NONPARAMETRIC ESTIMATION OF FINITE-MIXTURE MODELS

STÉPHANE BONHOMME[*] KOEN JOCHMANS[†] JEAN-MARC ROBIN[‡]

CEMFI Sciences Po Sciences Po and UCL

[First version: January 2012. This version: September 13, 2012.]

The aim of this paper is to provide simple nonparametric methods to estimate finite-mixture models from data with repeated measurements. Three measurements suffice for the mixtures to be fully identified and so our approach can be used even with very short panel data. We provide distribution theory for estimators of the number of mixture components, the mixing proportions, as well as of the mixture distributions and various functionals thereof. These estimators are found to perform well in a series of Monte Carlo exercises. We apply our techniques to document heterogeneity in log annual earnings using PSID data spanning the period 1969–1998.

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

I INTRODUCTION

Finite mixtures encompass a large class of models. Popular applications include modeling unobserved heterogeneity, performing factor analysis and ecological inference, as well as dealing with corrupted data and misclassified observations. We refer to Henry, Kitamura, and Salanié (2011) for illustrations of mixture models in economics and to McLachlan and Peel (2000) for a treatment at booklength. The conventional approach to inference in finite mixtures is parametric—i.e., by specifying the distribution function of the outcome of interest, say y, conditional on a discrete latent variable, say x, up to a finite-dimensional parameter—with identification being driven fully by functional form. Only few results on the nonparametric identifiability and estimability of mixtures have so far been obtained, and our aim here is to contribute to their development.

In line with the seminal work of Hall and Zhou (2003), our analysis relies on the availability of repeated measurements on y. These measurements—which could be collected either contemporaneously or over time, as with panel data, for example—are assumed to be independent and identically distributed conditional on x.[1] Our arguments yield point identification of the number of component mixtures, the component distributions, and the

[*] CEMFI, Casado del Alisal 5, 28014 Madrid, Spain; Tel. +34 914 290 551; bonhomme@cemfi.es.

[†] Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France; Tel. +33 1 45 49 85 99; koen.jochmans@sciences-po.org.

[‡] Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France; Tel. +33 1 45 49 72 43; jeanmarc.robin@sciences-po.fr.

We are grateful to Marc Henry and Bernard Salanié for comments and discussion.

[1] Our identification results continue to go through when the i.i.d. requirement is relaxed to allow for independent but non-identically distributed variables.

mixing proportions when three or more repeated measurements are available when a rank condition involving the component distributions is satisfied. Our approach to identification is constructive, and the resulting estimators are attractive from a computational point of view and have desirable large-sample properties.

For brevity, throughout, we focus on the case where the outcome variable exhibits continuous variation. Nevertheless, our methods can equally be applied when y is a discrete random variable, allowing for the number of support points to grow with the sample size. Our analysis is based on projections of the marginal densities and component mixtures into a basis of functions. We show that the mixture structure imposes a set of linear restrictions on the (generalized) Fourier coefficients of these densities. When the coefficients of the component densities are linearly independent, the number of component mixtures is identified as the rank of the matrix containing the Fourier coefficients of the bivariate marginal density. Further, the coefficients themselves are identified as the eigenvalues of a sequence of whitened matrices of Fourier coefficients of the trivariate marginal density. The mixing proportions are then readily pinned down as the unique minimizers of a minimum-distance criterion.

We propose estimating the number of components by sequentially testing the rank of an empirical analog of the matrix of bivariate Fourier coefficients and show consistency of the resulting point estimate. To estimate the component densities, we employ an algorithm for the joint approximate-diagonalization of a set of eigenmatrices developed by Cardoso and Souloumiac (1993). We derive both mean integrated square error (MISE) and uniform convergence rates, and provide conditions for pointwise asymptotic normality. Interestingly, the convergence rates coincide with those that would be obtained when data could be sampled directly from the component densities, and are optimal in the MISE sense for a variety of basis functions. A least-squares type estimator for the mixing proportions that converges at the parametric rate is then readily constructed. We also present distribution theory for GMM estimators of finite-dimensional parameters defined through conditional moment restrictions involving the component mixtures; prime examples of such estimands are their respective moments.

Our identification analysis is related to a recent independent contribution by Kasahara and Shimotsu (2010). However, first, we work with a projection of the relevant densities into a basis of functions while they employ a discretization argument to achieve identification when the outcome is continuous. Second, we explicitely use our identification arguments to construct estimators of the component mixtures and their functionals, and of the mixing proportions. Our estimators also complement and extend the work by Hall, Neeman, Pakyari, and Elmore (2005), who proposed a nonparametric approach to estimate bivariate mixtures. An advantage of our procedure is that it readily accommodates general K-variate mixtures, and this at no increase in computational cost.

The remainder of the paper is organized as follows. Section 2 formalizes our setup and presents our identification results. Section 3 contains an exposition of the resulting estimators and their large-sample properties. Section 4 provides simulation evidence on the performance of the various estimators. Section 5 investigates the presence of heterogeneity in male earnings. Two appendices collect auxiliary lemmata and technical proofs, respectively.

II IDENTIFICATION

Let x be a latent random variable with probability mass function (PMF) $\varpi : \mathscr{X} \mapsto [0,1]$, where $\mathscr{X} \equiv \{x_1, x_2, \ldots, x_K\}$ for an integer K. Let $\vec{y}_T \equiv (y_1, y_2, \ldots, y_T)'$ be a vector of repeated measurements on an observable outcome variable y whose marginal probability density function (PDF) takes the form

$$f(y_1, \ldots, y_T) = \sum_{k=1}^{K} \left\{ \prod_{t=1}^{T} f_k(y_t) \right\} \omega_k. \tag{2.1}$$

where $\omega_k \equiv \varpi(x_k)$ and $f_k : \mathscr{Y} \mapsto \mathscr{F}$ denotes the PDF of y conditional on $x = x_k$. Equalizing the support and image of the various f_k is without loss of generality.

Throughout this section we set $T = 3$, which suffices for identification.[2] Let $\mathsf{L}_2[\mathscr{S}]$ denote the set of functions that are square-integrable on the space $\mathscr{S}$, that is, the set of functions g for which $\|g\|_2 \equiv \int_{\mathscr{S}} |g(\epsilon)|^2 \, \mathrm{d}\epsilon < \infty$.[3] We restrict attention to the class of functions that satisfy the following regularity conditions.

Assumption 2.1 (Regularity)**.** *The intervals $\mathscr{Y}$ and $\mathscr{F}$ are, respectively, compact and bounded, and $f_k \in \mathsf{L}_2[\mathscr{Y}]$ for all $k = 1, 2, \ldots, K$.*

The demand for $\mathscr{Y}$ to be compact could be relaxed. However, if the outcome of interest, say y^*, takes values on the real line, for example, we may always consider a strictly-monotonic function $h : \mathscr{R} \mapsto \mathscr{Y}$ and work with the transformation $y = h(y^*)$. With $\mathscr{Y} = [0,1]$, for example, a simple choice for h would be a logistic CDF. Square-integrability validates a generalized Fourier expansion in a basis of functions, which is key for our subsequent developments.

[2] Non-trivial bounds on the component mixtures when $T = K = 2$ are given by Hall and Zhou (2003).

[3] We will use $|\epsilon|$ to denote the absolute value when ϵ is a real number and the cardinality when ϵ is a set. The notation $\|\epsilon\|$ will be reserved for the Euclidean norm when ϵ is a vector and for the matrix norm when ϵ is a matrix.

Let $\{\chi_j, j \geq 1\}$ be a complete orthonormal basis for $\mathsf{L}_2[\mathscr{Y}]$. For any integer J, the generalized Fourier approximation of $f_k(y)$ is given by the projection of the density onto the subspace spanned by $\chi_1, \chi_2, \ldots, \chi_J$, that is,

$$f_k(y; J) \equiv \sum_{j=1}^{J} \gamma_{kj} \chi_j(y), \qquad \gamma_{ki} \equiv \int_{\mathscr{Y}} \chi_j(\epsilon) f_k(\epsilon) \, \mathrm{d}\epsilon, \tag{2.2}$$

and satisfies $\lim_{J\to\infty} \| f_k(y; J) - f_k(y) \|_2 = 0$.

Let $\chi_{i_1 \ldots i_D} \equiv \Pi_{d=1}^{D} \chi_{i_d}$. Then $\{\chi_{i_1 \ldots i_D}, i_1, \ldots, i_D \geq 1\}$ forms a complete orthogonal (tensor-product) basis for $\mathsf{L}_2[\mathscr{Y}^D]$. Hence, we may define a truncated series expansion of the univariate marginal density as

$$f(y; I) \equiv \sum_{i=1}^{I} \sigma_i \chi_i(y), \qquad \sigma_i \equiv \int_{\mathscr{Y}} \chi_i(\epsilon) f(\epsilon) \, \mathrm{d}\epsilon,$$

as well as of the bivariate and trivariate marginal densities, as

$$f(y_1, y_2; I) \equiv \sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \sigma_{i_1 i_2} \chi_{i_1 i_2}(y_1, y_2),$$
$$f(y_1, y_2, y_3; I, J) \equiv \sum_{i_1=1}^{I} \sum_{i_2=1}^{I} \sum_{j=1}^{J} \sigma_{i_1 i_2 j} \chi_{i_1 i_2 j}(y_1, y_2, y_3),$$

where, for example, $\sigma_{i_1 i_2} \equiv \iint_{\mathscr{Y}^2} \chi_{i_1 i_2}(\epsilon_1, \epsilon_2) f(\epsilon_1, \epsilon_2) \, \mathrm{d}\epsilon_1 \mathrm{d}\epsilon_2$. The distinction between I and J is not important for identification but will matter for estimation, where I will be kept fixed and J will grow with the sample size.

Observe that, given a choice of basis functions, knowledge of the sequence of Fourier coefficients implies identification of the corresponding PDF, and vice versa. In our setup, while the data does not directly nonparametrically identify $\{\gamma_{kj}, j \geq 1\}$, they do reveal the sequences $\{\sigma_i, i \geq 1\}$, $\{\sigma_{i_1 i_2}, i_1, i_2 \geq 1\}$, and $\{\sigma_{i_1 i_2 j}, i_1, i_2, j \geq 1\}$, while the mixture structure implies the set of linear restrictions

$$\sigma_{i_1} = \sum_{k=1}^{K} \gamma_{k i_1} \; \omega_k, \qquad \sigma_{i_1 i_2} = \sum_{k=1}^{K} \gamma_{k i_1} \gamma_{k i_2} \; \omega_k, \qquad \sigma_{i_1 i_2 j} = \sum_{k=1}^{K} \gamma_{k i_1} \gamma_{k i_2} \gamma_{kj} \; \omega_k, \tag{2.3}$$

for all i_1, i_2, j. The relations in (2.3) imply identification of the component mixtures and associated mixing proportions (up to an arbitrary relabeling) under a weak restriction.

To discuss identification, it is useful to write the restrictions in matrix form. For any I, let

$$\Gamma \equiv \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{I1} & \gamma_{I2} & \cdots & \gamma_{IK} \end{pmatrix},$$

be the $I \times K$ matrix whose kth column contains the first I Fourier coefficients of f_k. Similarly, let $\sigma \equiv (\sigma_1, \sigma_2, \ldots, \sigma_I)'$ and introduce the $J+1$ symmetric $I \times I$ matrices

$$\Sigma_0 \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1I} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{I1} & \sigma_{I2} & \cdots & \sigma_{II} \end{pmatrix}, \qquad \Sigma_j \equiv \begin{pmatrix} \sigma_{11j} & \sigma_{12j} & \cdots & \sigma_{1Ij} \\ \sigma_{21j} & \sigma_{22j} & \cdots & \sigma_{2Ij} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{I1j} & \sigma_{I2j} & \cdots & \sigma_{IIj} \end{pmatrix},$$

containing the Fourier coefficients of the bivariate and of the trivariate marginal PDF, respectively. Then (2.3) can be equivalently expressed as

$$\Sigma_0 = \Gamma\Omega\Gamma', \qquad \Sigma_j = \Gamma\Omega^{1/2}\Delta_j\Omega^{1/2}\Gamma', \qquad j = 1, 2, \ldots, J, \tag{2.4}$$

where $\Omega \equiv \text{diag}[\omega] = \text{diag}[\omega_1, \omega_2, \ldots, \omega_K]$ and $\Delta_j \equiv \text{diag}[\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{Kj}]$ for each $j = 1, 2, \ldots, J$.

Identification rests on the following assumption.

Assumption 2.2 (Rank condition)**.** rank$[\Gamma] = K$ *and* det$[\Omega] > 0$.

The assumption of maximal column rank imposes absence of multicolinearity among the Fourier coefficients of the component mixtures and is intuitive. Clearly, it requires that $I \geq K$. Note, however, that the rank condition implicitly limits the support set of the PMF ϖ. To see why, note that the L_2-convergence of $f_k(y; I)$ in (2.2) requires the γ_{ki} to shrink to zero as $i \to \infty$ (by Parseval's identity; see below), so that I cannot be set without bound to make rank$[\Gamma] = K$ hold. The demand for Ω to be invertible is equivalent to imposing ϖ to have K support points, and thus ensures that (2.1) is a proper (multivariate) K-component mixture.

