Nonparametric estimation of finite mixtures from repeated measurements

Stéphane Bonhomme
University of Chicago
Koen Jochmans†
Sciences Po, Paris
Jean-Marc Robin
Sciences Po, Paris and University College London

[Revised July 22, 2014]

Abstract. This paper provides methods to nonparametrically estimate finite mixtures from data with repeated measurements. We present a constructive identification argument and use it to develop simple two-step estimators of the component distributions and all their functionals. We discuss a computationally-efficient method for estimation and derive asymptotic theory. Simulation experiments suggest that our theory provides confidence intervals with good coverage in small samples.

Keywords: finite mixture, repeated-measurement data, re-weighting, two-step estimation.

1. Introduction

Finite-mixture models are widely used in statistical analysis. Popular applications include modeling unobserved heterogeneity in structural behavioral models, learning about individual behavior from grouped data, and dealing with corrupted data. McLachlan and Peel (2000) discuss many examples. Mixture models are most often parametrically specified, and estimation may be done by maximum likelihood or indirect inference in a frequentist setting, or via MCMC techniques in a Bayesian approach.

There is a growing literature on nonparametric identification of finite mixtures. Univariate mixtures are generally not identified nonparametrically.[1] In contrast, data on multiple measurements can represent a powerful source of identification. Hettmansperger and Thomas (2000), Hall and Zhou (2003), and Allman, Matias, and Rhodes (2009) provide identification results on component distributions and mixing proportions under the assumption that measurements are conditionally independent and the number of components is known. Hu (2008) and Kasahara and Shimotsu (2009) establish related results in econometrics.

†*Address for correspondence:* Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France. *E-mail:* koen.jochmans@sciencespo.fr.

[1] Additional restrictions may lead to identification. Bordes, Mottelet, and Vandekerkhove (2006) and Hunter, Wang, and Hettmansperger (2007) study two- and three-component mixtures of symmetric location families. Kitamura (2004) and Henry, Jochmans, and Salanié (2013) study models with conditioning variables.

To nonparametrically estimate multivariate mixtures, computational procedures akin to the EM algorithm (Dempster, Laird, and Rubin 1977) have recently been introduced by Benaglia, Chauveau, and Hunter (2009) and Levine, Hunter, and Chauveau (2011). These approaches are applicable more generally than are the earlier proposals of Hettmansperger and Thomas (2000) and Elmore, Hettmansperger, and Thomas (2004), and of Hall and Zhou (2003) and Hall, Neeman, Pakyari, and Elmore (2005). Although simulation evidence suggests that these estimators work well in finite samples, their statistical properties are currently unknown. Chauveau, Hunter, and Levine (2014) provide an account of these developments.

The aim of this paper is to contribute to the development of practical procedures to nonparametrically estimate finite mixtures from data on repeated measurements, and to advance statistical inference on such models. Like Hettmansperger and Thomas (2000) and Elmore, Hettmansperger, and Thomas (2004), we will work in a framework where the measurements are independent and identically distributed conditional on knowing from which component they have been drawn. Restricting the conditional distributions to be the same across measurements is not necessary for identification (Allman, Matias, and Rhodes 2009). However, it facilitates the construction of simple estimators to which standard asymptotic theory can be applied. In addition, throughout the paper we will assume that the number of components is known. Determining the number of components nonparametrically is a difficult issue which we do not address.[2]

When three or more measurements are available and a rank condition is satisfied, linear functionals of component distributions are identified as suitably re-weighted versions of the same functionals of the marginal distribution. Notably, the component-specific cumulative distribution functions are themselves identified as weighted averages. Given this, identification of the mixing proportions follows readily from a minimum-distance argument.

Our identification argument is constructive. Moreover, it suggests a convenient two-step approach to estimate any linear functional of the component distributions. In the first step the weights are estimated. This step is generic, in the sense that it does not depend on the particular functional of interest. So, the weights need to be estimated only once; we discuss a computationally-efficient method for doing so. In the second step it then suffices to average the data with respect to the estimated weights to obtain an estimator of the functional of interest.

Under suitable regularity conditions, our estimators are consistent and asymptotically normal, and asymptotically-valid confidence sets can be constructed using plug-in estimators of their asymptotic variances. To cover nonlinear functionals and semiparametric estimators, we also extend our results to deal with minimum-distance estimators of Euclidean parameters. We also provide consistent estimators of the mixing proportions.

[2] Kasahara and Shimotsu (2014) present an argument based on the non-negative rank of two-way contingency tables. Even in parametric models, the problem of inferring the number of components is non-standard and has not been fully resolved; see, e.g., Leroux (1992), Zhu and Zhang (2004), and Woo and Sriram (2006).

Our approach works for both discrete and continuous component distributions. In the continuous case, a nonparametric density estimator may be of interest. We construct re-weighted kernel estimators of component densities, using the same weights as before. The large-sample properties of our density estimators are standard. Furthermore, the selection of the bandwidth can be done by standard data-driven methods like least-squares cross-validation (Rudemo 1982; Bowman 1984), for example. We emphasize, however, that estimation of functionals or other Euclidean parameters of interest can be done without first estimating these densities.

We study the performance of our estimators in a Monte Carlo simulation used in Levine, Hunter, and Chauveau (2011). The simulations show that our procedure yields estimates with comparable bias and only slightly higher standard deviations than the parametric maximum-likelihood estimator and the smoothed nonparametric likelihood estimator of Levine, Hunter, and Chauveau (2011). The results further show that our asymptotic theory provides confidence intervals that yield reliable inference in small samples. As an application we take our methods to a data set from cognitive psychology due to Thomas, Lohaus, and Brainerd (1993), and obtain similar point estimates as do Elmore, Hettmansperger, and Thomas (2004), Benaglia, Chauveau, and Hunter (2009), and Levine, Hunter, and Chauveau (2011).

We proceed as follows. In Section 2 we formally introduce the model and present a constructive identification argument. In Section 3 we use this argument to construct estimators, and we derive distribution theory. That section also discusses minimum-distance estimators of Euclidean parameters and kernel density estimation. In Section 4 we collect numerical evidence. Technical proofs, some further discussion, and additional simulation results are available as supplementary material.

2. Identification

Finite mixtures provide a paradigm for analyzing group data when group membership is unobserved. Let x be the random variable, supported on $\mathcal{X} \subseteq \mathcal{R}$, whose distribution is of interest. The set $\mathcal{X}$ need not be countable or finite; we allow for both discrete and continuous probability distributions. We will work in a framework where repeated measurements $x_1, x_2, \ldots, x_M$ on x have joint cumulative distribution function

$$F(x_1, x_2, \ldots, x_M) = \sum_{k=1}^{K} \omega_k \prod_{m=1}^{M} F_k(x_m). \tag{2.1}$$

The number of groups, K, is taken as known throughout. The parameters of interest in this multivariate finite mixture are the group-specific distribution functions $F_k : \mathcal{X} \to [0, 1]$ and the mixing proportions $\omega_k > 0$. Our goal in this section is to provide a method to recover the F_k and ω_k nonparametrically.

