NONPARAMETRIC ESTIMATION OF FINITE MIXTURES

STÉPHANE BONHOMME† KOEN JOCHMANS‡ JEAN-MARC ROBIN§

We study the identification and estimation of component distributions and mixture weights in nonparametric finite mixture models with discrete and continuous outcomes. Assuming that at least three repeated measurements on the outcome are available, we obtain simultaneous-diagonalization systems that point identify all mixture components provided an injectivity condition holds. We propose plug-in estimators that are easy to implement and analyze their large-sample behavior. With discrete outcomes, we obtain asymptotically-normal estimators that converge at the parametric rate. With continuous outcomes, the limit distributions remain normal but the convergence speed becomes that of a nonparametric estimator of a trivariate density function. We provide simulation evidence on the finite-sample performance of our methods. Our main application is the estimation of a static factor model with conditional heteroskedasticity.

Keywords: finite mixture model, nonparametric estimation, heteroskedastic factor model, series expansion, simultaneous-diagonalization system.

INTRODUCTION

Finite mixtures constitute an important modeling tool in statistics. They decompose the marginal probability law of an outcome variable into a weighted sum of laws that are conditional on a latent factor. The mixture weights are the point probabilities of the latent factor. While the factor is restricted to take on a finite number of values, the outcome may be either discrete or continuous. Mixture formulations encompass a large class of problems. They are particularly well suited to address the presence of unobserved heterogeneity. For example, factor models, regime-switching models, latent class models, and measurement-error formulations may all be viewed as mixtures. Fixed- and random-effect panel-data models with discrete unobserved heterogeneity, too, fit this framework.

Application of mixtures in economics, as well as in other sciences, are vast. For example, Heckman and Singer (1984a, 1984b) used such an approach in their analysis of the duration of unemployment spells. Type heterogeneity is also commonly introduced in structural discrete-choice models à la Keane and Wolpin (1997). Panel-data models are mixture models with repeated measurements if the data is stationary. A recent strand in the panel literature has focused on inferring the distribution of fixed effects, rather than simply controlling for them in estimation; examples include Browning and Carro (2007), Evdokimov (2010), and Arellano and Bonhomme (2011). Related formulations of

*Preliminary and incomplete; comments are welcomed. First version January 2012. This version February 3, 2012.
†CEMFI, Madrid; bonhomme@cemfi.es.
‡Sciences Po, Paris; koen.jochmans@sciences-po.org.
§Sciences Po, Paris and University College London; jeanmarc.robin@sciences-po.fr.
interest are error-component models for, say, analyzing the dynamics of earnings processes (see, e.g., Lilard and Willis, 1978, and many others since then), and factor models for disentangling common sources of variation from a time series of economic variables (see, e.g., Fama and French, 1993; Heckman, Stixrud, and Urzua, 2006). Mixture models may also be of use when data collection is imperfect. Choice-based sampling and contaminated-control schemes are obvious applications. Hsieh, Manski, and McFadden (1985), Imbens (1992) and Horowitz and Manski (1995), among others, considered methods for conducting inference from such samples. Learning distributions involving latent variables may also serve auxiliary purposes. Schennach (2004) and Hu (2008), for example, considered inference on finite-dimensional parameters in models with measured covariates. There, identifying restrictions depend on moments involving the latent regressor, which have to be estimated from the data in a prior step. As another example, Chernozhukov, Hahn, and Newey (2006) and Honoré and Tamer (2006) studied partial identification of index parameters in correlated random-effect models with discrete covariates.

The traditional approach to estimating mixtures has been parametric; McLachlan and Peel (2000) provide an overview at booklength. Popular implementation amounts to specifying the component distributions to be, say, multinomial or normal and viewing the problem as a missing data problem, which is subsequently estimated by means of the EM algorithm (Dempster, Laird, and Rubin, 1977). The need to impose strong parametric restrictions can be problematic. For example, one would like the data to reveal the shape of the heterogeneity, rather than having to impose it from the outset; see, e.g., Heckman and Singer (1984a) for a discussion.

Nonparametric identification, as well as estimation, of finite mixtures is the topic of a much more recent literature. In seminal work, Hall and Zhou (2003) established nonparametric identification in a fully discrete two-component mixture, assuming three or more repeated measurements are available. Their positive results have been extended to discrete-choice models with multiple mixture types by Kasahara and Shimotsu (2009). The availability of repeated measurements is important. The mixture problem imposes considerable structure on lower-dimensional submodels. This structure lead to restrictions that, in turn, imply point identification of the mixture components when an injectivity condition is satisfied. This latter condition takes the form of a familiar rank condition, which requires the outcome variable to exhibit no less variation than the latent factor. Two repeated measurements, in general, provide an insufficient amount of restrictions for point identification, although non-trivial bounds may be obtained; see Henry, Kitamura, and Salanié (2010). For situations where the outcome is continuous, Hu (2008) provided constructive identification results in the context of nonlinear models where a regressor is subject to misclassification error. His identification conditions impose a shape restriction but are qualitatively similar as above. Practical estimation methods are required to reconcile these identification results with the data. So far, only few nonparametric estimation approaches for finite mixtures have been proposed. In the discrete case, Hall and Zhou (2003)
provided a weighted bootstrap for their setup but their proposal does not readily generalize to \(K\)-component mixtures. Kasahara and Shimotsu (2009) suggested the possibility for estimation based on their analysis but did not work out details. Hu’s (2008) estimator of the component mixtures in the continuous-outcome case takes the form of a nonlinear transformation of kernel-based density and conditional-mean estimators and may be difficult to implement.

In this paper we consider nonparametric identification and estimation of finite mixtures with discrete and continuous outcomes. Like the references above, we assume the availability of repeated measurements to achieve identification. Moreover, the lower-dimensional submodels imply a set of linear restrictions that give rise to a simultaneous-diagonalization system. This system identifies the component distributions and the mixture weights provided a rank condition is satisfied. With discrete outcomes, our approach is similar in spirit as Kasahara and Shimotsu (2009), and the restrictions relate marginal probabilities to point probabilities of the conditional distributions and the mixture weights. When outcomes exhibit continuous variation, we utilize the observation that well-behaved densities admit series expansions in terms of generalized Fourier coefficients and orthonormal basis functions; see, e.g., Cencov (1962) and, more recently, Horowitz (2011). The joint diagonalization system then involves the generalized Fourier coefficients of the marginal densities and the conditional densities, as well as the mixture weights. Here, the rank condition guaranteeing point identification is stated in terms of differences in the Fourier coefficients of the various component mixtures. Formulating the identification problem as a simultaneous-diagonalization system is constructive as it leads to natural plug-in estimators of the mixture components. Moreover, we base our estimation on fast joint approximate-diagonalization algorithms as originally proposed by Cardoso and Souloumiac (1993) and recently utilized in a different context by Bonhomme and Robin (2009).

The remainder of the paper is organized as follows. The fully discrete case is treated first. We find estimators of the mixture components that converge at the parametric rate and that are asymptotically normal. The case with continuous outcomes is treated next. We first motivate the series expansions of the densities at hand and then show how our approach for the discrete case can be modified to handle continuous variation in the outcome. Here, we again obtain asymptotically-normal estimators of the mixture components although, not surprisingly, their convergence rate will be slower than root-\(N\). Moreover, the rate is analogous to the convergence speed of a nonparametric estimator of a trivariate density. Monte Carlo experiments to assess the small-sample performance of our methods follow. Our main application in the Monte Carlo section deals with a heteroskedastic formulation of a static factor model. Existing approaches to inferring component mixtures in factor models are based on deconvolution methods; see, e.g. Horowitz and Markatou (1996), Meister (2009), and Bonhomme and Robin (2010). Besides being computationally difficult, these approaches require full independence, which may be hard to justify. Overall, we find good performance of our methods, even in small samples. Throughout, we abstract away from covariates, although our results could be
extended to this more complicated setting.

