Nonparametric estimation of finite mixtures

Stéphane Bonhomme, Koen Jochmans, Jean-Marc Robin

SciencesPo. | Department of Economics

Discussion paper 2013-09

Nonparametric estimation of finite mixtures

Stéphane Bonhomme
Koen Jochmans
Jean-Marc Robin

Sciences Po Economics Discussion Papers

Nonparametric estimation of finite mixtures

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

[Revised March 28, 2013; First draft January 2012]

Abstract. 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 mixture 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 mixing proportions and the mixture distributions, and various functionals thereof. We also discuss inference on the number of components. 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.

Introduction

Finite-mixture models are widely used in statistical analysis. Popular applications include modeling unobserved heterogeneity in structural models, learning about individual behavior from grouped data, and dealing with corrupted data. As univariate mixtures are not nonparametrically identified from cross-sectional data, the conventional approach to inference is parametric; see, e.g., McLachlan and Peel (2000) for an overview. Recently, following the seminal work of Hall and Zhou (2003), a number of studies have shown that data on repeated measurements provide a powerful source of identification in mixture models; see Hall et al. (2005), Hu (2008), Allman et al. (2009), and Kasahara and Shimotsu (2009), among others.[1]

In this paper we present a simple and constructive identification argument and develop a practical procedure to nonparametrically estimate finite mixtures from data on repeated

†*Address for correspondence:* Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France. *E-mail:* koen.jochmans@sciences-po.org. Replication material is available at econ.sciences-po.fr/staff/koen-jochmans.

[1] In related work, Henry et al. (2013) study the identifying power of exclusion restrictions in conditional models.

measurements. We show that finite mixtures are generically identified from panel data when three or more measurements are available, and the component distributions are stationary and linearly independent. Our conditions are of a similar nature as those assumed by Allman et al. (2009). However, our approach is constructive. The resulting estimators are attractive from a computational point of view, even when the number of components is large, and have desirable large-sample properties.

We will focus on the case where outcomes are continuous and component densities are square-integrable. Our analysis is based on projections of the marginal densities and mixture components onto an orthonormal basis of functions. The mixture structure imposes a set of multilinear restrictions on the generalized Fourier coefficients of these densities, which takes the form of a joint-diagonalization problem for a set of commuting symmetric matrices. The eigenvalues of these matrices identify the Fourier coefficients of the component densities up to arbitrary relabeling. The component densities are therefore identified almost everywhere on their support. The mixing proportions are then identified as the solution to a linear system of equations.

To turn this identification result into an operational estimation procedure we appeal to the literature on blind source separation, which has extensively investigated numerical methods for the joint diagonalization of a set of matrices. Specifically, we propose estimating the Fourier coefficients of the component densities via the joint approximate-diagonalization algorithm developed by Cardoso and Souloumiac (1993). This procedure is straightforward to implement, highly stable, and computationally extremely fast. Given estimates of the projection coefficients, we then construct bona fide series estimators of the component densities and a least-squares estimator of the mixing proportions. In addition, while most of the analysis is conducted under the assumption that the number of components is known, we also construct an estimator of the number of components using a sequential-testing approach.

We derive integrated squared-error and uniform convergence rates of the estimator of the component densities, and provide conditions for pointwise asymptotic normality. The convergence rates coincide with those that would be obtained if the data could be sampled directly from the component distributions. We further show that our estimator of the mixing proportions converges at the parametric rate, and present distribution theory for two-step semiparametric GMM estimators of finite-dimensional parameters defined through moment restrictions involving the mixture components. Extensive numerical experimentation, which we report on through a series of Monte Carlo illustrations, provides encouraging evidence on the small-sample performance of our procedures.

As an empirical application, we investigate the presence of unobserved heterogeneity in earnings data from the PSID for the period 1969–1998. This application illustrates the usefulness of our approach as a tool to decompose earnings inequality into within-group and between-group components, where group membership is unobserved and inferred from the longitudinal dimension of the data. We find evidence of substantial unobserved

heterogeneity that goes beyond what is captured by the location-scale models commonly used in this literature. We further document the evolution of earnings inequality over time and compare it to the recent work by Moffitt and Gottschalk (2012).

The paper is organized as follows. Section 1 formalizes the setup and presents our identification results. Section 2 contains an exposition of the resulting estimators while Section 3 is devoted to a derivation of their large-sample properties. Section 4 presents simulation evidence on the performance of the various estimators, and Section 5 contains our application to earnings dynamics. Three appendices collect auxiliary theorems and technical proofs.

1. Identification

Let x be a latent discrete random variable, normalized to take on value $k \in \{1, 2, ..., K\}$ with probability $\omega_k > 0$. Let y be an observable outcome variable with probability density function (PDF)

$$f(y) = \sum_{k=1}^{K} f_k(y)\ \omega_k,$$

where f_k denotes the PDF of y conditional on $x = k$. We assume that the component densities f_k are supported on the interval $[-1, 1]$. This is without loss of generality because we can always transform the outcome by means of a suitably chosen strictly-monotonic function.

Let ρ be an almost everywhere positive and integrable function on the interval $[-1, 1]$. Let $\{\chi_i, i \geq 0\}$ be a complete system of functions that are orthonormal on this interval with respect to the weight function ρ.[2]

Assume that the f_k are square-integrable with respect to ρ on $[-1, 1]$. The projection of f_k onto $\{\chi_0, \chi_1, ..., \chi_J\}$ is given by

$$\text{Proj}_J[f_k] \equiv \sum_{j=0}^{J} b_{jk}\chi_j, \qquad b_{jk} \equiv \int_{-1}^{1} \chi_j(y) f_k(y)\rho(y)\,\mathrm{d}y = \mathbb{E}[\chi_j(y)\rho(y)|x = k],$$

and converges to f_k in L^2_ρ-norm, that is,

$$\|\text{Proj}_J[f_k] - f_k\|_2 \equiv \left(\int_{-1}^{1} \big[\,\text{Proj}_J[f_k](y) - f_k(y)\big]^2 \rho(y)\,\mathrm{d}y\right)^{1/2} \longrightarrow 0$$

as $J \to \infty$.

Let $y_1, y_2, \ldots, y_T$ be a set of repeated measurements on y that are independent and identically distributed conditional on x. In a similar fashion as before, the joint density of

[2] Orthonormality with respect to ρ means that, for all pairs (i, j), $\int_{-1}^{1} \chi_i(y)\chi_j(y)\rho(y)\,\mathrm{d}y = \delta_{ij}$, where δ_{ij} is Kronecker's delta.

$y_1, y_2, \ldots, y_H$ for any $H \leq T$ can be projected onto the tensor basis $\{\chi_0, \chi_1, ..., \chi_I\}^{\otimes H}$, with associated Fourier coefficients

$$a_{i_1 i_2 \cdots i_H} \equiv \mathbb{E}\left[\chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_2)\rho(y_2)\cdots\chi_{i_H}(y_H)\rho(y_H)\right],$$

where $(i_1, i_2, \ldots, i_H)$ ranges over all H-tuples from the set $\{0, 1, \ldots, I\}$. The data reveal the coefficients $\{a_{i_1 i_2 \cdots i_H}, i_1, i_2, \ldots, i_H \geq 0\}$, which are linked to the Fourier coefficients of component densities $\{b_{jk}, j \geq 0\}$ and to the mixing proportions ω_k through the set of multilinear restrictions

$$a_{i_1 i_2 \cdots i_H} = \sum_{k=1}^{K} b_{i_1 k}\, b_{i_2 k} \cdots b_{i_H k}\, \omega_k, \tag{1.1}$$

for each $H \leq T$ and all $i_1, i_2, \ldots, i_H \geq 0$. Below we will show that, under weak restrictions, the relations in (1.1) can be uniquely solved for the Fourier coefficients of the component densities b_{jk} and their associated mixing proportions ω_k up to arbitrary relabeling of the components, which we maintain as our definition of identification for the remainder of the paper. It is well known that nonparametric point-identification fails when $T < 3$—see Hall and Zhou (2003), for example—and so we set $T = 3$ throughout the remainder of this section.

For any non-negative integer I, let

$$B \equiv \begin{pmatrix} b_{01} & b_{02} & \cdots & b_{0K} \\ b_{11} & b_{12} & \cdots & b_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{I1} & b_{I2} & \cdots & b_{IK} \end{pmatrix}$$

be the $(I+1) \times K$ matrix whose kth column contains the leading $I+1$ Fourier coefficients of f_k. Impose the following condition.

Assumption 1 (rank). *For sufficiently large I,* $\operatorname{rank}[B] = K$.

Assumption 1 states that the component densities are linearly independent.

Moving on, let $a \equiv (a_0, a_1, \ldots, a_I)'$, and introduce the symmetric $(I+1) \times (I+1)$ matrices

$$A_* \equiv \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0I} \\ a_{10} & a_{11} & \cdots & a_{1I} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I0} & a_{I1} & \cdots & a_{II} \end{pmatrix}, \quad \text{and} \quad A_j \equiv \begin{pmatrix} a_{00j} & a_{01j} & \cdots & a_{0Ij} \\ a_{10j} & a_{11j} & \cdots & a_{1Ij} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I0j} & a_{I1j} & \cdots & a_{IIj} \end{pmatrix}, \quad j \geq 0,$$

which contain the Fourier coefficients of the bivariate and of the trivariate joint densities of pairs and triples of measurements, respectively. The restrictions in (1.1) can then be written as

$$a = B\omega, \qquad A_* = B\Omega B', \qquad A_j = B\Omega^{1/2} D_j \Omega^{1/2} B', \tag{1.2}$$

for $\omega \equiv (\omega_1, \omega_2, \ldots, \omega_K)'$, $\Omega \equiv \operatorname{diag}[\omega]$, and $D_j \equiv \operatorname{diag}[b_{j1}, b_{j2}, \ldots, b_{jK}]$.

The next theorem shows that Assumption 1 suffices for identification of the mixture structure on the basis of repeated measurements.

THEOREM 1 (IDENTIFICATION). *Let I be chosen in such a way that Assumption 1 is satisfied. Then (i) the number of components K, (ii) the component densities f_k, and (iii) the mixing proportions ω_k are all identified.*

The proof of Theorem 1 is constructive, and is therefore given here. By Assumption 1, the matrix A_* has rank K. As this matrix is known, so is its rank and, hence, the number of mixture components. This proves Theorem 1(i). To establish Theorem 1(ii), note that A_* is real and symmetric, and that it has rank K. Therefore, it admits the spectral decomposition

$$A_* = V\Lambda V',$$

where V is the $(I+1) \times K$ orthonormal matrix containing the eigenvectors of A_*, and Λ is the $K \times K$ diagonal matrix containing the associated eigenvalues. Construct the $K \times (I+1)$ *whitening* matrix $W \equiv \Lambda^{-1/2}V'$, and subsequently form the set of $K \times K$ matrices $C_j \equiv WA_jW'$, for $j \geq 0$. Then, using (1.2), we obtain the system

$$C_j = UD_jU', \tag{1.3}$$

where $U \equiv WB\Omega^{1/2}$. As $WA_*W' = UU' = \mathrm{I}_K$, where I_K denotes the $K \times K$ identity matrix, U is a full-rank orthonormal matrix. Observe that V, and thus W, is not unique if the eigenvalues of A_* are multiple. Nevertheless, in this case (1.3) holds irrespective of the choice of V. Now, by (1.3) the set of matrices $\{C_j, j \geq 0\}$ are simultaneously diagonalizable in the same orthonormal basis; namely, the columns of U. The eigenvalues of C_j are given by the diagonal coefficients of D_j, that is, the Fourier coefficients of the component densities f_k. Because B and U have full column rank and eigenvectors are orthonormal, the decomposition in (1.3) is unique up to relabeling of the eigenvectors and eigenvalues, and up to the directions of eigenvectors (see, e.g., De Lathauwer et al. 2004). Thus, for each k, $\mathrm{Proj}_J[f_k]$ is identified for all J. Because $\|f_k - g_k\|_2 = 0$ if and only if the function g_k has the same Fourier coefficients as f_k, this implies that the densities f_k are identified almost everywhere on $[-1, 1]$. Finally, to show Theorem 1(iii), note that identification of the matrices D_j implies identification of B. As B has maximal column rank, (1.2) identifies $\omega = (B'B)^{-1}B'a$. This concludes the proof of Theorem 1.

We make the following remarks on the proof of Theorem 1. First, identification of the mixture components implies identification of all their functionals. Furthermore, Bayes' rule implies that the posterior latent-class probabilities $\Pr\{x = k|y_t = y\}$, as a function of y, are also identified. This is a key object of interest in latent-class analysis, as it allows to classify observations based on marginal information.