To introduce our approach, let $\chi = (\chi_1, \chi_2, \ldots, \chi_I)'$ be a set of I univariate functions $\chi_i : \mathcal{X} \to \mathcal{R}$. Several choices for χ_i are possible. For example, given I values $v_1, v_2, \ldots, v_I$

in $\mathcal{X}$, we could set $\chi_i(x) = 1\{x \leq v_i\}$. Alternative choices for the χ_i when x is continuously distributed include orthogonal polynomials or splines. Consider, then, for all $x \in \mathcal{X}$, the $I \times I$ matrix

$$A(x) \equiv \mathbb{E}[\chi(x_{m_1})\chi(x_{m_2})'|x_{m_3} = x],$$

for any triple of distinct measurements (m_1, m_2, m_3). Throughout this section, we assume existence of the relevant expectations. Let $\Omega \equiv \text{diag}[\,\omega_1, \omega_2, \ldots, \omega_K]$ denote the diagonal matrix of mixing proportions. Introduce the vectors

$$b_k \equiv \mathbb{E}_k[\chi(x_m)] = \int_{\mathcal{X}} \chi(x)\,\mathrm{d}F_k(x),$$

and collect them in the $I \times K$ matrix $B \equiv (b_1, b_2, \ldots, b_K)$. Then Equation (2.1) implies that

$$A(x) = \left(B\Omega^{1/2}\right) D(x) \left(B\Omega^{1/2}\right)', \quad D(x) \equiv \text{diag}\left[\frac{f_1(x)}{f(x)}, \frac{f_2(x)}{f(x)}, \ldots, \frac{f_K(x)}{f(x)}\right],$$

where $f_1, f_2, \ldots, f_K$ and f denote the density functions of the group-specific distributions and of the marginal distribution, each defined with respect to the appropriate measure, respectively.

The diagonal entries of $D(x)$ have the important property that

$$\mathbb{E}\left[\frac{f_k(x_m)}{f(x_m)}\,\varphi(x_m)\right] = \mathbb{E}_k[\varphi(x_m)] \tag{2.2}$$

for any function φ. For example, if we set $\varphi(x_m) = 1\{x_m \leq x\}$ for some $x \in \mathcal{X}$, (2.2) gives

$$\mathbb{E}\left[\frac{f_k(x_m)}{f(x_m)}\,1\{x_m \leq x\}\right] = \mathbb{E}_k[1\{x_m \leq x\}] = F_k(x),$$

the component distributions at x. If the functional forms of f_k and f were known, this would be a standard moment calculation via importance sampling from f_k based on f (see, e.g., Robert and Casella, 2004, Section 3.3). Therefore, knowledge of the diagonal matrix $D(x)$ implies that the component distributions, and thus also all of their functionals, are identified.

We now show that the matrix $D(x)$ can be recovered—up to permutation of its diagonal entries—under the following condition.

Assumption 1 (rank). *The matrix B has maximal column rank.*

Assumption 1 is similar to the identification condition of Allman, Matias, and Rhodes (2009, Theorem 8), which requires component distributions to be linearly independent. Indeed, with $\chi_i(x_m) = 1\{x_m \leq v_i\}$ for a set of chosen values $v_1, \ldots, v_I$, we have $\mathbb{E}_k[\chi_i(x_m)] = F_k(v_i)$ and Assumption 1 demands linear independence of the component distributions on the grid $v_1, \ldots, v_I$. Note that Assumption 1 is testable. Indeed,

$$A \equiv \mathbb{E}[A(x_m)] = B\Omega B'.$$

Hence, for given I, Assumption 1 is equivalent to A having rank K, which is a testable restriction. See the supplementary material for details.

The following lemma establishes that the diagonal entries of $D(x)$ are identified. The proof will be instrumental for the construction of our estimators in the next section and is given below.

LEMMA 1 (IDENTIFICATION). *Let Assumption 1 hold. Then $D(x)$ can be recovered up to permutation of its diagonal entries.*

The $I \times I$ matrix A is real and symmetric and so admits an eigendecomposition of the form

$$A = V \Lambda V',$$

where V is the $I \times I$ orthonormal matrix containing the eigenvectors and Λ is the diagonal matrix containing the corresponding eigenvalues. By Assumption 1, A has rank K. Let Λ_K be the $K \times K$ submatrix of Λ containing the K non-zero eigenvalues of A, and write V_K for the $I \times K$ submatrix of V containing the associated eigenvectors. The matrix $W \equiv \Lambda_K^{-1/2} V_K'$ is such that, for all $x \in \mathcal{X}$,

$$W A(x) W' = U D(x) U', \tag{2.3}$$

where $U \equiv W B \Omega^{1/2}$. Because $WAW' = UU' = \mathrm{I}_K$, where I_K denotes the $K \times K$ identity matrix, U is a full-rank orthonormal matrix. Thus, the matrices $WA(x)W'$ share the same eigenvectors, which are given by the columns of U. Observe that V_K and, hence, W are not unique if the non-zero eigenvalues of A are multiple. Nevertheless, the joint eigendecomposition in (2.3) holds irrespective of the choice of V_K in such a case. Furthermore, because B and U have full column rank and the eigenvectors are orthonormal, the decomposition is unique up to relabelling of the eigenvectors and eigenvalues, and up to the directions of the eigenvectors (see, e.g., De Lathauwer, De Moor, and Vandewalle 2004). We therefore have established that

$$D(x) = \big(U'W\big) A(x) \big(U'W\big)' \tag{2.4}$$

is identified up to permutation of its diagonal entries. This concludes the proof of Lemma 1.

Given Lemma 1, identification of component distributions and their functionals follows immediately from (2.2). Here and in the following, identification is to be understood as to hold up to label-swapping of the various components. The possibility of relabelling is an ambiguity that is inherent to mixtures (see McLachlan and Peel 2000, Section 4.9), and we will henceforth leave it implicit.

An alternative formula that will be particularly useful when turning to estimation is obtained on recalling that $A(x)$ is a conditional expectation. Let

$$\tau_k(x_{m_1}, x_{m_2}) \equiv u_k' W \, \chi(x_{m_1}) \chi(x_{m_2})' \, W' u_k, \tag{2.5}$$

where u_k denotes the kth column of matrix U. Then (2.4) implies that the diagonal entries of $D(x) = \mathrm{diag}[d_1(x), d_2(x), \ldots, d_K(x)]$ can be written as

$$d_k(x) = \frac{f_k(x)}{f(x)} = \mathbb{E}[\tau_k(x_{m_1}, x_{m_2}) | x_{m_3} = x].$$

Iterating expectations then yields Theorem 1.

THEOREM 1 (IDENTIFICATION). *Let Assumption 1 hold. Then*

$$\mathbb{E}_k[\varphi(x_m)] = \mathbb{E}[\tau_k(x_{m_1}, x_{m_2})\,\varphi(x_{m_3})]$$

is identified for any function φ.

An application of Theorem 1 with $\varphi(x_m) = 1\{x_m \leq x\}$ for chosen $x \in \mathcal{X}$ leads to the following important corollary.

COROLLARY 1 (COMPONENT DISTRIBUTIONS). *Let Assumption 1 hold. Then*

$$F_k(x) = \mathbb{E}_k[1\{x_m \leq x\}] = \mathbb{E}[\tau_k(x_{m_1}, x_{m_2})\,1\{x_{m_3} \leq x\}]$$

is identified for all $x \in \mathcal{X}$.