Theorem 2.1 follows.

Theorem 2.1 (Identification)**.** *(i) The number of mixture components, (ii) the component mixtures, and (iii) the mixing proportions are all nonparametrically identified, where (ii) and (iii) are up to arbitrary relabeling of the components.*

The proof to Theorem 2.1 is constructive. By Assumption 2.2, the matrix Σ_0, which is real and symmetric, has rank K. As this matrix is nonparametrically identified, so is its rank and, hence, the number of mixture components. This establishes Theorem 2.1(i). Continuing on, Σ_0 admits the spectral decomposition

$$\Sigma_0 = \Upsilon\Lambda\Upsilon', \qquad \Upsilon \equiv (\upsilon_1, \upsilon_2, \ldots, \upsilon_K), \qquad \Lambda \equiv \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_K],$$

where the K eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_K$ are all positive, real, and have multiplicity one, and where the corresponding eigenvectors $\upsilon_1, \upsilon_2, \ldots, \upsilon_K$ are linearly independent. By construction, Σ_0 can be transformed to the identity matrix through pre- and post multiplication with the matrices $\Theta \equiv \Lambda^{-1/2}\Upsilon'$ and Θ', respectively. Moreover, from (2.4), we obtain

$$\Theta\Sigma_0\Theta' = \mathsf{I}_K = \Xi\Xi', \qquad \Xi \equiv \Theta\Gamma\Omega^{1/2}.$$

Note that, by Assumption 2.2, Ξ is a $K \times K$ orthonormal matrix of full rank. Now, on transforming Σ_j in the same way, we obtain

$$\daleth_j \equiv \Theta\Sigma_j\Theta' = \Xi\Delta_j\Xi'.$$

That is, the sequence of matrices $\{\daleth_j, j \geq 1\}$ is simultaneously diagonalizable in the same basis (namely, the columns of Ξ). The eigenvalue of $\daleth_j$ corresponding to its kth eigenvector equals γ_{kj}, the jth Fourier coefficient of f_k. Hence, $f_k(y; J)$ is identified for any J, which implies that f_k is identified on $\mathscr{Y}$ for all $k = 1, 2, \ldots, K$; recall the convergence results under (2.2). Identification is achieved up to arbitrary relabeling only because the eigenvectors can be rearranged without affecting the argument. This establishes Theorem 2.1(ii). Finally, consider the metric $\mathcal{D}_{\mathcal{F}}(g_1, g_2) \equiv \int_{-\infty}^{+\infty} [g_1(\epsilon) - g_2(\epsilon)]^2 \, \mathrm{d}\mathcal{F}(\epsilon)$ for some weight function $\mathcal{F}$. Then it is easy to show that ω uniquely minimizes $\mathcal{D}_{\mathcal{F}}(f(\cdot; I), \sum_k f_k(\cdot; I) w_k)$ with respect to $w_1, w_2, \ldots, w_K$. For example, on introducing $w = (w_1, w_2, \ldots, w_K)'$ and setting $\mathcal{F}$ to the uniform measure on $\mathscr{Y}$,

$$\mathcal{D}_{\mathcal{F}}(f(\cdot; I), \textstyle\sum_k f_k(\cdot; I) w_k) = \| f(y) - \sum_k f_k(y; I) w_k \|_2^2 = (\sigma - \Gamma w)'(\sigma - \Gamma w),$$

where the last transition follows from orthogonality of the basis functions. As the columns of Γ are linearly independent, $\Gamma'\Gamma$ has full rank, and $\omega = (\Gamma'\Gamma)^{-1}\Gamma'\sigma$ follows. Because Γ is identified, so is ω. This establishes Theorem 2.1(iii).[4]

Knowledge of the component mixtures implies identification of all their functionals. Theorem 2.2 provides one possible representation of this result that is particularly useful for the purpose of estimation. To state it, we introduce the functions

$$\tau_k(\epsilon) \equiv \frac{f_k(\epsilon)}{f(\epsilon)} = \frac{f_k(\epsilon)}{\sum_{k'=1}^{K} f_{k'}(\epsilon) \, \omega_{k'}} \qquad k = 1, 2, \ldots, K,$$

identification of which follows immediately from Theorem 2.1.

[4] We note that Theorem 2.1(iii) may also be established from the equality $\Xi = \Theta\Gamma\Omega^{1/2}$, which implies that $\Omega^{-1/2} = \Xi'\Theta\Gamma$. However, we prefer our exposition here, as the minimum-distance argument allows the construction of a variety of different estimators, by varying the weight function $\mathcal{F}$.

Theorem 2.2 (Functionals)**.** *For any measurable function g of y whose expectation under f_k exists,*

$$\mathbb{E}[g(y)|x = x_k] = \mathbb{E}[g(y)\tau_k(y)]$$

for each $k = 1, 2, \ldots, K$. Furthermore, the function τ_k and, therefore, all the functionals are nonparametrically identified.

Theorem 2.2 is instrumental in inferring a parameter ϑ_0 defined through moment conditions of the form

$$\mathbb{E}[g(y; \vartheta_0)|x = x_k] = 0$$

for a known measurable function g.

The weight function τ_k is interesting in its own right. More precisely, $\pi_k(\epsilon) \equiv \omega_k \tau_k(\epsilon) = \mathbb{E}[1\{x = x_k\}|y = \epsilon]$, the probability that $x = x_k$ given that $y = \epsilon$. This is a key object of interest in latent-class analysis, allowing classification of observations based on marginal information.

III ESTIMATION

In this section we construct estimators based on the identification arguments laid out above. For $\vec{y}_{nT} \equiv (y_{n1}, y_{n2}, \ldots, y_{nT})'$, let $\{\vec{y}_{nT}, n = 1, 2, \ldots, N\}$ denote a sample of size N drawn at random from $f(\vec{y}_T)$. Throughout, we assume that $T \geq 3$, keep I fixed to a chosen value, and let the truncation parameter J grow with the sample size. Admissible rates on J will be stated below.

3.1 Number of component mixtures

Let ϱ_D be the set of D-tuples of distinct integers from the set $\{1, 2, \ldots, T\}$. On invoking symmetry, an estimator of $\sigma_{i_1 \ldots i_D}$ for any D is given by the sample average

$$\widehat{\sigma}_{i_1 \ldots i_D} \equiv \frac{1}{N} \frac{1}{|\varrho_D|} \sum_{n=1}^{N} \sum_{(t_1, t_2, \ldots, t_D) \in \varrho_D} \chi_{i_1 \ldots i_D}(y_{nt_1}, y_{nt_2}, \ldots, y_{nt_D}), \tag{3.1}$$

and is $\sqrt{N}$-consistent and asymptotically normal.

Use (3.1) with $D = 2$ to construct an estimator of Σ_0, say $\widehat{\Sigma}_0$. Then $\widehat{\Sigma}_0$ is $\sqrt{N}$-consistent and asymptotically normal, that is,

$$\sqrt{N} \mathrm{vec}[\widehat{\Sigma}_0 - \Sigma_0] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Sigma_0}), \qquad \mathscr{V}_{\Sigma_0} \equiv \mathbb{E}[\psi_{\Sigma_0}(\vec{y}_T)\psi_{\Sigma_0}(\vec{y}_T)'],$$

with a typical element of the vector $\psi_{\Sigma_0}(\vec{y}_T)$ being $|\varrho_2|^{-1} \sum_{(t_1, t_2) \in \varrho_2} \chi_{i_1 i_2}(y_{t_1}, y_{t_2}) - \sigma_{i_1 i_2}$. Inferring K then boils down to testing the rank of this matrix, which can be done by any of a number of approaches. The procedure by Kleibergen and Paap (2006) has several attractive features and therefore carries our preference.[5] To describe it, fix $\varkappa \in \{0, 1, \ldots, I-1\}$ and let $\Lambda_{I-\varkappa}$ denote the lower-right $(I-\varkappa) \times (I-\varkappa)$ block of the matrix of eigenvalues of Σ_0. Write $(\mathscr{F}'_{\varkappa}, \mathscr{F}'_{I-\varkappa})'$ for the corresponding $I \times (I-\varkappa)$ matrix of eigenvectors, where $\mathscr{F}_{I-\varkappa}$ is $(I-\varkappa) \times (I-\varkappa)$. Let $\varphi_{\varkappa} \equiv \operatorname{vec}[\overline{\mathscr{F}}_{I-\varkappa}\Sigma_0\overline{\mathscr{F}}'_{I-\varkappa}]$ for $\overline{\mathscr{F}}_{I-\varkappa} \equiv (\mathscr{F}_{I-\varkappa}\mathscr{F}'_{I-\varkappa})^{1/2}\mathscr{F}^{-1\prime}_{I-\varkappa}[\mathscr{F}'_{\varkappa} \vdots \mathscr{F}'_{I-\varkappa}]$. Then, if $\operatorname{rank}[\Sigma_0] = K \leq \varkappa$, $\Lambda_{I-\varkappa} = 0$ and

$$\widehat{\mathrm{r}}_{\varkappa} \equiv N|\varrho_2|\widehat{\varphi}_{\varkappa}\widehat{\mathscr{V}}^{-1}_{\varphi_{\varkappa}}\widehat{\varphi}_{\varkappa} \xrightarrow{L} \chi^2\big((I-\varkappa)^2\big), \qquad \mathscr{V}_{\varphi_{\varkappa}} \equiv (\overline{\mathscr{F}}_{I-\varkappa} \otimes \overline{\mathscr{F}}_{I-\varkappa})\mathscr{V}_{\Sigma_0}(\overline{\mathscr{F}}'_{I-\varkappa} \otimes \overline{\mathscr{F}}'_{I-\varkappa}), \tag{3.2}$$

where $\widehat{\varphi}_{\varkappa}$ is the sample analog of $\varphi_{\varkappa}$ and $\widehat{\mathscr{V}}_{\varphi_{\varkappa}}$ is a consistent estimator of $\mathscr{V}_{\varphi_{\varkappa}}$. The rank statistic $\widehat{\mathrm{r}}_{\varkappa}$ can be used to test the null $H_0 : \operatorname{rank}[\Sigma_0] = \varkappa$ against the alternative $H_1 : \operatorname{rank}[\Sigma_0] > \varkappa$.

Furthermore, the statistic in (3.2) also leads to a consistent estimator of K, based on a sequential testing procedure. Following Robin and Smith (2000), let

$$\widehat{K} \equiv \min_{\varkappa \in \{0,1,\ldots,I-1\}} \big\{\varkappa : \widehat{\mathrm{r}}_k \geq p_{1-\alpha}(k), k = 0, 1, \ldots, \varkappa - 1, \widehat{\mathrm{r}}_{\varkappa} < p_{1-\alpha}(\varkappa)\big\}, \tag{3.3}$$

for $p_{1-\alpha}(\epsilon)$ the $100(1-\alpha)$th percentile of the $\chi^2((I-\epsilon)^2)$ distribution and $\alpha = \alpha(N)$ a chosen significance level. That is, $\widehat{K}$ is the first integer for which we fail to reject the null at significance level α.

Theorem 3.1 (Number of components)**.** *Let $\alpha \to 0$ and $-\log\alpha/N \to 0$ as $N \to \infty$. Then $\widehat{K} \xrightarrow{P} K$.*

Our estimator of K is similar to the recent proposal of Kasahara and Shimotsu (2010). Their approach is based on a partitioning of the Cartesian square $\mathscr{Y} \times \mathscr{Y}$ into a set of subplanes, say, $\{\mathscr{H}_{h_1} \times \mathscr{H}_{h_2}, h_1 = 1, 2, \ldots, H_1; h_2 = 1, 2, \ldots, H_2\}$ for some integer values H_1 and H_2. Their estimator is then constructed as discussed above, with $\widehat{\Sigma}_0$ replaced by a nonparametric estimator of the $H_1 \times H_2$ matrix whose (h_1, h_2)th entry is the cell probability $\mathbb{E}[1\{(y_1, y_2) \in \mathscr{H}_{h_1} \times \mathscr{H}_{h_2}\}]$.

3.2 Component mixtures

We next turn to estimation of the conditional densities (i.e., $f_k, k = 1, 2, \ldots, K$). Dealing with the pretesting problem that arises from utilizing $\widehat{K}$ rather than K is a problem beyond the scope of this paper and so, throughout the remainder of this section, we take K as given.

[5] Kleibergen and Paap (2006) [Section 3] discuss the relative merits of their test statistic. Prime advantages include its non-sensitivity to the ordering of variables and the fact that its limit distribution under the null is a Chi-squared distribution, which is free of nuisance parameters.