Second, the proof of Theorem 1 shows that identification does not require calculating the whole sequence $\{a_{i_1 i_2 i_3}, i_1, i_2, i_3 \geq 0\}$ to recover $\{b_{jk}, j \geq 0\}$. Only one dimension out of all three indices i_1, i_2, i_3 has to diverge. This observation will be important when deriving distribution theory, and is key in obtaining univariate convergence rates for our estimator of the component densities f_k.

Finally, note that simultaneous-diagonalization arguments have also been used elsewhere to establish identification in related contexts; see Hu (2008), Hu and Schennach (2008), and Kasahara and Shimotsu (2009), for example. Our approach here differs in two aspects. First, we work with a discretization of the component densities in the frequency domain rather than with a discretization of their support. Second, we explicitly construct an estimator based on the joint-eigenvector decomposition. The development of this estimator is the topic of the next section.

2. Estimation

Let $\{y_{n1}, y_{n2}, \ldots, y_{nT}, n = 1, 2, \ldots, N\}$ denote a random sample of size N. From now on, we set $T \geq 3$ and assume that an upper bound $I > K$ is available so that Assumption 1 is satisfied.

2.1. Number of components

Because the number of components equals the rank of A_*, a natural way to proceed is to estimate K by sequentially testing the rank of its empirical analog.[3] To describe the procedure, note that a plug-in estimator of A_*, say $\widehat{A}_*$, has (i_1, i_2)th-entry

$$\widehat{a}_{i_1 i_2} \equiv \frac{1}{N} \frac{(T-2)!}{T!} \sum_{n=1}^{N} \sum_{(t_1,t_2)} \chi_{i_1}(y_{nt_1})\rho(y_{nt_1})\chi_{i_2}(y_{nt_2})\rho(y_{nt_2}),$$

where (t_1, t_2) ranges over all ordered pairs from the set $\{1, 2, \ldots, T\}$. The averaging across all ordered pairs is done to exploit the stationarity restrictions across measurements. Note that $\widehat{A}_*$ is both an unbiased and a $\sqrt{N}$-consistent and asymptotically-normal estimator of A_* provided the basis functions have finite variance. We may then use the rank statistic of Kleibergen and Paap (2006), $\widehat{\mathrm{r}}_k$, to test the null $H_0 : \mathrm{rank}[A_*] = k$ against the alternative $H_1 : \mathrm{rank}[A_*] > k$ for any k. Moreover, following Robin and Smith (2000), a point estimator of K is given by

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

where $p_{1-\alpha}(k)$ is the $100(1-\alpha)$th percentile of the $\chi^2((I+1-k)^2)$ distribution and α is a chosen significance level (with the convention that $p_{1-\alpha}(I+1) = +\infty$). That is, the sequential-testing estimator is the first integer for which we fail to reject the null at significance level α. Asymptotically, this estimator will not underestimate the true rank of A_*. The probability of overestimation can be made to converge to zero by suitably

[3] This approach is similar in spirit to Kasahara and Shimotsu (2013), although their procedure only yields a consistent estimator of a lower bound on the number of components. Note that, in the absence of a known upper bound on K, our sequential-testing procedure, too, will only provide a consistent estimator of a lower bound on the number of components in general.

decreasing the significance level α as a function of the sample size; see Robin and Smith (2000) for details.

Although one could base the sequential testing procedure on a different test statistic, $\widehat{\mathrm{r}}_k$ has several attractive features and, therefore, carries our preference. Prime advantages include its non-sensitivity to the ordering of variables, and the fact that its limit distribution under the null is free of nuisance parameters.

2.2. Component densities and mixing proportions

2.2.1. Joint approximate diagonalization

We recover the Fourier coefficients from an empirical counterpart to the simultaneous-diagonalization system in (1.3). To this end, let I and J be two non-negative integers. In the asymptotic analysis we will let J tend to infinity, while keeping I fixed. We first construct the $(I+1)\times(I+1)$ matrices $\widehat{A}_j$, $j=0,1,\ldots,J$, whose typical (i_1,i_2)-entry takes the form

$$\widehat{a}_{i_1 i_2 j} \equiv \frac{1}{N}\frac{(T-3)!}{T!}\sum_{n=1}^{N}\sum_{(t_1,t_2,t_3)} \chi_{i_1}(y_{nt_1})\rho(y_{nt_1})\chi_{i_2}(y_{nt_2})\rho(y_{nt_2})\chi_j(y_{nt_3})\rho(y_{nt_3}),$$

where (t_1,t_2,t_3) ranges over all ordered triples from the set $\{1,2,\ldots,T\}$. We then form $\widehat{C}_j \equiv \widehat{W}\widehat{A}_j\widehat{W}'$, where $\widehat{W}$ is the sample counterpart to W, constructed from the spectral decomposition of $\widehat{A}_*$. Note that these matrices are all symmetric by construction. The restrictions in (1.3) then suggest estimating the matrices D_j by $\widehat{D}_j \equiv \mathrm{diag}\big[\widehat{U}'\widehat{C}_j\widehat{U}\big]$, where

$$\widehat{U} \equiv \arg\min_{U\in\mathcal{U}} \sum_{i=0}^{I} \Big\|\mathrm{off}\big[U'\widehat{C}_i U\big]\Big\|^2, \tag{2.1}$$

with $\mathcal{U}$ the set of $K\times K$ orthonormal matrices, $\mathrm{off}[A]\equiv A-\mathrm{diag}[A]$, and $\|\cdot\|$ the Euclidean norm. That is, $\widehat{U}$ is the *approximate* joint diagonalizer of the $\widehat{C}_i$.

Even though $\widehat{U}$ is a least-squares estimator, the first-order conditions of the minimization problem in (2.1) are highly nonlinear and difficult to solve using conventional gradient-based methods. Fortunately, this joint diagonalization problem has been extensively studied in numerical analysis (see, e.g., Bunse-Gerstner et al. 1993), and several numerical algorithms have been developed. Here, we use the JADE algorithm by Cardoso and Souloumiac (1993). This procedure is based on iteratively applying elementary Jacobi rotations. Its attractive computational properties have made JADE a workhorse technique in blind source separation (see, e.g., Comon and Jutten 2010). In extensive numerical experiments we found it to be very stable and computationally extremely fast.

2.2.2. Component densities and mixing proportions

Given an estimator of $\widehat{D}_j = \text{diag}[\widehat{b}_{j1}, \widehat{b}_{j2}, \ldots, \widehat{b}_{jK}]$, we estimate f_k by the orthogonal-series estimator

$$\widehat{f}_k \equiv \sum_{j=0}^{J} \widehat{b}_{jk} \chi_j.$$

Bona fide density estimates, that is non-negative functions that integrate to one, may then be obtained as $\overline{f}_k(y) \equiv \max\{0, \widehat{f}_k(y) - c_k\}$, where c_k is a constant chosen so that $\int_{-1}^{1} \overline{f}_k(z)\,\mathrm{d}z = 1$. Observe that $\overline{f}_k$ is the projection of $\widehat{f}_k$ onto the space of square-integrable functions that are non-negative and integrate to one, as shown by Gajek (1986) in a general setting.

Given an estimator of the component densities, an estimator of the mixing proportions is easily constructed. We use the leading $(I+1)$ matrices $\widehat{D}_i$ to construct a plug-in estimator $\widehat{B}$ of B, and compute $\widehat{a} \equiv (\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_I)'$, for

$$\widehat{a}_i \equiv \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \chi_i(y_{nt}) \rho(y_{nt}).$$

The vector ω is then readily estimated by

$$\widehat{\omega} \equiv (\widehat{B}'\widehat{B})^{-1} \widehat{B}'\widehat{a}.$$

This estimator is similar in spirit to the minimum-distance proposal of Titterington (1983), who worked in a framework where data can be sampled directly from the various mixture components.

2.3. Functionals

Given estimates of the component densities and mixing proportions, we next consider the problem of inferring a vector θ_0 defined as the unique solution to a moment condition of the form $\mathbb{E}[g(y;\theta_0)|x=k] = 0$ for some known function g. The first moment of f_k, for example, is defined through $\mathbb{E}[y - \theta_0 | x = k] = 0$.

Note that $\mathbb{E}[g(y;\theta)|x=k] = \mathbb{E}[g(y;\theta)\phi_k(y)]$ for $\phi_k(y) \equiv f_k(y)/f(y)$. Hence, a natural way to proceed is to consider a GMM estimator of the form

$$\widehat{\theta} \equiv \arg\min_{\theta \in \Theta} \; \widehat{m}(\theta)' \, \widehat{\Sigma} \, \widehat{m}(\theta), \tag{2.2}$$

where Θ is the parameter space, $\widehat{\Sigma}$ is a positive-definite weight matrix that converges in probability to a positive-definite and non-stochastic matrix Σ, and

$$\widehat{m}(\theta) \equiv \frac{1}{N} \sum_{n=1}^{N} \widehat{m}_n(\theta), \qquad \widehat{m}_n(\theta) \equiv \frac{1}{T} \sum_{t=1}^{T} g(y_{nt};\theta) \widehat{\phi}_k(y_{nt}),$$

for $\widehat{\phi}_k$ a suitable estimator of the weight function ϕ_k.

An alternative way to construct an estimator of θ_0 would be to replace $\widehat{m}(\theta)$ in (2.2) by

$$\widetilde{m}(\theta) \equiv \sum_{j=0}^{J} \widehat{b}_{jk} \overline{\chi}_j(\theta), \qquad \overline{\chi}_j(\theta) \equiv \int_{-1}^{1} g(y;\theta)\chi_j(y)\,\mathrm{d}y. \tag{2.3}$$

Quadrature methods can be used to numerically approximate the $\overline{\chi}_j(\theta)$ if the integrals are difficult to compute.

3. Distribution theory

The theory to follow uses orthonormal polynomials as basis functions. We first derive the large-sample properties of our estimators of mixture densities and mixing proportions assuming that K is known. We then present distribution theory for semiparametric two-step estimators of functionals. Technical details are collected in the appendix.

3.1. Component densities

3.1.1. Approximation to an infeasible estimator

To derive the large-sample properties of our estimator of f_k, it is instructive to link it to the infeasible estimator

$$\widetilde{f}_k \equiv \sum_{j=0}^{J} \widetilde{b}_{jk}\chi_j,$$

where the $\widetilde{b}_{jk}$ are the diagonal entries of $\widetilde{D}_j \equiv \mathrm{diag}[U'W\widehat{A}_jW'U]$. Note that this estimator is not feasible because it assumes knowledge of the whitening matrix W and of the joint eigenvectors U. We can write

$$\widetilde{f}_k(y) = \frac{1}{N}\frac{(T-3)!}{T!}\sum_{n=1}^{N}\sum_{(t_1,t_2,t_3)} \tau_k(y_{nt_1}, y_{nt_2})\ \kappa_J(y, y_{nt_3})\rho(y_{nt_3}),$$

where κ_J is the Christoffel-Darboux kernel associated with the system $\{\chi_j, j \geq 0\}$, that is,

$$\kappa_J(y_1, y_2) \equiv \sum_{j=0}^{J} \chi_j(y_1)\chi_j(y_2),$$

and where, using $U = (u_1, u_2, \ldots, u_K)$ and $W = (w_0, w_1, \ldots, w_I)$,

$$\tau_k(y_1, y_1) \equiv \sum_{i_1=0}^{I}\sum_{i_2=0}^{I} u_k' w_{i_1}\chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_2)\rho(y_2)w_{i_2}'u_k.$$

The function τ_k has an interpretation as a *tilt* function. Moreover, because the sample average of $\kappa_J(z, y_{nt})\rho(y_{nt})$ is just a conventional orthogonal-series estimator of $f(z)$, the infeasible estimator $\widetilde{f}_k$ can be seen as a re-weighting estimator based on an estimator of the marginal density.

Our estimator of f_k can be seen as an estimated counterpart to $\widetilde{f}_k$. Moreover, it equals

$$\widehat{f}_k(y) = \frac{1}{N}\frac{(T-3)!}{T!}\sum_{n=1}^{N}\sum_{(t_1,t_2,t_3)} \widehat{\tau}_k(y_{nt_1}, y_{nt_2})\ \kappa_J(y, y_{nt_3})\rho(y_{nt_3}),$$

where $\widehat{\tau}_k$ is an estimator of the tilt function based on the JADE program in (2.1), that is,

$$\widehat{\tau}_k(y_1, y_2) \equiv \sum_{i_1=0}^{I}\sum_{i_2=0}^{I} \widehat{u}_k'\widehat{w}_{i_1}\chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_2)\rho(y_2)\widehat{w}_{i_2}'\widehat{u}_k.$$

Under regularity conditions, replacing τ_k by its estimator $\widehat{\tau}_k$ will not affect the limit behavior of the density estimator.

3.1.2. *Global convergence rates and pointwise asymptotic normality*

To present distribution theory for the density estimator, let $\|\cdot\|_\infty$ denote the supremum norm and let $\alpha_J(y) \equiv \sum_{j=0}^{J}\chi_j(y)^2$. The following two assumptions suffice for the analysis of the infeasible estimator.

