Mixtures of Nonparametric Components and Hidden Markov Models

Elisabeth Gassiat

October 27, 2017

1 Introduction and General Ideas

The topic of this chapter is statistical inference of nonparametric finite mixtures. The latent variables (and thus the observations) will be mostly taken independent and identically distributed, but in some cases, they will be possibly non independently distributed. For each observation, the corresponding latent variable indicates from which population the observation comes from. In particular, when the latent variables form a Markov chain, the observation process will comme from a non parametric hidden Markov model (HMM) with finite state space. We would like to emphasise the fact that the nonparametric modeling will concern only the conditional distribution of the observations, conditional on the latent variables, not the mixing distribution. Nonparametric modeling of the mixing distribution (with possibly infinitely denumerable or continuous support) is considered in Chapter 6.

To fix ideas, assume that a random variable X follows a distribution

$$P = \sum_{g=1}^{G} \eta_g F_g. \tag{1}$$

In many problems, inference for the F_g is of interest in itself, for instance in genomic applications, signal analysis, or econometric modeling, see references in [7] or [15]. The distribution P may have density

$$p(x) = \sum_{g=1}^{G} \eta_g f_g(x). \tag{2}$$

Parametric modeling of the distributions F_g (or the densities f_g) constraints the distributions to have prescribed shapes. Because non parametric modeling leads to more flexibility, methods to deal with non parametric models were investigated in several applied papers, see for instance the references in [15] or [12] for HMMs. One major obstacle of using nonparametric modeling seems to be the very basic question of identifiability (apart from that due to label switching or to the identification of the number of hidden states).

However, it should be emphasized that nonparametric estimation of the distribution P of the observations (or its density p) is always possible, even if the model is not identifiable. What will be of interest for us is the estimation of the weights η_g and of the so-called emission distributions F_g (or the emission densities f_g) for $g = 1, \ldots, G$. If the model is not identifiable, inference of the weights and the emission distributions is hopeless. But though identifiability is impossible to get in widest generality, it has been shown recently that it is possible to get identifiability for particular classes of models. The aim of this chapter is to review situations where identifiability has been proved, and where inference thus can be meaningful.

Let us proceed with some general ideas about the situations we will consider. If the observation is one-dimensional, it is obvious that, to obtain identifiability of the weights and the emission distributions from the marginal distribution P, one needs to put some restrictions on the emission distributions. This will be the subject of Section 2 and Section 3. In Section 2, we will consider mixtures of two populations (that is $G = 2$) under specific restrictions on the emission densities. In Section 3, we will consider mixtures of translated densities. A first hint on the link between HMMs and multidimensional mixtures appears in this section. If one wants to achieve fully nonparametric modeling of the translated distribution, then one requires a block of two observations to be dependent, which is the case for consecutive observations in HMMs.

Section 4 deals with multivariate mixtures. It appears that when the observations are at least three-dimensional and the coordinates are conditionally independent, that is, the emission distributions are the tensor products of the marginal distributions of the coordinates (or of three blocks of coordinates), then identifiability holds under a simple linear independence assumption which will be discussed in detail in Section 4.1. Using the fact that, conditionally on the present state, past and future states of a Markov chain are independent, it is possible to write the distribution of three consecutive observations in a HMM as a multidimensional mixture where the coordinates of the three-dimensional observations (made of the consecutive observations) are conditionally independent, and identifiability is obtained for HMMs. One way to prove identifiability is constructive, by applying spectral methods to the tensor of the three-dimensional distribution. We subsequently present nonparametric inference based on spectral methods in Section 4.2. Once the model is proven to be identifiable, nonparametric estimation methods may be proposed based on frequentist model selection ideas or based on Bayesian methodology. Model selection methods usually lead to oracle inequalities for the risk of the estimation of the distribution of the observations. One has then to go back from the distribution of the observations to the weights and the emission densities. Development of results on these ideas is the aim of Section 4.3. A specific section, Section 4.4, is dedicated to HMMs.

We conclude the chapter by discussing some related questions and extensions to other nonparametric mixture models in Section 5.

2 Mixtures With One Known Component

We consider in this section mixtures of two populations on the real line. Obviously, any distribution may be split in any mixture of itself, so that one has to specify some assumptions on the emission densities. One way to do this is to fix one of the components, so that for a known fixed density g,

$$p(x) = (1 - \eta)g(x) + \eta f(x), \quad x \in \mathbb{R}. \tag{3}$$

Apart from the knowledge of the first component, one has still to restrict the set of possible densities f for the other component. Indeed, one has for instance $(1 - \eta)g(x) + \eta f(x) = (1 - \eta/2)g(x) + \eta/2[g(x) + 2f(x)]$ so that, with the proportion $\eta/2$ and the second component equal to $g(x) + 2f(x)$ one gets the same mixture as (3).

2.1 The other component is symmetric

One possibility to obtain consistency is to assume that the second component is symmetric around some unknown value μ, so that

$$p(x) = (1 - \eta)g(x) + \eta f(x - \mu), \tag{4}$$

where f is a symmetric density, that is for all $x \in \mathbb{R}$, $f(-x) = f(x)$. The parameter to be estimated is (f, θ) with $\theta = (\eta, \mu) \in [0, 1] \times \mathbb{R}$. As soon as one has an estimator $\widehat{\theta} = (\widehat{\eta}, \widehat{\mu})$ of θ, one can build an estimator of the unknown f from a non parametric estimator $\widehat{p}$ of p by taking a symmetrized version of $\hat{p}(\cdot + \widehat{\mu})/\widehat{\eta} - (1/\widehat{\eta} - 1)g(\cdot)$.

In [9], the authors prove, under a condition on the existence of moments, that identifiability holds constraining $f(\cdot)$ to symmetry only and almost everywhere on θ. They propose a minimum distance estimator for parameters θ, based on a symmetrization based distance, and prove that this estimator is consistent. [8] then established that the estimator is $\sqrt{n}$-consistent and asymptotically normal. Further, [21] studied the situation where the known component g is symmetric with a known center of symmetry. Identifiability is proven by comparing the tails of (and thus under assumptions on) the characteristic functions. The authors construct asymptotically normal estimators. [37] also proposed an asymptotically normal estimator based on the minimum profile Hellinger distance.

2.2 Mixture of a uniform and a non decreasing density

Model (3) is often considered in applications when dealing with multiple testing problems. When the considered data are the p-values, the model transfers into

$$p(x) = (1 - \eta) + \eta f(x), \; x \in [0, 1], \tag{5}$$

where the random variables take values in $[0, 1]$ and the first component of the mixture is the uniform distribution on $[0, 1]$.