Let $\widehat{\Theta}$ be the sample counterpart to Θ, constructed from the spectral decomposition of $\widehat{\Sigma}_0$. For $j = 1, 2, \ldots, J$, let $\widehat{\Sigma}_j$ be an estimator of Σ_j; $\widehat{\Sigma}_j$ can be computed as before, using (3.1) for $D = 3$. On subsequently forming $\widehat{\daleth}_j \equiv \widehat{\Theta}\widehat{\Sigma}_j\widehat{\Theta}'$, our estimator of Δ_j is

$$\widehat{\Delta}_j \equiv \widehat{\Xi}'\widehat{\daleth}_j\widehat{\Xi}, \qquad \widehat{\Xi} \equiv \arg \min_{X \in \mathscr{O}(K)} \sum_{j=1}^{J} \operatorname{off}\left[X'\widehat{\daleth}_j X\right], \tag{3.4}$$

where $\mathscr{O}(K)$ denotes the set of orthonormal matrices of dimension $K \times K$ and the function $\operatorname{off}[\cdot]$ maps a square matrix into the sum of its squared off-diagonal elements. For fixed $y \in \mathscr{Y}$, our estimator of $f_k(y)$ $(k = 1, 2, \ldots, K)$ then reads

$$\widehat{f}_k(y) \equiv \sum_{j=1}^{J} \widehat{\gamma}_{kj}\chi_j(y), \tag{3.5}$$

where $\widehat{\gamma}_{kj} \equiv q_k' \mathrm{vec}[\widehat{\Delta}_j]$ and q_k denotes the kth column of the selection matrix Q, which is defined as the $K^2 \times K$ matrix whose transpose turns a $K \times K$ matrix into a $K \times 1$ vector containing only its diagonal elements.

It is important to note that estimation is not based on the eigenspectra of the individual matrices $\{\widehat{\daleth}_j, j = 1, 2, \ldots, J\}$. Although the set $\{\daleth_j, j \geq 1\}$ all share the same eigenvectors, sampling error will imply that the eigenvectors of $\widehat{\daleth}_{j_1}$ and $\widehat{\daleth}_{j_2}$ will all be different for each $j_1 \neq j_2$. In small samples this is problematic. Rather, our estimator of Ξ in (3.4) is defined as that $X \in \mathscr{O}(K)$ that makes the set $\{X'\widehat{\daleth}_j X, j = 1, 2, \ldots, J\}$ as diagonal as possible.[6] It is attractive from an efficiency point of view, as it correctly utilizes the restriction that the $\{\daleth_j, j \geq 1\}$ do share the same eigenvectors. Nevertheless, in any finite sample, it will typically not correspond to the eigenvectors of any of the $\widehat{\daleth}_j$. The estimator $\widehat{\Xi}$ can easily be computed using the algorithm for the joint approximate-diagonalization of eigenmatrices (JADE) developed by Cardoso and Souloumiac (1993). JADE is fast and straightforward to implement.

The following high-level assumption is conventional in nonparametric curve estimation by series expansions and is compatible with a large variety of basis functions.

Assumption 3.1 (Smoothness)**.** *For each $k = 1, 2, \ldots, K$, f_k is absolutely continuous and of bounded variation on $\mathscr{Y}$, and $\|f_k(y; J) - f_k(y)\|_2 = O(J^{-\beta})$ for some $\beta \geq 1$.*

Assumption 3.1 can be motivated through a smoothness condition on the function f_k, demanding it to be β-smooth, and may be interpreted as a restriction on the shrinkage

[6] More precisely, minimizing the criterion defined in (3.4) is equivalent to solving the least-squares problem $\min_{X \in \mathscr{O}(K), L_j \in \mathcal{L}(K)} \sum_{j=1}^{J} \|\widehat{\daleth}_j - X L_j X'\|_F^2$, where $\mathcal{L}(K)$ is the set of $K \times K$ diagonal matrices and $\|\cdot\|_F$ is the Frobenius norm.

rate of the generalized Fourier coefficients, through Parseval's identity; see. e.g., Efromovich (1999) [Chapter 2] for an elegant exposition and Chen (2007) for illustrations with specific basis functions. We remark that demanding a certain degree of smoothness from the component mixtures is substantially less demanding than is requesting the same from the marginal density. The mixing will typically create nonsmoothness, even if the components are all smooth.

Consistent estimation requires the truncation parameter J to grow with the sample size. The (uniform) convergence speed depends on the basis functions used. To be able to proceed with a generic set of basis functions, let $\zeta(J)$ denote a sequence of constants satisfying $\sup_{y\in\mathscr{Y}}\|\mathcal{X}_J(y)\| \leq \zeta(J)$, where $\mathcal{X}_J(y) \equiv (\chi_1(y), \chi_2(y), \ldots, \chi_J(y))'$. For example, when orthonormal polynomials—e.g., Chebychev, Jacobi, or Legendre polynomials—are used, $\zeta(J) \propto J$. On using splines, $\zeta(J) \propto \sqrt{J}$. See, e.g., Newey (1997) or Chen (2007) for additional discussion.

Assumption 3.2 (Truncation)**.** *The integer sequence J grows so that (i) $\zeta(J)J^2/N \to 0$ and (ii) $\zeta(J)^2 J^{2\beta}/N \to \infty$ as $N \to \infty$.*

Assumption 3.2(i) is needed for consistency. Assumption 3.2(ii) will ensure that the limit distribution of the component mixtures at a fixed point is correctly centered. Of course, β in Assumption 3.1 could be allowed to vary with k, which would subsequently allow the truncation parameter associated with each component mixture to grow at a different rate.

Theorem 3.2 provides mean integrated squared error (MISE) and uniform convergence rates for the estimator of the component mixtures.

Theorem 3.2 (Component mixtures: convergence rates)**.** *The estimator of the component mixture satisfies*

$$\mathbb{E}\big\|\widehat{f}_k(y) - f_k(y)\big\|_2^2 = O_P\big(J/N + J^{-2\beta}\big), \quad \sup_{y\in\mathscr{Y}}\big|\widehat{f}_k(y) - f_k(y)\big| = O_P\big(\zeta(J)[\sqrt{J}/\sqrt{N} + J^{-\beta}]\big),$$

for each $k = 1, 2, \ldots, K$.

The rates in Theorem 3.2 equal the conventional univariate rates of nonparametric series estimators; see, e.g., Newey (1997). The MISE rate is known to be optimal for power series and splines in the sense that it achieves Stone's (1982) bound.

Theorem 3.3 below presents the pointwise limit distribution of the estimator of the kth component mixture. The proof to the theorem shows that

$$\widehat{f}_k(y) - f_k(y; J) = N^{-1}\sum_{n=1}^{N}\psi_{f_k}(\vec{y}_{nT}) + O_P(1/\sqrt{N}),$$

where $y \in \mathscr{Y}$ is a chosen value and where

$$\psi_{f_k}(\vec{y}_{nT}) \equiv |\varrho_3|^{-1} \sum_{(t_1,t_2,t_3)\in\varrho_3} \Big[\overline{\tau}_k(y_{nt_1}, y_{nt_2})\kappa_J(y_{nt_3}; y) - \mathbb{E}[\overline{\tau}_k(y_{t_1}, y_{t_2})\kappa_J(y_{t_3}; y)]\Big], \qquad (3.6)$$

whose dependence on y is kept implicit. Here,

$$\overline{\tau}_k(y_1, y_2) \equiv \sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \xi_k'\theta_{i_1}\chi_{i_1}(y_1)\chi_{i_2}(y_2)\theta_{i_2}'\xi_k, \qquad \kappa_J(y_1, y_2) \equiv \sum_{j=1}^{J}\chi_j(y_1)\chi_j(y_2).$$

The form of ψ_{f_k} in (3.6) is interesting. The function κ_j is known as the jth kernel of the system $\{\chi_j, j \geq 1\}$ and behaves much like its namesake in kernel-density estimation; e.g., $\mathbb{E}[\kappa_J(y, y)] = f(y; J)$. Concerning $\overline{\tau}_k$, observe that $\mathbb{E}[\overline{\tau}_k(y_{t_1}, y_{t_2})|x = x_{k'}] = \omega_k^{-1}1\{k' = k\}$ and so $\mathbb{E}[\overline{\tau}_k(y_{t_1}, y_{t_2})\kappa_J(y_{t_3}; y)] = f_k(y; J) = \mathbb{E}[\tau_k(y)\kappa_J(y; y)]$, as is readily verified by use of the law of iterated expectations. Hence, $\widehat{f}_k(y)$ can be viewed as a reweighting estimator based on an estimator of $f(y)$, the marginal density of y at y.

Theorem 3.3 (Component mixtures: normality)**.** *For each $k \in \{1, 2, \ldots, K\}$ and each $y \in \mathscr{Y}$, the estimated component mixture satisfies*

$$\sqrt{N}\mathscr{V}_{f_k}^{-1/2}[\widehat{f}_k(y) - f_k(y)] \xrightarrow{L} \mathcal{N}(0, 1), \qquad \mathscr{V}_{f_k} \equiv \mathbb{E}[\psi_{f_k}(\vec{y}_T)\psi_{f_k}(\vec{y}_T)],$$

as $N \to \infty$.

From Viollaz (1989), $\mathscr{V}_{f_k} = O(\wp_J(y))$ for $\wp_J(y) \equiv \int_{\mathscr{Y}} \kappa_J(\epsilon, y)^2 \, \mathrm{d}\epsilon$ and so the pointwise convergence speed is determined by the growth rate of $\wp_J(y)$. For example, for Jacobi and Legendre polynomials it is known that $\wp_J(y) = O(J)$; see, e.g., Hall (1982). The limit distribution is free of asymptotic bias provided that $\wp_J(y)J^{2\beta}/N \to \infty$ as $N \to \infty$. Because $\sqrt{\wp_J(y)} \leq \sup_{y\in\mathscr{Y}}\|\mathcal{X}_J(y)\| \leq \zeta(J)$, a weak bound is $|\widehat{f}_k(y) - f_k(y)| = O_P(\zeta(J)/\sqrt{N})$, and so a sufficient condition is that $\zeta(J)^2J^{2\beta}/N \to \infty$ as $N \to \infty$; cfr. Assumption 3.2(ii).

3.3 Mixing proportions

An estimator of the mixing proportions is easily constructed using the results obtained so far, by virtue of a minimum-distance procedure. Here, we provide results for the estimator

$$\widehat{\omega} \equiv (\widehat{\Gamma}'\widehat{\Gamma})^{-1}\widehat{\Gamma}'\widehat{\sigma},$$

where $\widehat{\Gamma}' = (Q'\text{vec}[\widehat{\Delta}_1], Q'\text{vec}[\widehat{\Delta}_2], \ldots, Q'\text{vec}[\widehat{\Delta}_I])$ and $\widehat{\sigma}$ is the I-vector whose ith element equals $(NT)^{-1}\sum_{n=1}^{N}\sum_{t=1}^{T}\chi_i(y_{nt})$; viz. (3.1). We could derive alternative estimators by considering a metric $\mathcal{D}_{\mathcal{F}}$ that uses a weight function that is different from the uniform one, but these are omitted for brevity. One possibility would be to set $\mathcal{F}$ to the marginal CDF of the data, which could be useful for downweighting regions of $\mathscr{Y}$ where the value of the component mixtures is small.

Denote by $\psi_\sigma(\vec{y}_{nT})$ the vector whose ith entry equals $T^{-1}\sum_{t=1}^{T}\chi_i(y_{nt})-\sigma_i$ and let $\psi_{\Gamma'}(\vec{y}_{nT})$ be the vector obtained on stacking $Q'\psi_{\Delta_i}(\vec{y}_{nT})$ $(i=1,2,\ldots,I)$. The function ψ_{Δ_i} is detailed in Theorem A.1 in Appendix A and is the influence function of $\widehat{\Delta}_i$. On letting

$$\psi_\omega(\vec{y}_{nT})\equiv(\Gamma'\Gamma)^{-1}\big[(\sigma'\otimes \mathsf{I}_K)\psi_\Gamma(\vec{y}_{nT})+\Gamma'\psi_\sigma(\vec{y}_{nT})\big],$$

Theorem 3.4 can be stated.

Theorem 3.4 (Mixing proportions)**.** *The minimum-distance estimator $\widehat{\omega}$ of the mixing proportions ω satisfies*

$$\sqrt{N}(\widehat{\omega}-\omega)\xrightarrow{L}\mathcal{N}(0,\mathscr{V}_\omega),\qquad \mathscr{V}_\omega\equiv\mathbb{E}[\psi_\omega(\vec{y}_T)\psi_\omega(\vec{y}_T)'],$$

as $N\to\infty$.

Our minimum-distance estimator is similar in spirit to the proposal of Titterington (1983), who worked in a framework where data could be sampled directly from the component mixtures.

