SciencesPo. | Department of Economics

Discussion paper 2014-18

Estimating Multivariate Latent-Structure Models

Stéphane Bonhomme
Koen Jochmans
Jean-Marc Robin

Sciences Po Economics Discussion Papers

ESTIMATING MULTIVARIATE LATENT-STRUCTURE MODELS

By Stéphane Bonhomme, Koen Jochmans and Jean-Marc Robin

University of Chicago; Sciences Po, Paris; and Sciences Po, Paris and University College, London

This version: December 12, 2014

A constructive proof of identification of multilinear decompositions of multiway arrays is presented. It can be applied to show identification in a variety of multivariate latent structures. Examples are finite-mixture models and hidden Markov models. The key step to show identification is the joint diagonalization of a set of matrices in the same non-orthogonal basis. An estimator of the latent-structure model may then be based on a sample version of this simultaneous-diagonalization problem. Simple algorithms are available for computation. Asymptotic theory is derived for this joint approximate-diagonalization estimator.

1. Introduction. Latent structures are a popular tool for modeling the dependency structure in multivariate data. Two important examples are finite-mixture models (see McLachlan and Peel 2000) and hidden Markov models (see Cappé, Moulines and Rydén 2005). Although these models arise frequently in applied work, the question of their nonparametric identifiability has attracted substantial attention only quite recently. Allman, Matias and Rhodes [2009] used algebraic results on the uniqueness of decompositions of multiway arrays due to Kruskal [1976; 1977] to establish identification in a variety of multivariate latent-structure models. Their setup covers both finite mixtures and hidden Markov models, among other models, and their findings substantially generalize the earlier work of Green [1951], Anderson [1954], Petrie [1969], Hettmansperger and Thomas [2000], Hall and Zhou [2003], and Hall et al. [2005].

Despite these positive identification results, direct application of Kruskal's method does not provide an estimator. Taking identification as given, some

AMS 2000 subject classifications: Primary, 15A69, 62G05; secondary, 15A18, 15A23, 62G20, 62H17, 62H30.

Keywords and phrases: hidden Markov model, finite mixture model, latent structure, multilinear restrictions, multivariate data, nonparametric estimation, simultaneous matrix diagonalization.

authors have developed EM-type approaches to nonparametrically estimate both multivariate finite mixtures (Benaglia, Chauveau and Hunter 2009; Levine, Hunter and Chauveau 2011) and hidden Markov models (Gassiat, Cleynen and Robin 2013). Numerical studies suggest that these estimators are well-behaved. Because they are not based directly on the identification argument, however, their statistical properties—that is, their consistency, convergence rates, and asymptotic distribution—are difficult to establish and are currently unknown.

There are some theoretical results on statistical inference in semi- and nonparametric finite-mixture models and hidden Markov models in more restrictive settings. These include location models (Bordes, Mottelet and Vandekerkhove 2006; Hunter, Wang and Hettmansperger 2007; Gassiat and Rousseau 2014), multivariate finite mixtures with identically distributed outcome variables, i.e., stationary data (Hettmansperger and Thomas 2000; Bonhomme, Jochmans and Robin 2014), and two-component mixtures and mixture models featuring exclusion restrictions (Hall and Zhou 2003; Henry, Jochmans and Salanié 2013). However, these models are too restrictive for many situations of practical interest and the arguments underlying these various methods are not easy to generalize.

In this paper we show that the multilinear structure underlying the results of Allman, Matias and Rhodes [2009] can be used to obtain a constructive proof of identification in a broad class of latent-structure models. We show that the problem of decomposing a multiway array can be reformulated as the problem of simultaneously diagonalizing a collection of matrices. This is a least-squares problem that has received considerable attention in the literature on independent component analysis and blind source separation (see Comon and Jutten 2010). Moreover, algorithms exist to recover the joint diagonalizer in a computationally-efficient manner; see Fu and Gao [2006], Iferroudjene, Abed-Meraim and Belouchrani [2009; 2010], and Luciani and Albera [2010; 2014].

We propose estimating the parameters of the latent-structure model by solving a sample version of the simultaneous-diagonalization problem. We provide distribution theory for this estimator below. Under weak conditions, it converges at the parametric rate and is asymptotically normal. Using this result, we obtain estimators of finite-mixture models and hidden Markov models that have standard asymptotic properties. Moreover, the fact that the dependency structure in the data is latent does not translate into a decrease in the convergence rate of the estimators. As such, this paper is

the first to derive the asymptotic behavior of nonparametric estimators of multivariate finite mixture models of the form in Hall and Zhou [2003] for more than two latent classes and of hidden Markov models of the form in Gassiat, Cleynen and Robin [2013]. Furthermore, our approach can be useful in the analysis of random graph models (Allman, Matias and Rhodes 2011) and stochastic blockmodels (Snijders and Nowicki 1997; Rohe, Chatterjee and Yu 2011), although we do not consider such models in detail in this paper. In a simulation study, we further find our approach to perform well in small samples.

There is a large literature on parallel factor analysis and canonical polyadic decompositions of tensors building on the work of Kruskal [1976; 1977]; see, e.g., De Lathauwer, De Moor and Vandewalle [2004], De Lathauwer [2006], and Anandkumar et al. [2014]. Although our identification strategy has some similarity with this literature, both our conclusions and our simultaneous-diagonalization problem are different. In the context of multivariate finite mixtures of identically distributed variables, Kasahara and Shimotsu [2009] and Bonhomme, Jochmans and Robin [2014] also used joint diagonalization to show nonparametric identification. However, the approaches taken there are different from the one developed in this paper and cannot be applied as generally.

We start out by motivating our approach via a discussion on the algebraic structure of multivariate finite-mixture models and hidden Markov models. We then present our identification strategy in a generic setting, and illustrate its usefulness by applying it to our two motivating examples. After this we turn to estimation and inference, and to the development of asymptotic theory. Next, the theory is used to set up orthogonal-series estimators of component densities in a finite-mixture model, and to show that these have the standard univariate convergence rates of series estimators. Finally, the orthogonal-series density estimator is put to work in a small simulation experiment. The proofs of all the main identification arguments are given in the text. The remaining proofs are collected in the Appendix.

2. Motivating examples.

We start by introducing three examples to motivate our subsequent developments.

2.1. *Finite-mixture models for discrete measurements.*

Let $Y_1, Y_2, \ldots, Y_q$ be observable random variables that are assumed independent conditional on realizations of a latent random variable Z. Suppose that Z has a finite state space of known cardinality r, which we set to $\{1, 2, \ldots, r\}$ without loss

of generality. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_r)'$ be the probability distribution of Z, so $\pi_j > 0$ and $\sum_{j=1}^r \pi_j = 1$. Then the probability distribution of $Y_1, Y_2, \ldots, Y_q$ is a multivariate finite mixture with mixing proportions $\pi_1, \pi_2, \ldots, \pi_r$. The parameters of interest are the r mixing proportions and the distributions of $Y_1, Y_2, \ldots, Y_q$ given Z. The Y_i need not be identically distributed, so the model involves qr such conditional distributions.

