to 0, where $\theta^\star$ denotes the true value of the parameter and $\widehat{\theta}$ its penalized pseudo-likelihood estimate. One could use similar tricks as in Dumont and Le Corff (2012) when a regularization term is used as penalty, or Gassiat and Rousseau (2014) when a model selection penalty is used. The main points are to get deviation inequalities for additive functions of the observed variables, and to control the complexity of the space where the density of the observations is supposed to live (by the regularization penalty or the model selection penalty). This usually allows to obtain that the plugged-estimated density stays in a compact set. Coupled with the identifiability result, consistency of such $\widehat{\theta}$ would follow.

For full-likelihood based non parametric estimators, consistency might be difficult to prove. Indeed, it requires a deep understanding of the asymptotic behavior of the full likelihood not only as a point function but as a process on the infinite dimensional space parameters.

Bayesian estimation procedures can also be proposed, for which, as is usual in such Bayesian situations, the choice of the prior is of great importance. Based on our identifiability result, Vernet (2013) has obtained posterior consistency results for Bayesian estimation in this context. One could also propose histogram-based estimation procedures using Hsu et al. (2012). Here the question would be to choose adequately the partition of the observation space as a function of the sample size.

3 Application to some specific models

In this section we present and discuss a series of HMMs 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 HMMs (see e.g. Cappé et al. 2005). The most widely used is probably the EM algorithm proposed by Dempster et al. (1977), 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

$$\begin{aligned} F^h(Q, M) &= \mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}(Y_{1:n}, X_{1:n})|Y_{1:n}\right] \\ &= \mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}(X_{1:n})|Y_{1:n}\right] \\ &\quad +\mathbb{E}_{Q^h,M^h}\left[\log p_{n,(Q,M)}(Y_{1:n}|X_{1:n})|Y_{1:n}\right] \end{aligned} \quad (3)$$

w.r.t. Q and M. This expectation involves the current estimates of the conditional probabilities: $\tau_{ij}^h := \mathbb{P}_{Q^h,M^h}(X_i = j|Y_{1:n})$ and $\mathbb{P}_{Q^h,M^h}(X_i = j, X_{i+1} = j'|Y_{1:n})$. These conditional probabilities are updated at the next (E) step, using the forward-backward recursion, which takes the current parameter estimates Q^h and M^h 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 HMM with discrete observations $(Y_i)_{i\geq 1}$ 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 and that the transition matrix has full rank, 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 Li et al. 2005). Taking $I(f) = \sum_y m(y)f(y)$ (e.g. $m(y) = y^\alpha$), the estimate of f_j satisfies

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

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 (Du et al. 2006; Zhai et al. 2010) 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 consensus distribution. However, only empirical arguments exist to motivate the use of the negative binomial for RNA-Seq data. Furthermore, the inference of negative binomial models raises several problems, especially for the over-dispersion parameter. The simulation study we perform in Sect. 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. Lin et al. 2007). Mixture distribution have recently been proposed to improve flexibility (see Baudry et al. 2010). 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 Baudry et al. 2010). However, its HMM 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.*

The proof is the following. 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 Volant et al. (2013), 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 hidden Markov chain, and the conditional probability $\xi_{i\ell} := \mathbb{P}(Z_i = \ell|(Y_j)_{j\geq 1})$ can be computed via the forward-backward recursion during the (E) step (see Volant et al. 2013).

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_ℓ^h resulting from the M step must satisfy

$$b'(\theta_\ell^h) = T_\ell^h / N_\ell^h$$

where $T_\ell^h = \sum_i \xi_{i\ell}^h t(Y_i)$ and $N_j^h = \sum_i \xi_{i\ell}^h$. 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 recalled in Titterington et al. (1985).

Convex emission distribution Discrete convex distributions are proved in Durot et al. (2013) to be mixtures of triangular discrete distributions. It may be proved, in the same way as in Theorem 8 of Durot et al. (2013) 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 Eq. (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 $\{\phi_j, 1 \leq j \leq k + 1\}$ are linearly independent to allow the use of Proposition 2, and give support to a vast literature (see DeSantis and Bandyopadhyay 2011; Olteanu and Ridgway

2012 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 Maugis and Michel (2011). See also Gassiat and Rousseau (2014) 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 Sect. 3.2 is a special case, and wavelet-based techniques. We refer to Donoho et al. (1996) for a complete description of wavelet-estimates properties, or Couvreur and Couvreur (2000) 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 bandwidth 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 j, $\sum_u \rho_{uj} = 1$. A similar estimate was proposed by Hall and Zhou (2003). 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

$$G^h(\rho) = \mathbb{E}_{Q^h, M^h}\left[\log p_{n,(Q,M)}(Y_{1:n}|X_{1:n})|Y_{1:n}\right],$$