3.4 Functionals

Now consider the problem of inferring a vector ϑ_0 defined as the unique solution to a moment condition of the form $m(\vartheta_0)=0$, where $m(\vartheta)\equiv\mathbb{E}[g(y;\vartheta)|x=x_k]$ for some known function g, which may be multivariate. From Theorem 2.2 we have that $m(\vartheta)=\mathbb{E}[g(y;\vartheta)\tau_k(y)]$, and so a natural way to proceed is to consider a GMM estimator of the form

$$\widehat{\vartheta}\equiv\arg\min_{\vartheta\in\mathscr{T}}\widehat{m}_N(\vartheta)'V_N\widehat{m}_N(\vartheta),\qquad \widehat{m}_N(\vartheta)\equiv(NT)^{-1}\sum_{n=1}^{N}\sum_{t=1}^{T}g(y_{nt};\vartheta)\;\widehat{\tau}_k(y_{nt}),$$

where $\mathscr{T}$ is the parameter space, V_N is a positive-definite weight matrix that converges in probability to a positive-definite and non-stochastic matrix V, and $\widehat{\tau}_k(y)$ is an estimator of the weight function τ_k at y.[7] We use

$$\widehat{\tau}_k(y)\equiv\frac{\widehat{f}_k(y)}{\sum_{k'=1}^{K}\widehat{f}_{k'}(y)\;\widehat{\omega}_{k'}},$$

[7] Some trimming will typically be warranted. With some work, Theorem 3.5 below may be generalized to hold under a trimming scheme where the tuning parameter converges to the identity slowly with N. We note that, if interest lies mainly in moments of the component mixtures, a fixed-trimming scheme would be inappropriate.

although other possibilities could be entertained. Prime examples of such ϑ_0 are the moments of the component mixtures; the first moment, for example, is defined through $\mathbb{E}[y\tau_k(y) - \vartheta_0] = 0$.

Impose the following conditions.

Assumption 3.3 (Regularity)**.** *The space $\mathscr{T}$ is compact and has ϑ_0 as an interior element. The function g is twice continuously differentiable in ϑ on $\mathscr{T}$. For each $k = 1, 2, \ldots, K$, the two moments $\sup_{\vartheta \in \mathscr{T}} \mathbb{E}[\|g(y;\vartheta)\tau_k(y)\|]$ and $\sup_{\vartheta \in \mathscr{T}} \mathbb{E}[\|\partial g(y;\vartheta)/\partial\vartheta'(y;\vartheta)\tau_k(y)\|]$ are finite, the matrix $\mathbb{E}[\partial g(y;\vartheta_0)/\partial\vartheta'\tau_k(y)]$ has full column rank, and the function τ_k is bounded away from zero and infinity on $\mathscr{Y}$. The truncation parameter grows so that $\zeta(J)^2 J^4/N \to 0$ as $N \to \infty$.*

Assumption 3.3 contains familiar conditions for GMM estimators to be asymptotically normal. It also imposes a slower growth rate on J than was required before. This is needed because our moment conditions are nonlinear in the estimated component mixtures. When power series are used as basis functions, for example, we now require that $J^6/N \to 0$—as opposed to $J^3/N \to 0$—as $N \to \infty$.

Let $M_\vartheta \equiv \mathbb{E}[\partial g(y;\vartheta_0)/\partial\vartheta'\tau_k(y)]$. Let M_ω be the K-vector that has $\mathbb{E}[g(y;\vartheta_0)\tau_k(y)\tau_{k'}(y)]$ as its k'th entry. Introduce

$$p_k(y_{nt}) \equiv g(y_{nt};\vartheta_0)\Big(\frac{1-f(y_{nt})}{f(y_{nt})}\Big)\tau_k(y_{nt}) - \mathbb{E}\Big[g(y;\vartheta_0)\Big(\frac{1-f(y)}{f(y)}\Big)\tau_k(y)\Big] - M_\omega'\psi_\omega(y_{nt}).$$

Then $\widehat{\vartheta} - \vartheta_0 = (NT)^{-1}\sum_{n=1}^N\sum_{t=1}^T \psi_\vartheta(y_{nt}) + o_P(1/\sqrt{N})$ with influence function

$$\psi_\vartheta(y_{nt}) \equiv [M_\vartheta' V M_\vartheta]^{-1} M_\vartheta' V \big[g(y_{nt};\vartheta_0)\tau_k(y_{nt}) + p_k(y_{nt})\big].$$

The influence function has the usual structure. The function p_k captures the impact of first-stage estimation error on the asymptotic variance of $\widehat{\vartheta}$. Let

$$\mathscr{V}_m \equiv \mathbb{E}[(g(y;\vartheta_0)\tau_k(y) + p_k(y))(g(y;\vartheta_0)\tau_k(y) + p_k(y))']$$

and assume this matrix to be positive definite. Then $\widehat{\vartheta}$ is $\sqrt{N}$-consistent and asymptotically normal.

Theorem 3.5 (Functionals)**.** *For any $k = 1, 2, \ldots, K$, the estimator $\widehat{\vartheta}$ of the estimand ϑ_0 satisfies*

$$\sqrt{N}(\widehat{\vartheta} - \vartheta_0) \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\vartheta), \qquad \mathscr{V}_\vartheta \equiv \mathbb{E}[\psi_\vartheta(y)\psi_\vartheta(y)'],$$

as $N \to \infty$.

Standard arguments show that the optimally-weighted estimator is obtained on setting $V \propto \mathscr{V}_m^{-1}$, in which case Theorem 3.5 implies that $\sqrt{N}(\widehat{\vartheta} - \vartheta_0) \xrightarrow{L} \mathcal{N}(0, [M_\vartheta' \mathscr{V}_m^{-1} M_\vartheta]^{-1})$.

IV MONTE CARLO ILLUSTRATIONS

We present numerical evidence on the small-sample performance of our estimators by means of two illustrations. In both examples we work with the family of generalized Beta distributions (see, e.g., McDonald, 1984), which is popular for modeling the distribution of income. On the interval $[\underline{y}, \overline{y}] \subset \mathscr{R}$, the generalized Beta distribution is

$$\mathsf{b}(y; \vartheta_1, \vartheta_2; \underline{y}, \overline{y}) \equiv \frac{1}{(\overline{y} - \underline{y})^{\vartheta_1 + \vartheta_2 - 1}} \frac{1}{\mathsf{B}(\vartheta_1, \vartheta_2)} (y - \underline{y})^{\vartheta_1 - 1} (\overline{y} - y)^{\vartheta_2 - 1},$$

where $\mathsf{B}(\vartheta_1, \vartheta_2) \equiv \int_0^1 \epsilon^{\vartheta_1} (1-\epsilon)^{\vartheta_2 - 1} \, \mathrm{d}\epsilon$ and ϑ_1 and ϑ_2 are positive real scale parameters. Its mean and variance are

$$\mu \equiv \underline{y} + (\overline{y} - \underline{y}) \frac{\vartheta_1}{\vartheta_1 + \vartheta_2}, \qquad \varsigma^2 \equiv (\overline{y} - \underline{y})^2 \frac{\vartheta_1 \vartheta_2}{(\vartheta_1 + \vartheta_2)^2 (\vartheta_1 + \vartheta_2 + 1)}, \tag{4.1}$$

respectively. Throughout, we use normalized Chebychev polynomials as basis functions. For $\epsilon \in [-1, 1]$, the ith such polynomial is

$$\chi_i(\epsilon) = \frac{2}{\pi} \frac{1}{2^{1\{i=1\}}} \frac{1}{\sqrt{1 - \epsilon^2}} \cos[(i-1) \arccos(\epsilon)].$$

In each illustration, we estimate the component mixtures and their associated CDFs, the mixing proportions, as well as the mean and variance of the component mixtures. To ensure bona fide density estimators we use

$$\widetilde{f}_k(y) \equiv \max\{0, \widehat{f}_k(y) - c_k\}, \tag{4.2}$$

where c_k is chosen so that $\int \widetilde{f}_k(\epsilon) \, \mathrm{d}\epsilon = 1$ (see Gajek, 1986). To infer the conditional CDFs, $F_k(y) \equiv \int_{-\infty}^{y} f_k(\epsilon) \, \mathrm{d}\epsilon$, we use Clenshaw-Curtis quadrature to approximate the integral $\int_{-\infty}^{y} \widetilde{f}_k(\epsilon) \, \mathrm{d}\epsilon$; our approximation uses 101 quadrature nodes. The moments are estimated using the GMM estimator from Theorem 3.5, without any trimming.

Experiment 1: A mixture of Betas. Our first experiment involves three generalized Beta distributions on the interval $[-1, 1]$. Moreover, we consider

$$\begin{aligned} f_1(y) &= \mathsf{b}(y; 2, 7; -1, 1), & \omega_1 &= .20, \\ f_2(y) &= \mathsf{b}(y; 5, 4; -1, 1), & \omega_2 &= .35, \\ f_3(y) &= \mathsf{b}(y; 6, 2; -1, 1), & \omega_3 &= .45. \end{aligned}$$

Using (4.1), the means of the component mixtures are $\mu_1 = -5/9 \approx -.556$, $\mu_2 = 1/9 \approx .111$, and $\mu_3 = 1/2 = .500$, while their respective variances are $\varsigma_1^2 = 28/405 \approx .069$, $\varsigma_2^2 = 8/81 \approx .099$, and $\varsigma_3^2 = 1/12 \approx .083$. Throughout, we set $T = 4$ and $I = J = 6$.

Table 1 presents simulation results for the estimator of K defined through the sequential testing procedure in (3.3) for various values of N and α. The table reports the frequency with which K was either underestimated, correctly estimated, or overestimated in $10,000$ Monte Carlo replications. The results show that $\widehat{K}$ performs well, correctly picking the true number of component mixtures in about 95% of the cases overall.

Table 1: Sequential rank test in Experiment 1 (Beta mixture)

	$\alpha = .100$			$\alpha = .050$			$\alpha = .025$		
N	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$
500	.005	.927	.068	.002	.959	.039	.001	.978	.021
750	.006	.926	.068	.002	.958	.040	.001	.976	.023
1000	.005	.924	.071	.002	.957	.041	.001	.979	.020
1500	.005	.929	.066	.002	.960	.038	.000	.976	.024
2000	.006	.930	.064	.002	.956	.042	.002	.980	.018
2500	.004	.929	.067	.002	.965	.033	.000	.975	.024

Note: $K = 3$, $T = 4$, $I = J = 6$; $10,000$ replications.

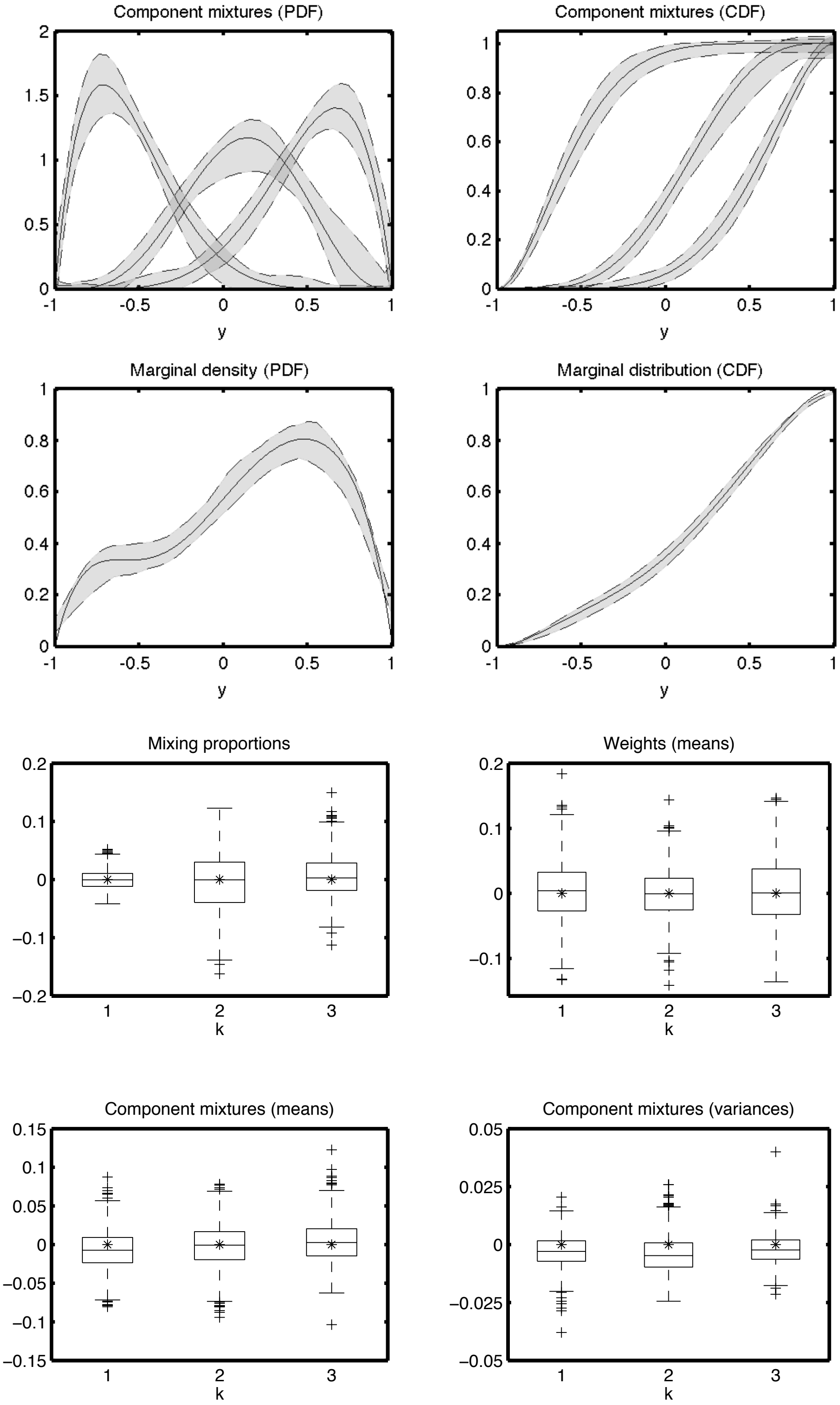
Taking K as given, we turn to estimation of the component mixtures and their moments, as well as the mixing proportions. For brevity, we report results only for $N = 1000$. The upper four panels in Figure 1 contain the upper and lower envelopes (dashed lines) over 1000 estimates of the component PDFs (upper left) and CDFs (upper right), as well as of the estimates of the corresponding marginal PDF (lower left) and CDF (lower right), together with their respective true values (solid lines).[8] The plots reveal the estimator to back out the underlying component mixtures quite well. In addition, the variability of the estimates of the f_k is not drastically different from the variability of the estimates of the marginal density.