Another consequence of Theorem 1 is that the matrix B is identified. Indeed, its kth column is $b_k = \mathbb{E}[\tau_k(x_{m_1}, x_{m_2})\chi(x_{m_3})]$. In addition, the mixture model in (2.1) implies that

$$a \equiv \mathbb{E}[\chi(x_m)] = B\omega,$$

for $\omega \equiv (\omega_1, \omega_2, \ldots, \omega_K)'$. By Assumption 1, the matrix $B'B$ has full rank. The mixing proportions are therefore identified.

COROLLARY 2 (MIXING PROPORTIONS). *Let Assumption 1 hold. Then*

$$\omega = (B'B)^{-1}B'a$$

is identified.

Combined, Corollaries 1 and 2 give identification of all the parameters in the mixture model in (2.1).

3. Estimation

In this section we construct estimators based on Theorem 1 and derive distribution theory.

3.1. Estimation by joint diagonalization

For a chosen function φ with $\dim \varphi$ components, consider the estimand

$$\theta_0 \equiv \mathbb{E}_k[\varphi(x_m)].$$

As an example, $\theta_0 = F_k(x)$ when $\varphi(x_m) = 1\{x_m \leq x\}$. Let $\{x_{nm}\}_{n,m}$ denote a sample of N observations drawn at random from the M-variate mixture in (2.1). Theorem 1 suggests estimating θ_0 based on the empirical analog

$$\widehat{\theta} \equiv \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)}\widehat{\tau}_k(x_{nm_1}, x_{nm_2})\,\varphi(x_{nm_3}), \tag{3.1}$$

where

$$\widehat{\tau}_k(x_{nm_1}, x_{nm_2}) \equiv \widehat{u}_k' \widehat{W}\, \chi(x_{nm_1}) \chi(x_{nm_2})'\, \widehat{W}' \widehat{u}_k, \tag{3.2}$$

is an estimator of $\tau_k(x_{nm_1}, x_{nm_2})$ using estimators $\widehat{W}$ and $\widehat{U} = (\widehat{u}_1, \widehat{u}_2, \ldots, \widehat{u}_K)$ of the transformation matrix W and of the matrix of eigenvectors U, respectively. In (3.1), (m_1, m_2, m_3) ranges over all ordered triples of distinct elements from the set $\{1, 2, \ldots, M\}$. The averaging across triples is done to exploit the fact that all measurements x_m have identical distributions in model (2.1).

Constructing the weights requires estimating W and U. The matrix $W = \Lambda_K^{-1/2} V_K'$ can be estimated by means of an eigendecomposition of

$$\widehat{A} \equiv \frac{1}{N} \frac{(M-2)!}{M!} \sum_{n=1}^{N} \sum_{(m_1, m_2)} \chi(x_{nm_1}) \chi(x_{nm_2})',$$

which is the empirical analog of A. In principle, by (2.3), the matrix U could be estimated by the eigenvectors of an empirical counterpart of $WA(x)W'$ for any choice of x. Proceeding in this way, however, would require a nonparametric estimation step. Moreover, efficiency gains should arise from imposing the constraint that the same matrix U jointly diagonalizes all matrices $A(x)$ for all x. We therefore proceed in a different, more constructive, manner, by averaging the restrictions and performing approximate joint diagonalization of the resulting matrices.

Let

$$A_i \equiv \mathbb{E}[A(x_m)\, \chi_i(x_m)], \qquad D_i \equiv \mathbb{E}[D(x_m)\, \chi_i(x_m)],$$

define averages of $A(x)$ and $D(x)$ with respect to $\chi_i(x)$ (note that other choices of functions are possible). Each A_i can be estimated by

$$\widehat{A}_i \equiv \frac{1}{N} \frac{(M-3)!}{M!} \sum_{n=1}^{N} \sum_{(m_1, m_2, m_3)} \chi(x_{nm_1}) \chi(x_{nm_2})'\, \chi_i(x_{nm_3}).$$

Moreover, (2.3) implies that

$$W A_i W' = U D_i U'. \tag{3.3}$$

While U is the joint diagonalizer of the matrices WA_iW', their estimates will generally not share the same eigenvectors due to sampling error. Our estimator $\widehat{U}$ is that matrix that makes $U'(\widehat{W}\widehat{A}_i\widehat{W}')U$ as close to diagonal as possible, in the sense of minimizing the sum of squares of their off-diagonal entries, that is,

$$\widehat{U} \equiv \arg\min_{U \in \mathcal{U}} \sum_{i=1}^{I} \left\| \text{offdiag}\left\{ U' \left(\widehat{W} \widehat{A}_i \widehat{W}' \right) U \right\} \right\|^2, \tag{3.4}$$

where $\mathcal{U}$ is the set of $K \times K$ orthonormal matrices, $\text{offdiag}(A) = A - \text{diag}(A)$, and $\|\cdot\|$ denotes the Frobenius norm.[3]

[3] In principle, the restrictions in (3.3) could also be enforced via a minimum-distance procedure involving separate estimates of the eigenvectors of each matrix A_i. However, label-swapping issues make this procedure difficult to implement in practice.

The first-order conditions to (3.4) are highly nonlinear and difficult to solve using conventional gradient-based methods. Fortunately, joint diagonalization problems of this kind have been extensively studied in numerical analysis and several algorithms have been developed (see, e.g., Bunse-Gerstner, Byers, and Mehrman 1993). Here, we shall use the JADE algorithm of Cardoso and Souloumiac (1993). This procedure is based on iteratively applying elementary Jacobi rotations. Its attractive computational properties have made JADE a workhorse technique in blind source separation (see, e.g., Shi 2011). In extensive numerical experiments we found this algorithm to be very stable and computationally extremely fast.

Lastly, estimators of the mixing proportions ω_k can be based on Corollary 2; see the supplementary material for details. Alternatively, given estimates of the component distribution functions, the ω_k could also be estimated via the procedures studied in Hall (1981) and Titterington (1983).

3.2. *Large-sample theory*

We next characterize the asymptotic distribution of $\widehat{\theta}$ in (3.1). All the proofs for this section are collected in the supplementary material. We start by noting that $\widehat{\theta}$ is a plug-in version of

$$\widetilde{\theta} \equiv \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)}\tau_k(x_{nm_1},x_{nm_2})\,\varphi(x_{nm_3}),$$

which is constructed using the true weight function τ_k. Because $\widetilde{\theta}$ is a sample average, distribution theory for this estimator is easy to establish. Therefore, in deriving asymptotic theory for $\widehat{\theta}$, the main task is to assess the impact of estimating the weight function. Given the form of the weight τ_k, this means establishing the asymptotic behavior of our estimator of $W'U$.

To be able to appeal to a central limit theorem we rely on the following condition.

Assumption 2 (moments). *For all i, $\chi_i^2 f$ is integrable.*

For example, this condition is trivially satisfied when the functions χ_i are bounded.

The integrability condition on χ_i^2 implies that our estimators of A, $A_1, A_2, \ldots, A_I$ are both unbiased and asymptotically linear, with influence functions

$$\upsilon_n \equiv \frac{(M-2)!}{M!}\sum_{(m_1,m_2)}\xi(x_{nm_1},x_{nm_2}) - \mathrm{vec}(A),$$
$$\upsilon_n^i \equiv \frac{(M-3)!}{M!}\sum_{(m_1,m_2,m_3)}\xi(x_{nm_1},x_{nm_2})\chi_i(x_{nm_3}) - \mathrm{vec}(A_i),$$

where $\xi(x_{nm_1},x_{nm_2}) \equiv \mathrm{vec}\;\chi(x_{nm_1})\chi(x_{nm_2})'$.