I DISCRETE OUTCOMES

Let $X$ be a latent random variable with probability mass function (pmf) $\omega : \mathcal{X} \mapsto [0, 1]$, where $\mathcal{X} \equiv \{x_1, \ldots, x_K\}$ for some finite and known $K$. Let $Y$ be an outcome variable whose pmf conditional on $X = x_k$ is $p_k : \mathcal{Y} \mapsto [0, 1]$, where $\mathcal{Y}$ is a set with cardinality $S$. Without loss of generality we may take $\mathcal{Y} \equiv \{1, \ldots, S\}$. It should be stresses that, in practice, the $p_k$ need not have the same support. We will work under the presumption that we have at our disposal $T \geq 3$ measurements on $Y$, $(Y_1, \ldots, Y_T)$. Assuming them to be i.i.d. given $X = x$ for all $x \in \mathcal{X}$, their joint (marginal) pmf takes the form

$$p(y_1, \ldots, y_T) = \sum_{k=1}^{K} \left\{ \prod_{t=1}^{T} p_k(y_t) \right\} \omega(x_k),$$

which is a finite mixture. Our aim is to recover the mixture components, that is, the conditional pmfs $p_k$ and the weights $\omega(x_k)$ for all $k = 1, \ldots, K$. Three measurements on $Y$ suffice for identification, but additional observations aid in estimation.

1.1 Identification

Write $P(y)$ for the $S \times S$ matrix whose $(s, r)$th entry is $p(s, r, y)$; note that this matrix is symmetric. Let $\Omega(y) \equiv \text{diag}[\omega(x_1)p_1(y), \ldots, \omega(x_K)p_K(y)]$ and define $P_S$ to be the $S \times K$ matrix whose $(s, k)$th entry is $p_k(s)$. Then Equation (1.1) yields

$$P(s) = P_S \Omega(s) P_S', \quad P \equiv \sum_{s=1}^{S} P(s) = P_S \Omega P_S', \quad \Omega \equiv \sum_{s=1}^{S} \Omega(s),$$

as $\sum_s p_k(s) = 1$ for each $k = 1, \ldots, K$. Observe that the matrix $P$ contains the bivariate probabilities $p(s, r)$, for $s, r = 1, \ldots, S$, while the matrix $\Omega = \text{diag}[\omega(x_1), \ldots, w(x_K)]$ bundles the mixture weights.

Identification of the mixture components rests on Assumption 1.1.

Assumption 1.1. The matrices $P_S$ and $\Omega$ satisfy $\text{rank}[P_S] = K$ and $\det[\Omega] > 0$, respectively.

The first condition states that, for each $(k, j) \in \{1, \ldots, K\}$, $p_k(s) \neq p_j(s)$ for at least one $s \in \{1, \ldots, S\}$, unless $k = j$. Obviously, it requires $Y$ to exhibit no less variation than $X$, in the sense that $S \geq K$ must hold. The second condition states that $\omega(x_k) > 0$ for all $k$, and thus ensures that (1.1) is a proper $K$-component mixture.

Assumption 1.1 implies that $P$, which is nonparametrically identified, factors as $P = Q\Lambda Q'$ for an orthogonal $S \times K$ matrix $Q$ and an associated diagonal $K \times K$ matrix $\Lambda$. Furthermore, as $CP' = I_K$ for $C \equiv \Lambda^{-1/2}Q'$,

$$M(y) \equiv CP(y)C' = V[\Omega^{-1}\Omega(y)]V', \quad V \equiv CP_S\Omega^{1/2}, \quad [4]$$
for each \( y \in \mathcal{Y} \), by the linear restrictions summarized in (1.2). Moreover, this decomposition is unique, thereby point-identifying \( V \) as well as the collection \( \{ p_k(y) \}_{k=1}^K \), as the diagonal elements of the matrix \( \Omega^{-1}\Omega(y) \). Now, with \( D \equiv QA^{1/2} = C''\Lambda \), \( DV = P_S\Omega^{1/2} \). As the columns of \( P_S \) sum up to one,

\[
\iota'_a DV = \iota'_K \Omega^{1/2}, \quad DV\Omega^{-1/2} = P_S,
\]

where \( \iota_a \) is a vector of ones of length \( a \). On squaring, the first equation in (1.3) provides a row vector containing the mixture weights. The second equation conveniently yields the point probabilities of the pmf \( p_k \) \((k = 1, \ldots, K)\) as the \( k \)th column of the matrix \( P_S \). Identification of the probability mass function \( p_k \) implies identification of all its functionals.

This identification argument is constructive. Estimators of the mixture components can be formed based on Equation (1.3), using plug-in estimators for \( C, D \) and \( V \). The matrices \( C \) and \( D \) can be estimated by means of an eigenvalue/eigenvector or a singular-value decomposition of a nonparametric estimator of \( P \). The matrix \( V \) can be estimated through a joint approximate-diagonalization estimation (JADE) procedure applied to sample counterparts of the \( S \) matrices \( M(s) = CP(s)C' \); see Cardoso and Souloumiac (1993) and Bonhomme and Robin (2009). JADE is fast and easy to implement.

### 1.2 Estimation

Throughout, let \( \{ Y_{1n}, \ldots, Y_{Tn} \}_{n=1}^N \) denote a random sample on \((Y_1, \ldots, Y_T)\). Let \( \hat{P}(y) \) indicate a nonparametric estimator of \( P(y) \). The simplest such estimator is just the matrix of all the sample proportions. Given the symmetry of the joint pmf implied by the mixture structure, however, a more efficient estimator has

\[
\hat{p}(s,r,y) \equiv \frac{1}{|\rho|N} \sum_{n=1}^N \sum_{(i,j,l) \in \rho} 1\{Y_{in} = s, Y_{jn} = r, Y_{ln} = y\}
\]

as its \((s,r)\)th entry. Here, \( \rho \) is the set of the \(|\rho| = T!/(T-3)! \) ordered triplets of distinct integers that can be formed from the set \( \{1, \ldots, T\} \). Besides an efficiency gain, this approach also enforces symmetry on the sample counterparts of the \( P(y) \). More advanced methods such as nonparametric smoothing methods, which may be a fruitful approach to follow when the data is sparse, could also be used; see, e.g., Aitchison and Aitken (1976) and Hall (1981), or Li and Racine (2007) [Chapter 4]. We leave the particular choice of \( \hat{P}(y) \) implicit, requiring only that it is asymptotically linear.

That is, for each \( s = 1, \ldots, S \), the estimator \( \hat{P}(s) \) is symmetric and satisfies

\[
\text{vec}[\hat{P}(s) - P(s)] = N^{-1} \sum_{n=1}^N \xi_n(s) + o_P(N^{-1}),
\]

\[5\]
where the influence functions, $\xi_n(s)$ \((s = 1, \ldots, S)\), are so that $E[\xi_n(s)] = 0$ and $E[\xi_n(s)\xi_n(s)']$ is finite.

Let $\hat{P}$ be a nonparametric estimator of $P$. For example, $\hat{P} \equiv \sum_{s=1}^{S} \hat{P}(s)$ when sample proportions are used. From Equation \((1.4)\), this estimator, too, is asymptotically linear. Estimators of $C$ and $D$—say, $\hat{C}$ and $\hat{D}$—can then be constructed by means of an eigenvalue/eigenvector decomposition or a singular-value decomposition of $\hat{P}$.

With the smallest eigenvalue of $P$ strictly positive, $\Lambda^{-1}$ is bounded and its estimator is well defined for large $N$. Then, like $\hat{P}$, the estimators $\hat{C}$ and $\hat{D}$ are both asymptotically linear. Let us write

$$\text{vec}(\hat{C} - C) = N^{-1} \sum_{n=1}^{N} \zeta_n^C + o_p(N^{-1}), \quad \text{vec}(\hat{D} - D) = N^{-1} \sum_{n=1}^{N} \zeta_n^D + o_p(N^{-1}),$$

\((1.5)\)

where $\zeta_n^C$ and $\zeta_n^D$ are the respective influence functions.

For each $s = 1, \ldots, S$, construct the matrix $\hat{M}(y) \equiv \hat{C}\hat{P}(y)\hat{C}'$. Using the linear representations under Equations \((1.4)\) and \((1.5)\) it is straightforward to show that

$$\text{vec}([\hat{M}(s) - M(s)]) = N^{-1} \sum_{n=1}^{N} \zeta_n(s) + o_p(N^{-1}),$$

\((1.6)\)

with influence function $\zeta_n(s) \equiv (CP \otimes I_K)\zeta_n^C + (C \otimes C)\xi_n(s) + (I_K \otimes CP)\zeta_n^{C'}$, for all $s = 1, \ldots, S$. Here, $\zeta_n^{C'}$ denotes the influence function of $\text{vec}[\hat{C}']$.