The lower four panels in Figure 1 provide box plots of the sampling distribution of the mixing proportions (upper left), the mean of the weight function (upper right), as well as of the mean (lower left) and variance (lower right) of the component mixtures. All box plots are centered around the respective true values. The results are encouraging. All estimators are virtually median unbiased, and the interquartile ranges indicate quite low variability of the point estimates.

Experiment 2: A location-scale model. Our methods contribute to the analysis of non-separable fixed-effect models, that is, models of the form $y_t = g(x, \varepsilon_t)$, for some

[8] The marginal PDF was estimated by means of a kernel estimator based on all NT data points, using a Gaussian kernel and $1.06 * \widehat{\varsigma}(NT)^{-1/5}$ for the bandwidth, where $\widehat{\varsigma}$ is the empirical standard deviation of the data.

Figure 1: Estimates in Experiment 1 (Beta mixture)



Note: $N = 1,000$, $T = 4$, $I = J = 6$; results obtained over 1000 replications.

unobservable ε_t. A location-scale version is

$$y_t = x + \eta_t, \qquad \eta_t = \varsigma(x)\varepsilon_t, \qquad x \perp \varepsilon_t,$$

for some function ς. Suppose that ε_t is drawn from the generalized Beta distribution on $[-1,1]$ with $\vartheta_1 = \vartheta_2 = \vartheta$. Then

$$\mathbb{E}[y|x = x_k] = x_k, \qquad \mathbb{V}[y|x = x_k] = \frac{\varsigma(x_k)^2}{2\vartheta + 1}.$$

The location-scale model can be seen as a stripped-down version of a linear fixed-effect model or as a one-factor model. Note, however, that the factor x and the error η_t are not independent.

Below we report simulation results for the simple design with $\mathscr{X} = \{-1, 0, 1\}$ and ϖ the PMF that spreads its mass uniformly over these three support points. To generate realizations of η_t we set $\vartheta = 2$ and $\varsigma(x) = 1 + .5||x| - 1|$. With these parameter constellations, $\mathscr{Y} = [-2, 2]$, the support of f_k is the subset $[-1 + x_k, 1 + x_k]$, its mean is simply x_k, while its variance equals $1/20$ if $k \in \{1, 3\}$ and $9/20$ if $k = 2$. We maintain $T = 4$ and choose $I = J = 8$.

Table 2 contains the simulation results for $\widehat{K}$ over $10,000$ replications and has the same structure as Table 1 above. Here, the estimator is somewhat more likely to overestimate K, although it still approaches K as $N \to \infty$ and $\alpha \to 0$.

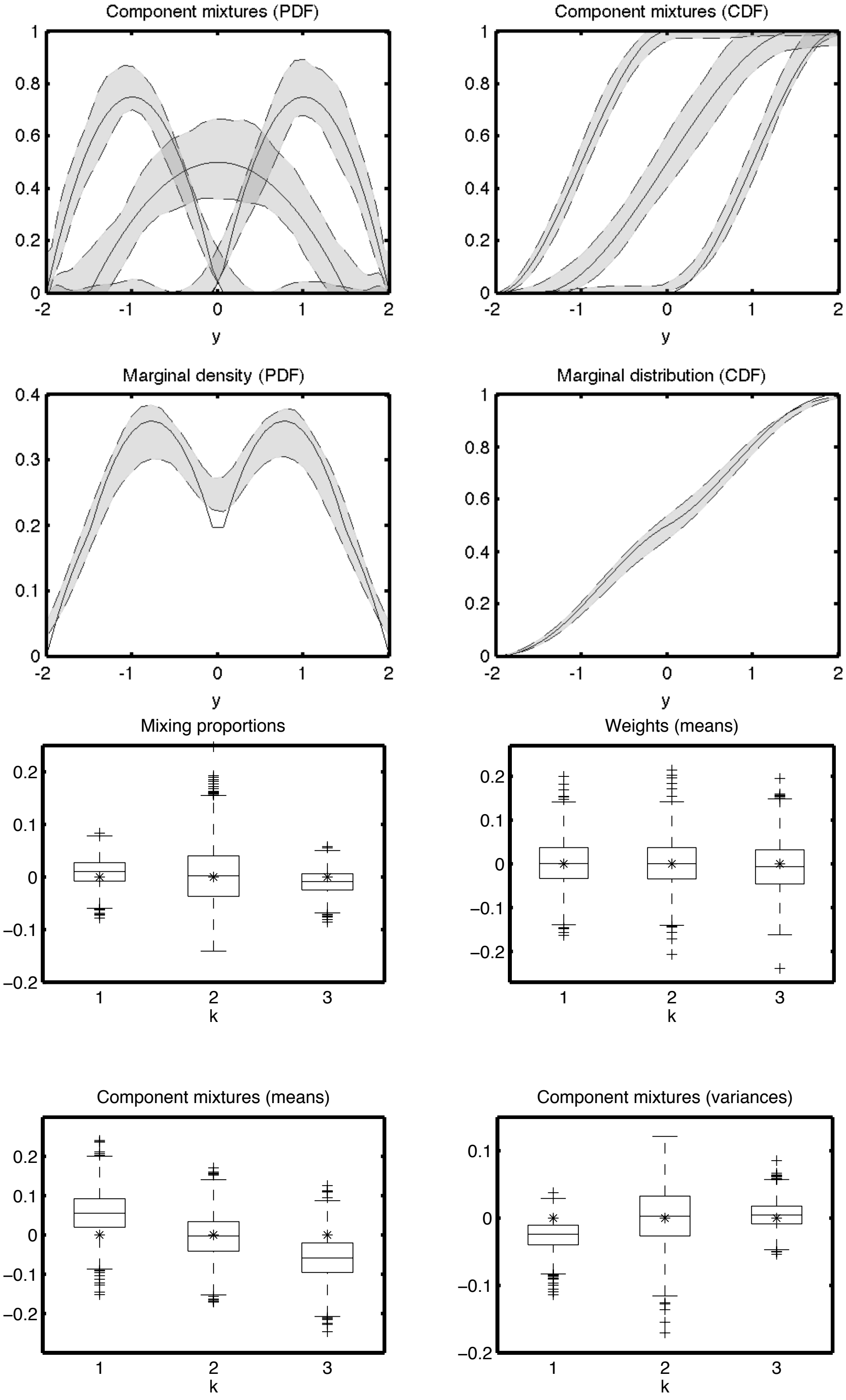
Table 2: Sequential rank test in Experiment 2 (Factor model)

	$\alpha = .100$			$\alpha = .050$			$\alpha = .025$		
N	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$	$\widehat{K}<K$	$\widehat{K}=K$	$\widehat{K}>K$
500	.000	.660	.340	.000	.719	.281	.000	.767	.233
750	.000	.733	.267	.000	.781	.219	.000	.815	.185
1000	.000	.769	.231	.000	.813	.187	.000	.853	.147
1500	.000	.810	.190	.000	.857	.143	.000	.885	.115
2000	.001	.830	.170	.000	.879	.121	.000	.904	.096
2500	.001	.856	.143	.000	.895	.105	.000	.923	.077

Note: $K = 3$, $T = 4$, $I = J = 8$; $10,000$ replications.

Figure 2, in turn, has the same layout as Figure 1, and similar conclusions may be drawn from it. It may be observed that the estimator of the second component mixture is somewhat more volatile than the others but, overall, we may conclude that the estimates reflect well the population densities and distributions. Further, the mixing proportions and various moments are all estimated well, although there tends to be somewhat more bias and

Figure 2: Estimates in Experiment 2 (Factor model)



Note: $N = 1,000$, $T = 4$, $I = J = 8$; results obtained over 1000 replications.

higher variability than in Experiment 1. Again, the distributions of the mixing proportion and variance associated with the second component mixture have a higher variance than those of the other components.

V EMPIRICAL APPLICATION

In this section we apply our methods to document heterogeneity in earnings dynamics of males using PSID[9] data from the period 1969–1998. The classic model (see, e.g., Hall and Mishkin, 1982) allows for heterogeneity in log-earnings levels through the inclusion of unit-specific intercepts. However, the recent literature has argued that unobserved heterogeneity in log earnings stretches beyond such additive fixed effects; see, e.g., Browning, Ejrnaes, and Alvarez (2010) for a discussion and an extensive empirical investigation in a parametric framework. Here, we adopt a fully nonparametric view on earnings. Our approach is very flexible in terms of unobserved heterogeneity. The cost of this is that we assume away the presence of state dependence.

From the PSID 1969–1998 we construct a set of three-period balanced subpanels, using a rolling window of length one.[10] This yields 28 subpanels. For each such subpanel, we obtain our measure of log (annual) earnings of unit n at time t, say y_{nt}, as the residual of a pooled regression of reported log earnings on a constant term, a set of time dummies, years of schooling, and a second-degree polynomial in experience. A graphical inspection of the marginal densities in each subpanel (not reported) does not suggest large disperion between the univariate marginal densities in a given subpanel, so that our smoothing policy seems reasonable.

Informal experimentation as well as testing for the number of components in the PSID data hints log earnings to decompose as a continuous mixture. In line with Heckman and Singer (1984) and many others since, we view the latent factor x as representing type heterogeneity and consider a discretization approach. Figures 3–4 provide plots of the estimated component mixtures in the PSID subpanels for the period 1969–1994 with K fixed to three. Larger values of K yielded a similar pattern in the component mixtures. We focus on a relatively small K for ease of exposition. In this way, one can think of the x as an indicator for low, intermediate, and high innate ability, for example. The plots were generated with Chebychev polynomials as basis functions, I set to five, and $J = .7 * \sqrt[3]{N}$ to

[9] Panel Study of Income Dynamics public use dataset. Produced and distributed by the University of Michigan with primary funding from the National Science Foundation, the National Institute of Aging, and the National Institute of Child Health and Human Development. Ann Arbor, MI.

[10] We excluded self-employed individuals and students, as well as individuals for whome earnings were top coded. We further restricted the sample to individuals between the ages of 20 and 60, with at most 40 years of experience.

accommodate the relatively strong increase in the number of cross-sectional observations in the subpanels spanning later time periods. N is reported below each plot. We used (4.2) to ensure non-negative density estimates that integrate to one.

The plots show well-separated unimodal component mixtures. The conditional densities are fairly symmetric about their respective modes, although the lowest one does have a significant right tail. This suggests that a flexible parametric specification involving location-scale densities may fit the data relatively well. Index the component mixtures $k = 1, 2, 3$ from left to right, i.e., the first component mixture has the lowest mode, etc.