When Y_i can take on any of a finite number κ_i of values the problem is effectively finite-dimensional. Write p_{ij} for the probability distribution of Y_i given $Z = j$, which is a $\kappa_i \times 1$ vector, and let $\otimes$ denote the outer (tensor) product. The probability distribution of $Y_1, Y_2, \ldots, Y_q$ given $Z = j$ then is the q-way table

$$\bigotimes_{i=1}^{q} p_{ij} = p_{1j} \otimes p_{2j} \otimes \cdots \otimes p_{qj},$$

which is of dimension $\kappa_1 \times \kappa_2 \times \cdots \times \kappa_q$. The outer-product representation follows from the conditional-independence restriction. Hence, the marginal probability distribution of $Y_1, Y_2, \ldots, Y_q$ equals

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} p_{ij}, \tag{2.1}$$

which is an r-linear decomposition of a q-way array. The parameters of the mixture model are all the vectors making up the outer-product arrays, $\{p_{ij}\}$ and the coefficients of the linear combination, $\{\pi_j\}$, transforming the conditional distributions into the marginal distribution $\mathbb{P}$.

The r-linear decomposition is not restricted to the contingency table. Indeed, any linear functional of $\mathbb{P}$ admits a decomposition in terms of the same functional of the p_{ij}. Moreover, for any collection of vector-valued transformations $y \mapsto \chi_i(y)$ we have

$$E\left[\bigotimes_{i=1}^{q} \chi_i(Y_i)\right] = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} E[\chi_i(Y_i)|Z = j], \tag{2.2}$$

provided the expectation exists. Equation (2.1) is recovered from (2.2) with $\chi_i(y) = (1\{y = y_{i1}\}, 1\{y = y_{i2}\}, \ldots, 1\{y = y_{i\kappa_i}\})'$ and $\{y_{ik}\}$ the support points of p_{ij}. Of course, identification of linear functionals follows from identification of the p_{ij}. However, other choices for the χ_i can be useful when turning to estimation. This idea was exploited in one way in a more restrictive setting in Bonhomme, Jochmans and Robin [2014]. Here, we wish to highlight another approach. To do so we turn to a model with continuous outcomes.

2.2. *Finite-mixture models for continuous measurements.* Suppose now that the Y_i are continuously-distributed random variables. Setting the χ_i in (2.2) to indicators that partition their state space yields a decomposition as in (2.1) for a discretized version of the mixture model. This can suffice for identification (see Allman, Matias and Rhodes 2009 and Kasahara and Shimotsu 2014) but it is not attractive for the estimation of the distributions and of the corresponding density functions. An alternative approach based on (2.2) is to work with a discretization in the frequency domain, which can subsequently be used to deliver smooth estimates.

Let f_{ij} be the density of Y_i given $Z = j$. Assume that $(Y_1, Y_2, \ldots, Y_q)$ lives in the q-dimensional compact set $[-1, 1]^q$; translation to generic compact sets is straightforward. Also, let $\{\phi_k, k > 0\}$ be a class of functions that form a complete orthonormal system with respect to some weight function ρ on $[-1, 1]$. Polynomials such as those belonging to the Jacobi class—e.g., Chebychev or Legendre polynomials—can serve this purpose.

Assume the f_{ij} to be square-integrable with respect to ρ. The projection of f_{ij} onto the subspace spanned by $\varphi_{\kappa_i} = (\phi_1, \phi_2, \ldots, \phi_{\kappa_i})'$ is

$$\mathrm{Proj}_{\kappa_i} f_{ij} = \varphi'_{\kappa_i} b_{ij}, \quad b_{ij} = \int_{-1}^{1} \varphi_{\kappa_i}(y)\rho(y) f_{ij}(y)\,\mathrm{d}y = E[\varphi_{\kappa_i}(Y_i)\rho(Y_i)|Z = j],$$

for any integer κ_i. The vector b_{ij} collects the (generalized) Fourier coefficients of f_{ij}. The projection converges to f_{ij} in L^2_ρ-norm, that is,

$$\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\|_2 \to 0$$

as $\kappa_i \to \infty$, where $\|f\|_2 = \big(\int_{-1}^{1} f(y)^2 \rho(y)\,\mathrm{d}y\big)^{1/2}$. Such projections are commonly-used tools in the approximation of functions (Powell 1981) and underly orthogonal-series estimators of densities and conditional-expectation functions (Efromovich 1999).

The b_{ij} are not directly observable. However, an application of (2.2) yields

$$\mathbb{B} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} b_{ij} \tag{2.3}$$

for $\mathbb{B} = E[\bigotimes_{i=1}^{q} \varphi_{\kappa_i}(Y_i)\rho(Y_i)]$. The latter expectation is a q-way array that can be computed directly from the data. It contains the leading Fourier coefficients of the q-variate density function of the data. Again, the array $\mathbb{B}$ factors into a linear combination of multiway arrays. In Section 6 we will use this representation to derive orthogonal-series density estimators that have standard large-sample properties.

2.3. *Hidden Markov models.* Let $\{Y_i, Z_i\}_{i=1}^q$ be a stationary sequence. Z_i is a latent variable with finite state space $\{1, 2, \ldots, r\}$, for known r, and has first-order Markov dependence. Denote by π the vector of stationary probabilities and write K for the $r \times r$ matrix of transition probabilities. Moreover, $K(z_1, z_2)$ equals the probability of moving from state z_1 to z_2. The observable random variables $Y_1, Y_2, \ldots, Y_q$ are independent conditional on realizations of $Z_1, Z_2, \ldots, Z_q$, and the distribution of Y_i only depends on the realization of Z_i. We write p_j for the distribution of Y_i given $Z_i = j$. Here, the parameters of interest are the emission distributions $\{p_j\}_{j=1}^r$, the stationary distribution of the Markov chain π, and the matrix of transition probabilities K.

Suppose that Y_i is discrete and that its state space contains κ points of support. Let $P = (p_1, p_2, \ldots, p_r)$, the $\kappa \times r$ matrix of emission distributions and write $\Pi = \text{diag}(\pi_1, \pi_2, \ldots, \pi_r)$. Then the matrix

$$B = PK' = (b_1, b_2, \ldots, b_r)$$

contains the probability distributions of Y_i given Z_{i-1} and, likewise, the matrix

$$A = P\Pi K \Pi^{-1} = (a_1, a_2, \ldots, a_r)$$

provides the distribution of Y_i given Z_{i+1}. To study identification it suffices to restrict attention to $q = 3$. In this case, it is easy to show that the contingency table of (Y_1, Y_2, Y_3) factors as

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \, (a_j \otimes p_j \otimes b_j); \tag{2.4}$$

see also Allman, Matias and Rhodes [2009] and Gassiat, Cleynen and Robin [2013]. When $q > 3$ we may bin several outcomes together and proceed as before.