Our JADE estimator of $V$ takes the form

$$\hat{V} \equiv \arg \min_{V \in \mathcal{V}} \sum_{s=1}^{S} \text{off}[V'\hat{M}(s)V]$$

\((1.7)\)

where $\mathcal{V}$ is the set of orthonormal $K \times K$ matrices and $\text{off}[\cdot]$ is the function that maps a square matrix to the sum of its squared off-diagonal elements. In light of Equation \((1.6)\), the large-sample distribution of $\hat{V}$ follows from an application of Theorem 5 in Bonhomme and Robin (2009). To state its asymptotic variance, introduce the $K \times K$ matrices $W$ and $U(1), \ldots, U(S)$, where the \((i, j)\)th elements of $W$ and $U(s)$ are

$$[W]_{i,j} \equiv \begin{cases} \left[ \sum_{s=1}^{S} (p_i(s) - p_j(s))^2 \right]^{-1} & \text{if } i \neq j, \\ 0 & \text{if } i = j, \end{cases} \quad [U(s)]_{i,j} \equiv p_i(s) - p_j(s),$$

respectively. The matrix $W$ is well defined as $p_i(s) - p_j(s)$ is bounded away from zero for each $s = 1, \ldots, S$ and all $i, j = 1, \ldots, K$ with $i \neq j$. For each $s = 1, \ldots, S$, let $H(s) \equiv \text{diag}\{\text{vec}[U(S)]\}$ and introduce $\chi_n \equiv \sum_{s=1}^{S} H(s)(V' \otimes V)\zeta_n(s)$. Then, as $N \to \infty$,

$$\gamma'_V^{-1/2} \sqrt{N} \text{vec}[\hat{V} - V] \overset{d}{\to} N(0, I_{K^2}), \quad \gamma'_V \equiv \mathbb{E}[\zeta_n^{V'}\zeta_n^V],$$

\((1.8)\)
where \( \zeta_n^V \equiv -(I_K \otimes V)(I_K \otimes W)\chi_n \).

We now turn to plug-in estimators of the mixture components based on the equalities in Equation (1.3). First, consider \( \tilde{w} \equiv (I_K \otimes \ell'_S \hat{D}) \text{vec} \hat{V} \) as an estimator of \( w \equiv \Omega^{1/2}r_K \). From Equations (1.5) and (1.8),

\[
\hat{w} - w = N^{-1} \sum_{n=1}^{N} \left[(I_K \otimes \ell'_S D) \zeta_n^V + (V' \otimes \ell'_S) \zeta_n^D \right] + o_p(N^{-1}),
\]

Let \( J \) be the Jacobian of the transformation from \( w \) to \( w^{-1} \) and write \( \hat{T} \) for the \( K^2 \times K \) matrix that turns \( w^{-1} \) into \( \Omega^{-1/2} \). Then

\[
\text{vec} \left[ \hat{\Omega}^{-1/2} - \hat{\Omega}^{-1/2} \right] = N^{-1} \sum_{n=1}^{N} \hat{T}J \left[(I_K \otimes \ell'_S D) \zeta_n^V + (V' \otimes \ell'_S) \zeta_n^D \right] + o_p(N^{-1}),
\]

by an application of the delta method. Plug-in estimators of the mixture components \( \Omega_K \) and \( P_S \) are then given by

\[
\hat{\Omega}_K \equiv \hat{w}^2, \quad \hat{P}_S \equiv \hat{D}\text{vec} \hat{\Omega}^{-1/2},
\]

respectively. These estimators are root-\( N \) consistent and asymptotically normal. Theorem 1 states the asymptotic distributions of \( \hat{\Omega}_K \) and \( \hat{P}_S \).

**Theorem 1.** Under our assumptions, as \( N \to \infty \),

\[
\mathcal{Y}_\Omega^{-1/2}\sqrt{N}[\hat{\Omega}_K - \Omega_K] \xrightarrow{d} \mathcal{N}(0, I_K), \quad \mathcal{Y}_P^{-1/2}\sqrt{N}\text{vec}[\hat{P}_S - P_S] \xrightarrow{d} \mathcal{N}(0, I_{SK}),
\]

where \( \mathcal{Y}_\Omega \equiv \mathbb{E}[\zeta_n^\Omega \zeta_n^\Omega] \) and \( \mathcal{Y}_P \equiv \mathbb{E}[\zeta_n^P \zeta_n^P] \) with

\[
\zeta_n^\Omega \equiv 2\Omega^{1/2} \left[(I_K \otimes \ell'_S D) \zeta_n^V + (V' \otimes \ell'_S) \zeta_n^D \right],
\]

\[
\zeta_n^P \equiv \left[\left(\Omega^{-1/2} \otimes I_K\right) + (I_K \otimes DV)TJ \left(V' \otimes \ell'_S\right)\right] \zeta_n^D + \left[(I_K \otimes DV)TJ (I_K \otimes \ell'_S D) + (\Omega^{-1/2} \otimes D)\right] \zeta_n^V,
\]

respectively. Note that an alternative estimator for the \( s \)th row of \( P_S \) would be the diagonal elements of the matrix \( \hat{V}M(s)\hat{V} \). Asymptotic properties for this estimator are easily established given the results stated above.

Asymptotics for functionals of \( p_k (k = 1, \ldots, K) \) follow from Theorem 1. For example, the \( r \)th raw moment of \( p_k \) is

\[
\mu_r(k) \equiv \mathbb{E}[Y^r | X = x_k] = \sum_{s=1}^{S} s \ p_k(s).
\]

So, an estimator of \( \mu_r \equiv (\mu_r(k), \ldots, \mu_r(K))' \) is \( \tilde{\mu}_r \equiv (I_K \otimes \tau_r^S) \text{vec}[\hat{P}_S] \) for \( \tau_r^S \equiv (1^r, \ldots, S^r) \). From Theorem 1 we obtain

\[
\mathcal{Y}_{\mu_r^{-1/2}}\sqrt{N}(\hat{\mu}_r - \mu_r) \xrightarrow{d} \mathcal{N}(0, I_K),
\]

where the asymptotic variance takes the form \( \mathcal{Y}_{\mu_r} \equiv (I_K \otimes \tau_r^S) \mathcal{Y}_P (I_K \otimes \tau_r^S) \).

[7]
Suppose that \( Y \) is continuously distributed on a compact interval \( \mathcal{Y} \). Without loss of generality we set \( \mathcal{Y} = [-1, 1] \). Let \( f_k \) be the probability density function (pdf) of \( Y \) given \( X = x_k \). With \( T \geq 3 \) repeated measurements on \( Y \), our mixture model takes the form
\[
f(y_1, \ldots, y_T) = \sum_{k=1}^{K} \left\{ \prod_{t=1}^{T} f_k(y_t) \right\} \omega(x_k),
\]
where \( \omega(x_k) \) is the mixture weight for each \( k = 1, \ldots, K \).

The goal is again to recover the densities \( f_k \) and the mixture weights \( \omega(x_k) \) for each \( k = 1, \ldots, K \).

Our analysis is based on (generalized) Fourier expansions—and subsequent series estimation—of \( f_k \). For this to be fruitful, we restrict attention to the class of functions that satisfy the following assumption.

**Assumption 2.1.** The functions \( f_k, k = 1, \ldots, K \), are absolutely continuous and square-integrable.

Let \( \{\psi_\ell\}_{\ell=1}^\infty \) denote a complete orthonormal basis for \( L^2[\mathcal{Y}] \). Then Assumption 2.1 implies that
\[
\lim_{L \to \infty} \|f_k(y; L) - f_k(y)\| = 0, \quad f_k(y; L) \equiv \sum_{\ell=1}^{L} a_{\ell k} \psi_\ell(y),
\]
where \( L \) is a positive integer, \( \|\cdot\| \) denotes the \( L^2 \) norm, and
\[
a_{\ell k} \equiv \langle f_k, \psi_\ell \rangle = \int_{\mathcal{Y}} \psi_\ell(y) f_k(y) \, dy
\]
is the \( \ell \)th (generalized) Fourier coefficient of \( f_k(y) \). We impose the following high-level assumption.

**Assumption 2.2.** The basis functions \( \{\psi_\ell\}_{\ell=1}^\infty \) are bounded and absolutely integrable, and
\[
\|f_k(y; L) - f_k(y)\| = O(L^{-\alpha})
\]
for each \( k = 1, \ldots, K \).