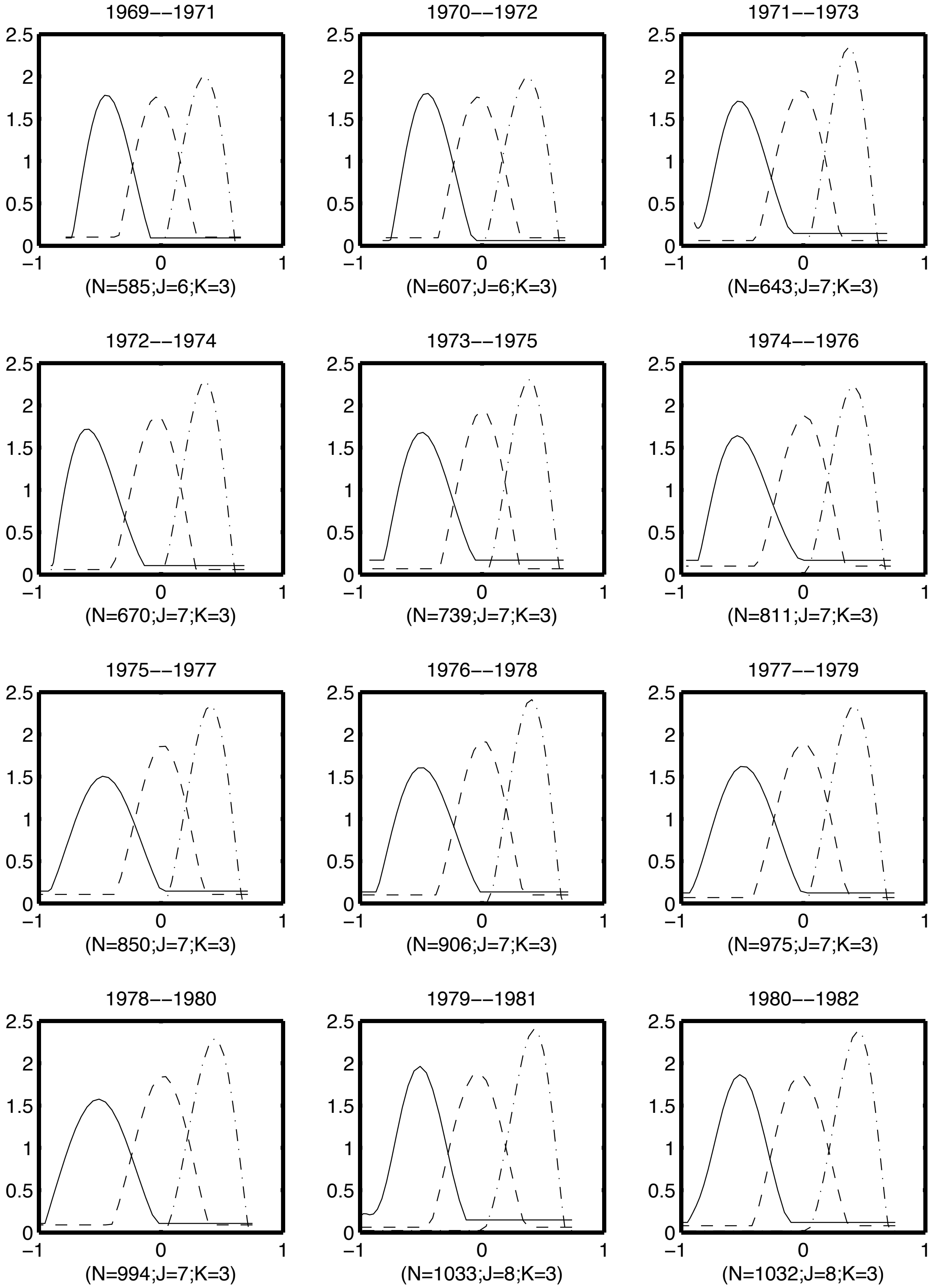
The plots further suggest the volatility of the component mixtures to be inversely related to the mode. To inspect the evolution of the location and variability over time, we estimated the mode and interquartile range of the component mixtures in each subpanel. The interquartile range was estimated by inverting the CDF of the components which, in turn, was computed using Clenshaw-Curtis quadrature (with 101 quadrature nodes, as before). The individual estimates are indicated by a $\bullet$ for $k = 1$, by an x for $k = 2$, and by a + for $k = 3$. To capture the overall time trend, regression lines through the individual points are also given. The regressors set consists a third-order polynomial in a time index and a constant term.

The left plot in Figure 5 shows the evolution of the mixture modes over the sampling period. As is clear from the regression lines, the dispersion between the modes tends to increase somewhat over time, with the mode for $k = 1$ decreasing, the mode for $k = 2$ remaining relatively constant, and the mode for $k = 3$ increasing slightly. In terms of location, this indicates the various component mixtures to move further away from each other, suggesting an increase in the between-group dispersion of log earnings.

The interquartile ranges are provided in the right plot of Figure 5. The plot confirms the intuition obtained from Figures 3–4 that the spread of the component mixtures is inversely related to their mode. Individuals at the lower end of the earnings distribution are exposed to higher volatility. Interestingly though, the difference in the interquartile ranges decreases over the sampling period. The regression lines show that the interquartile ranges of the first and second component mixture have decreased while the interquartile range of the high-end component mixture displays an upward trend.

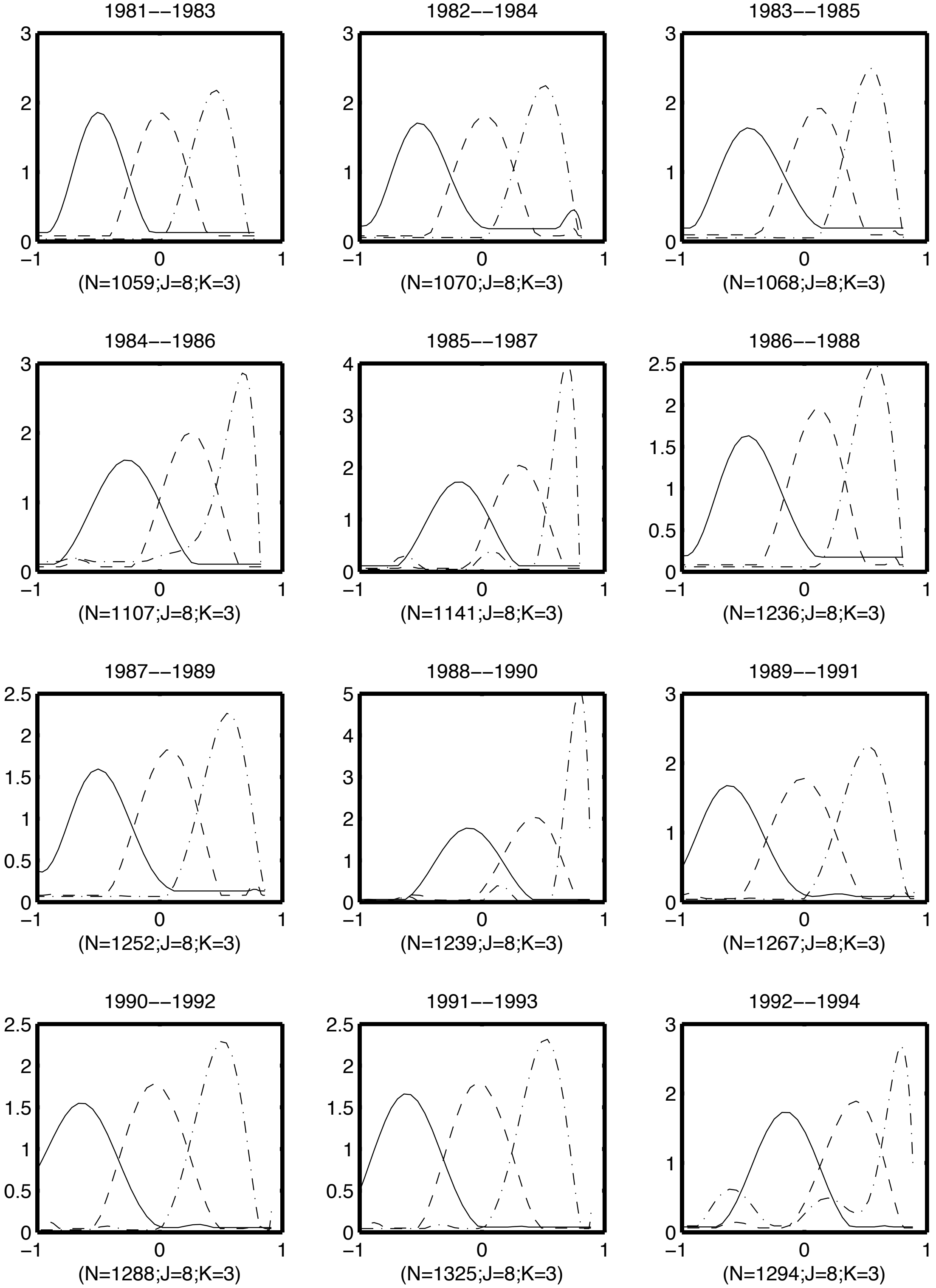
Overall, our analysis suggests the distribution of earnings to vary considerably with unobserved factors. The component distributions are well separated in terms of location. Individuals at the lower end of the (marginal) earnings distribution tend to be exposed to higher uncertainty, as measured by interquartile range. The increase in overall dispersion over time appears to be driven more by between-group than by within-group variation. Our results indicate the presence of unobserved heterogeneity beyond simple location shifts, as

Figure 3: Estimated conditional densities of log earnings



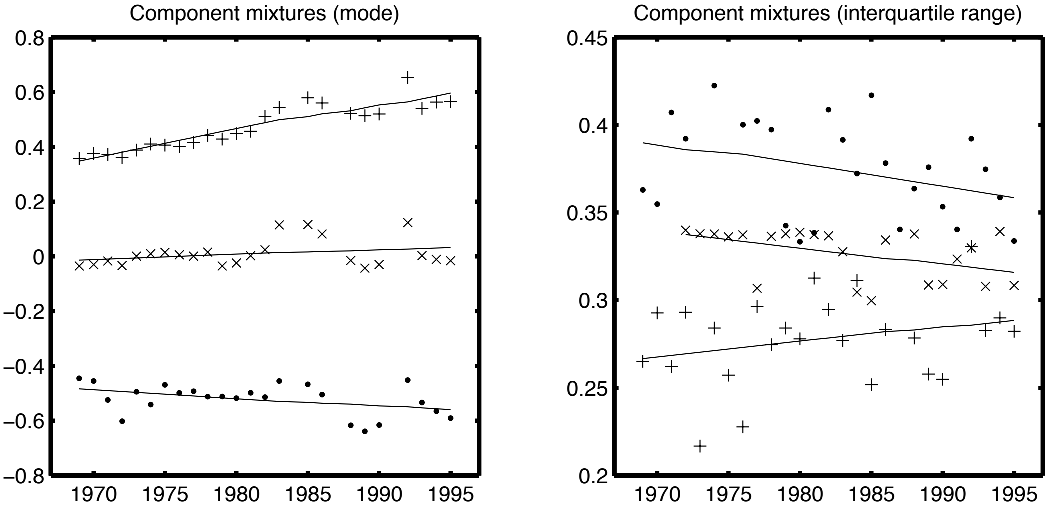
Note: In each plot, the densities are indicated as — ($k = 1$), −− ($k = 2$), and −· ($k = 3$). Conditional densities are ordered by mode.

Figure 4: Estimated conditional densities of log earnings (cont'd.)



Note: In each plot, the densities are indicated as — ($k = 1$), −− ($k = 2$), and −· ($k = 3$). Conditional densities are ordered by mode.

Figure 5: Estimated functionals of conditional log earnings densities



Note: In each plot, the functionals are indicated as • ($k = 1$), x ($k = 2$), and + ($k = 3$). Conditional densities are ordered by mode.

is evident from the different patterns in the interquartile ranges, for example. This implies the classic linear fixed-effect model with independent errors to be insufficiently flexible for modeling earnings processes.

VI CONCLUSION

We have discussed methods to estimate nonparametric finite-mixture models from short panel data. The estimators are straightforward to implement and have desirable large-sample properties. A Monte Carlo assessement further indicated good performance in small samples. As an empirical illustration we applied our approach to earnings data. We found evidence of substantial nonlinear heterogeneity in earnings processes, and documented their evolution over time.

In future work, we aim to provide a data-driven method for selecting the truncation parameter optimally, paving the path for adaptive estimation. Another potentially fruitful direction for further research is to investigate how our approach can be modified to allow K to grow with T. With some work, it should also be possible to extend our methodology to Markovian models. One area where this can be of use is in the estimation of dynamic discrete-choice models. Finally, our projection approach can potentially be used to estimate continuous mixtures. However, this leads to an ill-posed inverse problem, and the associated decrease in convergence rates.

APPENDIX A: AUXILIARY RESULTS

This appendix works toward Theorem A.1, which establishes the limit behavior of the $\widehat{\Delta}_j$. We do so by providing lemmata which, in turn, develop the large-sample behavior of the spectral-decomposition matrix $\widehat{\Theta}$ (Lemma A.1), the whitened matrices of Fourier coefficients, $\widehat{\daleth}_j$ (Lemma A.2), and the JADE estimator, $\widehat{\Xi}$ (Lemma A.3).

To state our results, some additional notation is needed. First, observe that $\mathrm{vec}[\widehat{\Sigma}_j - \Sigma_j]$ is asymptotically linear and that its influence function at $\vec{y}_{nT}$, $\psi_{\Sigma_j}(\vec{y}_{nT})$, say, has typical element

$$|\varrho_3|^{-1} \sum_{(t_1,t_2,t_3)\in\varrho_3} \chi_{i_1 i_2 j}(y_{nt_1}, y_{nt_2}, y_{nt_3}) - \sigma_{i_1 i_2 j},$$

for all $j = 1, 2, \ldots, J$. Let $\overline{\Upsilon} \equiv (\overline{\upsilon}_1, \overline{\upsilon}_2, \ldots, \overline{\upsilon}_K)'$ for $\overline{\upsilon}_k \equiv \upsilon_k \otimes \upsilon_k$ and define the selection matrix Q^{-1} through the relation $Q^{-1}\epsilon = \mathrm{vec}[\mathrm{diag}(\epsilon)]$, where ϵ is any K-vector. Use these expressions to introduce

$$\psi_\Theta(\vec{y}_T) \equiv \left[(\mathsf{I}_I \otimes \Lambda^{-3/2}\Upsilon') - 1/2(\Upsilon \otimes \mathsf{I}_K)Q^{-1}\Lambda^{-3/2}\overline{\Upsilon}\right]\psi_{\Sigma_0}(\vec{y}_T),$$
$$\psi_{\Theta'}(\vec{y}_T) \equiv \left[(\Lambda^{-3/2}\Upsilon' \otimes \mathsf{I}_I) - 1/2(\mathsf{I}_K \otimes \Upsilon)Q^{-1}\Lambda^{-3/2}\overline{\Upsilon}\right]\psi_{\Sigma_0}(\vec{y}_T).$$

The interpretation of these functions is that they are the respective influence functions of $\widehat{\Theta}$ and $\widehat{\Theta}'$, as stated in Lemma A.1.