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

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

where $R_{iu} = R((Y_i - Y_u)/w)$ and $\gamma_{iuj} = \rho_{uj}R_{iu}/\sum_v \rho_{vj}R_{iv}$, and the summations on i, u and v range from 1 to n while that on j ranges from 1 to k. We remind that the τ^h_{ij} are the current estimate of the conditional probability: $\mathbb{P}_{Q^h,M^h}(X_i = j|Y_{1:n})$.

Proposition 3 *The following recursion provides a sequence of increasing values of $G^h(\rho)$:*

$$\gamma^\ell_{iuj} = \rho^\ell_{uj}R_{iu}/\sum_v \rho^\ell_{vj}R_{iv},$$

$$\rho^{\ell+1}_{uj} = \sum_i \tau^h_{ij}\gamma^\ell_{iuj} \Big/ \sum_{i,v}\tau^h_{ij}\gamma^\ell_{ivj},$$

that is $G^h(\rho^{\ell+1}) \geq G^h(\rho^\ell)$.

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

$$\rho^{\ell+1} = \arg\max_\rho \sum_{i,u,j}\tau^h_{ij}\gamma^\ell_{iuj}\log\left(\rho_{uj}R_{iu}\right),$$

$$\text{s.t. } \forall j : \sum_u \rho_{uj} = 1.$$

It follows that

$$\begin{aligned} 0 &\leq \sum_{i,u,j}\tau^h_{ij}\gamma^\ell_{iuj}\log\left(\rho^{\ell+1}_{uj}R_{iu}\right) - \sum_{i,u,j}\tau^h_{ij}\gamma^\ell_{iuj}\log\left(\rho^\ell_{uj}R_{iu}\right) \\ &= \sum_{i,u,j}\tau^h_{ij}\gamma^\ell_{iuj}\log\frac{\rho^{\ell+1}_{uj}R_{iu}}{\rho^\ell_{uj}R_{iu}} \quad \leq \quad \sum_{i,j}\tau^h_{ij}\log\left(\sum_u \gamma^\ell_{iuj}\frac{\rho^{\ell+1}_{uj}R_{iu}}{\rho^\ell_{uj}R_{iu}}\right) \\ &= \sum_{i,j}\tau^h_{ij}\log\frac{\sum_u \rho^{\ell+1}_{uj}R_{iu}}{\sum_v \rho^\ell_{vj}R_{iv}} \quad = \quad G^h(\rho^{\ell+1}) - G^h(\rho^\ell) \end{aligned}$$

(where the second upper bounding results from Jensen's inequality) 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: Dempster et al. 1977).

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 Jin and Mokhtarian 2006) and in wavelet-based approaches (see Couvreur and Couvreur 2000). 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 Benaglia et al. (2009) the authors show through simulation studies that it outperforms other approximation algorithm but fail to obtain descent properties. Levine et al. (2011) proposes a very similar algorithm, based on the Majorization-Minimization principle, 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 $n = 4950$, defining a (known) sequence $X_{i:n}$ taking values in $1 : 4$ which is fixed in the simulation study. The count at each position 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 $(i = 1, \dots, n)$.

Model For each simulation, three HMM models were fitted with the observed read-counts at each location Y_i taking values in $\mathcal{Y} = \mathbb{N}$ the set of non-negative integers, and the hidden states X_i representing the expression level of the position. The emission distributions considered were respectively:

(a) negative binomial with a state-specific probability parameter,
(b) free non-parametric and
(c) regularized non-parametric as defined in Sect. 3.2, taking

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

Evaluation criteria 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 $\hat{X}_i \neq \hat{X}_{i'}$.