When the Y_i are continuously distributed we may again work from a factorization of linear functionals similar to the one considered in (2.2) and (2.3).

3. Algebraic structure and identification. Our approach can be applied to q-variate structures that decompose as q-ads.

DEFINITION 1. *A q-dimensional array $\mathbb{X} \in \mathbb{R}^{\kappa_1 \times \kappa_2 \times \cdots \times \kappa_q}$ is a q-ad if it can be decomposed as*

$$\mathbb{X} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} x_{ij} \tag{3.1}$$

for some integer r*, non-zero weights* $\pi_1, \pi_2, \ldots, \pi_r$*, and vectors* $x_{ij} \in \mathbb{R}^{\kappa_i \times 1}$.

Our interest lies in nonparametrically recovering $\{x_{ij}\}$ and $\{\pi_j\}$ from knowledge of $\mathbb{X}$ and r. Clearly, these parameters are not unique, in general. For example, a permutation of the x_{ij} and π_j leaves $\mathbb{X}$ unaffected, and a common scaling of the x_{ij} combined with an inverse scaling of the π_j, too, does not change the q-way array. However, the work of Kruskal [1976; 1977] and Sidiropoulos and Bro [2000] gives a simple sufficient condition for uniqueness of the decomposition up to these two indeterminacies. This condition, which cannot be satisfied when $q < 3$, was further established to be necessary by Derksen [2013].

While permutational equivalence of possible decompositions of $\mathbb{X}$ is a mostly trivial and inherently unresolvable ambiguity, indeterminacy of the scale of the vectors x_{ij} is undesirable in many situations. Indeed, in arrays of the general form in (2.2), recovering the scale of the x_{ij} and the constants π_j is fundamental. In some cases natural scale restrictions may be present. Indeed, in (2.1) the x_{ij} are known to be probability distributions, and so they have non-negative entries that sum to one. Suitably combining these restrictions with Kruskal's theorem, Allman, Matias and Rhodes [2009] were able to derive conditions under which the parameters in finite mixtures and hidden Markov models are uniquely determined up to relabelling of the latent classes.

We follow a different route to determine q-adic decompositions up to permutational equivalence. We require that, apart from the q-way array $\mathbb{X}$, lower-dimensional submodels are also observable. By lower-dimensional submodels we mean arrays that factor as

$$\sum_{j=1}^{r} \pi_j \bigotimes_{i \in \mathcal{Q}} x_{ij} \tag{3.2}$$

for sets $\mathcal{Q}$ that are subsets of the index set $\{1, 2, \ldots, q\}$. This is not a strong requirement in the models we have in mind. In the discrete multivariate mixture model in (2.1), for example, lower-dimensional submodels are simply the contingency tables of a subset of the outcome variables. There, going from a q-way table down to a $(q-1)$-table featuring all but the ith outcome boils down to summing the array in the ith direction. Such marginalizations are also inherent in the work on multivariate mixtures by Hall and Zhou [2003] and Kasahara and Shimotsu [2009]. In general settings such as (2.2), and in the multilinear equation involving Fourier coefficients in particular,

the advantage of working with submodels over marginalizations of the model is apparent.

Note that, throughout, we take r in (3.1) to be known. This ensures $\{x_{ij}\}$ and $\{\pi_j\}$ to be unambiguously defined. Indeed, for different r, there may exist a different set of weights and vectors so that $\mathbb{X}$ factors as a q-ad. The rank of $\mathbb{X}$ is the smallest integer r needed to arrive at a decomposition as in Definition 1.

3.1. *Unfolding.* We can state our main identification result for three-way arrays without loss of generality. This is so because any q-way array can be unfolded into a $(q-1)$-way array, much like any matrix can be transformed into a vector using the vec operator. Indeed, in any direction $i \in \{1, 2, \ldots, q\}$, a q-way array of dimension $\kappa_1 \times \kappa_2 \times \cdots \times \kappa_q$ is a collection of κ_i $(q-1)$-way arrays, each of dimension $\kappa_1 \times \kappa_2 \times \cdots \kappa_{i-1} \times \kappa_{i+1} \times \cdots \kappa_q$. This collection can be stacked in any of $i' \in \{1, 2, \ldots, i-1, i+1, \ldots\ldots, q\}$ directions—i.e., $(q-1)$ different ways—to yield a $(q-1)$-way array whose dimension will be $\kappa_1 \times \kappa_2 \times \kappa_i\kappa_{i'} \times \cdots \kappa_q$. This unfolding process can be iterated until it yields a three-way array (see, e.g., Sorensen et al. 2013).

3.2. *Identification via simultaneous diagonalization.* Let $\mathbb{X}$ be a three-way array of dimension $\kappa_1 \times \kappa_2 \times \kappa_3$ that factors as a tri-ad for known r, that is,

$$\mathbb{X} = \sum_{j=1}^{r} \pi_j \, (x_{1j} \otimes x_{2j} \otimes x_{3j}).$$

Let $X_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ and $\Pi = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_r)$. Also, for each pair (i_1, i_2) with $i_1 < i_2$ in $\{1, 2, 3\}^2$, let

$$\mathbb{X}_{\{i_1,i_2\}} = \sum_{j=1}^{p} \pi_j \, (x_{i_1 j} \otimes x_{i_2 j}).$$

Note that, from (3.2), $\mathbb{X}_{\{i_1,i_2\}}$ is the lower-dimension submodel obtained from $\mathbb{X}$ on omitting the index i_3.

Our first theorem concerns identification of the X_i and is the cornerstone result of our argument.

THEOREM 1 (Columns of X_i). *If X_{i_1} and X_{i_2} both have full column rank and $\mathbb{X}_{\{i_1,i_2\}}$ is observable, then X_{i_3} is identified up to a permutation matrix if all its columns are different.*

PROOF. Without loss of generality, fix $(i_1, i_2, i_3) = (1, 2, 3)$ throughout the proof. In each direction i, the three-way array $\mathbb{X}$ consists of a collection of κ_i matrices. Let $A_1, A_2, \ldots, A_{\kappa_3}$ denote these matrices for $i = 3$. So, the matrix A_k is obtained from $\mathbb{X}$ on fixing its third index to the value k, that is, $A_k = \mathbb{X}(:, :, k)$, using obvious array-indexing notation. Also, let $A_0 = \mathbb{X}_{\{1,2\}}$. Note that all of A_0 and $A_1, A_2, \ldots, A_{\kappa_3}$ are observable matrices of dimension $\kappa_1 \times \kappa_2$.