ASSUMPTION 2 (KERNEL). *The sequence $\{\chi_j, j \geq 0\}$ is dominated by a function π, which is continuous on $(-1,1)$ and positive almost everywhere on $[-1,1]$; $\pi\rho$ and $\pi^2\rho$ are integrable; and there exists a sequence of constants $\{\zeta_J, J \geq 0\}$ so that $\|\sqrt{\alpha_J}\|_\infty \leq \zeta_J$.*

These conditions are rather weak. They are satisfied for the class of Jacobi polynomials, for example, which are orthogonal to weight functions of the form $\rho(y) \propto (1-y)^{\vartheta_1}(1+y)^{\vartheta_2}$, where $\vartheta_1, \vartheta_2 > -1$, and are dominated by $\pi(y) \propto (1-y)^{-\vartheta_1'}(1+y)^{-\vartheta_2'}$, where $\vartheta_i' \equiv \max\{\vartheta_i, -1/2\}/2 + 1/4$. Further, with $\vartheta \equiv 1/2 + \max\{\vartheta_1, \vartheta_2, -1/2\}$, $\|\chi_j\|_\infty = j^\vartheta$, and so one can take $\zeta_J = J^{(1+\vartheta)/2}$; see, e.g., Viollaz (1989). Notable members of the Jacobi class are Chebychev polynomials of the first kind ($\vartheta_1 = \vartheta_2 = -1/2$), Chebychev polynomials of the second kind ($\vartheta_1 = \vartheta_2 = 1/2$), and Legendre polynomials ($\vartheta_1 = \vartheta_2 = 0$).

ASSUMPTION 3 (SMOOTHNESS). *For all k, f_k is continuous and $(\pi\rho)^4 f_k$ is integrable; and there exists a constant $\beta \geq 1$ such that $\|\mathrm{Proj}_J[f_k] - f_k\|_\infty = O(J^{-\beta})$ for all k.*

The integrability condition imposes existence of suitable moments of f_k. For consistency, integrability of $(\pi\rho)^2 f_k$, which is implied by integrability of $(\pi\rho)^4 f_k$, suffices. However, the fourth-order moments will be needed for establishing asymptotic normality. The condition on the rate at which the bias shrinks is conventional in nonparametric curve estimation by series expansions. Primitive conditions for it to hold depend on the orthogonal system used and on the differentiability properties of the f_k; see, e.g., Powell (1981).

The following assumption will allow us to extend the analysis to the feasible estimator.

ASSUMPTION 4 (EIGENVALUES). *The non-zero eigenvalues of A_* are all simple.*

This restriction is mainly done to facilitate the exposition. When Assumption 4 fails, the eigenvalues of A_* are no longer a continuous function of A_*, complicating the derivation of the asymptotic properties of $\widehat{\tau}_k$; see, e.g., Magnus (1985). Moreover, it is well known that the asymptotic distribution of eigenvalues depends in a complicated way on their multiplicity; see, e.g., Eaton and Tyler (1991).

Our first main result provides integrated squared-error and uniform convergence rates.

THEOREM 2 (CONVERGENCE RATES). *Let Assumptions 1–4 be satisfied. Then*

$$\left\|\widehat{f}_k - f_k\right\|_2^2 = O_P(J/N + J^{-2\beta}), \qquad \left\|\widehat{f}_k - f_k\right\|_\infty = O_P(\zeta_J\sqrt{J/N} + J^{-\beta}),$$

for all k.

The rates in Theorem 2 equal the conventional univariate rates of nonparametric series estimators; see, e.g., Schwartz (1967) and Newey (1997). Thus, the fact that x is latent does not affect the convergence speed of the density estimates. The integrated squared-error result is further known to be optimal, in the sense that it achieves the bound established by Stone (1982).

The next theorem states the pointwise limit distribution of the density estimator.

THEOREM 3 (ASYMPTOTIC NORMALITY). *Let Assumptions 1–4 be satisfied. Suppose that $N, J \to \infty$ so that $J^2/N \to 0$ and $NJ^{-2\beta} \to 0$. Then, for each y that lies in an interval on which f is of bounded variation,*

$$\sqrt{N}\mathscr{V}^{-1/2}\left[\widehat{f}_k(y) - f_k(y)\right] \xrightarrow{L} \mathcal{N}(0,1),$$

where $\mathscr{V}$ is the covariance of $\frac{(T-3)!}{T!}\sum_{(t_1,t_2,t_3)} \tau_k(y_{t_1}, y_{t_2})\kappa_J(y, y_{t_3})\rho(y_{t_3}) - \mathrm{Proj}_J[f_k](y)$.

The proof of Theorem 3 shows that $\mathscr{V} = O(\alpha_J(y))$, and so the pointwise convergence rate is determined by the growth rate of $\alpha_J(y)$. A weak bound is $O_P(\zeta_J/\sqrt{N})$ because $\|\sqrt{\alpha_J}\|_\infty \leq \zeta_J$.

3.2. *Mixing proportions*

Because a is estimated at the conventional parametric rate by $\widehat{a}$, the asymptotic behavior of the estimator of the mixing proportions is driven by the properties of $\widehat{B}$, that is, of our joint-diagonalization-based estimator of the leading Fourier coefficients. The analysis of the JADE estimator in the appendix shows that $\widehat{B}$ is asymptotically linear. It then readily follows that $\widehat{\omega}$ is a $\sqrt{N}$-consistent and asymptotically-normal estimator of ω.

To present the limit distribution of $\widehat{\omega}$, additional notation is needed. Let $\overset{\text{col}}{\otimes}$ and $\overset{\text{row}}{\otimes}$ denote the columnwise and rowwise Kronecker products, respectively. Under our maintained assumptions, $\widehat{W}$ is an asymptotically-linear estimator of the whitening matrix W. The

influence functions of $\widehat{W}$ and its transpose are

$$\begin{aligned}\psi_n^W &\equiv \left[\left(\mathrm{I}_I \otimes \Lambda^{-1} V\right) - \frac{1}{2}\left(\mathrm{I}_K \overset{\text{col}}{\otimes} W'\right)\left(W \overset{\text{row}}{\otimes} W\right)\right]\psi_n^{A_*}, \\ \psi_n^{W'} &\equiv \left[\left(\Lambda^{-1} V \otimes \mathrm{I}_I\right) - \frac{1}{2}\left(W' \overset{\text{col}}{\otimes} \mathrm{I}_K\right)\left(W \overset{\text{row}}{\otimes} W\right)\right]\psi_n^{A_*},\end{aligned} \tag{3.1}$$

respectively. Here, $\psi_n^{A_*}$ is the influence function of $\widehat{A}_*$. It has typical entry equal to $\frac{(T-2)!}{T!}\sum_{(t_1,t_2)} \chi_{i_1}(y_{nt_1})\rho(y_{nt_1})\chi_{i_2}(y_{nt_2})\rho(y_{nt_2}) - a_{i_1 i_2}$. The JADE estimators of U and its transpose, too, are asymptotically linear. Their influence functions, in turn, are

$$\begin{aligned}\psi_n^U &\equiv (\mathrm{I}_K \otimes HU')^{-1}\sum_{i=0}^{I} Q_i (U' \otimes U')\psi_n^{C_i}, \\ \psi_n^{U'} &\equiv (HU' \otimes \mathrm{I}_K)^{-1}\sum_{i=0}^{I} Q_i (U' \otimes U')\psi_n^{C_i},\end{aligned} \tag{3.2}$$

where $[H]_{k,\ell} \equiv \sum_{i=0}^{I}(b_{ik} - b_{i\ell})^2\ \delta_{k\ell}$, $Q_i \equiv \mathrm{diag}[\mathrm{vec}[D_i \iota_K \iota_K' - \iota_K \iota_K' D_i]]$ for ι_K a vector of ones of length K, and where

$$\psi_n^{C_i} \equiv (WA_i \otimes \mathrm{I}_K)\psi_n^W + (W \otimes W)\psi_n^{A_i} + (\mathrm{I}_K \otimes WA_i)\psi_n^{W'}, \tag{3.3}$$

with $\psi_n^{A_i}$ having typical element

$$\frac{(T-3)!}{T!}\sum_{(t_1,t_2,t_3)} \chi_{i_1}(y_{nt_1})\rho(y_{nt_1})\chi_{i_2}(y_{nt_2})\rho(y_{nt_2})\chi_i(y_{nt_3})\rho(y_{nt_3}) - a_{i_1 i_2 i}.$$

Note that the matrix H^{-1} is well defined for any $I + 1 \geq K$ because B has full column rank. It is readily shown that

$$\sqrt{N}\,\mathrm{vec}[\widehat{B} - B] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^B + o_P(1),$$

where $\psi_n^B \equiv \mathrm{vec}[(\psi_n^{D_0}, \psi_n^{D_1}, \ldots, \psi_n^{D_I})]$, with

$$\psi_n^{D_i} \equiv \left[\mathrm{I}_k \overset{\text{col}}{\otimes} \mathrm{I}_K\right]\left[(\mathrm{I}_K \otimes U'C_i)\psi_n^U + (U' \otimes U')\psi_n^{C_i} + (U'C_i \otimes \mathrm{I}_K)\psi_n^{U'}\right]. \tag{3.4}$$

The functions $\psi_n^{D_i}$ are the influence functions of the $\widehat{D}_i$, which contain the estimates of the Fourier coefficients on their main diagonal.

Note that we also have

$$\sqrt{N}(\widehat{a} - a) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^a + o_P(1),$$

where ψ_n^a has typical entry $T^{-1}\sum_t \chi_i(y_{nt})\rho(y_{nt}) - a_i$.

With these definitions at hand, we introduce

$$\psi_n^\omega \equiv (B'B)^{-1}\left[(a' \otimes \mathrm{I}_K)\psi_n^B + B'\psi_n^a\right]. \tag{3.5}$$

We can then state the following theorem.

THEOREM 4 (MIXING PROPORTIONS). *Let Assumptions 1–4 be satisfied. Then*

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

where $\mathscr{V}$ is the covariance of ψ_n^ω.

3.3. Functionals

With Theorems 2 to 4 at hand we turn to the analysis of two-step semiparametric estimators. Here we provide sufficient conditions for asymptotic linearity of the GMM estimator in (2.2), which estimates the moment function $\mathbb{E}[g(y;\theta)|x = k]$ by the re-weighting estimator $(NT)^{-1}\sum_{n,t} g(y_{nt};\theta)\widehat{\phi}_k(y_{nt})$ using

$$\widehat{\phi}_k(y) \equiv \widehat{f}_k(y) \Big/ \left(\sum_{\ell=1}^{K} \widehat{f}_\ell(y)\ \widehat{\omega}_\ell\right).$$

In practical applications, one may wish to trim observations for which the weight function is poorly estimated. Of course, other estimators of the moment function could be entertained. One example would be the series-based estimator in (2.3). Under regularity conditions, such estimators will all have similar properties, and so we omit a detailed analysis here for the sake of brevity.

Write G for the Jacobian of g. Let $g_0 \equiv g(\cdot;\theta_0)$ and $G_0 \equiv G(\cdot;\theta_0)$. We impose the following regularity conditions.

ASSUMPTION 5 (REGULARITY). *The value θ_0 lies in the interior of the compact set Θ; g_0 is square-integrable with respect to ρ; g is twice continuously-differentiable on Θ, and $\sup_{\theta\in\Theta}\mathbb{E}\|g(y;\theta)\|$ and $\sup_{\theta\in\Theta}\mathbb{E}\|G(y;\theta)\|$ are finite; the matrix $\mathbb{E}[G_0\phi_k(y)]$ has full column rank; for each k, ϕ_k is bounded away from zero and infinity on $[-1,1]$, and there exists an integer $\eta \geq 1$ so that $\|\mathrm{Proj}_J[g_0\phi_k] - g_0\phi_k\|_\infty = O(J^{-\eta})$.*

Assumption 5 contains familiar conditions for asymptotic normality of GMM estimators. It also postulates a shrinkage rate on the bias of orthogonal-series estimators of the $g_0\phi_\ell$ that is similar to Assumption 3.