Assumption 2.2 is conventional in nonparametric curve estimation by series expansions; see e.g., Efromovich (1999) for a treatment at booklength. It has recently been imposed in other contexts, such as in nonparametric instrumental-variable estimation; see, e.g., Hall and Horowitz (2005) and Horowitz (2007). Assumption 2.2 is compatible with a large variety of choices for the basis functions; Chen (2007) provides examples. The order of magnitude of the approximation error can be motivated through a smoothness condition on the function \( f_k \), demanding it to be \( \alpha \)-smooth. Alternatively, it may be interpreted as a restriction on the shrinkage rate of the Fourier coefficients, i.e., \( a_{\ell k} = O(\ell^{-\alpha}) \); see, e.g., Hall (1986) for more on this.
2.1 Identification

We can proceed as in Equation (2.2) and similarly define truncated versions of the bivariate and trivariate (marginal) pdfs $f(y_1, y_2)$ and $f(y_1, y_2, y_3)$, respectively. Moreover, for Fourier coefficients $b_{\ell_1\ell_2}$ and $b_{\ell_1\ell_2\ell_3}$ ($\ell_1, \ell_2, \ell_3 = 1, \ldots, L$), say,

\[
\begin{aligned}
    f(y_1, y_2; L) &\equiv \sum_{\ell_1=1}^L \sum_{\ell_2=1}^L b_{\ell_1\ell_2}\psi_{\ell_1}(y_1)\psi_{\ell_2}(y_2), \\
    f(y_1, y_2, y_3; L) &\equiv \sum_{\ell_1=1}^L \sum_{\ell_2=1}^L \sum_{\ell_3=1}^L b_{\ell_1\ell_2\ell_3}\psi_{\ell_1}(y_1)\psi_{\ell_2}(y_2)\psi_{\ell_3}(y_3),
\end{aligned}
\]

(2.3)

converge to $f(y_1, y_2)$ and $f(y_1, y_2, y_3)$, respectively, as $L \to \infty$.

Observe that knowledge of the sequence $\{a_{\ell k}\}_{\ell=1}^\infty$ implies identification of $f_k(y)$ for each $y \in \mathcal{Y}$. The key insight underlying our approach is that the mixture structure implies restrictions on the $a_{\ell k}$, $b_{\ell_1\ell_2}$, and $b_{\ell_1\ell_2\ell_3}$ that are akin to those in Equation (1.2). Moreover, for any chosen $L$, on recycling constructive notation, Equations (2.1)–(2.3) imply that

\[
B(\ell) = A\Omega(\ell)A', \quad B = A\Omega A',
\]

(2.4)

where now $\Omega(\ell) \equiv \text{diag}[w(x_1)a_{1\ell}, \ldots, w(x_K)a_{K\ell}]$, $\Omega \equiv \text{diag}[w(x_1), \ldots, w(x_K)]$, $A$ is the $L \times K$ matrix containing the first $L$ Fourier coefficients of the conditional pdfs

\[
A \equiv \begin{pmatrix}
    a_{11} & a_{12} & \ldots & a_{1K} \\
    a_{21} & a_{22} & \ldots & a_{2K} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{L1} & a_{L2} & \ldots & a_{LK}
\end{pmatrix},
\]

and $B(\ell)$ ($\ell = 1, \ldots, L$) and $B$ are the $L \times L$ symmetric matrices

\[
B(\ell) \equiv \begin{pmatrix}
    b_{11\ell} & b_{12\ell} & \ldots & b_{1L\ell} \\
    b_{21\ell} & b_{22\ell} & \ldots & b_{2L\ell} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{L1\ell} & b_{L2\ell} & \ldots & b_{LL\ell}
\end{pmatrix}, \quad B \equiv \begin{pmatrix}
    b_{11} & b_{12} & \ldots & b_{1L} \\
    b_{21} & b_{22} & \ldots & b_{2L} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{L1} & b_{L2} & \ldots & b_{LL}
\end{pmatrix}.
\]

The matrices $B(\ell)$ ($\ell = 1, \ldots, L$) and $B$ are all nonparametrically identified. Our strategy is to recover the matrices $A$ and $\Omega$ using the linear restrictions in Equation (2.4). Given the matrix $A$, we can construct

\[
F(y; L) \equiv (f_1(y; L), \ldots, f_K(y; L))' = \Gamma' A,
\]

for $\Gamma \equiv (\psi_1(y), \ldots, \psi_L(y))'$. By Assumption 2.2, $F(y; L)$ converges to $F(y) \equiv (f_1(y), \ldots, f_K(y))'$ as $L \to \infty$.

The following assumption serves a purpose analogous to Assumption 1.1 in the discrete case.
Assumption 2.3. The matrices $A$ and $\Omega$ satisfy $\text{rank}[A] = K$ and $\det[\Omega] > 0$, respectively.

The rank condition requires the functions $f_k$ to be sufficiently different across $k = 1, \ldots, K$. As such, it is a relevance condition in the sense that $(Y, X)$ need to be dependent. The condition again requires $L \geq K$ and thus places a restriction on the number of non-zero Fourier coefficients of the $f_k(y)$ vis-à-vis the cardinality of $\mathcal{X}$.

Identification of the matrices $\Omega$ and $A$ for any fixed $L$ may be established in a manner similar to in the discrete case. On re-using notation from above to keep the link as apparent as possible, we first factor $B$ as $Q\Lambda Q'$ to obtain the simultaneous-diagonalization system

$$M(\ell) \equiv CB(\ell)C' = V[\Omega^{-1}\Omega(\ell)]V', \quad C \equiv \Lambda^{-1/2}Q', \quad V \equiv CA\Omega^{1/2},$$

for each $\ell = 1, \ldots, L$. By Assumption 2.3, $V$ and $\Omega^{-1}\Omega(\ell) = \text{diag}[a_{\ell 1}, \ldots, a_{\ell K}]$ are identified. Thus, $DV = A\Omega^{1/2}$ for $D \equiv QA^{1/2}$ is identified. From this expression, we can recover the matrix of mixture weights $\Omega$.

2.2 Estimation

Let $\{Y_{1n}, \ldots, Y_{Tn}\}_{n=1}^N$ again be an i.i.d. sample of $T$ measurements on $Y$. Let $L = L(N)$ grow with $N$; rates are given below. For each $\ell = 1, \ldots, L$, the entries of the matrix $B(\ell)$ can be consistently estimated by means of samples averages over the basis functions. Again invoking symmetry to enhance efficiency, the $(\ell_1, \ell_2)$th entry of $\hat{B}(\ell)$ is

$$\hat{b}_{\ell_1\ell_2\ell} \equiv \frac{1}{|\rho|N} \sum_{n=1}^N \sum_{(i,j,\ell) \in \rho} \psi_{\ell_1}(Y_{in})\psi_{\ell_2}(Y_{jn})\psi_{\ell}(Y_{ln}),$$

where, recall, $\rho$ is the set of ordered triplets of distinct integers from the set $\{1, \ldots, T\}$. Clearly, each of the entries of this matrix is unbiased and root-$N$ consistent for the corresponding entry in $B(\ell)$, i.e., $\hat{b}_{\ell_1\ell_2\ell} = b_{\ell_1\ell_2\ell} + O_p(N^{-1/2})$. We estimate $B$ by $\hat{B}$, which is defined as $\hat{B}(\ell)$ but with $\psi_{\ell}(\cdot)$ omitted from each entry.