The next condition ensures that the asymptotic properties of $\widehat{A}$ carry over to the estimator of W.

ASSUMPTION 3 (EIGENVALUES). *The non-zero eigenvalues of A are all simple.*

Assumption 3 is imposed mainly to facilitate the exposition. If it is not satisfied, the eigenvalues of A are no longer a continuous function of A. This complicates the derivation of the asymptotic distribution of the eigenvalues of $\widehat{A}$, as this distribution depends in a complicated way on the multiplicity of the eigenvalues (Eaton and Tyler 1991).

Let $\otimes$ denote the Kronecker product, and let $\overset{\text{col}}{\otimes}$ and $\overset{\text{row}}{\otimes}$ denote the columnwise and rowwise Kronecker products, respectively. For any collection of matrices $M_1, M_2, \ldots, M_I$, let horzcat(M_i) and vertcat(M_i) denote the matrices obtained by horizontal and vertical concatenation, respectively. The matrices $\widehat{W}\widehat{A}_i\widehat{W}'$ in (3.4) are jointly asymptotically normal with influence function

$$\psi_n \equiv \text{vertcat}(WA_i \otimes \mathrm{I}_K)\, J_W \upsilon_n + (\mathrm{I}_I \otimes W \otimes W)\, \Upsilon_n + \text{vertcat}(\mathrm{I}_K \otimes WA_i)\, J_{W'} \upsilon_n.$$

Here, $\Upsilon_n \equiv \text{horzcat}(\upsilon_n^i)$, and J_W and $J_{W'}$ are the Jacobian matrices of the transformations from A to W and from A to W', respectively, namely

$$\begin{aligned} J_W &\equiv -(V \otimes \mathrm{I}_K)(\Lambda \ominus \Lambda_K)^+(V' \otimes W) - \frac{1}{2}(W' \overset{\text{col}}{\otimes} \mathrm{I}_K)(W \overset{\text{row}}{\otimes} W), \\ J_{W'} &\equiv \quad (\mathrm{I}_K \otimes V)(\Lambda_K \ominus \Lambda)^+(W \otimes V') - \frac{1}{2}(\mathrm{I}_K \overset{\text{col}}{\otimes} W')(W \overset{\text{row}}{\otimes} W), \end{aligned}$$

where $M_1 \ominus M_2 \equiv M_1 \otimes \mathrm{I}_{\dim M_2} - \mathrm{I}_{\dim M_1} \otimes M_2$ is the Kronecker difference between any two square matrices M_1 and M_2, and the $+$ superscript on a matrix indicates its Moore-Penrose pseudo-inverse. The estimators of W and U, too, are asymptotically linear. The influence function of $\widehat{W}$ is $J_W \upsilon_n$. The influence function of $\widehat{U}$ is $J_U \psi_n$, for

$$J_U \equiv -(\mathrm{I}_K \otimes U) \left(\sum_{i=1}^{I} (D_i \ominus D_i)^2 \right)^+ \text{horzcat}(D_i \ominus D_i)\, (\mathrm{I}_I \otimes U' \otimes U').$$

These results imply that our estimator of $W'U$ is consistent and asymptotically normal. Moreover, we have

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \sqrt{N}(\widetilde{\theta} - \theta_0) + \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \vartheta_0\, (e_k' \otimes \mathrm{I}_I)\, \iota_n + o_P(1),$$

where we use the notation $\mathrm{I}_K = (e_1, e_2, \ldots, e_K)$, and

$$\iota_n \equiv (U' \otimes \mathrm{I}_I)\, J_{W'} \upsilon_n + (\mathrm{I}_K \otimes W')\, \gamma_n, \qquad \vartheta_0 \equiv 2\, \mathbb{E}[\varphi(x_m) u_k' W A(x_m)].$$

The function ι_n is the influence function of our estimator of $W'U$. Premultiplication with $(e_k' \otimes \mathrm{I}_I)$ turns this into the influence function of our estimator of $W'u_k$. The $\dim \varphi \times I$ matrix ϑ_0, finally, translates estimation uncertainty in $\widehat{\tau}_k$ into an asymptotic-variance contribution for $\widehat{\theta}$.

Theorem 2 follows.

THEOREM 2 (ASYMPTOTIC DISTRIBUTION). *Let Assumptions 1–3 hold and let $\varphi\varphi' f$ be integrable. Then*

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{L} \mathcal{N}(0, \mathscr{V}),$$

where $\mathscr{V}$ is the covariance matrix of

$$\vartheta_0 \left(e_k' \otimes \mathrm{I}_I\right) \iota_n + \frac{(M-3)!}{M!} \sum_{(m_1,m_2,m_3)} \tau_k(x_{nm_1}, x_{nm_2}) \varphi(x_{nm_3}) - \theta_0,$$

as $N \to \infty$.

Inference on θ_0 can be performed by replacing $\mathscr{V}$ by a consistent plug-in estimator.

As a notable example, Theorem 2 states that

$$\widehat{F}_k(x) \equiv \frac{1}{N} \frac{(M-3)!}{M!} \sum_{n=1}^{N} \sum_{(m_1,m_2,m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2}) \, 1\{x_{nm_3} \leq x\},$$

is a consistent and asymptotically-normal estimator of the component distribution $F_k(x)$ for each $x \in \mathcal{X}$. The estimator is further uniformly consistent, that is,

$$\|\widehat{F}_k - F_k\|_\infty = o_P(1),$$

for $\|\cdot\|_\infty$ the supremum norm, provided the function class $\{\chi(x) : x \in \mathcal{X}\}$ is Euclidean for an envelope $G(x)$ such that $\mathbb{E}[G(x_m)] < \infty$. This follows from Pakes and Pollard (1989), who also provide a definition and many examples of Euclidean classes.

3.3. Extensions

3.3.1. Minimum-distance estimation

Consider a Euclidean parameter $\alpha_0 \in \mathcal{A}$ characterized as the unique solution to the moment equation

$$\mu(\alpha) \equiv \mathbb{E}_k[\varphi(x_m; \alpha)] = 0,$$

where φ is a known vector-valued function (with $\dim \alpha \leq \dim \varphi$). Examples of α_0 include linear or nonlinear functionals of f_k, and the solutions to semiparametric Z-estimation problems. The parameter α_0 can be characterized as the unique solution to the minimization problem

$$\min_{\alpha \in \mathcal{A}} \mu(\alpha)' \Sigma \, \mu(\alpha),$$

where Σ is a positive-definite matrix that defines the relevant metric when the number of moment equations exceeds the dimension of α (see Hansen 1982). Theorem 1 suggests a minimum-distance estimator of α_0 that takes the form

$$\widehat{\alpha} \equiv \arg\min_{\alpha \in \mathcal{A}} \widehat{\mu}(\alpha)' \Sigma \, \widehat{\mu}(\alpha), \quad \widehat{\mu}(\alpha) \equiv \frac{1}{N} \frac{(M-3)!}{M!} \sum_{n=1}^{N} \sum_{(m_1.m_2.m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2}) \, \varphi(x_{nm_3}; \alpha),$$

We now use Theorem 2 to derive the asymptotic distribution of $\widehat{\alpha}$ under the following regularity conditions.