To state the asymptotic distribution of $\widehat{\theta}$, let $M_\theta \equiv \mathbb{E}[G_0\phi_k(y)]$, and let M_ω be the matrix that has $\mathbb{E}[g(y;\theta_0)\phi_k(y)\phi_\ell(y)]$ as its ℓth column. Define

$$\psi_n^\theta \equiv (M_\theta'\Sigma M_\theta)^{-1}M_\theta'\Sigma\big[m_n(\theta_0) + \psi_n^\phi\big],$$

where $m_n(\theta) \equiv T^{-1}\sum_{t=1}^{T} g(y_{nt};\theta)\phi_k(y_{nt})$ and

$$\psi_n^\phi \equiv \frac{(T-3)!}{T!}\sum_{(t_1,t_2,t_3)} g_0(y_{nt_3})\Big[\tau_k(y_{nt_1},y_{nt_2}) - \phi_k(y_{nt_3})\sum_{\ell=1}^{K}\tau_\ell(y_{nt_1},y_{nt_2})\ \omega_\ell\Big]\rho(y_{nt_3}) - M_\omega\psi_n^\omega.$$

The function ψ_n^θ is the influence function of $\widehat{\theta}$ and has a familiar structure, as the term involving ψ_n^ϕ captures the impact of first-stage estimation error in $\widehat{\phi}_k$ on the asymptotic variance of $\widehat{\theta}$.

THEOREM 5 (FUNCTIONALS). *Let Assumptions 1–4 be satisfied. Suppose that* $N, J \to \infty$ *so that* $\zeta_J^4 J^2/N \to 0$, $NJ^{-2\beta} \to 0$, *and* $NJ^{-2\eta} \to 0$. *Then*

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

where $\mathscr{V}$ *is the covariance of* ψ_n^θ.

4. Monte Carlo Simulation

We present numerical evidence on the small-sample performance of our estimators by means of two illustrations. We will work with the family of Beta distributions on general intervals, which is popular for modeling the distribution of income; see, e.g., McDonald (1984). On the interval $[\underline{y}, \overline{y}]$, the Beta PDF 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 $\mathsf{B}(\vartheta_1, \vartheta_2) \equiv \int_0^1 z^{\vartheta_1 - 1}(1 - z)^{\vartheta_2 - 1} \,\mathrm{d}z$, and ϑ_1 and ϑ_2 are positive real scale parameters. Its mean and variance are

$$\mu \equiv \underline{y} + (\overline{y} - \underline{y}) \frac{\vartheta_1}{\vartheta_1 + \vartheta_2}, \quad \text{and} \quad \sigma^2 \equiv (\overline{y} - \underline{y})^2 \frac{\vartheta_1 \vartheta_2}{(\vartheta_1 + \vartheta_2)^2 (\vartheta_1 + \vartheta_2 + 1)}, \tag{4.1}$$

respectively.

Throughout this section and the next we use normalized Chebychev polynomials of the first kind as basis functions. For $y \in [-1, 1]$, the jth such polynomial is

$$\chi_j(y) = \frac{1}{2^{1\{j=0\}}} \cos[j \arccos(y)].$$

The system $\{\chi_j, j \geq 0\}$ is orthonormal with respect to the weight function $2/\sqrt{\pi^2(1 - y^2)}$, is uniformly bounded by the constant function 1, and is dominated in supremum-norm by $\zeta_J = \sqrt{J}$.

In each experiment, we estimate the number of components, the mixture densities and their associated CDFs, as well as the mixing proportions and means and variances of the mixture components. When estimating the component densities, we use $\overline{f}_k$ to ensure bona fide estimates. To infer the conditional CDFs, $F_k(y) \equiv \int_{-1}^{y} f_k(z) \,\mathrm{d}z$, we use Clenshaw-Curtis quadrature to approximate the integral $\int_{-1}^{y} \overline{f}_k(z) \,\mathrm{d}z$. The components are labeled according to the estimated modes, the smallest mode corresponding to $k = 1$. We also present simple kernel estimates of the marginal density of the data.[4]

[4] Our quadrature approximation uses 101 quadrature nodes. Means and variances are estimated using the GMM estimator from Section 3.3, without trimming. The estimates of the marginal densities are obtained using a Gaussian kernel and a bandwidth set according to Silverman's rule of thumb. When the support of the f_k is $[a, b]$, y is translated to $[-1, 1]$ through the transformation $(y - (a + b)/2)/((b - a)/2)$.

Table 1. Sequential rank test in Experiment 1

	$\alpha = .100$			$\alpha = .050$			$\alpha = .025$		
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

In all experiments we set $I = J$. In additional Monte Carlo exercises (not reported), we experimented with several other choices for the truncation parameters. We observed that the results were little affected by the choice of I. The choice of J has a larger impact, affecting the estimated densities in the usual manner. Providing an optimal data-dependent choice for J is a challenge that exceeds the scope of this paper.

Experiment 1. Our first experiment involves three generalized Beta distributions on the interval $[-1, 1]$. We set

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

Using (4.1), the means of the mixture components are $\mu_1 = -5/9 \approx -.556$, $\mu_2 = 1/9 \approx .111$, and $\mu_3 = 1/2$, while their respective variances are $\sigma_1^2 = 28/405 \approx .069$, $\sigma_2^2 = 8/81 \approx .099$, and $\sigma_3^2 = 1/12 \approx .083$. We set $T = 4$ and $I = J = 6$.

Table 1 presents simulation results for the estimator of K defined through the sequential-testing procedure based on the approach of Kleibergen and Paap (2006) 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. Overall, $\widehat{K}$ is found to perform well, correctly picking the true number of mixture components in more than $100(1 - \alpha)\%$ of the cases.

The first three panels in Figure 1 show upper 95% and lower 5% envelopes (dashed lines) over 1000 estimates from samples of size 1000 of the marginal PDF, the component PDFs, and the component CDFs, together with their respective true values (solid lines). The results reveal that our estimator reproduces the component densities very well, even though inspection of the marginal densities does not directly suggest the particular mixture structure.

The remaining three panels provide box plots of the sampling distribution of the mixing proportions, and means and variances of the mixture components. All box plots are centered around the respective true values. An asterisk marks zero. Overall, the plots suggest good performance. The sampling distribution of the mixing proportions and mixture means

Table 2. Sequential rank test in Experiment 2

	$\alpha = .100$			$\alpha = .050$			$\alpha = .025$		
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	.140	.642	.218	.201	.644	.155	.264	.617	.119
750	.053	.748	.199	.083	.781	.136	.120	.782	.098
1000	.015	.790	.193	.030	.836	.134	.044	.864	.092
1500	.001	.804	.194	.001	.864	.135	.004	.900	.097
2000	.000	.809	.189	.000	.864	.136	.000	.904	.096
2500	.000	.816	.181	.000	.866	.134	.000	.905	.095

are broadly correctly centered and have small interquartile ranges. The box plots for the variance show evidence of a slightly larger bias in the estimates of the mixture variances, in particular for the first and third mixture component. Nonetheless, the magnitude of the bias is small.

Experiment 2. 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 unobservable ε_t. A location-scale version is

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

for some function σ. 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 that here the factor x and the error η_t are allowed to be dependent.

Suppose that ε_t is drawn from the generalized Beta distribution on $[-1, 1]$ with $\vartheta_1 = \vartheta_2 = \vartheta$. Then its distribution is symmetric and $\mathrm{var}[\varepsilon_t] = 1/(2\vartheta+1)$ while f_k is supported on the interval $[k - \sigma(k), k + \sigma(k)]$. Below we report results for the scale-function specification $\sigma(x) = (2\vartheta + 1)/\sqrt{x}$. Here,

$$\mathbb{E}[y|x = k] = k, \qquad \mathrm{var}[y|x = k] = \frac{1}{k},$$

so that the variability is inversely related to the mean. Table 2 and Figure 2, which have the same structure as Table 1 and Figure 1 above, present simulation results for this location-scale model for $\vartheta = 2$ with $K = 5$ and $\omega = (.30, .20, .25, .15, .10)'$. This choice of mixing proportions implies that f is unimodal and virtually symmetric, hiding the underlying mixture structure well. The simulations were performed with $I = J = 8$. The table was generated over 10,000 replications. The plots are based on 1000 Monte Carlo runs.

Table 2 shows that the sequential rank test performs slightly worse than in the first design, although performance improves as N increases. The plots in Figure 2 display excellent estimates of the component PDFs and CDFs. The box plots of the mixing proportions are also broadly correctly centered, although interquartile ranges are higher than in Experiment 1. From the bottom panels in the figure it may be observed that the sampling distributions of the GMM estimators of the mean and variance of the mixture components are now more disperse. Variance estimates also suffer from non-negligible bias. This suggests that some amount of trimming may be desirable in applications.

Figure 1. Estimates in Experiment 1

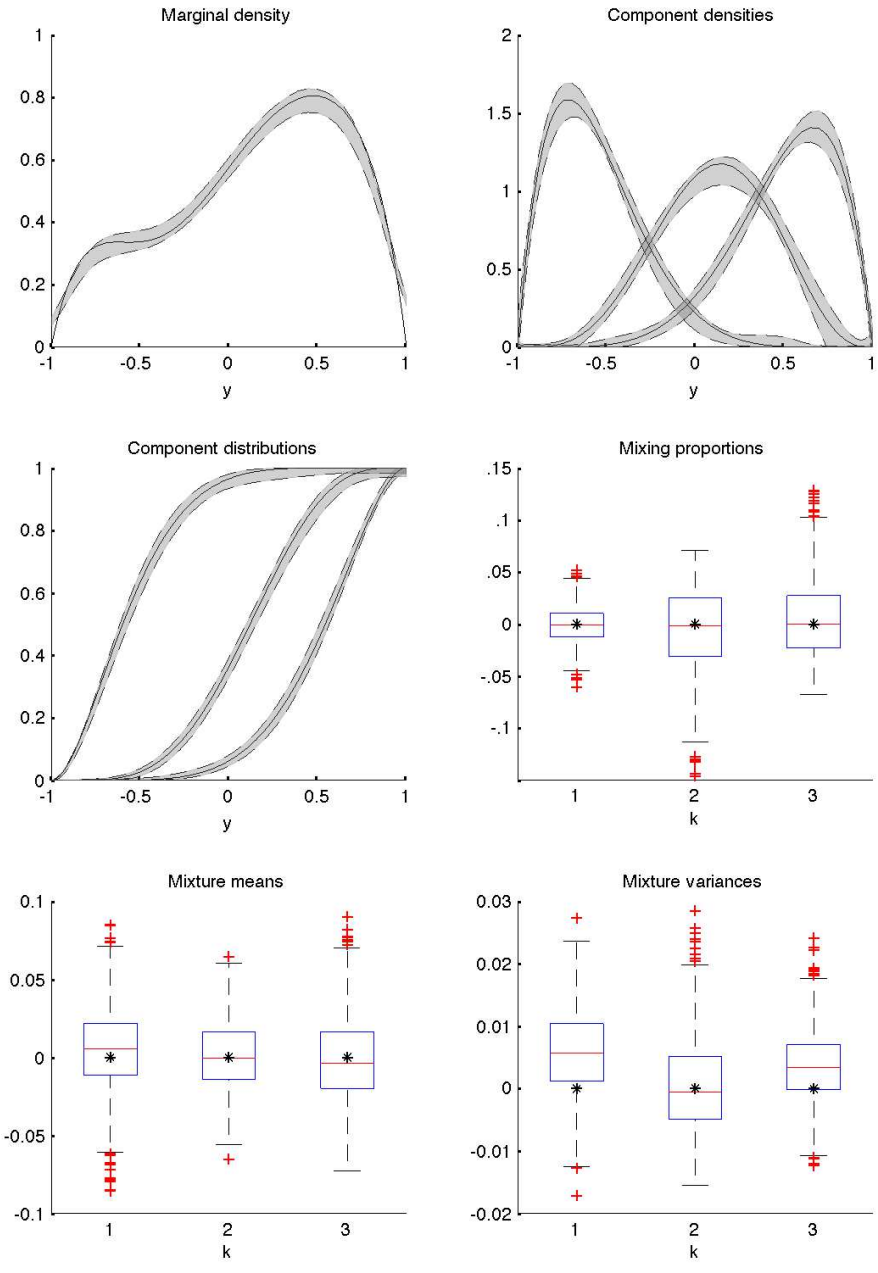
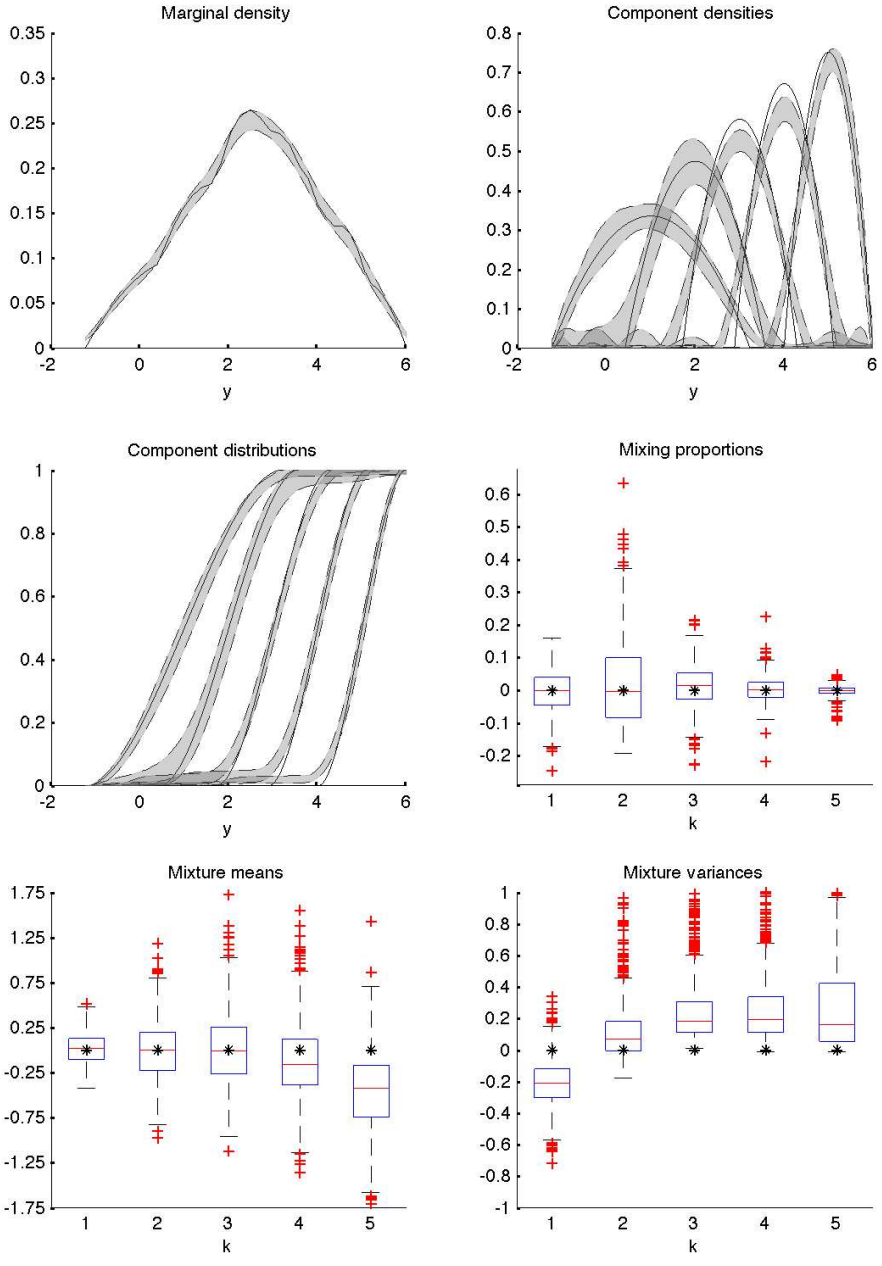