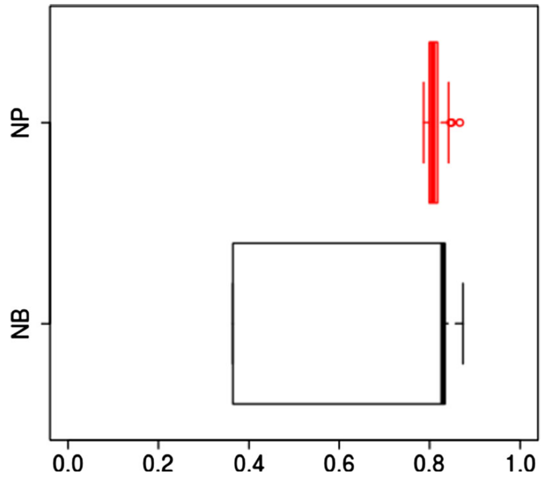

Fig. 1 Rand index for the two estimates for $k = 4$: parametric negative binomial (NB: *black*) and non-parametric (NP: *red*). (Color figure online)

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 Fig. 3 the fits obtained with different estimates for a typical simulation. For the regularized version we used $\lambda = 1$ as suggested by the preceding result. As expected, the unregularized non-parametric estimate (b) displays a better fit than the parametric estimate (a), the regularized non-parametric version (c) lying between the two.

4.2 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 tran-

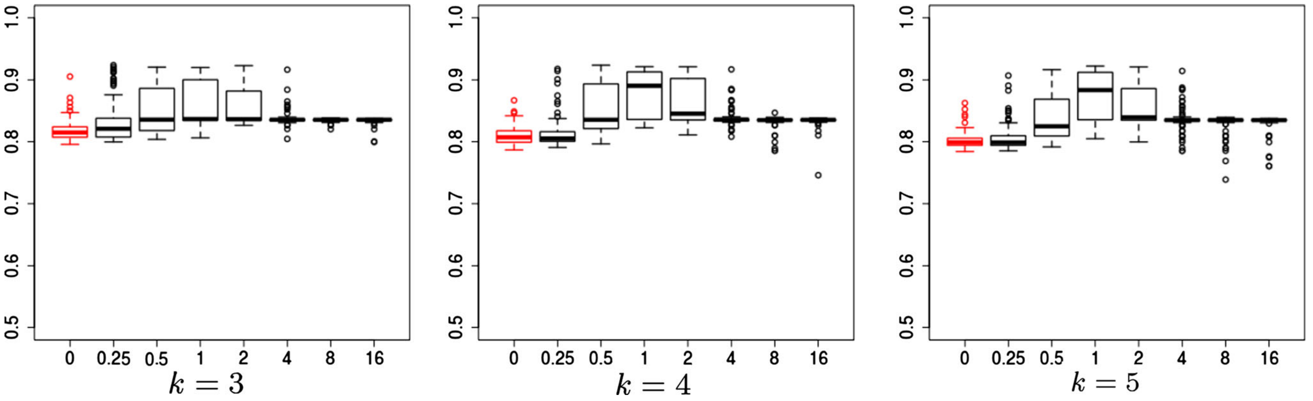

Fig. 2 Rand index as a function of the regularization parameter λ. λ = 0 (in *red*) corresponds to the non regularized estimate. (Color figure online)