Diagonalization of $\hat{B}$ provides estimators of $C$ and $D$, say $\hat{C}$ and $\hat{D}$, which are of dimension $K \times L$ and $L \times K$, respectively. We then proceed by estimating the matrix that approximately diagonalizes the $\hat{M}(\ell) \equiv \hat{C}\hat{B}(\ell)\hat{C}'$, $\ell = 1, \ldots, L$. That is, we compute the $K \times K$ orthonormal matrix

$$\hat{V} \equiv \arg \min_{V \in \mathcal{V}} \sum_{\ell=1}^L \text{off}[V'\hat{M}(\ell)V],$$

where $\mathcal{V}$ is as before, using fast diagonalization techniques. Because $L \to \infty$ as $N \to \infty$ and $\hat{M}(\ell)$ depends on $\ell$ only through $\hat{B}(\ell)$, the estimation noise in $\hat{C}$ will not contribute to the asymptotic
variance of $\hat{V}$. Moreover, proceeding along the lines of the proof to Theorem 5 in Bonhomme and Robin (2009), we find

$$\text{vec}[\hat{V} - V] = -(I_K \otimes V) (I_K \otimes W) \chi_L + O_p(LN^{-1/2}),$$

(2.6)

where

$$\chi_L \equiv \sum_{\ell=1}^L H(\ell) (V' \otimes V') (C \otimes C) \text{vec}[\hat{B}(\ell) - B(\ell)]$$

and, in line with before, $H(\ell) \equiv \text{diag}\{\text{vec}[U(\ell)]\}$ while $W$ and $U(1), \ldots U(L)$ are the $K \times K$ matrices whose $(i,j)$th elements are

$$[W]_{i,j} \equiv \begin{cases} \left[ \sum_{\ell=1}^L (a_{\ell i} - a_{\ell j})^2 \right]^{-1} & \text{if } i \neq j, \\ 0 & \text{if } i = j, \end{cases} \quad [U(\ell)]_{i,j} \equiv a_{\ell i} - a_{\ell j},$$

respectively.

First, observe that the matrix $\lim_{L \to \infty} W$ is well defined. In particular, by orthonormality of $\{\psi_\ell\}$,

$$\lim_{L \to \infty} \sum_{\ell=1}^L (a_{\ell i} - a_{\ell j})^2 = \int \|f_i(y) - f_j(y)\|^2 \, dy = \|f_i(y) - f_j(y)\|^2,$$

for each $i, j = 1, \ldots, K$. Assumption 2.3 implies that these quantities are bounded away from zero, and so the entries of $W$ are bounded away from infinity as $L \to \infty$.

Next, note that $\chi_L$ takes the form of the sample average $\chi_L = N^{-1} \sum_{n=1}^N \chi_{Ln}$, where the sequence $\{\chi_{Ln}\}$ is the triangular array of $K^2 \times 1$ vectors $\chi_{Ln} \equiv \text{vec}[\mathcal{X}_n]$, where the matrix $\mathcal{X}_n$ has as its $(i,j)$th entry

$$\mathcal{X}_{ijn} \equiv \frac{1}{|\rho|} \sum_{\ell_1=1}^L \sum_{\ell_2=1}^L \sum_{\ell_3=1}^L \sum_{(i,j,l) \in \rho} (a_{\ell_3 i} - a_{\ell_3 j}) v_i' c_{\ell_1} (\psi_{\ell_1}(Y_{in}) \psi_{\ell_2}(Y_{jn}) \psi_{\ell_3}(Y_{jn}) - b_{\ell_1 \ell_2 \ell_3}) c_{\ell_2}' v_j,$$

for $v_i$ and $c_i$ the $i$th column of $V$ and $C$, respectively. Impose the following restriction on the growth rate of $L$.

**Assumption 2.4.** The truncation parameter $L$ satisfies $L = o(N^{1/3})$.

As $\mathbb{E}[\mathcal{X}_{ijn}] = 0$ and $\sigma_{ijn}^2 \equiv \mathbb{E}[\mathcal{X}_{ijn}^2] = O(L^3)$ by orthonormality of the basis functions, a weak law of large numbers for triangular arrays yields $\chi_L \xrightarrow{p} 0$ as $N \to \infty$; see, e.g., Davidson (1994) [Theorem 19.11]. Next, impose Lyapounov’s condition.

**Assumption 2.5.** We have that $\sum_{n=1}^N \mathbb{E}|\mathcal{X}_{ijn}|^{2+\delta} = o(\sum_{n=1}^N \sigma_{ijn}^2)^{2+\delta}$ for some $\delta > 0$ and each $i,j = 1, \ldots, K$. 

[11]
This requirement implies primitive conditions in our specific setup but, for now, we simply state it as a high-level assumption. Assumption 2.5 allows a central limit theorem for triangular arrays to be applied to \( \bar{X}_L \); see, e.g., Davidson (1994) [Theorems 23.6 and 23.11]. Moreover, we have the convergence result

\[
\Sigma_L^{-1/2} \sqrt{N} \chi_L \xrightarrow{d} \mathcal{N}(0, I_{K^2}),
\]

where \( \Sigma_L \equiv \mathbb{E}[\chi_{Ln} \chi_{Ln}'] \).

From Equation (2.6) we then obtain that, as \( N \to \infty \),

\[
\mathcal{W}_V^{-1/2} \sqrt{N} \text{vec}[\bar{V} - V] \xrightarrow{d} \mathcal{N}(0, I_{K^2}),
\]

\[
\mathcal{W}_V \equiv \mathbb{E}[(\bar{\zeta}_n \zeta_n V)],
\]

for \( \zeta_n V \equiv -(I_K \otimes V)(I_K \otimes W) \chi_{Ln} \).

Several approaches may then be followed to infer the mixture components. A first approach would be to estimate the entries of \( A \) by the diagonal elements of \( \hat{V}^{'} \hat{M}(\ell) \hat{V} \) for each \( \ell = 1, \ldots, L \) and to bundle the results in the matrix \( \tilde{A} \), say. We could then proceed by estimating the diagonal elements of \( \Omega^{1/2} \) as \( (I_K \otimes \iota_L^{'})^{-1}(I_K \otimes \iota_L^{'}) \text{vec}[\hat{V}] \). However, this is computationally cumbersome and leads to an estimator of the mixture components that would converge like \( L^4 N^{-1} \).

A second strategy is more similar to the one followed in the discrete case. The first variant of this approach is valid for any set of basis functions satisfying our assumptions. As densities integrate to unity, we have the equality

\[
\int_{\mathcal{Y}} f_k(y) \, dy = \int_{\mathcal{Y}} \sum_{\ell=1}^{\infty} a_{\ell k} \psi_\ell(y) \, dy = \sum_{\ell=1}^{\infty} a_{\ell k} \phi_\ell = 1, \quad \phi_\ell \equiv \int_{\mathcal{Y}} \psi_\ell(y) \, dy.
\]

The \( \phi_\ell \) can be calculated. A second variant is based on approximating the above integral by quadrature methods. A computationally particularly convenient version of this procedure is obtained on specifying the basis functions to be Chebychev polynomials; see, e.g., Gil, Segura, and Temme (2007) [Chapter 3] for an overview of approximation methods using Chebychev polynomials. In our setting, the \( \ell \)th such polynomial takes the form

\[
\psi_\ell(y) = \frac{2}{\pi} \frac{1}{2^{1(\ell-1=0)}} \frac{1}{\sqrt{1 - y^2}} \cos[(\ell - 1) \cos^{-1}(y)].
\]

The Clenshaw and Curtis (1960) quadrature approximation to \( \int_{\mathcal{Y}} f_k(y) \, dy \) based on \( L \) Chebychev nodes is simply

\[
\sum_{\ell=1}^{L} \varphi_\ell a_{\ell k}, \quad \varphi_\ell = \begin{cases} \frac{2}{\pi(\ell-1)^2} & \text{if } \ell \text{ is uneven} \\ 0 & \text{otherwise} \end{cases},
\]

where \( L \) is taken to be an even integer. While Gaussian quadrature is theoretically superior to Clenshaw-Curtis in terms of approximation error, it is more difficult to implement. Furthermore, in practice, the Clenshaw-Curtis rule does not appear to be drastically dominated by Gaussian
quadrature; see, e.g., Trefethen (2008) for a discussion. Besides the choice between \( \phi \) and \( \varphi \), the remainder of our approach is identical. To describe it, let \( C \) be the \( L \times 1 \) vector collecting the first \( L \) coefficients, that is, \( C = (\phi_1, \ldots, \phi_L)' \) or \( C = (\varphi_1, \ldots, \varphi_L)' \).

Fix \( y \) to some value in \( Y \) and, throughout, leave the dependence of quantities on it implicit. Assumption 2.2 implies that \( C' A = \iota_K' + O(L^{-\alpha}) \), and also that \( F'_L - F' = O(L^{-\alpha}) \). Whence, from above,

\[
C' DV = w' + O(L^{-\alpha}), \quad \Gamma' DV \Omega^{-1/2} = F' + O(L^{-\alpha}).
\]

Plug-in estimators of \( w \) and \( F \) can then be formed as

\[
\hat{w} \equiv (I_K \otimes C' D) \text{vec}[\hat{V}], \quad \hat{F} \equiv \Omega^{-1/2} \hat{V}' \hat{D} \Gamma,
\]

where \( \hat{\Omega}^{-1/2} \equiv \text{diag}[\hat{w}_1, \ldots, \hat{w}_K]^{-1} \).