Lemma A.1 (Spectral decomposition)**.** *The estimators $\widehat{\Theta}$ and $\widehat{\Theta}'$ of the matrices Θ and Θ' satisfy*

$$\sqrt{N}\mathrm{vec}[\widehat{\Theta} - \Theta] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\Theta), \qquad \sqrt{N}\mathrm{vec}[\widehat{\Theta}' - \Theta'] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Theta'}),$$

for $\mathscr{V}_\Theta \equiv \mathbb{E}[\psi_\Theta(\vec{y}_T)\psi_\Theta(\vec{y}_T)']$ and $\mathscr{V}_{\Theta'} \equiv \mathbb{E}[\psi_{\Theta'}(\vec{y}_T)\psi_{\Theta'}(\vec{y}_T)']$, respectively, as $N \to \infty$.

Now use Lemma A.1 to construct

$$\psi_{\daleth_j}(\vec{y}_T) \equiv (\Theta\Sigma_j \otimes \mathsf{I}_K)\psi_\Theta(\vec{y}_T) + (\Theta \otimes \Theta)\psi_{\Sigma_j}(\vec{y}_T) + (\mathsf{I}_K \otimes \Theta\Sigma_j)\psi_{\Theta'}(\vec{y}_T),$$

for each $j = 1, 2, \ldots, J$. Then Lemma A.2 follows readily.

Lemma A.2 (Whitening)**.** *For each $j = 1, 2, \ldots, J$, the estimator $\widehat{\daleth}_j = \widehat{\Theta}\widehat{\Sigma}_j\widehat{\Theta}'$ of the whitened matrix $\daleth_j = \Theta\Sigma_j\Theta'$ satisfies*

$$\sqrt{N}\mathrm{vec}[\widehat{\daleth}_j - \daleth_j] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\daleth_j}), \qquad \mathscr{V}_{\daleth_j} \equiv \mathbb{E}[\psi_{\daleth_j}(\vec{y}_T)\psi_{\daleth_j}(\vec{y}_T)'],$$

as $N \to \infty$.

Moving on to the JADE estimator, let

$$\psi_{\Xi}(\vec{y}_{nT}) \equiv (\mathsf{I}_K \otimes \Xi\aleph_J)\sum_{j=1}^{J}\nabla_j(\Xi' \otimes \Xi')\psi_{\daleth_j}(\vec{y}_{nT}),$$

$$\psi_{\Xi'}(\vec{y}_{nT}) \equiv (\Xi\aleph_J \otimes \mathsf{I}_K)\sum_{j=1}^{J}\nabla_j(\Xi' \otimes \Xi')\psi_{\daleth_j}(\vec{y}_{nT}),$$

where $\aleph_J$ is the $K \times K$ matrix whose (k_1, k_2)th element is

$$[\aleph_J]_{k_1,k_2} \equiv \begin{cases} \left[\sum_{j=1}^{J}(\gamma_{k_1 j} - \gamma_{k_2 j})^2\right]^{-1} & \text{if } k_1 \neq k_2 \\ 0 & \text{if } k_1 = k_2 \end{cases},$$

and $\nabla_j \equiv \operatorname{diag}[\operatorname{vec}[\overline{\nabla}_j]]$ with $\overline{\nabla}_j \equiv \Delta_j \iota_K \iota_K' - \iota_K \iota_K' \Delta_j$ for $j = 1, 2, \ldots, J$. The asymptotic distribution of JADE is stated in Lemma A.3.

Lemma A.3 (JADE)**.** *The JADE estimator defined in* (3.4) *satisfies*

$$\sqrt{N}\operatorname{vec}[\widehat{\Xi} - \Xi] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Xi}), \qquad \sqrt{N}\operatorname{vec}[\widehat{\Xi}' - \Xi'] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Xi'}),$$

for $\mathscr{V}_{\Xi} \equiv \lim_{J\to\infty}\mathbb{E}[\psi_{\Xi}(\vec{y}_T)\psi_{\Xi}(\vec{y}_T)']$ *and* $\widetilde{\mathscr{V}}_{\Xi'} \equiv \lim_{J\to\infty}\mathbb{E}[\psi_{\Xi'}(\vec{y}_T)\psi_{\Xi'}(\vec{y}_T)']$*, respectively, as* $N \to \infty$*.*

To see why $\widehat{\Xi}$ converges at the parametric rate, note that, even though ψ_{Ξ} and $\psi_{\Xi'}$ are triangular arrays, $\sum_{j=1}^{J}\nabla_j = O(1)$, because $\gamma_{kj} \to 0$ as $j \to \infty$.

On combining Lemma A.2 and Lemma A.3, and introducing

$$\psi_{\Delta_j}(\vec{y}_T) \equiv (\mathsf{I}_K \otimes \Xi'\daleth_j)\psi_{\Xi}(\vec{y}_T) + (\Xi' \otimes \Xi')\psi_{\daleth_j}(\vec{y}_T) + (\Xi'\daleth_j \otimes \mathsf{I}_K)\psi_{\Xi'}(\vec{y}_T),$$

we obtain Theorem A.1.

Theorem A.1 (Fourier coefficients)**.** *For each* $j = 1, 2, \ldots, J$*, the estimator* $\widehat{\Delta}_j = \widehat{\Xi}'\widehat{\daleth}_j\widehat{\Xi}$ *of the matrix* $\Delta_j = \Xi'\daleth_j\Xi$ *satisfies*

$$\operatorname{vec}[\widehat{\Delta}_j - \Delta_j] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Delta_j}), \qquad \mathscr{V}_{\Delta_j} \equiv \mathbb{E}[\psi_{\Delta_j}(\vec{y}_T)\psi_{\Delta_j}(\vec{y}_T)']$$

as $N \to \infty$

Theorem A.1 shows that, although one can not draw data directly from the component mixtures, their Fourier coefficients can be estimated at the parametric rate.

Proof of Lemma A.1. Recall that Σ_0 is real, symmetric, and has rank K. Hence, it has K positive and distinct eigenvalues. Further, $\widehat{\Sigma}_0$—which, by construction, too, is symmetric—satisfies $\sqrt{N}\text{vec}[\widehat{\Sigma}_0 - \Sigma_0] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Sigma_0})$. From Eaton and Tyler (1991) [Theorem 4.2] and Magnus (1985) [Theorem 1], it then follows that $\sqrt{N}(\widehat{\lambda} - \lambda) \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\lambda)$, where $[\mathscr{V}_\lambda]_{k_1,k_2} \equiv (\upsilon_{k_1}' \otimes \upsilon_{k_1}')\mathscr{V}_{\Sigma_0}(\upsilon_{k_2} \otimes \upsilon_{k_2})$. Because $\text{vec}[\Lambda] = Q^{-1}\lambda$, the Jacobian associated with the transformation from λ to $\text{vec}[\Lambda^{-1/2}]$ is $-\frac{1}{2}Q^{-1}\,\text{diag}[\lambda_1^{-3/2}, \lambda_2^{-3/2}, \ldots, \lambda_K^{-3/2}]$. Hence, an application of the Delta method gives

$$\sqrt{N}\text{vec}[\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Lambda^{-1/2}}), \tag{B.1}$$

for $\mathscr{V}_{\Lambda^{-1/2}} \equiv 1/4 Q^{-1}\Lambda^{-3/2}\overline{\Upsilon}\mathscr{V}_{\Sigma_0}\overline{\Upsilon}'\Lambda^{-3/2}Q^{-1\prime}$. Moving on, from Bura and Pfeiffer (2008) [Corollary 1], we have that the estimated eigenvectors satisfy

$$\sqrt{N}\text{vec}[\widehat{\Upsilon} - \Upsilon] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\Upsilon), \qquad \sqrt{N}\text{vec}[\widehat{\Upsilon}' - \Upsilon'] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Upsilon'}), \tag{B.2}$$

where $\mathscr{V}_\Upsilon \equiv (\Lambda^{-1}\Upsilon' \otimes \mathsf{I}_I)\mathscr{V}_{\Sigma_0}(\Upsilon\Lambda^{-1} \otimes \mathsf{I}_I)$ and $\mathscr{V}_{\Upsilon'} \equiv (\mathsf{I}_I \otimes \Lambda^{-1}\Upsilon')\mathscr{V}_{\Sigma_0}(\mathsf{I}_I \otimes \Upsilon\Lambda^{-1})$. On combining (B.1) and (B.2) with the linearization $\widehat{\Theta} - \Theta = (\widehat{\Lambda} - \Lambda)^{-1/2}\Upsilon' + \Lambda^{-1/2}(\widehat{\Upsilon} - \Upsilon)' + o_P(1/\sqrt{N})$, and recalling that $\mathscr{V}_{\Sigma_0} = \mathbb{E}[\psi_{\Sigma_0}(\vec{y}_T)\psi_{\Sigma_0}(\vec{y}_T)']$, $\sqrt{N}\text{vec}[\widehat{\Theta} - \Theta] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_\Theta)$ and $\sqrt{N}\text{vec}[\widehat{\Theta}' - \Theta'] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Theta'})$ follow. □

Proof of Lemma A.2. Fix $j \in \{1, 2, \ldots, J\}$. From (3.1), the (i_1, i_2)th element of $\widehat{\Sigma}_j$ takes the form

$$\widehat{\sigma}_{i_1 i_2 j} = \frac{1}{N}\frac{1}{|\varrho_3|}\sum_{n=1}^{N}\sum_{(t_1,t_2,t_3)\in\varrho_3}\chi_{i_1 i_2 j}(y_{nt_1}, y_{nt_2}, y_{nt_3}),$$

and $\sqrt{N}\text{vec}[\widehat{\Sigma}_j - \Sigma_j] \xrightarrow{L} \mathcal{N}(0, \mathscr{V}_{\Sigma_j})$. Combining this with Lemma A.1 and the linearization

$$\text{vec}[\widehat{\daleth}_j - \daleth_j] = (\Theta\Sigma_j \otimes \mathsf{I}_K)\text{vec}[\widehat{\Theta} - \Theta] + (\Theta \otimes \Theta)\text{vec}[\widehat{\Sigma}_j - \Sigma_j] + (\mathsf{I}_K \otimes \Theta\Sigma_j)\text{vec}[\widehat{\Theta}' - \Theta'] + o_P(1/\sqrt{N})$$

yields the result. □

Proof of Lemma A.3. Taylor-expanding the JADE first-order conditions around Ξ and proceeding along the lines of the proof to Theorem 5 in Bonhomme and Robin (2009) gives

$$\text{vec}[\widehat{\Xi} - \Xi] = -(\mathsf{I}_K \otimes \Xi)(\mathsf{I}_K \otimes \aleph_J)\sum_{j=1}^{J}\nabla_j(\Xi' \otimes \Xi')\text{vec}[\widehat{\daleth}_j - \daleth_j] + o_P(1/\sqrt{N}). \tag{B.3}$$

Recall that $\aleph_J$ is the $K \times K$ matrix whose (k_1, k_2)th element is

$$[\aleph_J]_{k_1,k_2} = \left\{ \begin{array}{cl} \Big[\sum_{j=1}^J (\gamma_{k_1 j} - \gamma_{k_2 j})^2\Big]^{-1} & \text{if } k_1 \neq k_2 \\ 0 & \text{if } k_1 = k_2 \end{array} \right. .$$