ASSUMPTION 4 (REGULARITY CONDITIONS). *The parameter α_0 is an interior element of the compact set $\mathcal{A} \subset \mathcal{R}^{\dim \alpha}$. The function $\varphi(x;\alpha)$ is twice continuously differentiable in α on $\mathcal{A}$ with derivative $\varphi'(x;\alpha)$, and $\mathbb{E}[\sup_{\alpha\in\mathcal{A}}\|\varphi(x_m;\alpha)\|]$ and $\mathbb{E}[\sup_{\alpha\in\mathcal{A}}\|\varphi'(x_m;\alpha)\|]$ are finite.*

The conditions in Assumption 4 are conventional. The smoothness requirements on φ can be relaxed (see, e.g., Pakes and Pollard 1989).

Together with Assumption 4, Theorem 2 implies that $\widehat{\alpha}$ is consistent estimator of α_0. Furthermore,

$$\sqrt{N}\widehat{\mu}(\alpha_0) \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\mu).$$

where the covariance matrix $\mathscr{V}_\mu$ follows from applying Theorem 2 to $\varphi(\cdot;\alpha_0)$. Proposition 1 then follows readily.

PROPOSITION 1 (MINIMUM-DISTANCE ESTIMATION). *Let Assumptions 1–4 hold. Then*

$$\sqrt{N}(\widehat{\alpha} - \alpha_0) \xrightarrow{L} \mathcal{N}\big(0, (Q'\Sigma Q)^{-1}(Q'\Sigma\mathscr{V}_\mu\Sigma Q)(Q'\Sigma Q)^{-1}\big),$$

as $N \to \infty$, provided that $Q \equiv \mathbb{E}_k[\varphi'(x_m;\alpha_0)]$ has maximal rank and $\mathscr{V}_\mu$ is positive definite.

By standard arguments for minimum-distance estimators, the optimally-weighted estimator of α_0 satisfies

$$\sqrt{N}(\widehat{\alpha} - \alpha_0) \xrightarrow{L} \mathcal{N}\big(0, (Q'\mathscr{V}_\mu^{-1}Q)^{-1}\big).$$

This estimator is obtained from Proposition 1 by setting Σ equal to the inverse of a consistent estimator of $\mathscr{V}_\mu$. Consistent plug-in estimators of Q and $\mathscr{V}_\mu$, and thus of the asymptotic variance of $\widehat{\alpha}$ appearing in the proposition, are easily constructed.

3.3.2. Density estimation

When the component distributions are absolutely continuous, and so the f_k are proper density functions, Theorem 2 can be extended to characterize the asymptotic distribution of nonparametric estimators of the component densities.

We focus on a kernel estimator. Let κ denote a kernel function and let $h > 0$ be a bandwidth. A nonparametric density estimator of $f_k(x)$ is

$$\widehat{f}_k(x) \equiv \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)}\widehat{\tau}_k(x_{nm_1}, x_{nm_2})\,\frac{1}{h}\,\kappa\left(\frac{x_{nm_3} - x}{h}\right).$$

Besides the introduction of the weights, this estimator has the conventional form of a nonparametric density estimator.

We will work with standard kernel functions.

ASSUMPTION 5. *$\kappa : \mathcal{R} \to \mathcal{R}$ is a bounded function that is symmetric around zero, integrates to one, and satisfies $\int_{-\infty}^{+\infty} u^2\,\kappa(u)\,\mathrm{d}x < \infty$.*

Some smoothness and moment requirements are needed to derive asymptotic theory. We use the notation

$$q_k(x) \equiv \mathbb{E}\left[\left(\sum_{(m_1,m_2)}\tau_k(x_{m_1},x_{m_2})\right)^2 \,\middle|\, x_{m_3}=x\right]$$

in the following assumption.

ASSUMPTION 6 (REGULARITY CONDITIONS). *f_k is twice continuously differentiable on its support. f is bounded and $\chi_i^4 f$ is integrable. $q_k f$ is bounded and continuous on its support.*

Proposition 2 summarizes the asymptotic properties of the density estimator.

PROPOSITION 2. *Let Assumptions 1–3 and 5–6 hold. Then $|\widehat{f}_k(x) - f_k(x)| = o_P(1)$ and*

$$\sqrt{Nh}\big[\widehat{f}_k(x) - f_k(x)\big] \xrightarrow{L} \mathcal{N}(\sqrt{c}\,\mu_f, \mathscr{V}_f),$$

as $N\to\infty$, where $c\geq 0$ is a finite constant,

$$\mu_f \equiv \frac{1}{2} f_k''(x)\int_{-\infty}^{+\infty} u^2\kappa(u)\,\mathrm{d}u, \qquad \mathscr{V}_f \equiv M\left(\frac{(M-3)!}{M!}\right)^2 q_k(x)f(x)\int_{-\infty}^{+\infty}\kappa(u)^2\,\mathrm{d}u,$$

provided that $Nh\to\infty$ and $Nh^5\to c$.

In Proposition 2 we allow the bandwidth to vanish at the optimal rate of $N^{-1/5}$. In this case, the density estimator is asymptotically biased. Undersmoothing the estimator, that is, setting $h\propto N^{-r}$ for some $\frac{1}{5}<r<1$ yields

$$N^{(1-r)/2}\big[\widehat{f}_k(x) - f_k(x)\big] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_f).$$

Note that $\mathscr{V}_f$ is a tilted version of the asymptotic variance of a conventional kernel estimator of the marginal density function $f(x)$. The sample variance of

$$\frac{(M-3)!}{M!}\sum_{(m_1,m_2,m_3)}\frac{\widehat{\tau}_k(x_{nm_1},x_{nm_2})}{\sqrt{h}}\,\kappa\left(\frac{x_{nm_3}-x}{h}\right)$$

provides a consistent estimator of $\mathscr{V}_f$.

Bandwidth choice is important for the small-sample performance of kernel estimators. As emphasized by Benaglia, Chauveau, and Hunter (2011), in mixture models it is crucial to allow the bandwidth to be component specific, as the component distributions may be very different. In the current context, automated bandwidth selection can be achieved quite easily. For example, least-squares cross-validation (Rudemo 1982; Bowman 1984) is readily applicable to our estimator. The supplementary material contains a discussion of the implementation of this approach.

4. Numerical experiments

4.1. Implementation

Our approach requires choosing the number I of functions $\chi_1, \chi_2, \ldots, \chi_I$, as well as their functional form. These choices should be consistent with Assumptions 1 and 2, but are otherwise free to be decided. For example, one could set $\chi_i(x_m) = 1\{x_m \in \mathcal{I}_i\}$ for $\{\mathcal{I}_i\}_i$ any collection of intervals that forms a partition of $\mathcal{X}$ (as in Kasahara and Shimotsu 2014). Another choice would be $\chi_i(x_m) = 1\{x_m \leq v_i\}$ for a set of values $v_1, v_2, \ldots, v_I$. In the continuous case, another interesting option would be the leading I members of a class of orthonormal polynomials, such as Jacobi polynomials or Hermite polynomials. We experiment with both indicator functions and orthogonal polynomials in our simulations below.[4]

Although the $\widehat{\tau}_k(x_{nm_1}, x_{nm_2})$, averaged over observations, sum up to one by construction, these weights can be negative. This implies that our estimator of cumulative distribution functions is not necessarily monotonic, and that our kernel estimator needs not be bona fide. It is not straightforward to modify (3.4) in order to impose those restrictions. We therefore suggest to adjust the estimates ex post if a bona fide estimate is desired. Estimated cumulative distribution functions can be adjusted via the re-arrangement procedure of Chernozhukov, Fernández-Val, and Galichon (2009). Density estimates can be made to be bona fide by means of the correction procedure of Gajek (1986). The numerical results in this section were obtained without applying such corrections.