Figure 2. Estimates in Experiment 2



5. Empirical Application

In an influential paper, Gottschalk and Moffitt (1994) decompose earnings inequality into a permanent and a transitory component, and contrast earnings inequality in the 1970s with inequality in the 1980s. In a recent paper, Moffitt and Gottschalk (2012) re-estimate the trend in the transitory variance of male earnings in the United States using the PSID from 1970 to 2004. They estimated both a simple one-factor model and more sophisticated error-component models, and found that the transitory variance started to increase in the early 1970s, continued to increase through the mid-1980s, and then remained at this new higher level through the 1990s and beyond.

The literature on such decompositions of earnings dynamics builds on a representation of individual n's log earnings at time t as

$$y_{nt} = x_n + \eta_{nt}, \tag{5.1}$$

where y_{nt} is typically the residual from a standard Mincer regression, x_n is a fixed effect, and η_{nt} is an idiosyncratic white noise process. Extensions include replacing the fixed effect by a random walk with individual-specific drift or initial condition, and replacing the white noise by a stationary serially-correlated process. However, recent contributions have argued that it may be important to allow for more heterogeneity than in the additive model (5.1); see, e.g., Browning et al. (2010) for a parametric approach. In this section, we show how finite mixtures can be used to shed some new light on the anatomy of the rise of earnings inequality in the U.S.

From the PSID 1969–1998 we construct a set of five-year balanced subpanels, using a rolling window of length one. This yields 26 subpanels. For each such subpanel, we obtain our measure of log (annual) earnings 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.[5] Graphical inspection of the marginal densities in each subpanel (not reported) suggests that our stationarity assumption within subpanels is reasonable. We then estimate a trivariate mixture model for each subpanel. Experiments with up to ten components yielded similar patterns in the estimated component densities. We focus on a small number of components for ease of exposition. For example, one can think of x as a latent ability type with three categories: low, intermediate, and high. The densities were estimated as in the Monte Carlo experiments, setting $I = 5$ and $J = N^{1/5}$ to accommodate the increase in the number of cross-sectional observations in the subpanels spanning later time periods. Alternative choices for the truncation parameter J were found to have little effect on the estimated densities.

Figure 3 plots the estimated component densities for six non-overlapping subpanels. The cross-sectional sample size is indicated below each plot. The figure reveals well-separated

[5] We excluded self-employed individuals and students, as well as individuals for whom earnings were top coded. The sample was restricted to individuals between the ages of 20 and 60, with at most 40 years of experience. We computed experience as age − (years of schooling + 6).

Figure 3. Component densities in a selection of subpanels

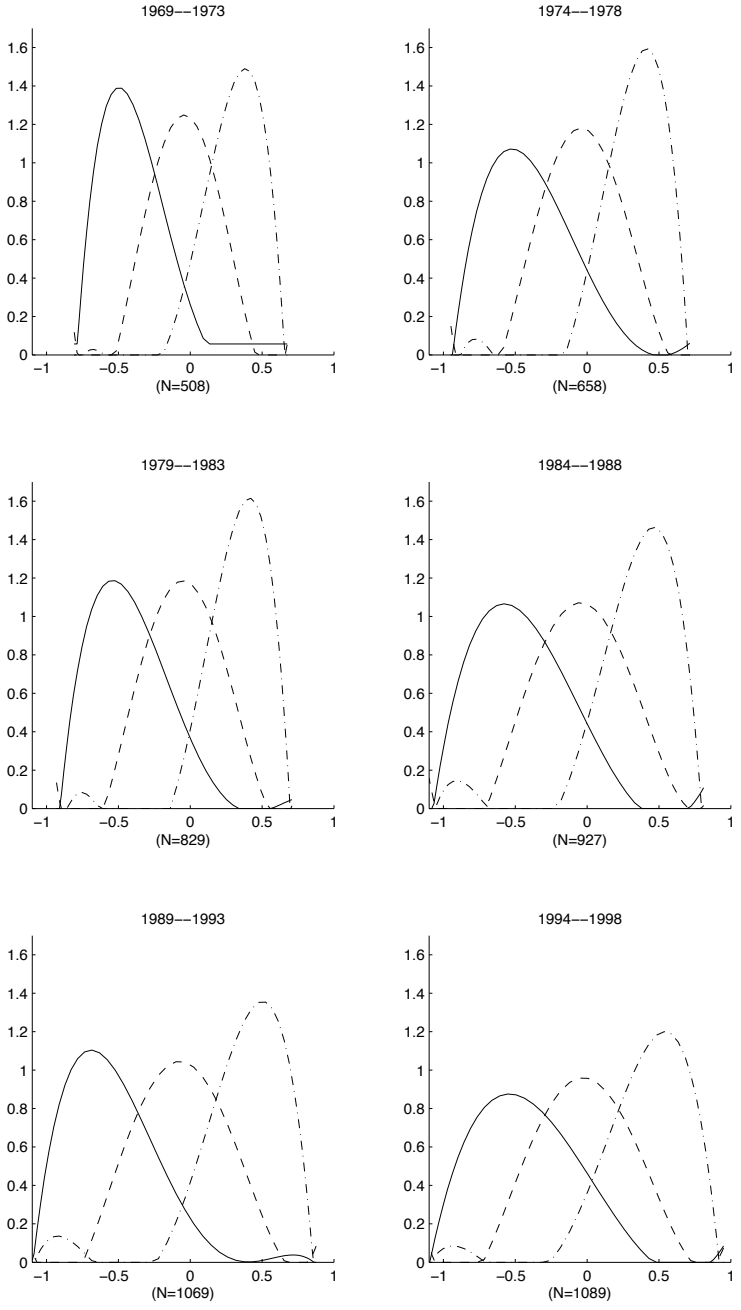


Figure 4. Evolution of functionals over the sampling period

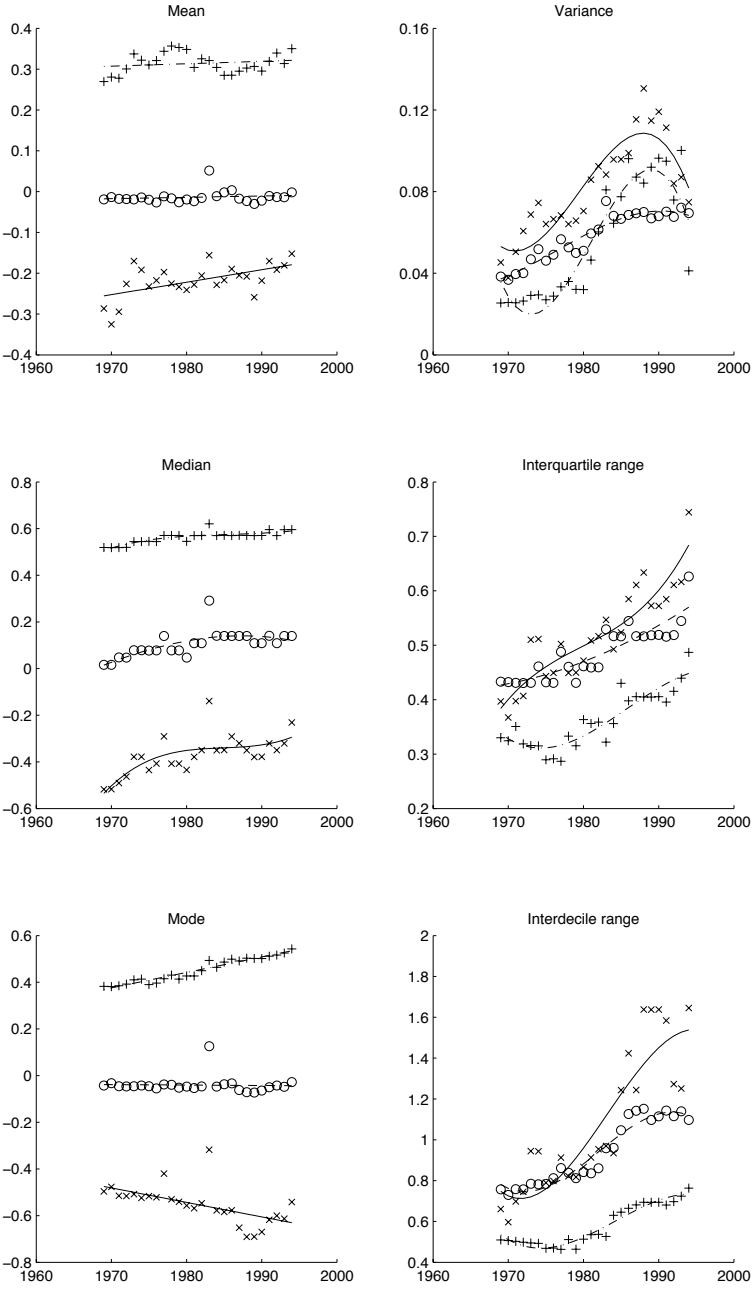
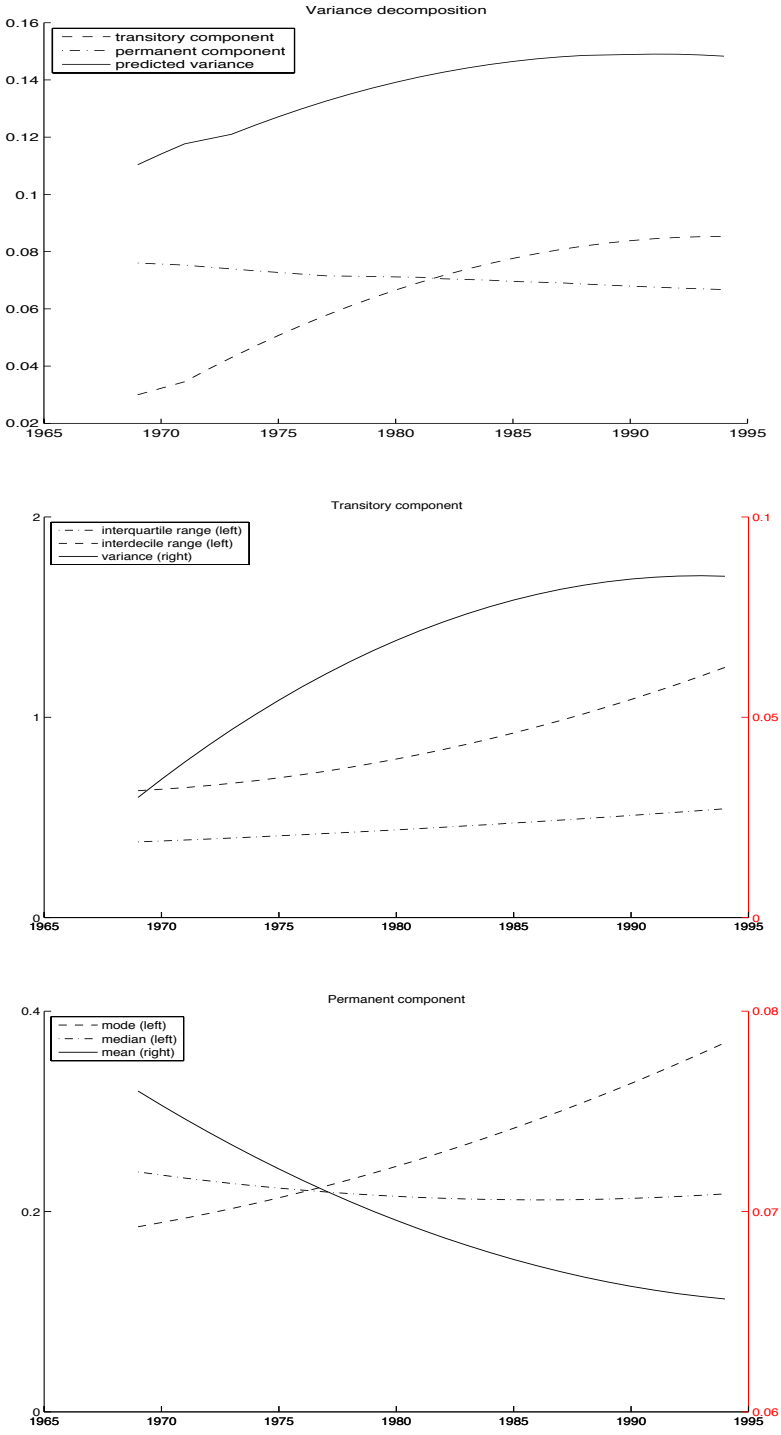