By construction, the lower-dimensional submodel A_0 has the structure

$$A_0 = X_1 \Pi X_2'.$$

Because the matrices X_1 and X_2 both have rank r and because all π_j are non-zero by definition, the matrix A_0, too, has rank r. Therefore, it has a singular-value decomposition

$$A_0 = USV'$$

for unitary matrices U and V of dimension $\kappa_1 \times r$ and $\kappa_2 \times r$, respectively, and a non-singular $r \times r$ diagonal matrix S. Now construct $W_1 = S^{-1/2}U'$ and $W_2 = S^{-1/2}V'$. Then,

$$W_1 A_0 W_2' = (W_1 X_1 \Pi^{1/2})(W_2 X_2 \Pi^{1/2})' = QQ^{-1} = \mathrm{I}_r,$$

where I_r denotes the $r \times r$ identity matrix and $Q = W_1 X_1 \Pi^{1/2}$.

Moving on, each of $A_1, A_2, \ldots, A_{\kappa_3}$ has the form

$$A_k = X_1 \Pi D_k X_2', \qquad D_k = \operatorname{diag}_k X_3,$$

where $\operatorname{diag}_k X$ denotes the diagonal matrix whose diagonal equals the kth row of matrix X. Applying the same transformation to $A_1, A_2, \ldots, A_{\kappa_3}$ yields the collection of $r \times r$ matrices

$$W_1 A_k W_2' = Q D_k Q^{-1}. \tag{3.3}$$

So, the matrices $\{W_1 A_k W_2'\}$ are diagonalizable in the same basis, namely, the columns of matrix Q. The associated eigenvalues $\{D_k\}$ equal the columns of the matrix X_3. These eigenvalues are unique up to a joint permutation of the eigenvectors and eigenvalues provided there exist no $k_1 \neq k_2$ so that the vectors of eigenvalues of $W_1 A_{k_1} W_2'$ and $W_1 A_{k_2} W_2'$ are equal (see, e.g., De Lathauwer, De Moor and Vandewalle 2004, Theorem 6.1). Now, this is equivalent to demanding that the columns of X_3 are all distinct. As this is true by assumption, the proof is complete. □

The proof of Theorem 1 shows that access to lower-dimensional submodels allows to disentangle the scale of the columns of the X_i and the weights on the diagonal of Π. This is so because the matrix Π equally shows up in the lower-dimensional submodels, and so transforming A_k to $W_1 A_k W_2'$ absorbs the weights into the joint diagonalizer Q in (3.3). Also note that the dimension of the matrices in (3.3) is $r \times r$, independent of the size of the original matrices X_i. On the other hand, larger matrices X_i could be beneficial for identification, as it becomes easier for them to satisfy the requirement of full column rank.

Theorem 1 can be applied to recover the tri-adic decomposition of $\mathbb{X}$ up to an arbitrary joint permutation matrix. We present the result in the form of two theorems.

THEOREM 2 (Vectors). *If X_1, X_2, and X_3 have full column rank and for each pair $(i_1, i_2) \in \{i_1, i_2 \in \{1,2,3\} : i_1 < i_2,\}$ $\mathbb{X}_{\{i_1,i_2\}}$ is observable, then X_1, X_2, and X_3 are all identified up to a common permutation matrix.*

PROOF. Theorem 1 can be applied to each direction of the three-way array $\mathbb{X}$. This yields the X_i up to permutation of their columns. However, as each X_i is recovered from a different simultaneous-diagonalization problem, the ordering of the columns of the X_i so obtained need not be the same. Hence, it remains to be shown that we can unravel the orderings. More precisely, application of Theorem 1 for each i identifies, say,

$$X_1, \qquad \overline{X}_2 = X_2\Delta_2, \qquad \overline{X}_3 = X_3\Delta_3,$$

where Δ_2 and Δ_3 are two permutation matrices.

Now, given X_1 and the lower-dimensional submodels $\mathbb{X}_{\{1,2\}}$ and $\mathbb{X}_{\{1,3\}}$, we observe the projection coefficients

$$M_2 = (X_1'X_1)^{-1}X_1'\,\mathbb{X}_{\{1,2\}} = \Pi X_2' = \Pi\Delta_2\overline{X_2}',$$
$$M_3 = (X_1'X_1)^{-1}X_1'\,\mathbb{X}_{\{1,3\}} = \Pi X_3' = \Pi\Delta_3\overline{X_3}',$$

where the first transition holds by the structure of the lower-dimensional submodels and the second transition follows from the fact that permutation matrices are orthogonal. Also,

$$\mathbb{X}_{\{2,3\}} = X_2\Pi X_3' = \overline{X}_2\Delta_2'\Pi\Delta_3\overline{X}_3' = M_2'\Delta_3\overline{X}_3', \quad \mathbb{X}_{\{2,3\}}' = M_3'\Delta_2\overline{X}_2'.$$

The latter two equations can be solved for the permutation matrices, yielding

$$\begin{aligned}\Delta_2 &= (M_3M_3')^{-1}M_3\;\mathbb{X}_{\{2,3\}}'\,\overline{X}_2(\overline{X}_2'\overline{X}_2)^{-1},\\ \Delta_3 &= (M_2M_2')^{-1}M_2\;\mathbb{X}_{\{2,3\}}\,\overline{X}_3(\overline{X}_3'\overline{X}_3)^{-1}.\end{aligned} \tag{3.4}$$

This concludes the proof. □

THEOREM 3 (Weights). *If X_i is identified up to a permutation matrix and has full column rank, and if $\mathbb{X}_{\{i\}}$ is observable, then π is identified up to the same permutation matrix.*

PROOF. The one-dimensional submodel $\mathbb{X}_{\{i\}}$ is the vector

$$\mathbb{X}_{\{i\}} = X_i \pi.$$

Given X_i, the one-dimensional submodel yields linear restrictions on the weight vector π. Moreover, if X_i is known and has maximal column rank, these equations can be solved for π, giving

$$\pi = (X_i' X_i)^{-1} X_i' \mathbb{X}_{\{i\}}, \tag{3.5}$$

which is the least-squares coefficient of a regression of $\mathbb{X}_{\{i\}}$ on the columns of X_i. □

In the field of multilinear algebra, the work of Kruskal [1976; 1977] has led to the development of parallel factor analysis, i.e., the canonical polyadic (CP) decomposition of tensors; see, e.g., Harshman [1970] and Carroll and Chang [1970] for early contributions. There is now a substantial literature on computational algorithms to arrive at CP decompositions; De Lathauwer [2006], De Lathauwer and Nion [2008], and Comon and De Lathauwer [2010] represent recent contributions. An important difference is that, unlike CP decompositions, our approach allows to learn the scale of the x_{ij}, which is particularly useful in the statistical applications we have in mind, such as those in (2.1) and (2.2).