In [32], efficient estimation of the proportion η is investigated when the non parametric component is assumed to be in $\mathcal{F}_\delta$, the set of non increasing probability densities that are positive on $[0, 1-\delta)$ and zero on $[1-\delta, 1]$. When $\delta > 0$, it is possible to compute the efficient Fisher information that gives a lower bound of the asymptotic variance of $\sqrt{n}$-consistent estimators. The authors prove that a histogram based estimator of η is $\sqrt{n}$-consistent and conjecture that there exist no asymptotically efficient estimators that can achieve the lower bound given by the efficient Fisher information. This lower bound explodes when δ goes to 0, which implies that, when $\delta = 0$, the quadratic risk of any estimator cannot converge to a finite value at a parametric rate.

When one has at hand a preliminary estimator of the unknown proportion η and a non parametric estimator $\widehat{p}$ of the mixture density, it is possible to estimate the non parametric component f from equation (5), after plugging-in the estimators for η and p. However, though this estimator, theoretically, may have good asymptotic properties, it does not behave well in practice. In [31], the authors propose two different estimators for f that exhibit better performance in simulation studies. The first estimator is a randomly weighted kernel estimator, while the second estimator is a maximum smoothed likelihood estimator.

3 Translation Mixtures

A case of particular interest is the situation where the components of the mixture are shifted versions of one (unknown) distribution, that is when, for some unknown distribution F, and some unknown parameters $\mu_1, \dots, \mu_G$ the mixture distribution is

$$P(\cdot) = \sum_{g=1}^{G} \eta_g F(\cdot - \mu_g). \tag{6}$$

When the observations $X_1, \dots, X_n$ are i.i.d. with distribution P, modeling the distribution as (6) is not enough to get identifiability, and one has to add some assumption on the unknown distribution F. This is the subject of Section 3.1. However, when the observations are no longer independent, it is possible to get identifiability without any assumption on F, as will be discussed in Section 3.2.

3.1 Translation of a symmetric density

The usual assumption made in the literature to get identifiability is the restriction of F to a symmetric cumulative function in the sense that, for all $x \in \mathbb{R}$, $F(-x) + F(x) = 1$. This situation was studied independently by [10] and [24].

Let $\mathcal{F}$ be the set of symmetric cumulative functions. Denote Ω_G the set of weights $(\eta_g)_{1\le g\le G}$ and translation parameters $(\mu_g)_{1\le g\le G}$ in $\mathbb{R}^G$. The weights $(\eta_g)_{1\le g\le G}$ have to be such that $\eta_g \ge 0$, $g = 1, \dots, G$, and $\sum_{g=1}^{G} \eta_g = 1$. Denote $\Omega_G^\star$ the subset of identifiable parameters of Ω_G, that is the parameters $(\eta_g)_{1\le g\le G}$, $(\mu_g)_{1\le g\le G}$ such that for any $F \in \mathcal{F}$, one may recover the parameters from $\sum_{g=1}^{G} \eta_g F(\cdot - \mu_g)$.

The main identifiability result (see Theorem 1 in [24]) says that the following two statements are equivalent.

- The set of parameters $(\eta_g)_{1\le g\le G}$, $(\mu_g)_{1\le g\le G}$ is in $\Omega_G^\star$
- For all $(\eta'_g)_{1\le g\le G}$, $(\mu'_g)_{1\le g\le G}$, the convolution $\sum_{g=1}^G \eta_g\delta_{\mu_g} \star \sum_{g=1}^G \eta'_g\delta_{-\mu'_g}$ is symmetric if and only if $\sum_{g=1}^G \eta'_g\delta_{-\mu'_g} = \sum_{g=1}^G \eta_g\delta_{-\mu_g}$.

From this result, [24] deduce that, when $G = 2$, identifiability holds if and only if $\eta_1 \notin \{0, 1/2, 1\}$, which is equivalent to the assumption that the mixture $\eta_1 F(\cdot - \mu_1) + (1-\eta_1)F(\cdot - \mu_2)$ is not symmetric.

In [10], the authors propose an iterative estimation procedure for which they prove that the resulting estimator of the parameters is $n^{-1/4+\alpha}$-consistent for any positive α. [24] build on their symmetry considerations, to develop an estimator which is proven to be $\sqrt{n}$-consistent and asymptotically normal under technical assumptions. Later, [11] propose a new class of M-estimators for these parameters based on a Fourier approach, and prove that these estimators are $\sqrt{n}$-consistent under mild regularity conditions.

3.2 Translation of any distribution and HMMs

We consider now observations with marginal distribution (6) but that are not independent. This section mainly bulids on [16]. To get identifiability, we need dependent two consecutive observations. Then to use identifiability for building estimators, we need repetitions of such consecutive observations. We may have independent repetitions of a block of two non-independent variables, or we may have a stationary sequence of random variables. Thus, HMMs enter in this setting.

Consider a pair (X_1, X_2) of random variables such that the marginal distribution of X_1 (resp. X_2) is (6), and the latent variables (Z_1, Z_2) have a distribution given by the $G \times G$ matrix ξ on $\{1,\dots,G\}^2$, that is for all g, g', $\xi(g,g') = P(Z_1 = g, Z_2 = g')$. It is obvious that, by translating the distribution F and translating all μ_g's reversely, one obtains the same distribution. We thus fix arbitrarily μ_1 to 0, and consider the set of parameters Θ_G for the matrix ξ and the translation parameters μ_g, $g = 1,\dots,G$ such that $\mu_1 = 0 < \mu_2 < \dots < \mu_G$ and such that ξ is full rank.

The main identifiability result (Theorem 1) of [16] is the following. If the parameters lie in Θ_G for some (possibly unknown) G and if F is any probability distribution, then one may recover G, ξ, μ_g, $g = 1,\dots,G$ and F from the distribution of (X_1, X_2). This is a strong result. Indeed, it requires very weak assumptions. In particular, no assumption is put on F. The assumption that the translation parameters are distinct is obviously a basic assumption (the order constraint just fixes the label switching). The only structural assumption is that ξ has full rank. Notice that, with only two latent states (that is, $G = 2$), assuming that ξ has full rank is the same as assuming that the variables X_1 and X_2 are *not* independent.