Figure 5. Decomposition of variance, interquartile range, and interdecile range



unimodal component densities, which we label $k = 1, 2, 3$ according to their mode (from left to right). The plots indicate that the component densities, and their dispersion in particular, vary with k, which provides evidence against the common assumption of independence between x and η in model (5.1).

The plots further suggest that the component densities evolve over time. To investigate this, we estimate their mean, median, and mode as measures of location, as well as their variance, interquartile range, and interdecile range as measures of dispersion. The results are presented in Figure 4, which plots the point estimates as well as regression lines over time, for $k = 1$ ($\times$; full), $k = 2$ ($\circ$; dashed), and $k = 3$ ($+$; dashed-dotted). Means and medians show a slightly upward-sloping trend, with the lower-end component density catching up over time. Modes, however, diverge over the period. The ranking of variances, interquartile ranges, and interdecile ranges of the component densities are inversely related to their measures of location, indicating that individuals at the lower end of the earnings distribution are subject to higher volatility in earnings. Interestingly, Figure 4 shows that all volatility measures increased over the sampling period. This suggests that earnings inequality within the three latent groups k has increased over time. Note also that, while the increase in within-group variance slows down at the end of the 1980s, this pattern is much less apparent when considering interquartile and interdecile ranges.

In the top panel of Figure 5 we present the evolution of the within-group (dashed) and the between-group (dashed-dotted) variances. In line with the literature, we interpret the former as measuring a transitory component and the latter as reflecting a permanent component. The transitory component is the average of the variances of the component densities; the permanent component is the variance of the means over the latent classes. We observe that the transitory component increased during the period, although it tended to flatten out in the late 1980s. This is in line with the findings of the more parametrized models of Moffitt and Gottschalk (2012). We also find that the permanent component has been decreasing steadily throughout the sampling period. This decrease is dominated by the evolution of the transitory component, however, as is apparent from the evolution of the total variance (solid line).

The remaining panels in Figure 5 document the evolution of other measures of within-group and between-group variation. The middle plot shows that, like the transitory variance (solid line), the averages over groups of within-group interquartile ranges (dashed-dotted) and of within-group interdecile ranges (dashed) both increased over time, although the former did so at a slower pace. Unlike the transitory variance, however, these other measures do not show signs of slowing down over the course of the sampling period. We tend to place more confidence in inter-quantile ranges as measures of dispersion because they are well known to be more robust than the variance (and other higher-order moments). When tails get fatter more observations are needed to obtain the same precision for the variance. Thus, the flattening-out of the transitory variance might just reflect the fattening tails of the within-group distributions after 1980.

Lastly, the bottom plot in Figure 5 shows the evolution of the variances (across the three latent classes) of the group-specific means, medians, and modes. The plot shows that the evolution of these permanent components is very different depending on the location parameter considered.

Conclusion

In this paper we have introduced simple nonparametric procedures to estimate finite-mixture models from short panel data. Our estimators rely on fast joint-diagonalization techniques for efficient computation. Their theoretical properties as well as our Monte Carlo experiments suggest good performance. In particular, the component densities are estimated at the usual univariate nonparametric rates, and mixing proportions as well as functionals converge at the parametric rate. Although we have focused on the case where the outcome variable is continuous, our methods can also be applied with discrete outcomes.

Several directions for future research present themselves. First, it would be interesting to extend our approach to continuous mixtures. Nonparametric estimation of continuous mixtures is challenging, however, as an ill-posed inverse problem arises. A second extension would be to assess the impact of estimating the number of components on the statistical properties of our estimators. In numerical experiments, we have found that working with a K that is too low yields density estimates that tend to aggregate the true component densities over groups. A final direction would be to relax the assumption that the component densities are stationary. These questions all raise challenges for computation and statistical analysis, and are currently under study.

Acknowledgments

We are grateful to Marc Henry and Bernard Salanié for comments and discussion. We also thank seminar and workshop participants at Columbia, CREST, Université de Montréal, UCL, and Vanderbilt, as well as attendees of the 18th International Conference on Panel Data in Paris, the 66th European Meetings of the Econometric Society in Malaga, and the conference celebrating the 65th birthday of Jean-Pierre Florens in Toulouse. Bonhomme gratefully acknowledges the support from the European Research Council (ERC) through the grant ERC-2010-StG-0263107-ENMUH. Jochmans gratefully acknowledges financial support from the Scientific Advisory Board through the project 'Nonparametric estimation of finite mixtures'. Robin gratefully acknowledges financial support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and from the ERC grant ERC-2010-AdG-269693-WASP.

Appendix A. Analysis of the infeasible density estimator

For notational simplicity, throughout the appendices, we set $T = 3$ and ignore permutations of observations when constructing estimators; we set $(t_1, t_2, t_3) = (1, 2, 3)$.

THEOREM A.1. *Let Assumptions 2–3 be satisfied. Then*

$$\left\|\widetilde{f}_k - f_k\right\|_2^2 = O_P(J/N + J^{-2\beta}), \qquad \left\|\widetilde{f}_k - f_k\right\|_\infty = O_P(\zeta_J\sqrt{J/N} + J^{-\beta}),$$

for all k.

PROOF. Let $b \equiv (b_{0k}, b_{1k}, \ldots, b_{Jk})'$, and define $\widetilde{b}$ in an analogous fashion. We begin by showing that $\|\widetilde{b} - b\| = O_P\big(\sqrt{J/N}\big)$. First observe that, for any j,

$$\begin{aligned}
\mathbb{E}\left\|\widehat{A}_j - A_j\right\|^2 &= \sum_{i_1=0}^{I}\sum_{i_2=0}^{I}\mathbb{E}\left[(\widehat{a}_{i_1i_2j} - a_{i_1i_2j})^2\right] \\
&= \sum_{i_1=0}^{I}\sum_{i_2=0}^{I}\mathbb{E}\left[\Big(\frac{1}{N}\sum_{n=1}^{N}\chi_{i_1}(y_{n1})\rho(y_{n1})\chi_{i_2}(y_{n2})\rho(y_{n2})\chi_j(y_{n3})\rho(y_{n3}) - a_{i_1i_2j}\Big)^2\right] \\
&= \sum_{i_1=0}^{I}\sum_{i_2=0}^{I}\frac{\mathbb{E}\left[\chi_{i_1}(y_1)^2\rho(y_1)^2\chi_{i_2}(y_2)^2\rho(y_2)^2\chi_j(y_3)^2\rho(y_3)^2\right] - a_{i_1i_2j}^2}{N} \\
&\leq \sum_{i_1=0}^{I}\sum_{i_2=0}^{I}\frac{\big(\int_{-1}^{1}\pi(y)^2\rho(y)^2 f(y)\,\mathrm{d}y\big)^3 - a_{i_1i_2j}^2}{N} \\
&\leq \frac{(I+1)^2\big(\int_{-1}^{1}\pi(y)^2\rho(y)^2 f(y)\,\mathrm{d}y\big)^3}{N}.
\end{aligned}$$

As $\pi^2\rho^2 f$ is integrable, we thus have that $\mathbb{E}\left\|\widehat{A}_j - A_j\right\|^2 = O(1/N)$ uniformly in j. Therefore,

$$\sum_{j=0}^{J}\left\|\widehat{A}_j - A_j\right\|^2 = O_P(J/N)$$

by Markov's inequality. Now, because $\widetilde{b}_{jk} - b_{jk}$ is the kth diagonal entry of $\widetilde{D}_j - D_j$ and $\widetilde{D}_j - D_j = \mathrm{diag}[U'W(\widehat{A}_j - A_j)W'U]$,

$$\|\widetilde{b} - b\|^2 \leq \sum_{j=0}^{J}\left\|\widetilde{D}_j - D_j\right\|^2 \leq \left\|U'W \otimes U'W\right\|^2 \sum_{j=0}^{J}\left\|\widehat{A}_j - A_j\right\|^2 = O_P(J/N)$$

follows by the Cauchy-Schwarz inequality and the fact that $\left\|U'W \otimes U'W\right\| = O(1)$. This establishes the rate result on the Fourier coefficients.

Now consider the integrated squared-error result. Using orthonormality of the basis functions, a small calculation shows that

$$\left\|\widetilde{f}_k - f_k\right\|_2^2 = \left\|\widetilde{f}_k - \mathrm{Proj}_J[f_k]\right\|_2^2 + \left\|\mathrm{Proj}_J[f_k] - f_k\right\|_2^2 = \|\widetilde{b} - b\|^2 + \left\|\mathrm{Proj}_J[f_k] - f_k\right\|_2^2.$$

From above, $\|\widetilde{b} - b\|^2 = O_P(J/N)$. Further, by Assumption 3,

$$\left\| \operatorname{Proj}_J[f_k] - f_k \right\|_2^2 \leq \int_{-1}^{1} \left\| \operatorname{Proj}_J[f_k] - f_k \right\|_\infty^2 \rho(y)\,\mathrm{d}y = O(J^{-2\beta})$$

because ρ is integrable. This proves the first part of Theorem A.1.

To derive the uniform convergence rate, finally, use the triangle inequality to obtain that

$$\left\| \widetilde{f}_k - f_k \right\|_\infty \leq \left\| \widetilde{f}_k - \operatorname{Proj}_J[f_k] \right\|_\infty + \left\| \operatorname{Proj}_J[f_k] - f_k \right\|_\infty.$$

By the Cauchy-Schwarz inequality in the first step and by the uniform bound on $\alpha_J(y)$ and the convergence rate of $\|\widetilde{b} - b\|$ in the second, the first right-hand side term is bounded as

$$\left\| \widetilde{f}_k - \operatorname{Proj}_J[f_k] \right\|_\infty \leq \|\sqrt{\alpha_J}\|_\infty \, \|\widetilde{b} - b\| = O\big(\zeta_J\big)\; O_P\big(\sqrt{J/N}\big).$$

By Assumption 3, $\left\| \operatorname{Proj}_J[f_k] - f_k \right\|_\infty = O(J^{-\beta})$. This yields the second part of Theorem A.1 and thus concludes the proof. □

THEOREM A.2. *Let Assumptions 2–3 be satisfied. Suppose that $N, J \to \infty$ so that $J/N \to 0$ and $J^{2\beta}/N \to \infty$. Then, for each y that lies in an interval on which f is of bounded variation,*

$$\sqrt{N}\mathscr{V}^{-1/2}\big[\widetilde{f}_k(y) - f_k(y)\big] \xrightarrow{L} \mathcal{N}(0,1),$$

where $\mathscr{V}$ is the covariance of $\tau_k(y_1, y_2)\kappa_J(y, y_3)\rho(y_{t_3}) - \operatorname{Proj}_J[f_k](y)$.