4. Applications. Before turning to estimation we apply our approach to obtain constructive identification results in our motivating examples from Section 2.

4.1. *Latent-class models.* First reconsider the finite-mixture model with discrete outcomes and a known number of components r in (2.1), that is, the q-way table

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} p_{ij}.$$

Let $P_i = (p_{i1}, p_{i2}, \ldots, p_{ir})$ and $\Pi = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_r)$. Our next theorem concerns identification of these parameters.

THEOREM 4 (Identification of finite mixtures). *The matrices* $\{P_i\}$ *and* Π *in the finite mixture model in* (2.1) *are all identified if* $\operatorname{rank} P_i = r$ *for all* i, $\pi_j > 0$ *for all* j, *and* $q \geq 3$.

PROOF. To show Theorem 4 it suffices again to set $q = 3$. The proof is then a direct application of our identification result. Theorem 2 yields the matrices of component distributions $\{P_i\}$ and Theorem 3 yields the vector of mixing proportions π, all up to a common permutation matrix. □

Theorem 4 requires that $\kappa_i \geq r$ for all $i = 1, 2, \ldots, q$, that is, that all distributions p_{ij} have more than r points of support but applies as soon as $q = 3$. The rank conditions can be weakened when $q > 3$. Our approach to proving Theorem 4 is a constructive version of the proof of Theorem 4 in Allman, Matias and Rhodes [2009].

4.2. *Hidden Markov models.* Now turn to the hidden Markov model in (2.4) with a known number of latent states, r. In this model, the parameters of interest are the $\kappa \times r$ matrix of emission distributions $P = (p_1, p_2, \ldots, p_r)$, the stationary distribution of the r latent states π, and the $r \times r$ matrix of transition probabilities K. The next theorem gives sufficient conditions for identification.

THEOREM 5 (Identification of hidden Markov models). *The matrices* P, K, *and* Π *in the hidden Markov model are all identified if* $\operatorname{rank} P = r$ *and* $\operatorname{rank} K = r$, *and* $\pi_j > 0$ *for all* j *provided* $q \geq 3$.

PROOF. Set $q = 3$. Then the contingency table of three measurements factors as

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \, (a_j \otimes p_j \otimes b_j);$$

see (2.4). Moreover, this states that appropriate conditioning allows to write the hidden Markov models as a finite-mixture model of the form in (2.1). Furthermore, the rank conditions on P and K imply that both $B = PK'$ and $A = P\Pi K \Pi^{-1}$ also have full column rank r. Hence, Theorem 4 immediately yields identification of the matrix of emission distributions P and of the stationary distribution π.

Finally, Theorem 4 also provides the matrix B and, because the model implies that $B = PK'$,

$$K = B'P(P'P)^{-1}.$$

The hidden Markov model is overidentified. Indeed, besides B we also have the matrix A, which yields the same type of restrictions on the matrix K. □

Theorem 5 states the same identification requirements as Theorem 2.1 of Gassiat, Cleynen and Robin [2013], but the method of proof followed here is constructive.

5. Estimation by joint approximate diagonalization. Theorem 1 shows that the key restrictions underlying our results take the form of a set of matrices being simultaneously diagonalizable in the same basis. The problem of simultaneous matrix diagonalization has received considerable attention in the field of independent component analysis. Computationally-efficient algorithms for it have been developed; see Fu and Gao [2006], Iferroudjene, Abed-Meraim and Belouchrani [2009; 2010], and Luciani and Albera [2010; 2014]. Such algorithms can be exploited here to construct easy-to-implement nonparametric estimators of multivariate latent-structure models. Moreover, we recommend to use the algorithm developed in Iferroudjene, Abed-Meraim and Belouchrani [2009; 2010], which minimizes a least-squares criterion (see (5.4) below). The formulation of their approach as an extremum estimator allows for easy derivation of asymptotic theory.

Thus, we propose estimating the latent-structure model in (3.1) as follows. Given an estimate of the array $\mathbb{X}$ and of its lower-dimensional submodels, first estimate all x_{ij} by solving a sample version of the joint diagonalization problem in (3.3), possibly after unfolding if $q > 3$. Next, back out the weights $\pi_1, \pi_2, \ldots, \pi_r$ by solving the sample analog of the minimum-distance problem in (3.5). Asymptotic theory for this second step follows readily by the delta method. If desired, a consistent labelling can be recovered by estimating the permutation matrices from a plug-in version of (3.4).

5.1. *Estimator.* Consider a generic situation in which a set of κ $r \times r$ matrices $C_1, C_2, \ldots, C_\kappa$ can be jointly diagonalized by an $r \times r$ invertible matrix Q_0, that is,

$$C_k = Q_0 D_k Q_0^{-1}, \tag{5.1}$$

for diagonal matrices $D_1, D_2, \ldots, D_\kappa$. Knowledge of the joint eigenvectors implies knowledge of the eigenvalues, as

$$D_k = Q_0^{-1} C_k Q_0. \tag{5.2}$$

The matrix Q_0 is not unique. Moreover, let $\operatorname{off} Q = Q - \operatorname{diag} Q$ and let $\|Q\|_F = \sqrt{\operatorname{trace}(Q'Q)}$ denote the Frobenius norm. Then any solution to the least-squares problem

$$\min_{Q} \sum_{k=1}^{\kappa} \|\operatorname{off}(Q^{-1} C_k Q)\|_F^2 \tag{5.3}$$

is a joint diagonalizer in the sense of (5.1). Each of these delivers the same set of eigenvalues in (5.2) (up to a joint permutation).

The statistical problem of interest in this section is to perform inference on the $D_1, D_2, \ldots, D_\kappa$ when we only observe noisy versions of the input matrices $C_1, C_2, \ldots, C_\kappa$, say $\widehat{C}_1, \widehat{C}_2, \ldots, \widehat{C}_\kappa$. Sampling noise in the $\widehat{C}_k$ prevents them from sharing the same set of eigenvectors. Indeed, in general, there does not exist a Q such that $Q^{-1}\widehat{C}_k Q$ will be exactly diagonal for all k. For this, the least-squares formulation in (5.2)–(5.3) is important as it readily suggests using, say $\widehat{Q}$, any solution to

$$\min_{Q \in \mathscr{Q}} \sum_{k=1}^{\kappa} \|\operatorname{off}(Q^{-1} \widehat{C}_k Q)\|_F^2, \tag{5.4}$$

where $\mathscr{Q}$ is an appropriately-specified space of matrices to search over; see below. The estimator $\widehat{Q}$ is that matrix that makes all these matrices as diagonal as possible, in the sense of minimizing the sum of their squared off-diagonal entries. It is thus appropriate to call the estimator $\widehat{Q}$ the joint approximate-diagonalizer of $\widehat{C}_1, \widehat{C}_2, \ldots, \widehat{C}_\kappa$. An estimator of the D_k (up to a joint permutation of their eigenvalues) then is

$$\widehat{D}_k = \operatorname{diag}(\widehat{Q}^{-1} \widehat{C}_k \widehat{Q}). \tag{5.5}$$

Distribution theory for this estimator is not available, however, and so we provide it here. Throughout, we work under the convention that estimates are computed from a sample of size n.