The proof of the identifiability result uses characteristic functions and relies on complex analysis arguments. Let us explain the main ideas.Let θ be a parameter vector, containing the parameters in the matrix ξ as well as the translation parameters $(\mu_g)_g$, and let F be a probability distribution. One may rewrite the model as

$$X_i = \mu_{Z_i} + U_i, \quad i = 1, 2, \tag{7}$$

where U_1 and U_2 are independent variables with distribution F, and independent from (Z_1, Z_2) which has distribution ξ. Then, the characteristic functions of the variables X_i are products of the characteristic functions of corresponding variables μ_{Z_i} and the characteristic functions of corresponding variables U_i. Consider (θ_1, F_1) and (θ_2, F_2) such that the distribution of the pair (X_1, X_2) is the same under both sets of parameters. Using the fact that the characteristic functions of (X_1, X_2), of X_1, and of X_2, are the same for both sets of parameters, one sees that it is possible to separate what comes from F_1, F_2 and what comes from θ_1, θ_2. This leads to the following equation, for all (s, t) in a neighborhood of $(0, 0)$:

$$\Phi_{\theta_1}(s,t)\phi_{\theta_2}(s)\phi_{\theta_2}(t) = \Phi_{\theta_2}(s,t)\phi_{\theta_1}(s)\phi_{\theta_1}(t), \tag{8}$$

where for any θ, Φ_θ is the characteristic function of (μ_{Z_1}, μ_{Z_2}) under θ, and ϕ_θ is the characteristic function of μ_{Z_1} (or μ_{Z_2}) under θ. The identifiability proof is completed through studying the set of zeros of those functions, using properties of the entire function. The fact that ξ has full rank is used to obtain that Φ_θ is not the null function.

The milestone of the work in [16] is that, as soon as (8) holds in a neighborhood of $(0, 0)$, then $\theta_1 = \theta_2$. Let θ_1 be the true (unknown) parameter and let θ_2 be any possible θ. Taking the modulus of the difference of both sides of (8) and integrating in a neighborhood of $(0, 0)$, one obtains a contrast function $M(\theta)$ which is non negative, and zero if and only if θ is the true (unknown) value.

Since characteristic functions may be estimated empirically, this allows to build an empirical contrast function $M_n(\theta)$ that estimates $M(\theta)$ well enough to define an estimator $\widehat{\theta}$ as a minimizer of $M_n(\theta)$. It is proven in [16] that such an estimator has good properties (parametric rate $\sqrt{n}$ of convergence and asymptotic Gaussian distribution).

Once an estimator of θ is given, one may use a model selection approach to estimate the distribution F. One possibility is described in [16] and works as follows. Assume that possible distributions are dominated, and have a density $f \in \mathcal{F}$, $\mathcal{F}$ being an infinite dimensional set of densities. Assume you are given a collection $(\mathcal{F}_k)_{k\geq 1}$ of finite dimensional approximating sets of $\mathcal{F}$ (the larger k, the better is the approximation and the larger is the dimension of $\mathcal{F}_k$). For instance, $\mathcal{F}_k$ may be the set of stepwise dentities defined by a partition which is refined when k gets larger, or $\mathcal{F}_k$ may be the finite dimensional space spanned by the first k elements of a basis. One may define for any density function f

$$\ell_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\sum_{g=1}^{G} \widehat{\eta}_g f\left(X_i - \widehat{\mu}_g\right)\right],$$

which could be seen as the log likelihood if the variables were independent (which is not the case) and where θ is replaced by a preliminary estimator. For any k, define $\widehat{f}_k$ to be the maximizer of $\ell_n(f)$ over $\mathcal{F}_k$. As usual one has to make a trade-off between complexity and variance so that the estimator will be chosen by selecting k using a penalized criterion such as

$$D_n\left(k\right) = -\ell_n\left(\widehat{f}_k\right) + \operatorname{pen}\left(k, n\right),$$

where $\operatorname{pen}\left(k, n\right)$ is some penalty term that has to be chosen. Then the estimator is defined by $\widehat{f} = \widehat{f}_{\widehat{k}}$, with $\widehat{k}$ being a minimizer of D_n. Doing so, adaptivity results are proven in [16] on the estimation of the marginal density of X_1 when the penalty is adequately chosen. Then, when it is possible to go back from the risk on $p(\cdot) = \sum_{g=1}^{G} \eta_g f\left(\cdot - \mu_g\right)$ to that on f, adaptivity results are proven for the estimation of f.

Let us briefly describe how this works. Following model selection theory, one gets an oracle inequality for the square of the Hellinger distance $h^2(p, \widehat{p})$ from which adaptivity for the estimation of p is deduced. Here, $\widehat{p}(\cdot) = \sum_{g=1}^{G} \widehat{\eta}_g \widehat{f}\left(\cdot - \widehat{\mu}_g\right)$, and for any densities p and q with respect to some measure ν, $h^2(p, q) = \int(\sqrt{p} - \sqrt{q})^2 d\nu(x)$. The problem is now to lower bound $h(p, \widehat{p})$ by some distance between f and $\widehat{f}$. This may be done using the $L_1(\nu)$-distance, when $\sup_g \eta_g > 1/2$. Indeed, one has $h(p, \widehat{p}) \geq \|\widehat{p} - p\|_1$, and then, using the triangular inequality, the fact that the L_1-norm of a density is 1 and that the weights add up to 1, we have on one hand

$$\|\sum_g \widehat{\eta}_g \widehat{f}(\cdot - \widehat{\mu}_g) - \eta_g f(\cdot - \mu_g)\|_1 \leq \sum_g |\widehat{\eta}_g - \eta_g| + \sup_g \|f(\cdot - \widehat{\mu}_g) - f(\cdot - \mu_g)\|_1.$$

On the other hand, using iteratively the triangle inequality one gets that

$$\|\sum_g \eta_g(\widehat{f}(\cdot - \widehat{\mu}_g) - f(\cdot - \widehat{\mu}_g))\|_1 \geq \left(2 \max_g \eta_g - 1\right) \|\widehat{f} - f\|_1.$$

Then, if one has parametric rates for the estimation of η_g and μ_g, $g = 1, \ldots, G$, if f satisfies a Lipschitz property and if moreover $\max_g \eta_g > 1/2$ (which is a weak assumption when $G = 2$), one can transfer adaptive results from $h^2(p, \widehat{p})$ to adaptive results on $\|\widehat{f} - f\|_1$. This is a first example of how model selection methods allow to get adaptive estimators: obtain oracle inequality for the estimation of the density of the observed variables, transfer it to the non parametric part of the model if you are able to prove an inequality linking both risks. Notice that doing so requires to have a preliminary estimator with parametric rate. Indeed, one could use the model selection strategy to estimate θ and f together, but usual methods do not give directly a parametric rate for the parametric part alone, so that it does not seem easy to go back to adaptive rates for the non parametric part.

Finally, Bayesian methods can also be used, see [36].