Note that Assumption 2.2 ensures that $\aleph_J$ is well defined. Moreover, as $N \to \infty$, $\aleph_J \to \aleph$, where $\aleph$ is the matrix whose (k_1, k_2)th element equals

$$[\aleph]_{k_1,k_2} \equiv \left\{ \begin{array}{cl} \|f_{k_1}(y) - f_{k_2}(y)\|_2^{-2} & \text{if } k_1 \neq k_2 \\ 0 & \text{if } k_1 = k_2 \end{array} \right. ,$$

because, for all $k_1, k_2 = 1, 2, \ldots, K$, $\lim_{J\to\infty} \sum_{j=1}^J (\gamma_{k_1 j} - \gamma_{k_2 j})^2 = \int_{\mathscr{Y}} |f_{k_1}(y) - f_{k_2}(y)|^2 \, \mathrm{d}y$ by orthonormality of the basis functions. Continuing on, establishing asymptotic normality of $\widehat{\Xi}$ requires verifying that the triangular array vector

$$\frac{1}{N}\sum_{n=1}^N \delta(\vec{y}_{nT}), \qquad \delta(\vec{y}_{nT}) \equiv \sum_{j=1}^J \nabla_j (\Xi' \otimes \Xi') \; \psi_{\neg_j}(\vec{y}_{nT}), \tag{B.4}$$

satisfies the conditions of Lyapunov's central limit theorem. Thus, it suffices to show that (i) $\mathbb{E}[c'\delta(\vec{y}_T)] = 0$, (ii) $\mathbb{V}[c'\delta(\vec{y}_T)] = O(1)$, and (iii) $\mathbb{E}[|c'\delta(\vec{y}_T)/\sqrt{N}|^{2+d}] = o(1)$ for some $d > 0$ and any vector of finite constants c that satisfies $c'c = 1$. Condition (i) follows readily from the fact that $\psi_{\Sigma_j}(\vec{y}_T)$ has zero mean. Condition (ii) follows from the fact that $\mathbb{E}[\psi_{\neg_{j_1}}(\vec{y}_T)\psi_{\neg_{j_2}}(\vec{y}_T)'] = O(1)$ and $\sum_{j=1}^J \nabla_j = O(1)$. To verify Condition (iii), observe that

$$\begin{aligned} \Big|\frac{c'\delta(\vec{y}_T)}{\sqrt{N}}\Big| \leq{} & \Big|\frac{c' \sum_{j=1}^J [\nabla_j (\Xi' \otimes \Xi')(\Theta\Sigma_j \otimes \mathsf{I}_K)]\psi_{\Theta}(\vec{y}_T)}{\sqrt{N}}\Big| + \Big|\frac{c' \sum_{j=1}^J [\nabla_j (\Xi'\Theta \otimes \Xi'\Theta)]\psi_{\Sigma_j}(\vec{y}_T)}{\sqrt{N}}\Big| \\ & + \Big|\frac{c' \sum_{j=1}^J [\nabla_j (\Xi' \otimes \Xi')(\mathsf{I}_K \otimes \Theta\Sigma_j)]\psi_{\Theta'}(\vec{y}_T)}{\sqrt{N}}\Big|. \end{aligned}$$

The first and third right-hand side terms are readily shown to have expectation $o(1)$. For the second term, note that

$$\begin{aligned} & \mathbb{E}\Big|\frac{c' \sum_{j=1}^J [\nabla_j (\Xi'\Theta \otimes \Xi'\Theta)]\psi_{\Sigma_j}(\vec{y}_T)}{\sqrt{N}}\Big|^{2+d} \\ & \leq \frac{\Big[\sum_{j=1}^J \Big\| c'\nabla_j (\Xi'\Theta \otimes \Xi'\Theta)\Big\|^2\Big]^{\frac{2+d}{2}} \mathbb{E}\Big[\sum_{j=1}^J \Big\|\psi_{\Sigma_j}(\vec{y}_T)\Big\|^2\Big]^{\frac{2+d}{2}}}{N^{\frac{2+d}{2}}} \\ & \leq O\Big(\frac{1}{N}\Big)^{\frac{2+d}{2}} \Big[\sum_{j=1}^J \Big(\mathbb{E}\Big\|\psi_{\Sigma_j}(\vec{y}_T)\Big\|^{2+d}\Big)^{\frac{2}{2+d}}\Big]^{\frac{2+d}{2}} \\ & \leq O\Big(\frac{1}{N}\Big)^{\frac{2+d}{2}} O\Big(J\zeta(J)^2\Big)^{\frac{2+d}{2}} = O\Big(\frac{J\zeta(J)^2}{N}\Big)^{\frac{2+d}{2}}, \end{aligned}$$

which is $o(1)$ by Assumption 3.2. Here, the first two inequalities follow from the Cauchy-Schwarz inequality and Minkowski's inequality, respectively, and the third inequality comes from boundedness of the marginal density and the fact that the basis functions are bounded in Euclidean norm by the sequence $\zeta(J)$. Hence, $\sum_{n=1}^N \delta(\vec{y}_{nT})/\sqrt{N}$ converges in law to a normal random variable. In tandem with (B.3), (B.4), and an application of the Delta method, this proves the result. □

Proof of Theorem A.1. The result follows on combining Lemma A.2 and Lemma A.3 with a linearization of $\text{vec}[\widehat{\Delta}_j - \Delta_j]$. □

Proof of Theorem 2.1. The proof is given in the main text. □

Proof of Corollary 2.2. The result follows immediately from Theorem 2.1. □

Proof of Theorem 3.1. The result follows from an application of Theorem 5.2 in Robin and Smith (2000). □

Proof of Theorem 3.2. Fix $k \in \{1, 2, \ldots, K\}$ throughout. Start with Theorem 3.2(i). Using orthonormality, the MISE decomposes as

$$\mathbb{E}\|\widehat{f}_k(y) - f_k(y)\|_2^2 = \sum_{j=1}^J \mathbb{E}[(\widehat{\gamma}_{kj} - \gamma_{kj})^2] + \|f_k(y;J) - f_k(y)\|_2^2.$$

The second term is $O(J^{-2\beta})$ by Assumption 3.1. For the first term, with $\widehat{\gamma}_{kj} - \gamma_{kj} = q_k' \text{vec}[\widehat{\Delta}_j - \Delta_j]$, Theorem A.1 gives $\widehat{\gamma}_{kj} - \gamma_{kj} = O_P(1/\sqrt{N})$. Hence, $\sum_{j=1}^J \mathbb{E}[(\widehat{\gamma}_{kj} - \gamma_{kj})^2] = O(J/N)$, which proves the MISE result. To obtain the uniform-convergence rate, we may proceed similarly. Let $\gamma_k \equiv (\gamma_{k1}, \gamma_{k2}, \ldots, \gamma_{kJ})'$ and define $\widehat{\gamma}_k$ similarly. Then we have that $\sup_{y\in\mathscr{Y}} |\widehat{f}_k(y) - f_k(y)|$ is bounded by $\sup_{y\in\mathscr{Y}} \|\mathcal{X}_J(y)\| [\|\widehat{\gamma}_k - \gamma_k\| + O(J^{-\beta})] = O(\zeta(J))\, O_P(\sqrt{J}/\sqrt{N} + J^{-\beta})$, which follows from repeated use of the triangle and Cauchy-Shwarz inequalities, together with the rate result from Theorem A.1 and the fact that $\sup_{y\in\mathscr{Y}} \|\mathcal{X}_J(y)\| = O(\zeta(J))$. □

Proof of Theorem 3.3. Fix $k \in \{1, 2, \ldots, K\}$, $y \in \mathscr{Y}$ throughout. On adding and subtracting $f_k(y;J)$, we have $\widehat{f}_k(y) - f_k(y) = N^{-1} \sum_{n=1}^N \sum_{j=1}^J q_k' \psi_{\Delta_j}(\vec{y}_{nT}) \chi_j(y) + O(J^{-\beta})$. Under our assumptions, the bias term is asymptotically negligible. Theorem A.1 and Lemmata A.1–A.3, together with a small calculation, allow the leading right-hand side term to be simplified as

$$\widehat{f}_k(y) - f_k(y;J) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} q_k'(\Xi'\Theta \otimes \Xi'\Theta)\psi_{\Sigma_j}(\vec{y}_{nT})\chi_j(y) + O_P(1/\sqrt{N}).$$

On recalling the form of the function ψ_{Σ_j}, a small calculation gives

$$q_k'(\Xi'\Theta \otimes \Xi'\Theta)\psi_{\Sigma_j}(\vec{y}_{nT}) = |\varrho_3|^{-1} \sum_{(t_1,t_2,t_3)\in\varrho_3} \Big[\sum_{i_1=1}^{I}\sum_{i_2=1}^{I} \xi_k'\theta_{i_1}\chi_{i_1 i_2 j}(y_{nt_1}, y_{nt_2}, y_{nt_3})\theta_{i_2}'\xi_k - \gamma_{kj}\Big],$$

On using the tensor-product structure of the multivariate basis functions, we arrive at the expression in (3.6). Establishing normality then requires verifying that (i) $\mathbb{E}[\psi_{f_k}(\vec{y}_T)] = 0$, (ii) $\mathbb{E}[\psi_{f_k}(\vec{y}_T)\psi_{f_k}(\vec{y}_T)'] = O(\wp_J(y))$, and (iii) $\mathbb{E}[|\psi_{f_k}(\vec{y}_T)/\sqrt{N\wp_J(y)}|^{2+d}] = o(1)$ for some $d > 0$. Condition (i) is immediate from the form of ψ_{f_k}. To verify Condition (ii), use the law of iterated expectations to see that it suffices to show that $\mathbb{E}[\overline{\tau}_k(y_1,y_2)^2|x = x_{k'}] = O(1)$ and that $\mathbb{E}[\kappa_J(y_3;y)^2|x = x_{k'}] = O(\wp_J(y))$ for any $k' = 1, 2, \ldots, K$. From Viollaz (1989) [Theorem 2.2.2], $\mathbb{E}[\kappa_J(y;y)^2|x = x_{k'}]/\wp_J(y) \to f_{k'}(y)$, which is $O(1)$ by Assumption 2.1. For the first term, on letting $\mho_{k'} \equiv \mathbb{E}[\mathcal{X}_I(y)\mathcal{X}_I(y)'|x = x_{k'}]$ and defining Q_k through the relation $q_k = \operatorname{vec}[Q_k]$, $\mathbb{E}[\overline{\tau}_k(y_1,y_2)^2|x = x_{k'}]$ equals

$$\big[q_k'(\Xi'\Theta \otimes \Xi'\Theta)\ \operatorname{vec}[\mho_{k'}]\big]^2 = \operatorname{tr}[\mho_{k'}(\Theta'\Xi Q_k'\Xi\Theta')(\Theta'\Xi Q_k'\Xi\Theta')'\mho_{k'}] = O(1),$$

where the first transition follows from elementary properties of the vec operator. Finally, using similar arguments as in the proof to Lemma A.3, we obtain

$$\mathbb{E}\Big[\Big|\frac{\psi_{f_k}(\vec{y}_T)}{\sqrt{N\wp_J(y)}}\Big|^{2+d}\Big] = O\Big(\frac{J\zeta(J)^2}{N}\Big)^{2+d} O\Big(\frac{1}{\wp_J(y)}\Big)^{2+d} = o(1),$$

with the conclusion following from the growth rates in Assumption 3.2. This establishes Condition (iii). □

Proof of Theorem 3.4. Theorem A.1 implies that $\|\widehat{\Gamma} - \Gamma\| = O_P(1/\sqrt{N})$. Assumption 2.2 ensures that $\Gamma'\Gamma$ has full rank. Hence, $(\widehat{\Gamma}'\widehat{\Gamma})^{-1} \xrightarrow{P} (\Gamma'\Gamma)^{-1}$. Further, a linearization yields $\widehat{\Gamma}'\widehat{\sigma} - \Gamma'\sigma = (\widehat{\Gamma}' - \Gamma')\sigma + \Gamma'(\widehat{\sigma} - \sigma) + o_P(1/\sqrt{N})$, In tandem with the asymptotic results on $\widehat{\Gamma}$ and $\widehat{\sigma}$, the Delta method then provides the result. □

Proof of Theorem 3.5. The result follows from standard arguments on two-step estimators using first-step series estimators, making use of Theorems 3.2–3.4 and Assumption 3.3; see, e.g., Newey (1994). □

REFERENCES

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics*, 149:12–25.

Browning, M., M. Ejrnæs, and J. Alvarez. (2010). modeling income processes with lots of heterogeneity. *Review of Economic Studies*, 77:1353–1381.

Bura, E. and R. Pfeiffer (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics & Probability Letters*, 78:2275–2280.

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

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (eds), *Handbook of Econometrics*, Volume VI, Part B. Elsevier.

Eaton, M. L. and D. E. Tyler (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Annals of Statistics*, 19:260–271.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications.* Springer.

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

Hall, P. (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval. *Journal of Multivariate Analysis*, 12:432–449.

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). Estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224.

Hall, R. and F. Mishkin (1982). The sensitivity of consumption to transitory income: Estimates from panel data on households. *Econometrica*, 50:461–481.

Heckman, J. J. and B. Singer (1984). A method of minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52:271–320.

Henry, M., Y. Kitamura, and B. Salanié (2011). Identifying finite mixtures in econometric models. Unpublished manuscript.

Kasahara, H. and K. Shimotsu (2010). Nonparametric identification of multivariate mixtures. Unpublished manuscript.

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

Magnus, J. (1985). On differentiating eigenvalues and eigenvectors. *Econometric Theory*, 1:179–191.

McDonald, J. B. (1984). Some generalized functions for the size distributions of income. *Econometrica*, 52:647–663.

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

Newey, W. K. (1994). Series estimation of regression functionals. *Econometric Theory*, 10:1–28.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168.

Robin, J.-M. and R. J. Smith (2000). Tests of rank. *Econometric Theory*, 16:151–175.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053.

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

Viollaz, A. J. (1989). Nonparametric estimation of probability density functions based on orthogonal expansions. *Revista Matematica de la Universidad Complutense de Madrid*, 2:41–82.