5.2. *Asymptotic theory.* For our problem to be well-defined we assume that the matrix of joint eigenvectors is bounded. In (5.4), we may therefore restrict attention to the set of $r \times r$ matrices $Q = (q_1, q_2, \ldots, q_r)$ defined as

$$\mathscr{Q} = \{Q : \det Q = 1, \|q_j\|_F = c \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m < \infty\}$$

for some $m > 0$. The restrictions on the determinant and the column norms are without loss of generality and only reduce the space of matrices to be

searched over when solving (5.4). Let Q_* be any solution to (5.3) on $\mathscr{Q}$ and let $\mathscr{Q}_0 \subset \mathscr{Q}$ be the set of all matrices $Q_*\Delta\Theta$ for permutation matrices Δ and diagonal matrices Θ whose diagonal entries are equal to 1 and -1 and have $\det\Theta = 1$. Then $\mathscr{Q}_0$ is the set of solutions to (5.3) on $\mathscr{Q}$.

Construct the $r \times r\kappa$ matrix $C = (C_1, C_2, \ldots, C_\kappa)$ by concatenation and define $\widehat{C}$ similarly.

THEOREM 6 (Consistency). *If the set $\mathscr{Q}_0$ belongs to the interior of $\mathscr{Q}$, $\widehat{C} = C + o_p(1)$, and $\widehat{Q} \in \mathscr{Q}$ satisfies*

$$\sum_{k=1}^{\kappa} \|\mathrm{off}(\widehat{Q}^{-1}\widehat{C}_k\widehat{Q})\|_F^2 = \inf_{Q\in\mathscr{Q}} \left\{ \sum_{k=1}^{\kappa} \|\mathrm{off}(Q^{-1}\widehat{C}_k Q)\|_F^2 \right\} + o_p(1),$$

then $\lim_{n\to\infty} \Pr(\widehat{Q} \in \mathscr{O}) = 1$ for any open subset $\mathscr{O}$ of $\mathscr{Q}$ containing $\mathscr{Q}_0$.

Each $Q \in \mathscr{Q}_0$ has associated with it a permutation matrix Δ and a diagonal matrix Θ as just defined so that $Q = Q_*\Delta\Theta$. Theorem 6 states that (up to a subsequence) we have that $\widehat{Q} \xrightarrow{p} Q_*\Delta_0\Theta_0$ for well-defined Δ_0 and Θ_0. We may then set $Q_0 = Q_*\Delta_0\Theta_0$ in (5.1). It then equally follows that

$$\widehat{D}_k \xrightarrow{p} D_k = \Delta_0' D_k^* \Delta_0,$$

where D_k is as in (5.2) and $D_k^* = Q_*^{-1}C_kQ_*$, both of which are equal up to a permutation. Thus, the consistency of the eigenvalues (up to a joint permutation) essentially follows from the consistency of the estimator of the input matrices C.

To provide distribution theory, let

$$D_{k_1} \ominus D_{k_2} = (D_{k_1} \otimes \mathrm{I}_{\dim D_{k_2}}) - (\mathrm{I}_{\dim D_{k_1}} \otimes D_{k_2})$$

be the Kronecker difference between square matrices D_{k_1} and D_{k_2}. Construct the $r^2 \times r^2\kappa$ matrix

$$T = ((D_1 \ominus D_1), (D_2 \ominus D_2), \ldots, (D_\kappa \ominus D_\kappa))$$

by concatenation and let

$$G = (\mathrm{I}_r \otimes Q_0) \left(\sum_{k=1}^{\kappa} (D_k \ominus D_k)^2 \right)^+ T\ (\mathrm{I}_\kappa \otimes Q_0' \otimes Q_0^{-1}),$$

where Q^+ is the Moore-Penrose pseudo inverse of Q. Theorem 7 contains distribution theory for our estimator of the matrix of joint eigenvectors $\widehat{Q}$ in (5.4).

THEOREM 7 (Asymptotic distribution). *If* $\|\widehat{C} - C\|_F = O_p(n^{-1/2})$, *then*

$$\sqrt{n}\,\mathrm{vec}(\widehat{Q} - Q_0) = G\,\sqrt{n}\,\mathrm{vec}(\widehat{C} - C) + o_p(1)$$

as $n \to \infty$.

If, further, $\sqrt{n}\,\mathrm{vec}(\widehat{C} - C) \xrightarrow{d} \mathcal{N}(0, V)$ for some covariance matrix V, Theorem 7 implies that

$$\sqrt{n}\,\mathrm{vec}(\widehat{Q} - Q_0) \xrightarrow{d} \mathcal{N}(0, G\,V\,G')$$

as $n \to \infty$. In our context, $\sqrt{n}$-consistency and asymptotic normality of the input matrices is not a strong requirement. Indeed, the proof of Theorem 1 showed that the input matrices are of the form $C_k = W_1 A_k W_2'$, where W_1 and W_2 follow from a singular-value decomposition of A_0. An estimator of C_k can thus be constructed using a sample analog of A_0 to estimate W_1 and W_2, together with a sample analog of A_k. If the estimators of A_0 and A_k are $\sqrt{n}$-consistent and asymptotically normal and all non-zero singular values of A_0 are simple, then $\sqrt{n}\,\mathrm{vec}(\widehat{C} - C) \xrightarrow{d} \mathcal{N}(0, V)$ holds. A detailed derivation of V is readily obtained from the argument on the estimation of eigendecompositions of normal matrices in the supplementary material to Bonhomme, Jochmans and Robin [2014, Lemma S.2].

We next present the asymptotic behavior of $\widehat{D} = (\widehat{D}_1, \widehat{D}_2, \ldots, \widehat{D}_\kappa)$, our estimator of the eigenvalues $D = (D_1, D_2, \ldots, D_\kappa)$. To state it, let $\mathrm{S}_r = \mathrm{diag}(\mathrm{vec}\,\mathrm{I}_r)$ be an $r^2 \times r^2$ selection matrix; note that $\mathrm{S}_r\,\mathrm{vec}\,Q = \mathrm{vec}(\mathrm{diag}\,Q)$. Let

$$H = (\mathrm{I}_\kappa \otimes \mathrm{S}_r)\,(\mathrm{I}_\kappa \otimes Q_0' \otimes Q_0^{-1}).$$

Theorem 8 follows.