4 Multivariate Mixtures

Let us consider now multidimensional mixtures for which the coordinates may be blocked in at least $d \geq 3$ blocks that constitute random variables that are conditionally independent knowing the population.That is, the observation has distribution P given by

$$P = \sum_{g=1}^{G} \eta_g \otimes_{j=1}^{d} F_{g,j}, \tag{9}$$

where for each $g = 1, \dots, G$, $F_{g,1}, \dots, F_{g,d}$ are probability distributions on d spaces (that may have different dimensions). When the spaces are equal and $F_{g,1} = \dots = F_{g,d}$ for all $g = 1, \dots, G$, this may be seen as modeling independently repeated measurements in unknown several populations. We shall first consider the situation where the observations $X_1, \dots, X_n$ are i.i.d., then we will exhibit a structural link with HMMs via the fact that, when $(Z_t)_{t\geq 1}$ is a Markov chain, conditionally on Z_t, Z_{t-1} and Z_{t+1} are independent, so that what has been understood for independent variables will be used to understand finite state space non parametric HMMs in Section 4.4.

4.1 Identifiability

In the statistical literature, the first results may be found in [20] where the case $G = 2$ and $d = 2$ or 3 is addressed, and in [19] who do not fix G and discuss $d = 2$ and $d > 2$. General insights on identifiability for various latent models is developed in [2]. Here, the authors point out a fundamental algebraic result by [28], which may be stated as an identifiability result of model (9) when $d = 3$ and the probability distributions are on finite sets.

Building upon this result, the authors of [2] prove that, when $d \geq 3$, model (9) is identifiable as soon as, for $j = 1, \dots, d$, the probability measures $F_{1,j}, \dots, F_{G,j}$ are linearly independent (which reduces to $F_{1,g}$ and $F_{2,G}$ being distinct when $G = 2$). Results in the parametric literature about spectral methods such as [3] and [33] may be used to get the same result; see also [7] and [6].

Let us present the spectral methods argument of [3] in more details. For sake of simplicity, we assume that $d = 3$ and that the $F_{g,j}$'s are distributions on $\mathbb{R}$ so that an observation

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \in \mathbb{R}^3.$$

Let $\phi_1, \dots, \phi_M$ be M real valued functions and denote $A^{(j)}$ for $j = 1, 2, 3$ the $M \times G$ matrix such that $A^{(j)}_{l,g} = \int \phi_l dF_{g,j}$, $l = 1, \dots, M$, $g = 1, \dots, G$. For instance, when the ϕ_l's are indicator functions of a partition of $\mathbb{R}$, then $A^{(j)}$ has the conditional distributions of the associated discretized coordinate X_j as columns. More generally, when the ϕ_l's are such that $(\phi_l)_{l\geq 1}$ forms a basis of the space of densities or of the space of distributions, then for large enough M, $A^{(j)}$

has rank M as soon as $F_{1,j},\dots,F_{G,j}$ are linearly independent, which we now assume for $j=1,2,3$. Let D be the diagonal $G\times G$ matrix having the η_g's on the diagonal and denote S the $M\times M$ matrix such that $S_{l,m}=E[\phi_l(X_1)\phi_m(X_2)]$. Then, one has

$$S=A^{(1)}D(A^{(2)})^\top,$$

so that as soon as $A^{(1)}$ and $A^{(2)}$ have full rank, which occurs for large enough $M\geq G$, one has $\operatorname{rank}(S)=G$. Thus, G is identifiable based on the joint distribution of (X_1,X_2). We now fix M such that $A^{(1)}$, $A^{(2)}$ and $A^{(3)}$ have full rank.

Let T be the $M\times M\times M$ tensor such that

$$T(l_1,l_2,l_3)=E[\phi_{l_1}(X_1)\phi_{l_2}(X_2)\phi_{l_3}(X_3)]$$

and let U_1 and U_2 be $M\times G$ matrices such that $U_1^\top SU_2$ is invertible (such matrices may be found by singular value decomposition of S). Let V be a vector in $\mathbb{R}^M$, and define $T[V]$ to be the $M\times M$ matrix given by applying the tensor T to V, that is

$$T[V]_{l,m}=E[\phi_l(X_1)\phi_m(X_2)\langle V,\Phi(X_3)\rangle],\ l,m=1,\dots,M,$$

where $\Phi(X_3)=(\phi_h(X_3))_{1\leq h\leq M}$. Define now for all V the matrix $B(V)$ (which may be computed as soon as one knows the distribution of X) as follows

$$B(V)=(U_1^\top T[V]U_2)(U_1^\top SU_2)^{-1}.$$

Then, denoting $\Delta(V)$ the diagonal $G\times G$ matrix with the coordinates of $(A^{(3)})^\top V$ on the diagonal, one has

$$B(V)=(U_1^\top A^{(1)})\Delta(V)(U_1^\top A^{(1)})^{-1}.$$

Thus, all matrices $B(V)$ have the same eigenvectors, and their eigenvalues are the coordinates of $(A^{(3)})^\top V$. This means that, by exploring various vectors V, one may recover $A^{(3)}$. The eigenvectors stay also the same when permuting coordinates 2 and 3 of the observed variable, so that one may recover $A^{(2)}$, and thus also $A^{(1)}$. Recovering D is then also possible. Finally, by taking M to infinity, one may recover the whole distributions $F_{1,g}$, $F_{2,g}$ and $F_{3,g}$, $g=1,\dots,G$.

4.2 Estimation with spectral methods

As seen in the spectral proof of identifiability, for large enough M, one may recover all parameters by spectral analysis (singular value decompositions and eigenvalue decompositions) using the matrix S and the tensor T. Given a sample of observation of X, S and T may be estimated empirically, by taking empirical means as estimators of the involved expectations. Thus, by spectral analysis using the estimators $\widehat{S}$ of S and $\widehat{T}$ of T, one get estimators of $A^{(1)}$, $A^{(2)}$, $A^{(3)}$, and D for fixed M. Such an algorithm is studied in [3] and [33].

To achieve the right rate of convergence for the estimators of the non parametric part, that is the distributions $F_{1,g}$, $F_{2,g}$ and $F_{3,g}$, $g = 1, \ldots, G$, one has to choose M appropriately as a function of the number of observations n, typically in a way that depends on the smoothness of the densities $f_{1,g}$, $f_{2,g}$ and $f_{3,g}$, $g = 1, \ldots, G$, of the distribution. This is studied in [7] for repeated measurements and in [6], where asymptotic results are given at the minimax asymptotic rate. However, choosing the right M when using spectral methods for estimation requires prior knowledge on the smoothness of densities.