In all our numerical exercises we implemented our density estimator using a standard normal kernel, and we selected the bandwidth by least-squares cross-validation.

4.2. Simulations

We present Monte Carlo results for a design from Levine, Hunter, and Chauveau (2011). Additional results are available in the supplementary material. Data are generated from a two-component mixture of normal location models,

$$f_1(x) = \phi(x), \qquad f_2(x) = \phi(x-3), \qquad \omega = (.30, .70)'.$$

In each of $1,000$ Monte Carlo replications we draw three measurements on 500 observations and estimate the component means $\mu_1 = 1$ and $\mu_2 = 3$, the component distributions $F_1(x) = \Phi(x)$ and $F_2(x) = \Phi(x-3)$, and the component densities. To deal with label swapping in our simulation study we proceed as follows. In each replication we first estimate the mean of each component. We then label the estimated component with the smallest estimated mean as the first mixture component. We carefully checked for label-swapping, and found no mix up. We note that, although it must be taken care of in our simulations, label swapping does not cause any complications for estimation and inference based on our

[4] Because the χ_i affect the asymptotic variance of $\widehat{\theta}$, their choice could, in principle, also be guided by asymptotic-efficiency considerations. We postpone a detailed analysis to future work.

Table 1. Results for component means. Design from Levine, Hunter, and Chauveau (2011). $N = 500, M = 3$. Statistics obtained over $1,000$ replications.

ESTIMATOR	I	BIAS		SD		SE/SD		CR(95%)	
		μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
Two-step estimation									
indicators	5	.0039	.0031	.0595	.0669	1.0080	.9768	.9470	.9470
polynomials	5	−.0024	.0017	.0688	.0345	.9954	1.0274	.9470	.9560
indicators	10	.0038	.0021	.0567	.0414	1.0114	1.0176	.9530	.9570
polynomials	10	.0021	.0011	.0554	.0360	1.0268	.9893	.9490	.9530
EM estimation									
parametric		.0028	.0003	.0501	.0323	—	—	—	—
nonparametric		.0033	.0006	.0501	.0322	—	—	—	—

approach. Label swapping does present an important challenge for inference methods based on resampling algorithms such as bootstrap or jackknife, and on MCMC procedures (see, e.g., Stephens 2000).

The top panel in Table 1 contains the bias and the standard deviation (SD) of our point estimates of the component means μ_1 and μ_2. We also compute the average across simulations of estimated standard errors (SE), and report the ratio of SE to the standard deviation (SE/SD), as well as the coverage rate of 95% confidence intervals (CR(95%)). Standard errors and coverage rates are based on the plug-in estimator of $\mathscr{V}$ in Theorem 2. We provide results for two choices of functions χ_i, namely indicator functions $1\{x_m \leq v_i\}$ for I equidistant points v_i on $[-4, 4]$, and the I leading normalized Chebychev polynomials of the first kind $\cos\{(i-1)\arccos(t\{x_m\})\}/2^{1\{i=1\}}$, where $t\{x\} \equiv (x - (x_{\min} + x_{\max})/2)/((x_{\max} - x_{\min})/2)$ for $x_{\min}$ and $x_{\max}$ the minimum and maximum value of x_{nm} observed in the sample. For each of these choices, we provide results for both $I = 5$ and $I = 10$.

Table 1 shows that point estimates have negligible biases relative to their standard deviations. The table also shows that no choice for χ_i uniformly outperforms the other. Furthermore, the plug-in estimator of the asymptotic variance closely mimics the Monte Carlo variability of the point estimates. As a consequence, the confidence intervals have near perfect coverage. The adequacy of the large-sample approximation is further confirmed by inspecting the empirical distribution of the point estimates. Figure 1 plots the smoothed density of the Studentized point estimates (full line) of the component means constructed using $I = 5$ Chebychev polynomials, together with the reference standard-normal density (dashed line). For both component means, the approximation is very close.

The lower panel in Table 1 is reproduced from Levine, Hunter, and Chauveau (2011). It provides the mean and the standard deviation (over 300 Monte Carlo replications) of the parametric maximum-likelihood estimator and of the nonparametric smoothed likelihood estimator of Levine, Hunter, and Chauveau (2011) of the component means. In the latter case, the component means are estimated as the mean of a kernel density estimator. Levine, Hunter, and Chauveau (2011) used a standard-normal density as kernel and set the bandwidth according to Silverman's rule of thumb (Silverman 1998, Section 3.4). Note that the

Figure 1. Normal approximation to the sampling distribution of estimated component means. Design from Levine, Hunter, and Chauveau (2011). $N = 500, M = 3$. Statistics obtained over $1,000$ replications.

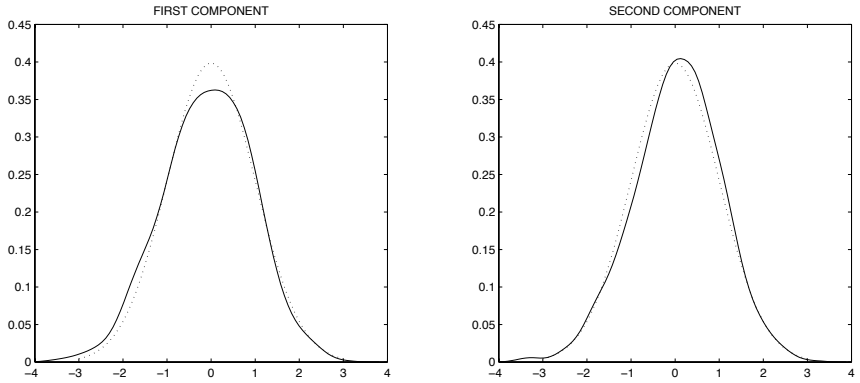


Table 2. Results for component distributions. Design from Levine, Hunter, and Chauveau (2011). $N = 500, M = 3$. Statistics obtained over $1,000$ replications.