PROOF. Fix the evaluation point y throughout the proof. We first show that $\widetilde{f}_k$ is an unbiased estimator of $\operatorname{Proj}_J[f_k]$.

$$\mathbb{E}\left[\widetilde{f}_k(y)\right] = \mathbb{E}\left[\tau_k(y_1, y_2)\kappa_J(y, y_3)\rho(y_3)\right] = \sum_{\ell=1}^{K} \mathbb{E}\left[\tau_k(y_1, y_2)\middle| x = \ell\right] \mathbb{E}\left[\kappa_J(y, y_3)\rho(y_3)\middle| x = \ell\right] \omega_\ell,$$

where the expectations are with respect to (y_1, y_2, y_3) and we have used the conditional independence of the measurements. Now,

$$\mathbb{E}\left[\tau_k(y_1, y_2)\middle| x = \ell\right] = \sum_{i_1=0}^{I} \sum_{i_2=0}^{I} u_k' w_{i_1} b_{i_1\ell} b_{i_2\ell} w_{i_2}' u_k = \omega_k^{-1}\delta_{k\ell}$$

because $U'WBB'W'U = \Omega^{-1}$, which follows from (1.3). Further, the Christoffel-Darboux kernel satisfies

$$\mathbb{E}\left[\kappa_J(y, y_3)\rho(y_3)\middle| x = \ell\right] = \sum_{j=0}^{J} b_{j\ell}\chi_j(y) = \operatorname{Proj}_J[f_\ell](y)$$

for each ℓ. Thus, $\mathbb{E}\left[\widetilde{f}_k(y)\right] = \operatorname{Proj}_J[f_k](y)$, as claimed.

Centering the estimator around its expectation gives

$$\widetilde{f}_k(y) - \operatorname{Proj}_J[f_k](y) = \frac{1}{N}\sum_{n=1}^{N} \psi_n, \qquad \psi_n \equiv \tau_k(y_{n1}, y_{n2})\kappa_J(y, y_{n3})\rho(y_{n3}) - \operatorname{Proj}_J[f_k](y).$$

Because Assumption 3 yields $|\operatorname{Proj}_J[f_k](y) - f_k(y)| \leq \|\operatorname{Proj}_J[f_k] - f_k\|_\infty = O(J^{-\beta})$ and we require that $\sqrt{N}J^{-\beta} \to 0$, the bias induced by truncating the projection is asymptotically negligible. It thus suffices to derive the limit distribution of the sample average of the ψ_n. For this we verify that the conditions of Lyapunov central limit theorem for triangular arrays are satisfied. We have already demonstrated that $\mathbb{E}[\psi_n] = 0$ and so, if we can show that

$$\text{(i)}\ \ \mathbb{E}\left[\psi_n^2/\alpha_J(y)\right] = O(1); \qquad \text{(ii)}\ \ \mathbb{E}\left[\left(\psi_n^2/\operatorname{var}[\psi_n]\right)^2\right] = o(N),$$

the result will be proven.

To show Condition (i), use the conditional independence of the measurements to obtain

$$\operatorname{var}[\psi_n] = \sum_{\ell=1}^{K} \mathbb{E}\left[\tau_k(y_1, y_2)^2 | x = \ell\right]\ \mathbb{E}\left[\kappa_J(y, y_3)^2 \rho(y_3)^2 | x = \ell\right]\ \omega_\ell - \operatorname{Proj}_J[f_k](y)^2.$$

Exploiting conditional independence, a direct calculation shows that, for each ℓ, the second moment of the weight function is

$$\mathbb{E}[\tau_k(y_1, y_2)^2 | x = \ell] = \left(\sum_{i_1=0}^{I} \sum_{i_2=0}^{I} u_k' w_{i_1}\, \mathbb{E}[\chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_1)\rho(y_1) | x = \ell] w_{i_2}' u_k \right)^2 = O(1),$$

with boundedness following from the fact that the $u_k' w_i$ are $O(1)$ and the observation that

$$\mathbb{E}[\chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_1)\rho(y_1) | x = \ell] \leq \int_{-1}^{1} \pi(y_1)^2 \rho(y_1)^2 f_\ell(y_1)\, \mathrm{d}y_1 = O(1),$$

which holds because $\pi^2\rho^2 f_\ell$ is integrable. Next, under the conditions of Theorem A.2, we can apply Theorem 2.2.3 in Viollaz (1989) to get

$$\frac{\mathbb{E}\left[\kappa_J(y, y_3)^2 \rho(y_3)^2 | x = \ell\right]}{\alpha_J(y)} \longrightarrow f_\ell(y)\rho(y),$$

which exists for all ℓ. Finally, $\operatorname{Proj}_J[f_k](y)/\sqrt{\alpha_J(y)} = o(1)$ because $\operatorname{Proj}_J[f_k](y) \to f_k(y)$, which is finite. We therefore conclude that $\operatorname{var}[\psi_n]/\alpha_J(y)$ tends to a positive constant, and that Condition (i) is satisfied.

Finally, to verify Condition (ii), let X_{nj} be the $(I+1) \times (I+1)$ matrix with typical entry $[X_{nj}]_{i_1,i_2} \equiv \chi_{i_1}(y_{n1})\rho(y_{n1})\chi_{i_2}(y_{n2})\rho(y_{n2})\chi_j(y_{n3})\rho(y_{n3})$. This allows us to write ψ_n as

$$\psi_n = \sum_{j=0}^{J} \left(u_k' W [X_{nj} - A_j] W' u_k\right) \chi_j(y) = \sum_{j=0}^{J} \operatorname{tr}\left([W' u_k u_k' W] [X_{nj} - A_j] \right) \chi_j(y).$$

By repeatedly applying the Cauchy-Schwarz inequality to this expression we then establish

$$\psi_n^2 \leq \|W' u_k u_k' W\|^2 \ \sum_{j=0}^{J} \|X_{nj} - A_j\|^2\ \ \alpha_J(y).$$

Because $\|W'u_k u_k' W\|^2 = O(1)$ and $\mathrm{var}[\psi_n] \asymp \alpha_J(y)$ we then obtain

$$\mathbb{E}\left[\left(\psi_n^2/\,\mathrm{var}[\psi_n]\right)^2\right] \leq O(1)\ \mathbb{E}\left[\Big(\sum_{j=0}^{J}\|X_{nj}-A_j\|^2\Big)^2\right] \leq O(J)\ \sum_{j=0}^{J}\mathbb{E}\,\|X_{nj}-A_j\|^4,$$

where the last transition follows by the Cauchy-Schwarz inequality. Using similar arguments as in the proof of Theorem A.1, it is straightforward to show that $\mathbb{E}\,\|X_{nj}-A_j\|^4 = O(1)$ uniformly in j, because $\pi^4\rho^4 f$ is integrable. Therefore,

$$\frac{1}{N}\,\mathbb{E}\left[\left(\frac{\psi_n^2}{\mathrm{var}[\psi_n]}\right)^2\right] = O\left(\frac{J^2}{N}\right),$$

which converges to zero as $N\to\infty$. This shows that Condition (ii) is satisfied. With the requirements of the central limit theorem verified, it follows that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\frac{\psi_n}{\sqrt{\mathrm{var}[\psi_n]}} \xrightarrow{L} \mathcal{N}(0,1).$$

This completes the proof. □

Appendix B. Distribution theory for the JADE estimator

The first step is to derive the asymptotic distribution of the estimated whitening matrix.

Lemma B.1. *Let Assumptions 2–4 be satisfied. Then*

$$\sqrt{N}\,\mathrm{vec}[\widehat{W}-W] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^W + o_P(1), \qquad \sqrt{N}\,\mathrm{vec}[\widehat{W}'-W'] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{W'} + o_P(1),$$

where ψ_n^W *and* $\psi_n^{W'}$ *are given in* (3.1).

Proof. Recall that A_* is real, symmetric, and has rank K. Also, its K non-zero eigenvalues are all distinct by Assumption 4. Further, $\widehat{A}_*$ satisfies $\sqrt{N}\mathrm{vec}[\widehat{A}_*-A_*] \xrightarrow{L} \mathcal{N}(0,\mathscr{V}_{A_*})$ for $\mathscr{V}_{A_*} \equiv \mathbb{E}[\psi_n^{A_*}\psi_n^{A_*\prime}]$. From Theorem 4.2 in Eaton and Tyler (1991) and Theorem 1 in Magnus (1985), it then follows that $\sqrt{N}(\widehat{\lambda}-\lambda) \xrightarrow{L} \mathcal{N}(0,\mathscr{V}_\lambda)$, where $[\mathscr{V}_\lambda]_{k,\ell} \equiv (v_k'\otimes v_k')\mathscr{V}_{A_*}(v_\ell\otimes v_\ell)$. With $\Lambda = \mathrm{diag}[\lambda]$, the Jacobian associated with the transformation from λ to $\mathrm{vec}[\Lambda^{-1/2}]$ is $-\frac{1}{2}(\mathrm{I}_K \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)\mathrm{diag}[\lambda_1^{-3/2},\lambda_2^{-3/2},\ldots,\lambda_K^{-3/2}]$, where $(\mathrm{I}_K \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)$ is the $K^2\times K$ selection matrix that transforms λ into $\mathrm{vec}[\Lambda]$. Hence, by an application of the delta method,

$$\sqrt{N}\mathrm{vec}[\widehat{\Lambda}^{-1/2}-\Lambda^{-1/2}] = -\frac{1}{2}(\mathrm{I}_K \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)\Lambda^{-1/2}(W \overset{\mathrm{row}}{\otimes} W)\ \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{A_*} + o_P(1), \qquad \text{(B.1)}$$

using $W = \Lambda^{-1/2}V'$. Moreover, from Corollary 1 in Bura and Pfeiffer (2008), we have that the estimated eigenvectors satisfy

$$\begin{aligned}\sqrt{N}\mathrm{vec}[\widehat{V} - V] &= (\Lambda^{-1}V' \otimes \mathrm{I}_I)\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{A_*} + o_P(1),\\ \sqrt{N}\mathrm{vec}[\widehat{V}' - V'] &= (\mathrm{I}_I \otimes \Lambda^{-1}V')\ \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{A_*} + o_P(1).\end{aligned} \tag{B.2}$$

Combined with the linearization

$$\widehat{W} - W = (\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})V' + \Lambda^{-1/2}(\widehat{V} - V)' + o_P(1/\sqrt{N}),$$

(B.1) and (B.2) yield the influence functions as given in (3.1). This completes the proof. □

The following lemma contains distributional results for the JADE estimator of U.

LEMMA B.2. *Let Assumptions 2–4. Then*

$$\sqrt{N}\,\mathrm{vec}[\widehat{U} - U] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{U} + o_P(1), \qquad \sqrt{N}\,\mathrm{vec}[\widehat{U}' - U'] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{U'} + o_P(1),$$

where ψ_n^U *and* $\psi_n^{U'}$ *are given in* (3.2).

PROOF. Because $\widehat{C}_i = \widehat{W}\widehat{A}_i\widehat{W}'$ and both $\widehat{W}$ and $\widehat{A}_i$ are asymptotically linear, we have

$$\sqrt{N}\,\mathrm{vec}[\widehat{C}_i - C_i] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^{C_i} + o_P(1)$$

for $\psi_n^{C_i}$ as in (3.3). Theorem 5 in Bonhomme and Robin (2009) then yields the result. □

Appendix C. Proofs of theorems in the main text

Proof of Theorem 2. Write $b = (b_{0k}, b_{1k}, \ldots, b_{Jk})'$ and define $\widehat{b}$ and $\widetilde{b}$ in an analogous fashion, as before. We first show that $\|\widehat{b} - \widetilde{b}\| = O_P(1/\sqrt{N}) + O_P(\sqrt{J}/N)$. The theorem will then follow easily. By the Cauchy-Schwarz inequality,

$$\|\widehat{b} - \widetilde{b}\|^2 \le \sum_{j=0}^{J}\|\widehat{D}_j - \widetilde{D}_j\|^2 \le \|\widehat{U}'\widehat{W} \otimes \widehat{U}'\widehat{W} - U'W \otimes U'W\|^2 \ \sum_{j=0}^{J}\|\widehat{A}_j\|^2.$$

Lemmas B.1 and B.2 imply that $\|\widehat{U}'\widehat{W} \otimes \widehat{U}'\widehat{W} - U'W \otimes U'W\|^2 = O_P(1/N)$. Further,

$$\sum_{j=0}^{J}\|\widehat{A}_j\|^2 \le 2\sum_{j=0}^{J}\|A_j\|^2 + 2\sum_{j=0}^{J}\|\widehat{A}_j - A_j\|^2 = O(1) + O_P(J/N),$$

which follows from the proof of Theorem A.1. Thus, $\|\widehat{b} - \widetilde{b}\| = O_P(1/\sqrt{N}) + O_P(\sqrt{J}/N)$.