4.3 Estimation with nonparametric methods

When identifiability holds, one may use model selection methods for the estimation of the parameters of the model (parametric weights and non parametric probability distributions), as described in Section 3.2. This leads to oracle inequalities for the risk of the estimator of the density of the observed variables, and one can deduce results for the risk on the parameters if it is possible to relate both risks. We shall describe some possible way of using such ideas in multivariate mixtures below.

Let us first mention that [20] propose an estimator of the weight and of the repartition functions of the distributions that are $\sqrt{n}$- consistent , for d-dimensional mixtures with $d \geq 3$ and $G = 2$. The method is to minimize some distance between the empirical d-dimensional repartition function and the set of repartition functions belonging to the model (9) with $G = 2$, and then to take the minimizer as an estimator. In [5], the authors propose an EM-type algorithm for semi- and non-parametric estimation in multivariate mixtures, but do not provide theoretical properties for the obtained estimator.

We shall now describe possible model selection methods such as penalized maximum likelihood and penalized least squares. Assume that on each of the d spaces, possible probability distributions are dominated and denote $\mathcal{F}_j$ the (non parametric) sets of possible densities, $j = 1, \ldots, d$. For each $j = 1, \ldots, d$, denote $(\mathcal{F}_{j,k})_{k \geq 1}$ a collection of finite dimensional approximating sets of $\mathcal{F}_j$. The log likelihood is given by

$$\ell_n\left((\eta_g)_{1 \leq g \leq G}, (f_{g,j})_{1 \leq g \leq G, 1 \leq j \leq d}\right) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \eta_g \prod_{j=1}^{d} f_{g,j}\left((X_i)_j\right) \right],$$

and for any $k = (k_1, \ldots, k_d) \in (\mathbb{N}^*)^d$, define $((\widehat{\eta}_g)_{1 \leq g \leq G}, (\widehat{f}_{g,j,k})_{1 \leq g \leq G, 1 \leq j \leq d})$ the maximizer of $\ell_n(\cdot)$ for $(\eta_g)_{1 \leq g \leq G}$ and for $f_{g,j} \in \mathcal{F}_{j,k_j}$, $1 \leq g \leq G, 1 \leq j \leq d$. To achieve a trade-off between complexity and variance, the estimator will be chosen by selecting $\widehat{k}$ as minimizing a penalized criterion such as

$$D_n(k) = -\ell_n\left((\widehat{\eta}_g)_{1 \leq g \leq G}, (\widehat{f}_{g,j,k})_{1 \leq g \leq G, 1 \leq j \leq d}\right) + \text{pen}(k_1, \ldots, k_d, n).$$

Denote p the density of $X_1 = ((X_1)_j)_{1 \leq j \leq d}$ and $\widehat{p}$ the estimator of p obtained with $\widehat{k}$. Then, following methods of [30], it may be possible to prove some oracle inequality with a statement as follows. Assume that the penalty is larger than some quantity related to the complexity of the model times $\log n/n$. Then, the risk $E\left[h^2\left(\widehat{p},p\right)\right]$ for the estimation of p may be upper bounded, up to some universal constant C, by the best (over all approximation spaces) Kullback-Leibler divergence of p to its best approximation plus the penalty, plus some negligible term.

Instead of using maximum likelihood, one may use least squares by minimizing, on each approximating space, the contrast function γ_n given by

$$\gamma_n\left((\eta_g)_{1\leq g\leq G},(f_{g,j})_{1\leq g\leq G,1\leq j\leq d}\right)=\int p^2(x)dx-\frac{2}{n}\sum_{i=1}^{n}p\left(X_i\right)$$

where, for $x=(x_j)_{1\leq j\leq d}$,

$$p(x)=\sum_{g=1}^{G}\eta_g\prod_{j=1}^{d}f_{g,j}(x_j),$$

and by selecting k with a penalized criterion as before. Then the oracle inequality that may be proven is now based on the L_2-risk $E\|\widehat{p}-p\|_2^2$.

A very important feature of such model selection methods is that the choice of the approximation space is data-driven, and often leads to adaptive minimax rates of estimation. Moreover, on a more practical side, one may apply the so-called slope heuristic to calibrate the penalty, see [4] for details.

If one is interested in risks about the parameters of the mixture, one has to follow the same route as in Section 3.2. First, find a preliminary estimator of the weights η_g, $g=1,\ldots,G$, with convergence rate $\sqrt{n}$, and use model selection to obtain an estimator for the non parametric part by plugging in the estimation criterion the estimator of the weights. Second, obtain an oracle inequality (this requires slightly more elaborate analysis due to the plugged estimator in the criterion). Third, go back from the risk on the density of the observations to the risk of the emission densities.

Let us show an example of such inequality in the simple case of repeated measurements models, that is when for all g, the $f_{g,j}$'s are the same, that is $f_{g,j}=f_g$, $j=1,\ldots,d$. We assume that all f_g's are in $L^2(I)$ for some subset I of $\mathbb{R}$. In what follows, norms of functions are L_2-norms and norms of vectors are euclidian norms. We denote by $\langle\cdot,\cdot\rangle$ the L_2-inner product between functions. For $\eta=(\eta_1,\ldots,\eta_G)$ and $\mathbf{f}=(f_1,\ldots,f_G)$ denote

$$p_{\eta,\mathbf{f}}(x_1,x_2,x_3)=\sum_{g=1}^{G}\eta_g f_g(x_1)f_g(x_2)f_g(x_3).$$

Let $\mathcal{H}$ be a closed bounded subset of $L^2(I)$. Let $\mathcal{F}$ be the subspace spanned by $f_1,\ldots,f_G$, and define, for $\mathbf{q}\in\mathbb{R}^G$ and $\mathbf{u}=(u_1,\ldots,u_G)\in\mathcal{F}^G$

$$D(\mathbf{q},\mathbf{u})=3\sum_{g,j}q_g q_j\langle f_g,f_j\rangle^2\langle u_g,u_j\rangle+6\sum_{g,j}q_g q_j\langle f_g,f_j\rangle\langle u_g,f_j\rangle\langle f_g,u_j\rangle.$$

When writing $u_1, \dots, u_G$ in the basis $f_1, \dots, f_G$, then $D(\mathbf{q}, \mathbf{u})$ is a quadratic form in the coordinates of $u_1, \dots, u_G$, with $G^2 \times G^2$ matrix $A(\mathbf{q}, \mathbf{f})$. We shall assume, to obtain the inequality, that the determinant $DetA(\eta, \mathbf{f}) \neq 0$. Notice that $DetA(\mathbf{q}, \mathbf{f})$ is a polynomial in the q_j's and the $\langle f_g, f_j \rangle$'s. Thus, if it is not the null function, the set of zeros of $DetA(\mathbf{q}, \mathbf{f})$ is negligible. But by taking functions f_g's with non intersecting supports, we get easily that $D(\mathbf{q}, \mathbf{u}) \geq 3 \sum_{g=1}^{G} q_g^2 \|f_g\|^4 \|u_g\|^2$ so that for such $\mathbf{f}$, $DetA(\mathbf{q}, \mathbf{f}) \neq 0$ and $DetA(\mathbf{q}, \mathbf{f})$ is not the null function. The assumption $DetA(\eta, \mathbf{f}) \neq 0$ is thus generically satisfied. Moreover by continuity, under this assumption, there exists a neighborhood of η such that for $\mathbf{q}$ in this neighborhood, $DetA(\mathbf{q}, \mathbf{f}) \neq 0$. We now state our theorem.