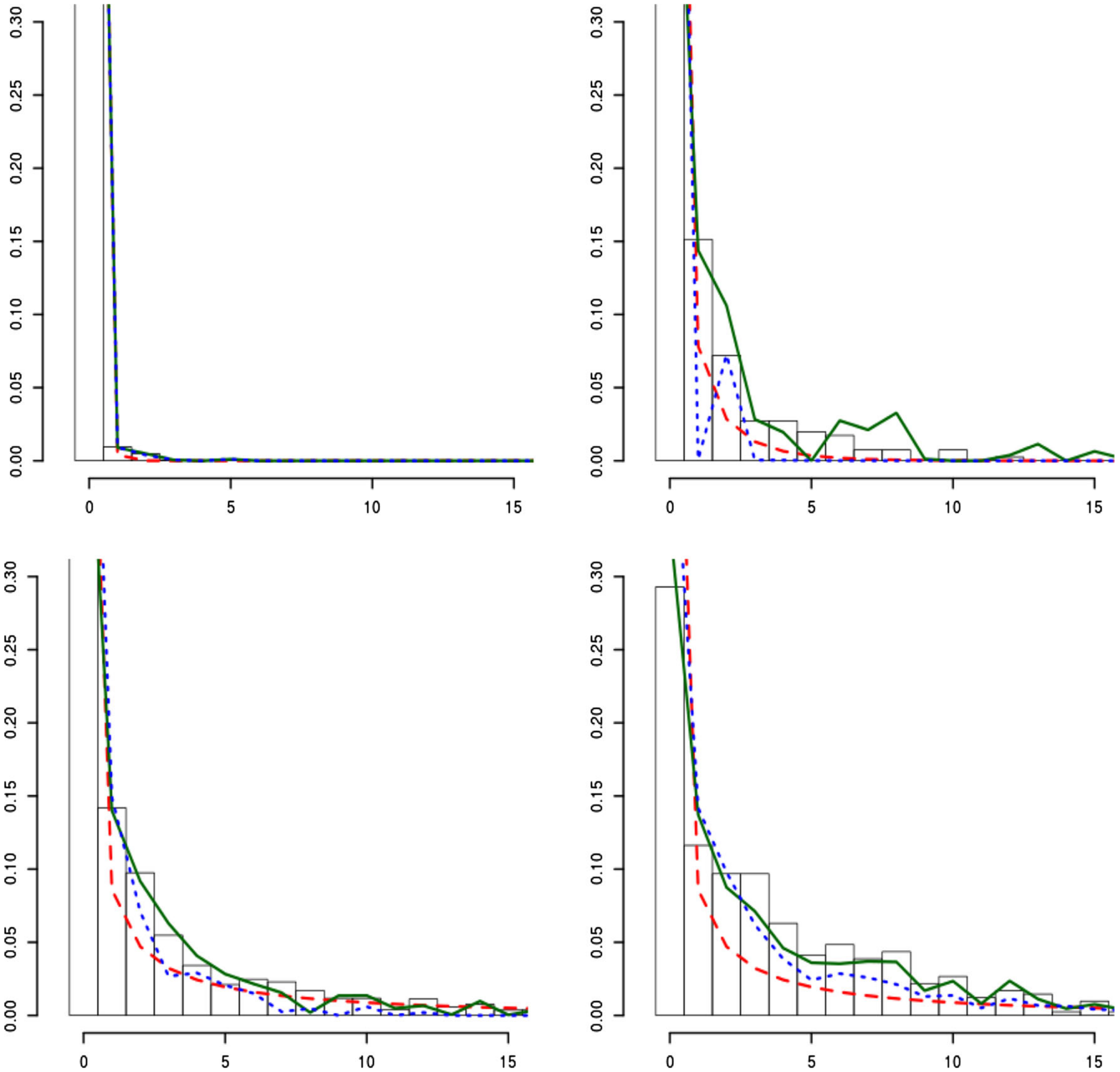

Fig. 3 Fit of the estimated distributions with the three estimates: *bars* represent the empirical distribution from the RNA-Seq data while lines show the estimates from the negative binomial (NB: *dashed red*), non-parametric (NP: *solid green*) and regularized non-parametric (λ = 1, rNP: *dotted blue*). *Top left*: intronic regions (almost no expression), *top right*: weakly expressed regions, *bottom left*: expressed regions, *bottom right*: highly expressed regions. The x-axis has been truncated for legibility. (Color figure online)