Now turn to the integrated squared-error result. Arguing as in the proof of Theorem A.1, we have

$$\|\widehat{f}_k - f_k\|_2^2 \leq 2\|\widehat{b} - \widetilde{b}\|^2 + 2\|\widetilde{f}_k - f_k\|_2^2.$$

From above, the first right-hand side term is $O_P(1/N) + O_P(J/N^2)$. By Theorem A.1, the second right-hand side term is $O_P(J/N + J^{-2\beta})$. Therefore, the difference between $\widehat{b}$ and $\widetilde{b}$ is asymptotically negligible, and $\|\widehat{f}_k - f_k\|_2^2 = O_P(J/N + J^{-2\beta})$. This establishes the first part of the theorem.

To show the second part, we use the triangle inequality to obtain the bound

$$\|\widehat{f}_k - f_k\|_\infty \leq \|\widehat{f}_k - \widetilde{f}_k\|_\infty + \|\widetilde{f}_k - f_k\|_\infty$$

Here, by the Cauchy-Schwarz inequality and using Assumption 2 and the argument above,

$$\|\widehat{f}_k - \widetilde{f}_k\|_\infty \leq \left\|\sqrt{\alpha_J}\right\|_\infty \left\|\widehat{b} - \widetilde{b}\right\| = O_P(\zeta_J/\sqrt{N}) + O_P(\zeta_J\sqrt{J}/N),$$

while $\|\widetilde{f}_k - f_k\|_\infty = O_P(\zeta_J\sqrt{J/N} + J^{-\beta})$, as was shown in the proof of Theorem A.1. Thus, $\|\widehat{f}_k - f_k\|_\infty = O_P(\zeta_J\sqrt{J/N} + J^{-\beta})$ and, again, the difference between the feasible and infeasible density estimators is asymptotically negligible. This concludes the proof. □

Proof of Theorem 3. Fix the evaluation point y throughout the proof. We will show that

$$\left|\widehat{f}_k(y) - \widetilde{f}_k(y)\right| = o_P(\sqrt{\alpha_J(y)/N}).$$

The result will then follow from the analysis of the infeasible estimator, which is provided in Theorem A.2. Introduce the shorthand $\gamma_{i_1 i_2}(y_1, y_2) \equiv \chi_{i_1}(y_1)\rho(y_1)\chi_{i_2}(y_2)\rho(y_2)$. Then we can write

$$\widehat{\tau}_k(y_1, y_2) - \tau_k(y_1, y_2) = \sum_{i_1=0}^{I}\sum_{i_2=0}^{I} \left(\widehat{u}_k'\widehat{w}_{i_1}\widehat{u}_k'\widehat{w}_{i_2} - u_k' w_{i_1} u_k' w_{i_2}\right) \gamma_{i_1 i_2}(y_1, y_2).$$

This allows us to bound $|\widehat{f}_k(y) - \widetilde{f}_k(y)|$ by

$$\sum_{i_1=0}^{I}\sum_{i_2=0}^{I} \left(\left|\widehat{u}_k'\widehat{w}_{i_1}\widehat{u}_k'\widehat{w}_{i_2} - u_k' w_{i_1} u_k' w_{i_2}\right| \left| \frac{1}{N}\sum_{n=1}^{N} \gamma_{i_1 i_2}(y_{n1}, y_{n2})\ \kappa_J(y, y_{n3})\rho(y_{n3}) \right| \right).$$

We will handle each of the terms inside the brackets in turn. First, by Lemmas B.1 and B.2,

$$|\widehat{u}_k'\widehat{w}_{i_1}\widehat{u}_k'\widehat{w}_{i_2} - u_k' w_{i_1} u_k' w_{i_2}| = O_P(1/\sqrt{N}) \tag{C.1}$$

for all (i_1, i_2), and so the first term converges at the parametric rate. It is readily verified that $\mathbb{E}[\gamma_{i_1 i_2}(y_1, y_2)\ \kappa_J(y, y_3)\rho(y_3)] = O(1)$, and so

$$\frac{\mathrm{var}[\gamma_{i_1 i_2}(y_1, y_2)\ \kappa_J(y, y_3)\rho(y_3)]}{\alpha_J(y)} = \frac{\mathbb{E}[\gamma_{i_1 i_2}(y_1, y_2)^2\ \kappa_J(y, y_3)^2\rho(y_3)^2]}{\alpha_J(y)} + o(1) = O(1),$$

which follows from the same arguments as those that were used in the proof of Theorem A.2. Therefore, for all (i_1, i_2),

$$\Big|\frac{1}{N}\sum_{n=1}^{N}\gamma_{i_1 i_2}(y_{n1}, y_{n2})\ \kappa_J(y, y_{n3})\rho(y_{n3})\Big| = O_P(\sqrt{\alpha_J(y)/N}). \tag{C.2}$$

Combining (C.1) and (C.2) then gives $|\widehat{f}_k(y) - \widetilde{f}_k(y)| = O_P(\sqrt{\alpha_J(y)}/N)$, which implies that

$$\sqrt{\frac{N}{\alpha_J(y)}}\big[\widehat{f}_k(y) - \widetilde{f}_k(y)\big] = o_P(1).$$

Together with Theorem A.2, this yields the result and concludes the proof. □

Proof of Theorem 4. Because $b_{ik} = u_k' C_i u_k$ for all i, Lemmas B.1 and B.2 directly imply asymptotic linearity of $\widehat{B}$. To derive the form of its influence function, first consider the linearization

$$\widehat{U}'\widehat{C}_i\widehat{U} - U'C_iU = \big(\widehat{U} - U\big)'C_iU + U'\big(\widehat{C}_i - C_i\big)U + U'C_i\big(\widehat{U} - U\big) + o_P(1/\sqrt{N}).$$

Recalling the $K^2 \times K$ selection matrix $(\mathrm{I}_K \overset{\text{col}}{\otimes} \mathrm{I}_K)$ used in the proof of Lemma B.1, we can further linearize the rows of $\widehat{B} - B$ as

$$\operatorname{vec}\big[\widehat{U}'\widehat{C}_i\widehat{U} - U'C_iU\big]'(\mathrm{I}_K \overset{\text{col}}{\otimes} \mathrm{I}_K).$$

We thus have that $\operatorname{vec}[\widehat{B}' - B'] = N^{-1}\sum_{n=1}^{N}\psi_n^B + o_P(1/\sqrt{N})$ for the random variables $\psi_n^B = \operatorname{vec}[(\psi_n^{D_0}, \psi_n^{D_1}, \ldots, \psi_n^{D_I})]$, where the form of $\psi_n^{D_i}$, which is the influence function of $\widehat{D}_i$, is given in (3.4). Now, by a linearization,

$$\widehat{B}'\widehat{a} - B'a = \frac{1}{N}\sum_{n=1}^{N}(a' \otimes \mathrm{I}_K)\psi_n^B + B'\psi_n^a + o_P(1/\sqrt{N}).$$

Further, as $\|\widehat{B} - B\| = o_P(1)$ and $B'B$ has full rank, $(\widehat{B}'\widehat{B})^{-1} \xrightarrow{P} (B'B)^{-1}$. It thus follows that $\widehat{\omega}$ is asymptotically linear with influence function ψ_n^ω, as given in (3.5). This proves the result. □

Proof of Theorem 5 Given our assumptions, standard arguments show that consistency of $\widehat{\theta}$ follows from uniform convergence of the empirical moment to the population moment. Let $m(\theta) \equiv N^{-1}\sum_n m_n(\theta)$. By a mean-value expansion around the first-step estimators, Assumption 5, in tandem with Theorems 2–4, yields

$$\sup_{\theta\in\Theta}\big\|\widehat{m}(\theta) - \mathbb{E}[m(\theta)]\big\| = O_P(\zeta_J\sqrt{J/N} + J^{-\beta}), \tag{C.3}$$

which converges to zero as $N \to \infty$. Turning to asymptotic normality, a Taylor expansion around θ_0 yields

$$\sqrt{N}(\widehat{\theta} - \theta_0) = (M_\theta'\Sigma M_\theta)^{-1}M_\theta'\Sigma\sqrt{N}\widehat{m}(\theta_0) + o_P(1),$$

because $\widehat{\Sigma} \xrightarrow{P} \Sigma$ and $\sup_{\theta\in\Theta}\|\partial_{\theta'}\widehat{m}(\theta) - M_\theta(\theta)\| = o_P(1)$, which follows as (C.3). It then remains to show that

$$\sqrt{N}\widehat{m}(\theta_0) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\big(m_n(\theta_0) + \psi_n^\phi\big) + o_P(1),$$

which amounts to quantifying the impact of the first-stage estimation error in the weight function. First, note that a second-order Taylor expansion around the $\widehat{f}_\ell(y_{nt})$ and ω_ℓ yields

$$\sqrt{N}\big[\widehat{m}(\theta_0) - m(\theta_0)\big] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\sum_{\ell=1}^{K}\Big[\xi_{n\ell} + \partial_{\omega_\ell} m_n(\theta_0)(\widehat{\omega}_\ell - \omega_\ell)\Big] + o_P(1),$$

where the order of the remainder term stems from the fact that $\sqrt{N}|\widehat{\omega}_\ell - \omega_\ell|^2 = o_P(1)$ and that $\sqrt{N}\|\widehat{f}_\ell - f_\ell\|_\infty^2 = \sqrt{N}(\zeta_J^2 J/N + J^{-2\beta}) \to 0$ as $N \to \infty$ for all ℓ by Theorems 2 and 4, and where we have introduced

$$\xi_{n\ell} = \frac{1}{T}\sum_{t=1}^{T} g(y_{nt};\theta_0)\partial_{f_\ell(y_{nt})}\phi_k(y_{nt})\big[\widehat{f}_\ell(y_{nt}) - f_\ell(y_{nt})\big].$$

Using the linear representation of the $\widehat{f}_\ell$ in Theorem 3, the sample average of $\xi_{n\ell}$ can be expressed as a symmetric U-statistic of order two. The projection of this U-statistic equals

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[g_0(y)\partial_{f_\ell(y)}\phi_k(y)\big[\tau_\ell(y_{n1},y_{n2})\kappa_J(y,y_{n3})\rho(y_{n3}) - \mathrm{Proj}_J[f_\ell](y)\big]\Big],$$

where the expectation is taken with respect to y. Computing this expectation for each ℓ and invoking Lemma A.3 in Ahn and Powell (1993) then shows that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\sum_{\ell=1}^{K}\xi_{n\ell} = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} g_0(y_{n3})\Big[\tau_k(y_{n1},y_{n2}) - \phi_k(y_{n3})\sum_{\ell=1}^{K}\tau_\ell(y_{n1},y_{n2})\omega_\ell\Big]\rho(y_{n3}) + o_P(1),$$

where we have used the assumption that $\sqrt{N}\|\mathrm{Proj}_J[g_0\phi_\ell] - g_0\phi_\ell\|_\infty = \sqrt{N}O(J^{-\eta}) = o(1)$, as well as the fact that $\partial_{f_\ell(y_{nt})}\phi_k(y_{nt}) = (\delta_{k\ell} - \phi_k(y_{nt})\omega_\ell)/f(y_{nt})$. Also, by essentially the same argument,

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\sum_{\ell=1}^{K}\partial_{\omega_\ell} m_n(\theta_0)(\widehat{\omega}_\ell - \omega_\ell) = -\frac{1}{\sqrt{N}}\sum_{n=1}^{N} M_\omega\psi_n^\omega + o_P(1),$$

were we have used that $\partial_{\omega_\ell}\phi_k(y_{nt}) = -\phi_k(y_{nt})\phi_\ell(y_{nt})$ in defining the Jacobian matrix M_ω. Putting together the result gives

$$\sqrt{N}\big[\widehat{m}(\theta_0) - m(\theta_0)\big] = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\psi_n^\phi + o_P(1).$$

This completes the proof. □

References

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*, 3–29.

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

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.

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

Bura, E. and R. Pfeiffer (2008). On the distribution of 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. *IEEE-Proceedings, F 140*, 362–370.

Comon, P. and C. Jutten (2010). *Handbook of Blind Source Separation*. Academic Press.

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

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

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

Gottschalk, P. and R. A. Moffitt (1994). The growth of earnings instability in the U.S. labor market. *Brookings Papers on Economic Activity 25*, 217–272.

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

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

Henry, M., Y. Kitamura, and B. Salanié (2013). Partial identification of finite mixtures in econometric models. Unpublished manuscript.

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

Hu, Y. and S. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*, 195–216.

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

Kasahara, H. and K. Shimotsu (2013). Nonparametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society, Series B*, forthcoming.

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

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

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

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

Moffitt, R. A. and P. Gottschalk (2012). Trends in the transitory variance of male earnings. *Journal of Human Resources 47*, 204–236.

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

Powell, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press.

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

Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *Annals of Mathematical Statistics 38*, 1261–1265.

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.