THEOREM 8 (Asymptotic distribution). *If* $\|\widehat{C} - C\|_F = O_p(n^{-1/2})$, *then*

$$\sqrt{n}\,\mathrm{vec}(\widehat{D} - D) = H\,\sqrt{n}\,\mathrm{vec}(\widehat{C} - C) + o_p(1)$$

as $n \to \infty$.

Again, if $\sqrt{n}\,\mathrm{vec}(\widehat{C} - C) \xrightarrow{d} \mathcal{N}(0, V)$, then

$$\sqrt{n}\,\mathrm{vec}(\widehat{D} - D) \xrightarrow{d} \mathcal{N}(0, H\,V\,H')$$

as $n \to \infty$.

6. Application to finite mixtures of continuous measures. With discrete outcomes, both the finite-mixture model in (2.1) and the hidden Markov model in (2.4) are finite dimensional. Further, the input matrices to be simultaneously diagonalized are contingency tables. These tables can be estimated by simple empirical cell probabilities and are $\sqrt{n}$-consistent and asymptotically normal. Hence, the theory from the previous section can directly be applied to deduce the large-sample behavior of the parameter estimates.

Now consider a mixture model with continuous outcomes. Here we derive convergence rates for an orthogonal-series estimator of the densities in the mixture model based on (2.3). The results readily extend to a hidden Markov model with continuous emission distributions.

The q-variate finite-mixture model with r latent classes implies that the joint density of the outcomes $Y_1, Y_2, \ldots, Y_q$ factors as

$$\sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} f_{ij}$$

for mixing proportions π_j and conditional densities f_{ij}. With the density supported on the compact set $[-1,1]^q$ and the f_{ij} square-integrable with respect to a weight function ρ on $[-1,1]$, the orthogonal-series approximation

$$\mathrm{Proj}_{\kappa_i} f_{ij} = \varphi'_{\kappa_i} b_{ij}$$

yields the multilinear restrictions

$$\mathbb{B} = E[\bigotimes_{i=1}^{q} \varphi_{\kappa_i}(Y_i)\rho(Y_i)] = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} E[\varphi_{\kappa_i}(Y_i)\rho(Y_i)|Z=j] = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} b_{ij},$$

where, recall, φ_{κ_i} is the vector containing the κ_i leading polynomials from $\{\phi_k, k>0\}$.

The entries of the array $\mathbb{B}$ can be estimated as simple sample averages over the orthogonal polynomials, weighted by ρ. Moreover, its sample analog is

$$\widehat{\mathbb{B}} \equiv n^{-1} \sum_{m=1}^{n} \bigotimes_{i=1}^{q} \varphi_{\kappa_i}(Y_{im})\rho(Y_{im}), \tag{6.1}$$

where $\{Y_{1m}, Y_{2m}, \ldots, Y_{qm}\}_{m=1}^{n}$ is a size-n sample drawn at random from the mixture model. Simultaneous diagonalization of $\widehat{\mathbb{B}}$ as in (3.3) above

then yields estimators $\widehat{b}_{ij}$ of the Fourier coefficients, and the nonparametric density estimator

$$\widehat{f}_{ij} = \varphi'_{\kappa_i} \widehat{b}_{ij}. \tag{6.2}$$

This nonparametric problem is not directly covered by the arguments from the previous section. Consistency of the estimator in (6.2) requires that $\kappa_i \to \infty$ as $n \to \infty$. The remaining $\kappa_{i'}$ $(i' \neq i)$ only show up in the estimator of the Fourier coefficients, via the dimension of the q-way array in (6.1). This means that, to estimate the Fourier coefficients of f_{ij}, we solve a joint diagonalization system where the number of matrices to be diagonalized, κ_i, diverges, but the size of the matrices remains fixed. That is, in the notation of the proof of Theorem 1, the size of each matrix A_k is fixed but the number of such matrices diverges with the sample size. So, throughout this section, $\kappa_{i'}$ for $i' \neq i$ is treated as fixed.

Under mild regularity conditions, the estimator in (6.2) exhibits standard large-sample behavior.

Let $\|\cdot\|_\infty$ denote the supremum norm.

ASSUMPTION 1 (Regularity). *The sequence $\{\phi_k, k > 0\}$ is dominated by a function ψ that is continuous on $(-1, 1)$ and positive almost everywhere on $[-1, 1]$. ρ, $\psi\rho$, and $\psi^2\rho$ are integrable. There exists a sequence of constants $\{\zeta_\kappa, \kappa > 0\}$ so that $\|\sqrt{\varphi'_\kappa \varphi_\kappa}\|_\infty \leq \zeta_\kappa$.*

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 $\psi(y) \propto (1-y)^{-\vartheta'_1}(1+y)^{-\vartheta'_2}$, where $\vartheta'_i \equiv \max\{\vartheta_i, -1/2\}/2 + 1/4$. Further, with $\vartheta \equiv 1/2 + \max\{\vartheta_1, \vartheta_2, -1/2\}$, $\|\phi_k\|_\infty = k^\vartheta$, and so one can take $\zeta_\kappa = \kappa^{(1+\vartheta)/2}$; 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 2 (Smoothness). *The f_{ij} are continuous, the $(\psi\rho)^2 f_{ij}$ are integrable, and $\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\|_\infty = O(\kappa_i^{-\beta})$ for some constant $\beta \geq 1$.*

Convergence in L^2_ρ-norm implies that $\lim_{\kappa_i \to \infty} \|b_{ij}\|_F$ is finite, and so that the Fourier coefficient associated with ϕ_k shrinks to zero as $k \to \infty$. The constant β in Assumption 2 is a measure of how fast the Fourier coefficients

shrink. In general, β is larger the smoother the underlying function that is being approximated.

Under these conditions we obtain integrated squared-error and uniform convergence rates.

THEOREM 9 (Convergence rates). *Let Assumptions 1–2 hold. Then*

$$\|\widehat{f}_{ij} - f_{ij}\|_2^2 = O_p(\kappa_i/n + \kappa_i^{-2\beta}), \quad \|\widehat{f}_{ij} - f_{ij}\|_\infty = O_p(\zeta_{\kappa_i}\sqrt{\kappa_i/n} + \kappa_i^{-\beta}),$$

for all i, j.

The rates in Theorem 9 equal the conventional univariate rates of series estimators; see, e.g., Schwartz [1967] and Newey [1997]. Thus, the fact that Z 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]. It may be possible to improve on the uniform-convergence rate using recent results by Belloni et al. [2013] but we do not pursue such a refinement here for reasons of conciseness. Indeed, our series estimator differs from the standard estimator only in the way the Fourier coefficients are estimated. Once we have shown that $\|\widehat{b}_{ij} - b_{ij}\|_F = O_p(\sqrt{\kappa_i/n})$, standard series arguments can be applied in the usual way.