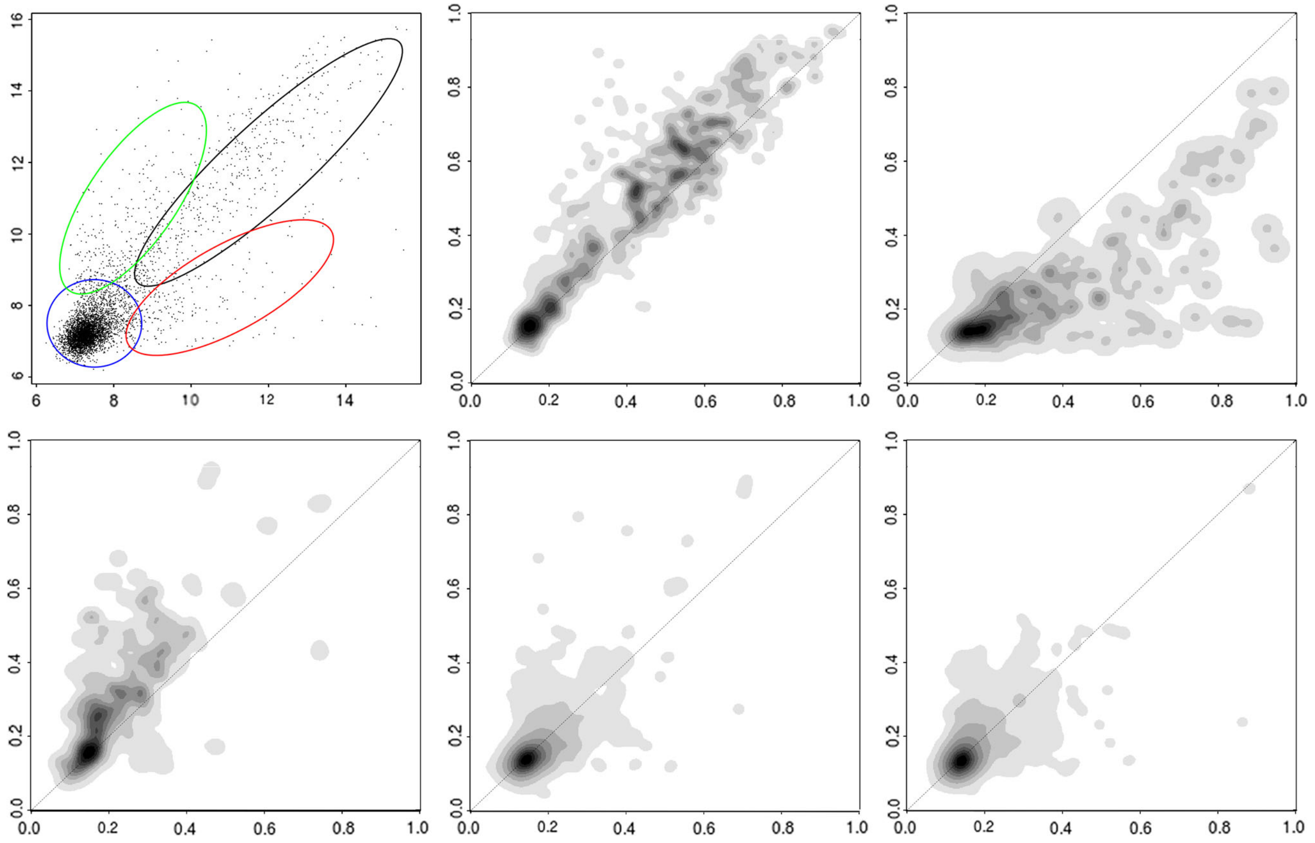

Fig. 4 *Top left panel*: raw tiling array data from chromosome 4 + idealized groups. *Other panels*: contour plots of the kernel estimate of each emission distribution for the 5-state non-parametric HMM. The *idealized blue group* is split into two HMM states (*bottom center* and *bottom right*). (Color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

scriptional 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 expression measurements of probes located on chromosome 4 in both seed and leaf, and takes its values in $\mathcal{Y} = \{\mathbf{R}^+ \times \mathbf{R}^+\}$. The top left panel of Fig. 4 is an idealization of the expected reconstruction of the hidden classes. 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 four corresponding hidden classes were drawn arbitrarily on the plot to illustrate these four behaviors.

Mixture as emission distributions The same data has already been analyzed in Volant et al. (2013) and Bérard et al. (2011), 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 Baudry et al. (2010) to HMMs. 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 Sect. 3.4. We used a spherical Gaussian kernel for which we first estimated the bandwidth w via cross-validation on the whole data set. The model with $k = 5$ provided the expected structure, splitting the "null" group containing probes expressed in none of the organs (in blue in the idealized plot) into two, resulting in the two bottom left figures in Fig. 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 Volant et al. (2013) (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 shown that non-parametric HMMs 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 use of kernel-based estimation of emission densities in an application to 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. Part of this work was supported by the ABS4NGS ANR project (ANR-11-BINF-0001-06).

References

Allman, E.S., Matias, C., Rhodes, J.: Identifiability of parameters in latent structure models with many observed variables. Ann. Stat. **37**, 3099–3132 (2009)

An, Y., Hu, Y., Hopkins, J., Shum, M.: Identifiability and inference of hidden Markov models. Technical report (2013)

Baudry, J.-P., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R.: Combining mixture components for clustering. J. Comput. Gr. Stat. **19**(2), 332–353 (2010)

Benaglia, T., Chauveau, D., Hunter, D.R.: An EM-like algorithm for semi-and nonparametric estimation in multivariate mixtures. J. Comput. Gr. Stat. **18**(2), 505–526 (2009)

Bérard, C., Martin-Magniette, M.L., Brunaud, V., Aubourg, S., Robin, S.: Unsupervised classification for tiling arrays: ChIP-chip and transcriptome. Stat. Appl. Genet. Mol. Biol. **10**(1), 1–22 (2011)

Bordes, L., Mottelet, S., Vandekerkhove, P.: Semiparametric estimation of a two components mixture model. Ann. Stat. **34**, 1204–1232 (2006a)