Theorem 1. *Assume $\eta_g > 0$, $g = 1, \dots, G$ and that $f_1, \dots, f_G$ are linearly independent. Assume moreover that $DetA(\eta, \mathbf{f}) \neq 0$. Let $\mathcal{K}$ be a compact neighborhood of η in $\mathbb{R}^G$ such that if $\mathbf{q} = (q_1, \dots, q_G) \in \mathcal{K}$, then $DetA(\mathbf{q}, \mathbf{f}) \neq 0$ and $q_g > 0$, $g = 1, \dots, G$. Then, there exists a constant $c(\mathcal{K}, \mathbf{f}) > 0$ such that for all $\mathbf{h} \in \mathcal{H}^G$ and all $\mathbf{q} = (q_1, \dots, q_G) \in \mathcal{K}$,*

$$\|p_{\mathbf{q},\mathbf{f}+\mathbf{h}} - p_{\mathbf{q},\mathbf{f}}\|^2 \geq c(\mathcal{K}, \mathbf{f}) \left(\sum_{g=1}^{G} \|h_g\|^2 \right).$$

Notice that for large enough n the estimator of η will be in $\mathcal{K}$ with large probability.

Let us now prove Theorem 1.

Denote $N(\mathbf{q}, \mathbf{h}) = \|p_{\mathbf{q},\mathbf{f}+\mathbf{h}} - p_{\mathbf{q},\mathbf{f}}\|^2$. Notice first that the identifiability proof uses only spectral arguments so that it may be extended to get:

$$\mathbf{h} \in \mathcal{H}^G \text{ is such that } N(\mathbf{q}, \mathbf{h}) = 0 \Longleftrightarrow \mathbf{h} = 0.$$

One may compute

$$\begin{aligned} N(\mathbf{q}, \mathbf{h}) &= \textstyle\sum_{g,j} \quad q_g q_j \langle f_g + h_g, f_j + h_j \rangle^3 - q_g q_j \langle f_g, f_j + h_j \rangle^3 \\ &\qquad\quad - q_g q_j \langle f_g + h_g, f_j \rangle^3 + q_g q_j \langle f_g, f_j \rangle^3. \end{aligned}$$

Let $\mathcal{F}^\perp$ be the orthogonal of $\mathcal{F}$. For $g = 1, \dots, G$, let u_g be the projection of h_g on $\mathcal{F}$ and $h_g^\perp$ its projection on $\mathcal{F}^\perp$. Then

$$N(\mathbf{q}, \mathbf{h}) = N(\mathbf{q}, \mathbf{u}) + M(\mathbf{q}, \mathbf{u}, \mathbf{h}^\perp)$$

where

$$\begin{aligned} M(\mathbf{q}, \mathbf{u}, \mathbf{h}^\perp) &= \textstyle\sum_{g,j} q_g q_j \quad \left\{ \langle h_g^\perp, h_j^\perp \rangle^3 + 3 \langle h_g^\perp, h_j^\perp \rangle^2 \langle f_g + u_g, f_j + u_j \rangle \right. \\ &\qquad \left. + 3 \langle h_g^\perp, h_j^\perp \rangle \langle f_g + u_g, f_j + u_j \rangle^2 \right\}. \end{aligned}$$

One may now write $N(\mathbf{q}, \mathbf{u})$ as a finite sum of homogeneous functions in the q_g's and the u_g's. The constant and the linear terms are zero, and denote the homogeneous term of degree 2 as $D(\mathbf{q},\mathbf{u})$. Using $N(\mathbf{q},\mathbf{u}) = D(\mathbf{q},\mathbf{u}) + o\left(\sum_{i=1}^{G}\|h_i\|^2\right)$ we easily get that for all $\mathbf{q}\in\mathcal{K}$ and all $\mathbf{u}\in\mathcal{F}^G$, $D(\mathbf{q},\mathbf{u})\geq 0$.

Denote $\|\mathbf{h}\|^2 = \sum_{g=1}^{G}\|h_g\|^2$ and define

$$c_1(\mathcal{K},\mathbf{f}) := \inf_{\mathbf{q}\in\mathcal{K},\mathbf{u}\in(\mathcal{F}\cap\mathcal{H})^G}\frac{N(\mathbf{q},\mathbf{u})}{\|(\mathbf{q},\mathbf{u})\|^2}.$$

Let $(\mathbf{q}_n,\mathbf{u}_n)_n$ be a sequence realizing $c_1(\mathcal{K},\mathbf{f})$. By compacity (closed and bounded subset in a finite dimensional space), $(\mathbf{q}_n,\mathbf{u}_n)_n$ has a limit point $(\bar{\mathbf{q}},\bar{\mathbf{h}})$. If $(\bar{\mathbf{q}},\bar{\mathbf{h}})\neq 0$, one has $N(\bar{\mathbf{q}},\bar{\mathbf{h}})\neq 0$ and we get $c_1(\mathcal{K},\mathbf{f})>0$. Else,

$$c_1(\mathcal{K},\mathbf{f}) = \lim_{n\to+\infty}\frac{D(\mathbf{q_n},\mathbf{u_n})}{\|(\mathbf{q}_n,\mathbf{u}_n)\|^2}.$$