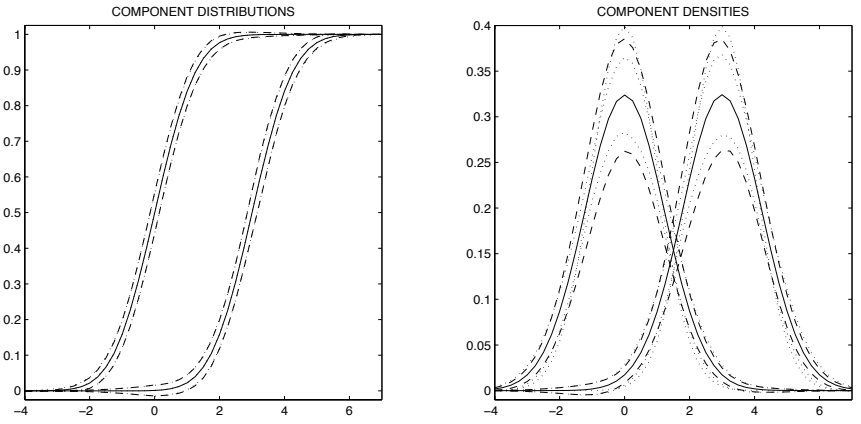
	BIAS		SD		SE/SD		CR(95%)	
x	$F_1(x)$	$F_2(x)$	$F_1(x)$	$F_2(x)$	$F_1(x)$	$F_2(x)$	$F_1(x)$	$F_2(x)$
-1	.0000	.0000	.0199	.0036	.9953	1.0030	.9360	.9570
0	.0000	.0000	.0284	.0078	1.0310	1.0161	.9600	.9600
1	.0000	.0000	.0207	.0106	1.0092	1.0456	.9550	.9540
2	.0000	.0000	.0094	.0206	.9997	.9998	.9390	.9500
3	.0000	.0000	.0039	.0306	.9807	1.0166	.9260	.9420

parametric maximum-likelihood estimator rests on the knowledge of the functional form of the component densities, in contrast with both our approach and the one of Levine, Hunter, and Chauveau (2011). The results show that our procedure produces similar bias as the smoothed likelihood estimator of Levine, Hunter, and Chauveau (2011), but that it is slightly less efficient. A formal comparison of the relative efficiency of the two approaches would require distribution theory for the smoothed likelihood estimator, which is currently not available.

In Table 2 we provide estimation and inference results for the component cumulative distribution functions at a grid of evaluation points x. Biases are again very small relative to standard deviations, and confidence intervals provide excellent coverage. The left plot in Figure 2 summarizes the Monte Carlo results over the whole support. It contains the mean of the point estimates (full lines) and pointwise 95% confidence bands, together with the true distribution functions (dashed lines) and the .025 and .975 quantiles of the empirical distribution of the point estimates (dotted lines). Dashed and dotted lines are virtually identical, reflecting accurate coverage.

The right panel in Figure 2 contains the corresponding results for the density estimator. The estimator is biased downwards around the mode, in line with Proposition 2. At the same

Figure 2. Component distributions and densities. Design from Levine, Hunter, and Chauveau (2011). $N = 500, M = 3$. Statistics obtained over $1,000$ replications.



time, the confidence bands based on a normal asymptotic approximation are comparable to, though slightly wider than, the quantiles of the finite-sample distributions.

4.3. Empirical illustration

We applied our methods to a data set from an experiment in cognitive psychology. These data have also been used by Elmore, Hettmansperger, and Thomas (2004), Benaglia, Chauveau, and Hunter (2009), and Levine, Hunter, and Chauveau (2011) to illustrate their respective approaches. Hence, they provide a useful means of comparison.

The experiment involved 405 children aged between 11 and 16 years that aims to assess childrens' understanding of the physical world. The water-level task they were given is as follows. Each child is presented with rectangular shaped two-dimensional vessels on a sheet of paper, each tilted to a clock-hour orientation. The child is then asked to draw a line representing the surface of still liquid in each of these vessels. The outcome variable is the deviation of the child's line from a horizontal line, in degrees. Drawn lines with a positive slope are recorded as positive deviations and negative slopes are recorded as negative deviations.

We work with four measurements, each corresponding to a clock-hour orientation of the vessel. We use clock orientations one, four, seven, and ten. Histograms of each of the measurements are provided in the supplementary material. Although these plots suggest that our assumption that the four measurements have identical distributions is not unreasonable, there might be a concern that the component distributions corresponding to clock-hour orientations one and seven are different from those associated with clock-hour orientations four and ten. Benaglia, Chauveau, and Hunter (2009) and Levine, Hunter, and Chauveau (2011) fitted a mixture model to these data that allows for the distribution of these two

Table 3. Component means of a three-component mixture fitted to the water-level data of Thomas, Lohaus, and Brainerd (1993).

component	mean estimate	s.e.	p-value
first	−.3336	.6509	.6083
second	6.2642	2.7267	.0216
third	−4.4121	12.8515	.7314

blocks of measurements to be different. Below we discuss similarities and differences between our findings and theirs.

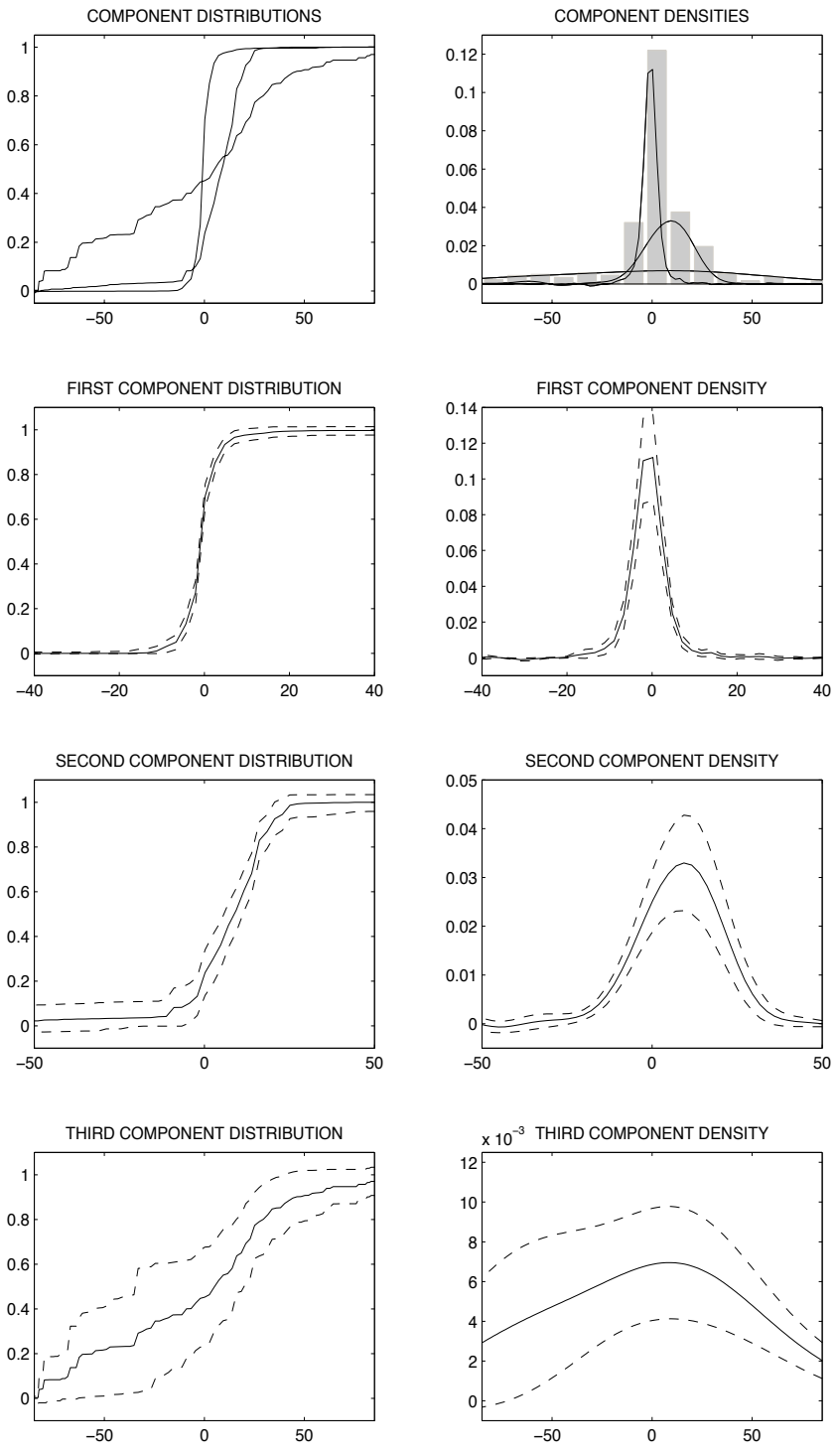
The discussion in Elmore, Hettmansperger, and Thomas (2004) suggests that children can be broadly classified into three latent groups. Therefore, we fit a three-component mixture to these data. We use seven indicator functions for the χ_i. Experimentation with a different number of indicator functions gave qualitatively similar results. Moreover, we set $\chi_i(x_m) = 1\{x_m \leq v_i\}$ where the v_i are Chebychev nodes that have been translated to cover the support of the measurements. The p-value of the Kleibergen and Paap (2006) rank test is .9539, giving strong support in favor of our Assumption 1.