Bordes, L., Delmas, C., Vandekerkhove, P.: Semiparametric estimation of a two-component mixture model where one component is known. Scand. J. Stat. **33**(4), 733–752 (2006b)

Butucea, C., Vandekerkhove, P.: Semiparametric mixtures of symmetric distributions. Scand. J. Stat. **41**(1), 227–239 (2014)

Cappé, O., Moulines, E., Rydén, T.: Inference Hidden Markov Models. Springer, New York (2005)

Couvreur, L., Couvreur, C.: Wavelet based non-parametric HMMs: theory and methods. In: ICASSP '00 Proceedings, pp. 604–607 (2000)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)

DeSantis, S.M., Bandyopadhyay, D.: Hidden Markov models for zero-inflated Poisson counts with an application to substance use. Stat. Med. **30**(14), 1678–1694 (2011)

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. Ann. Stat. **24**(2), 508–539 (1996)

Du, J., Rozowsky, J.S., Korbel, J.O., Zhang, Z.D., Royce, T.E., Schultz, M.H., Snyder, M., Gerstein, M.: 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)

Dumont, T., Le Corff, S.: Nonparametric regression on hidden phi-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure. Technical report, arXiv:1209.0633D (Sep., 2012)

Durot, C., Huet, S., Koladjo, F., Robin, S.: Least-squares estimation of a convex discrete distribution. Comput. Stat. Data Anal. **67**, 282–298 (2013)

Gassiat, E., Rousseau, J.: Non parametric finite translation hidden Markov models and extensions. Bernoulli. to appear (2014)

Hall, P., Zhou, X.-H.: Nonparametric estimation of component distributions in a multivariate mixture. Ann. Stat. **31**(1), 201–224 (2003)

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

Jin, N., Mokhtarian, F.: A non-parametric HMM learning method for shape dynamics with application to human motion recognition. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 2, pp. 29–32. IEEE (2006)

Lambert, M., Whiting, J., Metcalfe, A.: A non-parametric hidden Markov model for climate state identification. Hydrol. Earth Syst. Sci. **7**(5), 652–667 (2003)

Lefèvre, F.: Non-parametric probability estimation for HMM-based automatic speech recognition. Comput. Speach Lang. **17**, 113–136 (2003)

Levine, M., Hunter, D.R., Chauveau, D.: Maximum smoothed likelihood for multivariate mixtures. Biometrika. **98**(2), 403–416 (2011)

Li, H., Zhang, K., Jiang, T.: The regularized EM algorithm. In: Proceedings of the National Conference on Artificial Intelligence, vol. 20, p. 807. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2005)

Lin, T.I., Lee, J.C., Yen, S.Y.: Finite mixture modelling using the skew normal distribution. Stat. Sin. **17**(3), 909–927 (2007)

Massart, P.: Concentration inequalities and model selection. Volume 1896 of Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Springer, Berlin (2007)

Maugis, C., Michel, B.: A non asymptotic penalized criterion for Gaussian mixture model selection. ESAIM Probab. Stat. **15**, 41–68 (2011)

Olteanu, M., Ridgway, J., et al.: Hidden Markov models for time series of counts with excess zeros. Proc. ESANN **2012**, 133–138 (2012)

Petrie, T.: Probabilistic functions of finite state Markov chains. Ann. Math. Stat. **40**, 97–115 (1969)

Shang, L., Chan, K.: Nonparametric discriminant HMM and application to facial expression recognition. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 2090–2096 (2009)

Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, Chichester (1985)

Tune, P., Nguyen, H. X., Roughan, M.: Hidden Markov model identifiability via tensors. In: 2013 IEEE International Symposium on Information Theory Proceedings (ISIT), pp. 2299–2303. IEEE (2013)

van de Geer, S.A.: Empirical processes in M-estimation. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2000)

Vernet, E.: Posterior consistency for nonparametric Hidden Markov Models with finite state space. Technical report, arXiv:1311.3092V (2013)

Volant, S., Bérard, C., Martin-Magniette, M.-L., Robin, S.: Hidden Markov models with mixtures as emission distributions. Stat. Comput. 1–12 (2013). doi:10.1007/s11222-013-9383-7

Yakowitz, S.J., Spragins, J.D.: On the identifiability of finite mixtures. Ann. Math. Stat. **39**, 209–214 (1968)

Zhai, Z., Ku, S.-Y., Luan, Y., Reinert, G., Waterman, M.S., Sun, F.: The power of detecting enriched patterns: an HMM approach. J. Comput. Biol. **17**(4), 581–592 (2010)