To obtain correctly-centered asymptotic distributions for these estimators, the following rate on \( L \) is required.

**Assumption 2.6.** \( L \) is such that, as \( N \to \infty \), (i) \( L^3 N^{-1} \to 0 \) and (ii) \( NL^{-3-2\alpha} \to 0 \)

The first restriction implies the variance of the estimators to converge to zero. It is a repetition of Assumption 2.4. The second restriction ensures truncation bias to be asymptotically negligible. If we set \( L \propto N^\beta \) (rounded to the nearest integer value) for some \( \beta \), then Assumption 2.6 requires that

\[
\beta \in \left( \frac{1}{3+2\alpha}, \frac{1}{3} \right).
\]

Observe that this is the conventional interval for nonparametric estimation of a trivariate density function that is \( \alpha \)-smooth. Smoother functions allow for a slower growth rate. Obviously, in our setting, \( \alpha \) relates to the least smooth of the \( f_k \) (\( k = 1, \ldots, K \)). Stone’s (1982) optimal rate is obtained on setting \( \beta = 1/(3+2\alpha) \), but this would lead to an asymptotic bias. So, Assumption 2.6 demands undersmoothing, and implies a convergence rate that can be arbitrarily close to \( N^{-\alpha/(3+2\alpha)} \).

Note that the entries of \( C \) and \( \Gamma \) remain well defined as \( L \to \infty \). Observe that estimation noise in \( \hat{D} \) is again asymptotically negligible to obtain

\[
\hat{w} - w = -(I_K \otimes C' D) \zeta_n^V + O(L^{-\alpha}) + o_p(L^{3/2} N^{-1/2}).
\]

Let \( J \) be the \( K \times K \) Jacobian of the transformation of \( w \) to \( w^{-1} \) and write \( T \) for the \( K^2 \times K \) matrix that turns \( w \) into \( \text{vec}[\Omega^{1/2}] \). Then,

\[
\text{vec}[\Omega^{1/2} - \Omega^{-1/2}] = -J (I_K \otimes C' D) \zeta_n^V + O(L^{-\alpha}) + o_p(L^{3/2} N^{-1/2})
\]

by an application of the delta method. These representations, together with Equation (2.7) provide us with the following theorem.

[13]
Theorem 2. Under our assumptions, for each \( y \in \mathcal{Y} \), as \( N \to \infty \),
\[
\mathcal{V}^{-1/2}_\Omega \sqrt{N} [\hat{\Omega}_K - \Omega_K] \xrightarrow{d} \mathcal{N}(0, I_K), \quad \mathcal{V}^{-1/2}_F \sqrt{N} [\hat{F} - F] \xrightarrow{d} \mathcal{N}(0, I_K),
\]
where \( \mathcal{V}_\Omega \equiv \mathbb{E}[\zeta_{\Omega n}^\Omega \zeta_{\Omega n}^\Omega] \) and \( \mathcal{V}_F \equiv \mathbb{E}[\zeta_{F n}^F \zeta_{F n}^F] \) for
\[
\zeta_{\Omega n}^\Omega \equiv 2\Omega^{1/2} (I_K \otimes C'D) \zeta_{\hat{V} n}^V,
\]
\[
\zeta_{F n}^F \equiv [(\Gamma' D \otimes \Omega^{-1/2}) + (\Gamma' D V \otimes I_K) \mathcal{J} (I_K \otimes C'D)] \zeta_{\hat{V} n}^V,
\]
respectively.

Estimators of the moments of the \( f_k \) follow. For example, an estimator of the \( r \)th raw moment,
\[
\mathbb{E}[Y^r | X = x_k] = \int_\mathcal{Y} y^r f_k(y) \, dy,
\]
can be obtained by replacing \( f_k(y) \) by \( \hat{f}_k(y) \) and approximating the integral by quadrature methods. With \( \overline{L} \) quadrature nodes, the additional bias induced by this approximation is \( O(\overline{L}^{-\alpha}) \). Other methods for obtained estimators of moments may exist. For example, one could start from a series expansion of the moment of interest directly. Because
\[
\mathbb{E}[Y^r] = \sum_{k=1}^K \mathbb{E}[Y^r | X = x_k] \omega(x_k),
\]
this would again lead to a diagonalization system, now in terms of the Fourier coefficients of the moments.

III Monte Carlo Illustrations

We experimented with our methods on simulated data. Here, we report on the performance of our procedures in a three-variate mixture of Beta distributions and in a linear factor model with heteroskedasticity.

3.1 Three-component beta mixture

Our first experiment has the following structure. For \( k \in \{1, 2, 3\} \) and \( y \in \mathcal{Y} \) we set \( f_k(y) = .5f(.5 + .5y; \theta_k, \varrho_k) \), where
\[
f(y; \theta, \varrho) \equiv \frac{1}{\mathcal{B}(\theta, \varrho)} y^{\theta-1}(1 - y)^{\varrho-1}, \quad \mathcal{B}(\theta, \varrho) \equiv \int_0^1 u^\theta (1 - u)^{\varrho-1} \, du,
\]

[14]
is the Beta distribution with positive real shape parameters \((\theta, \vartheta)\). Our design has the parameter choices
\[
(\theta_1, \vartheta_1) = (2, 9), \quad \omega(x_1) = .35,
(\theta_2, \vartheta_2) = (5, 5), \quad \omega(x_2) = .45,
(\theta_3, \vartheta_3) = (5, 2), \quad \omega(x_3) = .20.
\]
This makes \(f_1\) skewed to the right, \(f_2\) symmetric, and \(f_3\) skewed to the left. A plot of these densities is provided in Figure 1 below.

Figure 1: Population densities in the three-component mixture

Marginal density of the data (left) and the individual component mixtures (right).

We used (normalized) Chebychev polynomials as basis functions; see Equation (2.8) above. The growth rate on the truncation parameter was governed by setting \(L = 1.90 * N^{1/6}\) and subsequently rounding off the result to the nearest integer. This scheme is in line with our theoretical demands as postulated in Assumption 2.6. For fixed \(N\), larger values of \(L\) lead to curves that are less smooth. The matrices \(\hat{C}\) and \(\hat{D}\) were computed through a singular-value decomposition of \(\hat{B}\). This was found to be more stable than a simple eigenvalue/eigenvector decomposition. For now, we have no formal way of establishing the number of mixture components, \(K\), in practice. We suggest inspecting the size of the diagonal entries of \(\hat{A}\) as an informal way of selecting \(K\). Under our assumptions, the \((K + 1)\)th entry should be close to zero in reasonably sized samples. In our experiments, this worked well. Clenshaw-Curtis quadrature was used to approximate \(\int f_k(y) \, dy\) in the computation of the mixture weights.

To get an overall grasp of how well our methods perform, Figure 2 below provides 10 estimates of \(f_k\) \((k = 1, 2, 3)\) for \(N \in \{500; 1,000; 2,500; 10,000\}\); \(T\) was set to three throughout. The value of the truncation parameter used is indicated above each plot. The estimated densities were computed at the points \(\cos(i * \pi/51)\) \((i = 0, \ldots, 50)\), which forms a dense set of Chebychev nodes. The plots
reveal that the approach is able to capture the true underlying mixture densities well, even for small $N$ and $L$. The upper-left plot still reveals a fair amount of bias in $\hat{f}_1$. Notably the region around the mode is underestimated while the right tail is too curved, rendering negative density estimates over a subset of $\mathcal{Y}$. The fact that series estimators of densities can produce negative values in small samples is well known; see, e.g., Schwartz (1968). Nevertheless, as the sample grows the performance rapidly improves uniformly over $\mathcal{Y}$. With $N = 10,000$, the estimates virtually co-incide with the true conditional densities.