But using the fact that $D(\mathbf{q},\mathbf{u})$ is non negative and non degenerate by the assumption $DetA(\mathbf{f})\neq 0$, we obtain that $c_1(\mathcal{K},\mathbf{f})>0$.
Using now Schur's theorem (which says that the Hadamard product of two positive matrices gives a positive matrix) and the fact that Gram matrices are non negative, we easily get

$$M(\mathbf{q},\mathbf{u},\mathbf{h}^{\perp}) \geq 3\lambda_{\min}(G(\mathbf{f}+\mathbf{u}))\left(\sum_{g=1}^{G}q_g^2\|h_g^{\perp}\|^2\right),$$

where $G(\star)$ denotes the Gram matrix of the function $\star$ and $\lambda_{\min}(\Box)$ is the minimum eigenvalue of $\Box$. Then

$$\begin{aligned} N(\mathbf{q},\mathbf{h}) &\geq 3\lambda_{\min}(G(\mathbf{f}+\mathbf{u}))\left(\sum_{g=1}^{G}q_g^2\|h_g^{\perp}\|^2\right) + c_1(\mathcal{K},\mathbf{f})\left(\sum_{g=1}^{G}\|u_g\|^2\right), \\ &\geq 3(\lambda_{\min}(G(\mathbf{f}+\mathbf{u}))(\inf_{\mathbf{q}\in\mathcal{K}}q_g^2)\sum_{g=1}^{G}\|h_g^{\perp}\|^2 + c_1(\mathcal{K},\mathbf{f})\left(\sum_{g=1}^{G}\|u_g\|^2\right). \end{aligned}$$

Let now $(\mathbf{q}_n,\mathbf{h}_n)_n$ be such that

$$c(\mathcal{K},\mathbf{f}) := \inf_{(\mathbf{q},\mathbf{h})}\frac{N(\mathbf{q},\mathbf{h})}{\|(\mathbf{q},\mathbf{h})\|^2} = \lim_{n\to+\infty}\frac{N(\mathbf{q}_n,\mathbf{h}_n)}{\|(\mathbf{q}_n,\mathbf{h}_n)\|^2}.$$

If $c(\mathcal{K},\mathbf{f})=0$, and since $c_1(\mathcal{K},\mathbf{f})>0$, we have that $\mathbf{u}_n$ tends to 0. But using the fact that$\lambda_{\min}(G(\mathbf{f}+\mathbf{u}))$ is a continuous function of $\mathbf{u}$ (in the finite dimensional space F) we get the contradiction

$$c(\mathcal{K},\mathbf{f}) = 0 \geq \left(\lambda_{\min}(G(\mathbf{f}))(\inf_{\mathbf{q}\in\mathcal{K}}q_g^2)\wedge c_1(\mathbf{K},\mathbf{f})\right) > 0,$$

so that we may conclude that $c(\mathcal{K},\mathbf{f})>0$.

4.4 Hidden Markov models

We shall now consider stationary finite state space non parametric HMMs. In this section, observations $(X_t)_{t\geq 1}$ are independent conditionally on the latent variables $(Z_t)_{t\geq 1}$, and the conditional distribution of X_t depends only on Z_t. The latent variables $(Z_t)_{t\geq 1}$ form a stationary Markov chain on the finite set $\{1,\ldots,G\}$, we shall denote ξ the transition matrix of the chain, and assume that it is irreducible and aperiodic. Let H_g denote the distribution of X_t conditional on $Z_t = g$, $g = 1,\ldots,G$. As discussed in Chapter **??**, finite state space HMMs are widely used as extensions of finite mixture models to model dependent variables coming from different populations. If $(\eta_g)_{g=1,\ldots,G}$ denotes the stationary probability mass function of ξ, then the marginal distribution of each variable X_t is the finite mixture

$$\sum_{g=1}^{G} \eta_g H_g.$$

Now, let us exhibit the structural link between HMMs and multivariate mixtures given by (9). Consider X to be the vector of 3 consecutive observations X_{t-1}, X_t, X_{t+1}. One may write the probability distribution of X as

$$\sum_{g=1}^{G} \left(\sum_{g_1=1}^{G} \eta_{g_1} \xi_{g_1,g} H_{g_1} \right) \otimes H_g \otimes \left(\sum_{g_3} \xi_{g,g_3} H_{g_3} \right),$$

which, since all weights η_g are positive, is the same as

$$\sum_{g=1}^{G} \eta_g \left(\sum_{g_1=1}^{G} \frac{\eta_{g_1} \xi_{g_1,g}}{\eta_g} H_{g_1} \right) \otimes H_g \otimes \left(\sum_{g_3} \xi_{g,g_3} H_{g_3} \right),$$

and may thus be seen as coming from the fact that, when $(Z_t)_{t\geq 1}$ is a Markov chain, the past Z_{t-1} and the future Z_{t+1} are independent conditional on the present Z_t. Thus the distribution of X is a 3-dimensional mixture given by (9) with

$$F_{g,1} = \sum_{g_1=1}^{G} \frac{\eta_{g_1} \xi_{g_1,g}}{\eta_g} H_{g_1},\ F_{g,2} = H_g,\ F_{g,3} = \sum_{g_3} \xi_{g,g_3} H_{g_3},$$

and one may apply the identifiability results of Section 4.1. This is what is done in [23] for parametric HMMs where the observation can take finitely many values and in [15] for non parametric HMMs.

When ξ is moreover full rank and the probability distributions $H_1,\ldots,H_G$ are linearly independent, then $F_{1,1},\ldots,F_{G,1}$ are linearly independent, $F_{1,3},\ldots,F_{G,3}$ are linearly independent, and the finite state space non parametric HMM model is identifiable. A more general identifiability result is proven in [1] by using Kruskal's algebraic result and the methods of [2]. They get that, for a large enough (more than 3) and depending on G number of consecutive variables, one may identify the HMM as soon as ξ is full rank and the distributions $H_1,\ldots,H_G$

are distinct. The proof is however not constructive in contrast to proofs using spectral methods.

In [15], several estimation methods are proposed and real data results are presented to support the conclusion that clustering using non parametric HMMs may lead to better results than by using conventional parametric HMMs algorithms. Spectral algorithms are proposed in [23] and their application in a non parametric setting is investigated in [13] where theoretical results on the rates of convergence are given. [36] gives assumptions under which Bayesian methods lead to consistent posterior distributions, moreover the rates of convergence are studied in [35]. In [12] the authors propose model selection based on penalized least squares estimators for the emission distributions which is statistically optimal and practically tractable. They prove a non asymptotic oracle inequality for the nonparametric estimator of the emission distributions H_g, $g = 1, \dots, G$. This requires an inequality similar to Theorem 1 which is proven to hold under a very weak assumption. A consequence is that this estimator is rate minimax adaptive up to a logarithmic term. The results hold under the assumption that the transition matrix Q is full rank and in the setting where the emission distributions have square integrable densities that are linearly independent. Simulations are given that show the improvement obtained when applying the least squares minimization consecutively to the spectral estimation. In [29], simulations study the sensitivity of the method to the linear independence assumption. It is shown that when the smallest eigenvalueof the Gram matrix of the scalar products of the densities is very small, more observations are needed for a good performance.