The upper left plot in Figure 3 presents our estimates of the component distributions. The upper right plot contains the corresponding estimates of the component densities, as well as the histogram of the measurements. The estimated distributions are similar to those reported in Elmore, Hettmansperger, and Thomas (2004). One component density is almost degenerate at zero. This first component captures the children who understand that the orientation of water is independent of the orientation of the vessel. A second component has larger variance and is centered slightly above zero. This component seems to capture the children who have a grasp that water has surface and mass, but have some difficulty in consistently getting its orientation right. Finally, the third component distribution is close to a uniform distribution on the whole interval. This last component corresponds to children who do not comprehend the behavior of liquid in a vessel and randomly draw lines at all possible angles.

Compared to Elmore, Hettmansperger, and Thomas (2004), our results allow us to go beyond point estimation and consider inference. The remaining plots in Figure 4.3 contain the point estimates of one of the components (full lines), together with a pointwise 95% confidence band (dashed lines). The confidence bands associated with the first two components indicate that they are fairly accurately estimated. The last component has a somewhat wider confidence band.

In Table 3 we also provide point estimates and standard errors of the mean of each of the fitted component distributions, along with the p-value for the null hypothesis that the mean equals zero, that is, that the orientation of the lines is correct on average. The t-test for the null hypothesis that the second distribution is centered at zero has a p-value of .0216, giving some statistical evidence for an upward bias. We do not reject the null for the other two distributions at any conventional significance level.

Figure 3. A three-component mixture fitted to the water-level data of Thomas, Lohaus, and Brainerd (1993).



Our estimation results differ from those of Benaglia, Chauveau, and Hunter (2009) and Levine, Hunter, and Chauveau (2011) mostly with respect to the third component. Their estimation routine does not impose that the blocks of clock-hour orientations one and seven, and four and ten have the same distribution. They are thus able to estimate a different component for each block. In particular, they find a left-skewed and an antisymmetric right-skewed distribution for the third component of these two blocks (see Panels 2 and 4 in Figure 2 of Levine, Hunter, and Chauveau 2011). Our procedure enforces that all four measurements have the same distribution, which, as a result, seems to bundle up the two distributions corresponding to the third component into one flat one. Extending our approach to allow the measurements to have different distributions is an interesting next step for research.

Acknowledgments

Earlier versions of this paper circulated under the title 'Nonparametric estimation of finite mixtures'. We are grateful to the editor, an associate editor, two referees, and Marc Henry and Bernard Salanié for comments and suggestions.

Bonhomme acknowledges support from the European Research Council through grant ERC-2010-StG-0263107-ENMUH. Jochmans acknowledges support from Sciences Po's SAB. Robin acknowledges support from the Economic and Social Research Council through grant RES-589-28-0001 and from the European Research Council through grant ERC-2010-AdG-269693-WASP.

Supplementary material

Proofs, additional results, and replication files are available as supplementary material and can be downloaded from `econ.sciences-po.fr/staff/koen-jochmans`.

References

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

Benaglia, T., T. Chauveau, and D. R. Hunter (2009). An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics 18*, 505–526.

Benaglia, T., T. Chauveau, and D. R. Hunter (2011). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In D. R. Hunter, D. S. P. Richards, and J. L. Rosenberg (Eds.), *Nonparametrics and Mixture Models: A Festschrift Dedicated to Thomas P. Hettmansperger*. World Scientific.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics 34*, 1204–1232.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*, 353–360.

Bunse-Gerstner, A., R. Byers, and V. Mehrman (1993). Numerical methods for simultaneous diagonalization. *SIAM Journal of Matrix Analysis and Applications 14*, 927–949.

Cardoso, J.-F. and A. Souloumiac (1993). Blind beamforming for non-Gaussian signals. *IEEE-Proceedings, F 140*, 362–370.

Chauveau, D., D. Hunter, and M. Levine (2014). Semi-parametric estimation for conditional independence multivariate finite mixture models. Technical Report 14-02, Department of Statistics, Purdue University.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*, 559–575.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2004). Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. *SIAM Journal of Matrix Analysis and Applications 26*, 295–327.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Eaton, M. L. and D. E. Tyler (1991). On Wielandt's inequality and its applications. *Annals of Statistics 19*, 260–271.

Elmore, R. T., T. P. Hettmansperger, and H. Thomas (2004). Estimating component cumulative distribution functions in finite mixture models. *Communications in Statistics: Theory and Methods 33*, 2075–2086.

Gajek, L. (1986). On improving density estimators which are not bona fide functions. *Annals of Statistics 14*, 1612–1618.

Hall, P. (1981). On the non-parametric estimation of mixing proportions. *Journal of the Royal Statistical Society, Series B 43*, 147–156.

Hall, P., A. Neeman, R. Pakyari, and R. Elmore (2005). Nonparametric inference in multivariate mixtures. *Biometrika 92*, 667–678.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics 31*, 201–224.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Henry, M., K. Jochmans, and B. Salanié (2013). Inference on mixtures under tail restrictions. Discussion Paper No 2014-01, Department of Economics, Sciences Po.

Hettmansperger, T. P. and H. Thomas (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 811–825.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*, 27–61.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *Annals of Statistics 35*, 224–251.

Kasahara, H. and K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica 77*, 135–175.

Kasahara, H. and K. Shimotsu (2014). Nonparametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society, Series B 76*, 97–111.

Kitamura, Y. (2004). Nonparametric identifiability of finite mixtures. Mimeo.

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

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics 20*, 1350–1360.

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

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley-Blackwell.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1057.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (2 ed.). Springer.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*, 65–78.

Shi, X. (2011). *Blind Signal Processing: Theory and Practice*. Springer.

Silverman, B. W. (1998). *Density Estimation for Statistics and Data Analysis*. CRC Press.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

Thomas, H., A. Lohaus, and C. J. Brainerd (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development 58*, 1–191.

Titterington, D. M. (1983). Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B 45*, 37–46.

Woo, M.-J. and T. N. Sriram (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association 101*, 1475–1486.

Zhu, H.-T. and H. Zhang (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society, Series B 66*, 3–16.