Figure 2: Realizations of $\hat{f}_k$ in the three-component mixture

Table 1 presents summary statistics obtained over 10,000 Monte Carlo replications. Moreover, it contains the mean bias (Bias), the standard deviation (STD), the median bias (M-bias), and interquartile range of estimators of $f_k(m_k)$ and $\omega(x_k)$, where $m_k$ denotes the mode of $f_k$. Start with the upper panel, which contains the results for $N = 500$. For $L = 5$, in line with Figure 2, there is some bias in the $\hat{f}_k$, in particular for $k = 1$. The mixture weights are estimated with much less bias. Overall, both the STD and the IQR are not particularly large. Increasing $L$ to 10 results
in a substantially smaller bias in the conditional densities. The bias in the weights appears fairly insensitive to this design change. For both \( \hat{f}_k \) and \( \hat{\omega}(x_k) \), setting \( L = 10 \) does lead to an increase in the variability measures. These findings are no surprise, as increasing \( L \) for a given \( N \) makes the estimated conditional densities more spiked, allowing \( \hat{f}_k \) to better fit all data points.

### Table 1: Summary statistics for the three-component mixture

<table>
<thead>
<tr>
<th>( N = 500 )</th>
<th>( \hat{f}_k )</th>
<th>( \hat{\omega}(x_k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L = 5 )</td>
<td>Bias</td>
<td>STD</td>
</tr>
<tr>
<td>( f_1 )</td>
<td>-.3268</td>
<td>.0483</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>-.1458</td>
<td>.0422</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>.0834</td>
<td>.0155</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( L = 10 )</th>
<th>Bias</th>
<th>STD</th>
<th>M-bias</th>
<th>IQR</th>
<th>Bias</th>
<th>STD</th>
<th>M-bias</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_1 )</td>
<td>.0131</td>
<td>.1247</td>
<td>.0092</td>
<td>.1666</td>
<td>-.0141</td>
<td>.0625</td>
<td>-.0129</td>
<td>.0717</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>.0013</td>
<td>.0441</td>
<td>.0006</td>
<td>.0557</td>
<td>.0647</td>
<td>.0673</td>
<td>.0630</td>
<td>.0800</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>-.0041</td>
<td>.0345</td>
<td>-.0028</td>
<td>.0335</td>
<td>-.0044</td>
<td>.0294</td>
<td>-.0032</td>
<td>.0363</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( N = 1,000 )</th>
<th>( \hat{f}_k )</th>
<th>( \hat{\omega}(x_k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L = 6 )</td>
<td>Bias</td>
<td>STD</td>
</tr>
<tr>
<td>( f_1 )</td>
<td>-.0500</td>
<td>.0587</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>-.1035</td>
<td>.0299</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>.0580</td>
<td>.0082</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( L = 12 )</th>
<th>Bias</th>
<th>STD</th>
<th>M-bias</th>
<th>IQR</th>
<th>Bias</th>
<th>STD</th>
<th>M-bias</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_1 )</td>
<td>.0087</td>
<td>.1080</td>
<td>.0081</td>
<td>.1434</td>
<td>-.0054</td>
<td>.0400</td>
<td>-.0059</td>
<td>.0488</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>-.0010</td>
<td>.0415</td>
<td>-.0009</td>
<td>.0418</td>
<td>.0560</td>
<td>.0437</td>
<td>.0565</td>
<td>.0540</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>-.0013</td>
<td>.0160</td>
<td>-.0007</td>
<td>.0165</td>
<td>-.0025</td>
<td>.0195</td>
<td>-.0023</td>
<td>.0263</td>
</tr>
</tbody>
</table>

Statistics computed over 10,000 Monte Carlo replications.

The lower panel of Table 1 shows the effect of increasing \( N \) from 500 to 1,000. We also increase \( L \), from 5 and 10 to 6 and 12, respectively, which is in line with our asymptotic requirements. Bias and volatility measures both go down, as expected. The decrease in bias is most apparent for the \( \hat{f}_k \). Some bias remains in the middle component distribution and in the associated mixture weight but, overall, the performance is very good. We computed summary statistics for \( \hat{f}_k \) evaluated at other points than the mode and found very similar results. This could have been expected from an inspection of the plots above. We also ran experiments with other shape parameters in the Beta distributions and with triangular density functions. Again, the results were much in line with those shown here. Hence, for brevity, these results are not reported on further here.
3.2 A heteroskedastic factor model

For our second set of experiments we focus on models where the outcome is a linear convolution of $X$ and a latent variate $U$, that is,

$$ Y = X + \sigma(X) U, \quad U \perp X, \quad \text{(3.1)} $$

for some function $\sigma$. It follows that the mean and variance of the component mixtures are given by

$$ \mathbb{E}[Y|X=x] = x, \quad \mathbb{V}[Y|X=x] = \sigma^2(x) \mathbb{V}[U]. \quad \text{(3.2)} $$

Equation (3.1) constitutes a linear one-factor model with (conditional) heteroskedasticity. Given the availability of repeated measurements, it can also be seen as a linear fixed-effect model from which covariates have been partialled out. As the within-group estimator delivers root-$N$ consistent estimates of common parameters in linear panel-data models, abstracting away from any covariates is without loss of generality. Horowitz and Markatou (1996) and Bonhomme and Robin (2010) have both developed deconvolution-based methods to recover the distribution of latent variables in models of this form under the assumption that $\sigma(x) = 1$ for all $x$.

Note that Equation (3.1) implies the support (as well as the shape) of the component mixtures to change with $x$. The results reported on below specified $U$ to be symmetrically distributed on the interval $[-1, 1]$, following the Beta distribution from above with parameter $\theta = \vartheta$. Thus, the pdf of $U$ at $u$ is $0.5f(0.5 + 0.5u, \theta, \theta)$, has mean zero, and variance $\mathbb{V}[U] = 1/(2\theta+1)$. We worked with $K$ uneven, uniformly distributing the factor $X$ over equidistant points in the interval $[-(K-1)/2, (K-1)/2]$. These choices, imply that the $k$th component mixture is

$$ f_k(y) = \frac{1}{2\sigma(x)} f\left(\frac{1}{2} \left( 1 - \frac{x}{\sigma(x)} \right) + \frac{1}{2} y, \theta, \theta \right) = \frac{1}{2\sigma(x)B(\theta, \theta)} \left[ \frac{1}{4} \left( 1 - \frac{x}{\sigma(x)} + y \right) \left( 3 + \frac{x}{\sigma(x)} - y \right) \right]^{\theta-1}, $$

which is centered at $x_k$ and supported on the interval $[x_k - \sigma(x_k), x_k + \sigma(x_k)]$. Below we report only results for $\theta = 2$, although other choices gave similar results. For the variance choices, finally, we set

$$ \sigma(x) = \begin{cases} 1 & \text{homoskedastic model} \\ 1 + \frac{1}{4}||x|-(\bar{y} - 1)| & \text{heteroskedastic model} \end{cases}, $$

where we denote by $[\bar{y}, \bar{y}]$ the support of the marginal density of the outcome. The chosen variance formulation in the heteroskedastic case enforces a certain form of symmetry in the heteroskedasticity, with $\sigma(x_1) = \sigma(x_K) = 1$ and the magnitude of the heteroskedasticity subsequently increasing as $X$ moves toward the center of its support (which is zero for any $K$).

Throughout, we fixed $N = 1000$ and $T = 3$, and set $L = 11$. Chebychev polynomials were again used as basis functions. To deal with the fact that $[\bar{y}, \bar{y}]$ differs from $[-1, 1]$, the support of
the Chebyshev polynomials, one needs to transplant the observations on $Y$ through the conversion formula

$$
\tilde{y} = 2\frac{y - \bar{y}}{\bar{y} - \underline{y}} - 1
$$

when computing the matrices $\hat{B}(\ell) \ (\ell = 1, \ldots, L)$. The remainder of the implementation is as before. Note that $\bar{y}$ and $\underline{y}$ are nonparametrically identified from the data. Knowledge of the support of the component mixtures is not needed for estimation. To estimate the conditional moments, we employed Clenshaw-Curtis quadrature. Moreover, the conditional mean and variance of $f_k$ were estimated by

$$\hat{E}[Y | X = x_k] = \sum_{i=0}^{L-1} c_i y_i \hat{f}_k(y_i), \quad \hat{V}[Y | X = x_k] = \sum_{i=0}^{L-1} c_i \left( y_i - \hat{E}[Y | X = x_k] \right)^2 \hat{f}_k(y_i),$$

where the $y_i$ are the set of Chebyshev roots $\cos(i \pi / L) \ (i = 0, \ldots, L - 1)$ appropriately transplanted from $[-1, 1]$ to the interval $[\underline{y}, \bar{y}]$, and the $c_i$ are the associated quadrature weights. Below, $L$ was set to 101.