5 Related Questions

5.1 Clustering

Identification of the parameters of the mixture may be the first step for model-based clustering, see the introductory section of Chapter 8. In [13], we consider the filtering and smoothing recursions in nonparametric HMMs when the parameters of the model are unknown and replaced by estimators. We provide an explicit and time uniform control of the filtering and smoothing errors in total variation norm as a function of the parameter estimation errors, so that performances on estimation may be transferred to performances in clustering via a posteriori probabilities.

5.2 Order estimation

In a parametric context, the question of order estimation has been discussed in Chapter 7. General principles for order identification as described in [17, Section 4.1] remain valid, but are far more difficult to apply. In [27], the authors develop a procedure to estimate consistently a lower bound on the number of components in a multivariate finite mixture with conditionally independent coordinates. Recently, [29] provided two different methods to estimate consistently the number of hidden states in a non parametric HMM, one using a tresholding method on the singular values of the estimated distribution of two consecutive variables, the other using model selection techniques such as developed in [12]. The theoretical results are completed by a very interesting simulation study to compare the methods.

5.3 Semi-parametric estimation

As may be understood and proved for instance using spectral methods, the weights η_g, $g = 1, \ldots, G$, of the mixture may be estimated at a parametric $\sqrt{n}$ rate. Computing the efficient Fisher information allows to understand what loss occurs due to the fact that the non parametric emission distributions are unknown. In the case of finite mixtures of multidimensional distributions which are tensor products of at least three one-dimensional distributions, using step functions to approximate the densities, one obtains parametric models in which the weights may be asymptotically efficiently estimated. Of course, the finer the approximation, the better the asymptotic efficient variance of the estimator. However, choosing the right degree of approximation with a finite number of observations is a non trivial problem. Such questions are investigated in [18].

5.4 Regressions with random (observed or non observed) design

One may consider situations where the model is a mixture of regression models with unknown regression functions. The regressor variable may be random or not random, observed or not observed.

Finite mixtures of regressions with observed design are discussed in Chapter 13, see also [26], [34], [25], [22].

When the regressor variable is random and observed with noise,one can relate the model to the so-called errors-in-variables model. But when a regression model with random design is considered and the regressor variable is not observed, then one faces a model which is a mixture of regressions. When the design follows a continuous distribution, then the mixture is no longer finite, and difficult identifiability problems occur. A recent situation where identifiability can be solved may be found in [14].

6 Concluding remarks

We have presented in this chapter several mixture models where identifiability is verified with non parametric modeling of the population distributions. In such cases, one may use non parametric strategies such as model selection or non parametric Bayesian methods, with provable guarantees. There is still a lot to investigate both on the applied and theoretical sides. It is for instance fascinating that mixture models for which identifiability was obviously not true became identifiable when considering that the observations are dependent variables.

References

[1] G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

[2] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. *inCOLT 2012*, 2012.

[4] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[5] Tatiana Benaglia, Didier Chauveau, and David R. Hunter. An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *J. Comput. Graph. Statist.*, 18(2):505–526, 2009.

[6] Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Estimating multivariate latent-structure models. *Ann. Statist.*, 44(2):540–563, 2016.

[7] Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Nonparametric estimation of finite mixtures from repeated measurements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):211–229, 2016.

[8] L. Bordes and P. Vandekerkhove. Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Math. Methods Statist.*, 19(1):22–41, 2010.

[9] Laurent Bordes, Céline Delmas, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statist.*, 33(4):733–752, 2006.

[10] Laurent Bordes, Stéphane Mottelet, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, 34(3):1204–1232, 2006.

[11] Cristina Butucea and Pierre Vandekerkhove. Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.*, 41(1):227–239, 2014.

[12] Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 17:Paper No. 111, 43, 2016.

[13] Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Trans. Info. Th.*, 2017.

[14] Thierry Dumont and Sylvain Le Corff. Nonparametric regression on hidden Φ-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure. *Bernoulli*, 23(2):990–1021, 2017.

[15] E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71, 2016.

[16] E. Gassiat and J. Rousseau. Non parametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016.

[17] Élisabeth Gassiat. *Codage universel et identification d'ordre par sélection de modèles*, volume 21 of *Cours Spécialisés [Specialized Courses]*. Société Mathématique de France, Paris, 2014.

[18] Elisabeth Gassiat, Judith Rousseau, and Elodie Vernet. Semi-parametric estimation in nonparametric hidden Markov models. *submitted*, 2016.

[19] Peter Hall, Amnon Neeman, Reza Pakyari, and Ryan Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678, 2005.

[20] Peter Hall and Xiao-Hua Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, 31(1):201–224, 2003.

[21] Daniel Hohmann and Hajo Holzmann. Semiparametric location mixtures with distinct components. *Statistics*, 47(2):348–362, 2013.

[22] Daniel Hohmann and Hajo Holzmann. Two-component mixtures with independent coordinates as conditional mixtures: nonparametric identification and estimation. *Electron. J. Stat.*, 7:859–880, 2013.

[23] 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.

[24] David R. Hunter, Shaoli Wang, and Thomas P. Hettmansperger. Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251, 2007.

[25] David R. Hunter and Derek S. Young. Semiparametric mixtures of regressions. *J. Nonparametr. Stat.*, 24(1):19–38, 2012.

[26] Hiroyuki Kasahara and Katsumi Shimotsu. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175, 2009.

[27] Hiroyuki Kasahara and Katsumi Shimotsu. Non-parametric identification and estimation of the number of components in multivariate mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):97–111, 2014.

[28] Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977.

[29] Luc Lehéricy. Order estimation for non-parametric hidden Markov models. *Bernoulli*, to appear.

[30] 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.

[31] Van Hanh Nguyen and Catherine Matias. Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. *ESAIM Probab. Stat.*, 18:584–612, 2014.

[32] Van Hanh Nguyen and Catherine Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat.*, 41(4):1167–1194, 2014.

[33] L. Song, A. Anandkumar, B. Dai, and B.. Xie. Nonparametric estimation of multiview latent variable models. *in ICML2014*, 2014.

[34] Pierre Vandekerkhove. Estimation of a semiparametric mixture of regressions model. *J. Nonparametr. Stat.*, 25(1):181–208, 2013.

[35] E. Vernet. Non parametric hidden Markov models with finite state space: Posterior concentration rates. *submitted*, 2015.

[36] Elodie Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electron. J. Stat.*, 9:717–752, 2015.

[37] Sijia Xiang, Weixin Yao, and Jingjing Wu. Minimum profile Hellinger distance estimation for a semiparametric mixture model. *Canad. J. Statist.*, 42(2):246–267, 2014.