**Homoskedastic design**   When $\sigma(x) = 1$, the outcome is just a shifted version of the latent variate $U$. From Equation (3.2) the variance of the component mixtures is simply $0.20$. Figure 3 provides a plot of the component mixtures and their respective cumulative distribution functions for the cases $K = 3$ and $K = 5$. With this design, a graphic exploration would suggest the marginal distribution of the outcome to be uniform on $[\underline{y}, \bar{y}]$.

Ten realizations of the estimators of the component mixtures are presented in Figure 4. The estimated densities mimic well the true component mixtures for both choices of $K$ shown. Similar pictures are obtained on augmenting $K$. For the sample size considered the tails of the estimated densities tend to extend somewhat over the actual support, although this bias disappears in larger samples. For inferring the overall shape of the component mixtures this tail bias presents little problems. However, since sample averages are sensitive to outlying observations, it can bias the estimators of the moments of the $f_k$. Depending on the situation at hand, one may wish to consider some form of trimming in such cases. A graphical inspection of the $\hat{f}_k$ can aid in deciding whether this is necessary.

Figures 5 and 6 contain plots of the estimated mixture weights and the mean and variance of the corresponding component mixtures for $K = 3$ and $K = 5$, respectively. In each figure, the upper left panel plots the estimated mixture weights for each conditional density. The horizontal lines indicate the true $\omega(x_k)$, which all equal $1/K$ in our design. The spread of the points around the truth differs across $k$ but, overall, we find estimated mixture weights that are close to their true values. The upper right panels plot the estimated means against the true means of the $f_k$, which equal $x_k$. For both $K = 3$ and $K = 5$ the first moment is accurately estimated for all component mixtures. The
Figure 3: Component mixtures in the homoskedastic model

Left: Conditional densities. Right: Conditional distributions. Upper panels contain the plots for $K = 3$; lower panels contain the plots for $K = 5$.

Figure 4: Realizations of $\hat{f}_k$ in the homoskedastic model

Each plot contains ten realizations of $\hat{f}_k$ (dotted), together with $f_k$ (solid). Left panel: results for $K = 3$. Right panel: results for $K = 5$. 
lower left panel displays the estimated variances of the $f_k$ as a function of $x_k$. Horizontal lines are provided at the true variances, which all equal 1/5. Some dispersion in the point estimates may be observed but, in all but one case, the estimates are fairly evenly spread around the truth. A downward bias is observed for $k = 2$ ($x_2 = 0$) when $K = 3$. The symmetry of the design leads to a symmetry in the estimated variances. This is best observed in the lower right panel of Figures 5 and 6, which plots the estimated variances against the value obtained by applying Clenshaw-Curtis quadrature using the unobserved $f_k$.

Figure 5: Mean and variance estimates in the homoskedastic model ($K = 3$)

Heteroskedastic design We next present parallel results for the factor model with heteroskedastic disturbances. Recall that, here,

$$
\sigma(x) = 1 + \frac{1}{4} |x| - (\bar{y} - 1)|.
$$

This implies the component mixtures to have support sets of a different size. The variances of the more inward densities are higher; the most outward densities have variance .20 and support
Figure 6: Mean and variance estimates in the homoskedastic model ($K = 5$)
\([-1 + x_1, 1 + x_1]\) and \([-1 + x_K, 1 + x_K]\), respectively. When \(K = 3\), \(y = 2\) and the conditional-variance factors are \(\sigma(x_1) = 1\), \(\sigma(x_2) = 1.5\), and \(\sigma(x_3) = 1\). When \(K = 5\), \(y = 3\) and the factors, in turn, become \(\sigma(x_1) = 1\), \(\sigma(x_2) = 1.25\), \(\sigma(x_3) = 1.5\), \(\sigma(x_4) = 1.25\), and \(\sigma(x_5) = 1\). To illustrate, Figure 7 provides the density and distribution functions of the conditional models for both \(K = 3\) and \(K = 5\).

**Figure 7: Component mixtures in the heteroskedastic model**

![Component mixtures](image)

Left: Conditional densities. Right: Conditional distributions. Upper panels contain the plots for \(K = 3\); lower panels contain the plots for \(K = 5\).

Figure 8 again provides ten realizations of the \(\hat{f}_k\) for \(K = 3\) (left panel) and for \(K = 5\) (right panel). While the introduction of heteroskedasticity complicates the structure of our model, the estimates of \(f_k\) remain well behaved. Inspection of the plots reveals that the underlying true component mixtures are picked out well by the estimation algorithm. The simulation results do suggest that densities near the middle are estimated with somewhat more noise compared to those that are positioned more outward. This is reasonable, as their supports overlap much more with those of the other mixture components. As the various conditional densities become more similar, the difference between their Fourier coefficients decreases which, in turn, leads to more volatile estimators of the component mixtures.
Figure 8: Realizations of $\hat{f}_k$ in the heteroskedastic model

Each plot contains ten realizations of $\hat{f}_k$ (dotted), together with $f_k$ (solid). Left panel: results for $K = 3$. Right panel: results for $K = 5$.

Figure 9: Mean and variance estimates in the heteroskedastic model ($K = 3$)
The performance of the estimators of the mixture weights, and means and variances is reported on in Figure 9 ($K = 3$) and Figure 10 ($K = 5$), which have the same layout as Figures 5 and 6 above. The largest difference compared to the homoskedastic design is that there tends to be a somewhat larger dispersion in the point estimates of all estimands. There is also more variability across the estimates when $K = 5$ compared to when $K = 3$. This is in line with the overall density estimates from Figure 8 above. The estimates of the conditional means are spread around their respective true values. The variance estimators are all biased upwards, although the magnitude of the bias tends to be small.

Figure 10: Mean and variance estimates in the heteroskedastic model ($K = 5$)

Overall, we find good performance of our procedures in the various problems considered. The estimated conditional densities capture well the underlying mixture structure. The mixture weights, too, reflect well the relative contributions of the individual component mixtures to the marginal distribution of the data. It may also be useful to note that generating the graphs above is virtually immediate, even with samples in the order of 10,000 observations and polynomials of a reasonable order.
CONCLUDING REMARKS

We have proposed nonparametric estimation methods for finite mixture models and derived there asymptotic properties. We conclude with mentioning some possible extensions of interest. First, it would be worthwhile to investigate how one could incorporate semiparametric restrictions in our setup. This may be useful to study bound estimation in random-coefficient panel data models such as in Chernozhukov, Hahn, and Newey (2006), for example. Second, and related, it should be relatively easy to derive asymptotics for finite-dimensional parameters $\theta$ satisfying moment restrictions of the form

$$\mathbb{E}[m(Y, X; \theta)|X = x_k] = \int_Y m(y, x; \theta) f_k(y) \, dy = 0$$

for some measurable function $m$. An obvious application of such a framework is in measurement-error models such as considered by Hu (2008). A simulated method-of-moment estimator could be developed, using $\hat{f}_k$ as an input. Under suitable conditions, this should lead to estimators of $\theta$ that converge at the parametric rate. Third, our methods have the potential to be extended to the continuous $X$ case, by means of a discretization argument. We experimented with this and found that, on the population level, this works well. However, it would require a substantial change in the asymptotic analysis, as continuity induces an ill-posed inverse type of problem involving the number of discretization points. Hence, we leave this extension to future research.

On the more practical side, too, some questions are left unanswered. First, a practical data-driven method to choose the truncation parameter would be useful. Wahba (1981) and Hall (1986), among others, have developed procedures for doing so in the traditional series-based estimation of univariate densities. Whether their approaches can be adapted to the current setting currently remains unestablished. Second, it would be useful to derive formal procedures to estimate $K$ from the data. One obvious approach would be based on inferring the rank of the matrices $P$ and $B$; see, e.g., Robin and Smith (2000) or Kleibergen and Paap (2006). However, such procedures tend to underestimate the true rank. In addition, they would not utilize the equivalent restrictions imposed on the lower dimensional submodels, i.e., the $P(s)$ ($s = 1, \ldots, S$) and $B(\ell)$ ($\ell = 1, \ldots, L$), respectively. In light of this, one attractive avenue to pursue may be based on the value of the JADE criterion, as a function of $K$.

REFERENCES


[26]