The orthogonal-series estimator requires choosing the number of Fourier coefficients—$\kappa_1, \kappa_2 \ldots, \kappa_q$—to include in (6.1). This is a more complicated problem than in the conventional univariate context, where one only needs to select κ_i to estimate f_{ij}; see Diggle and Hall [1986], among others. It would be interesting to see how these approaches can be modified to the current setting.

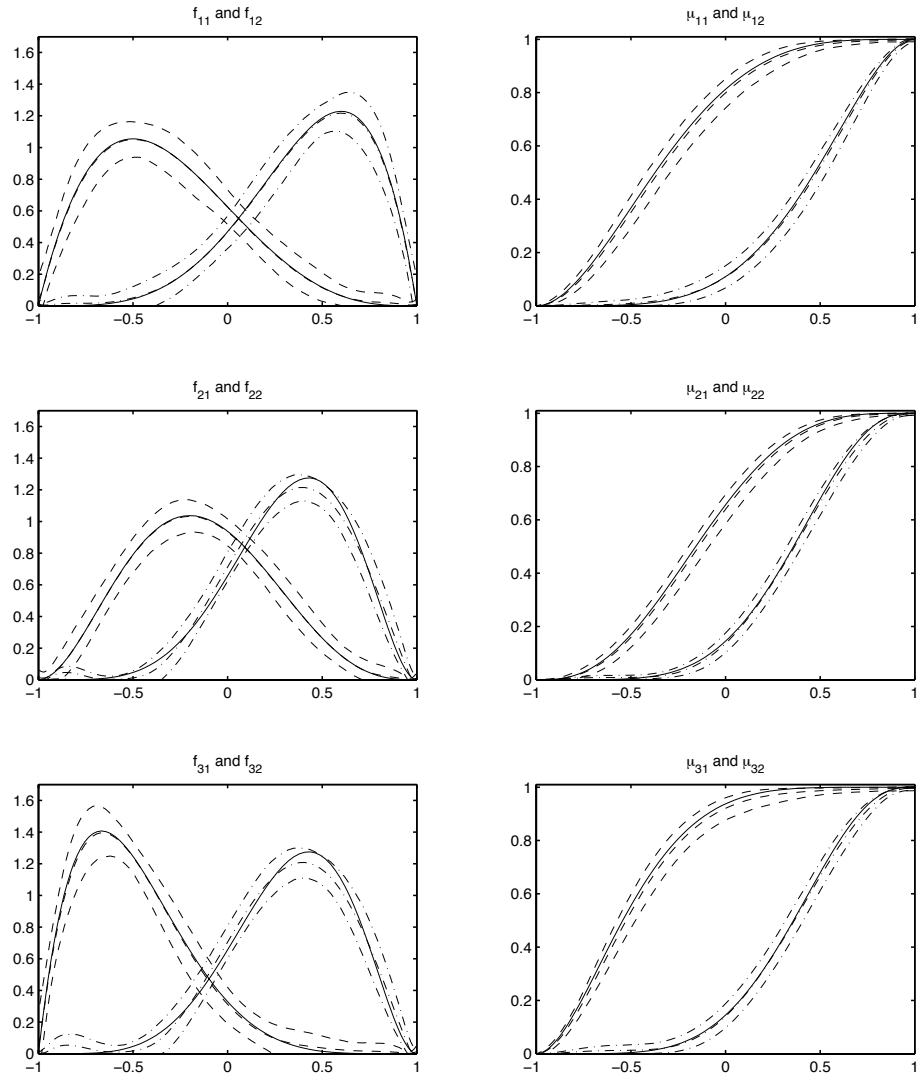
7. Monte Carlo illustration. The orthogonal-series estimator was applied to simulated data from a three-variate two-component mixture of beta distributions on the interval $[-1, 1]$. The family of Beta densities on $[-1, 1]$ has members

$$f(y; \vartheta_1, \vartheta_2) = \frac{1}{2^{\vartheta_1+\vartheta_2-1}} \frac{1}{\mathsf{B}(\vartheta_1, \vartheta_2)} (1+y)^{\vartheta_1-1}(1-y)^{\vartheta_2-1},$$

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. We generated $n = 500$ observations from a mixture with components

$$\begin{aligned} f_{11}(y) &= f(y; 5, 2), \quad f_{12}(y) = f(y; 2, 4), \\ f_{21}(y) &= f(y; 6, 3), \quad f_{22}(y) = f(y; 3, 4), \\ f_{31}(y) &= f(y; 6, 3), \quad f_{32}(y) = f(y; 2, 6), \end{aligned}$$

FIG 1. *Component densities and distributions*



and mixing proportions $\pi_1 = \pi_2 = 1/2$. For each i, f_{i1} is skewed to the left and f_{i2} is skewed to the right but all component densities vary with i. We estimated the densities by means of our joint approximate-diagonalization estimator using the five leading Chebychev polynomials of the first kind as basis functions for each i on data sets of size $n = 500$. Experimentation with different choices yielded similar results. After estimating the component densities we used Clenshaw-Curtis quadrature to construct an estimator of $\mu_{ij}(y) = \int_{-1}^{y} f_{ij}(z)\,\mathrm{d}z$.

The solid lines in the left plots of Figure 1 represent the densities f_{ij}. The solid lines in the right plots are the corresponding distribution functions μ_{ij}. In each plot, dashed lines are given for the mean and for the upper and lower envelopes of $1,000$ replications of our estimation procedure. The plots show our approach is effective at recovering the component densities as well as the component distributions.

Acknowledgements. Previous versions of this paper circulated under the title 'Nonparametric spectral-based estimation of latent structures'. We thank Xiaohong Chen, Marc Henry, Laurent Albera, and Xavier Luciani. Bonhomme was supported by the European Research Council through grant ERC-2010-StG-0263107-ENMUH. Jochmans was supported by Sciences Po's Scientific Advisory Board. Robin was supported by the Economic and Social Research Council through the Centre for Microdata Methods and Practice grant RES-589-28-0001, and by the European Research Council through the grant ERC-2010-AdG-269693-WASP.

References.

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

Allman, E. S., Matias, C. and Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference* **141** 1719–1736.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal Of Machine Learning Research* **15** 2773–2832.

Anderson, T. W. (1954). On estimation of parameters in latent structure analysis. *Psychometrika* **19** 1–10.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). On the asymptotic theory for least squares series: Pointwise and uniform results. CeMMAP Working Paper 73/13.

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

Bonhomme, S., Jochmans, K. and Robin, J. M. (2014). Nonparametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society, Series B*, forthcoming.

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

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models. Springer Series in Statistics*. Springer.

Carroll, J. D. and Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika* **35** 283–319.

Comon, P. and De Lathauwer, L. (2010). Algebraic identification of under-determined mixtures. In *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (P. Comon and C. Jutten, eds.) 9 325–365. Academic Press.

Comon, P. and Jutten, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press.

De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications* **28** 642-666.

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

De Lathauwer, L. and Nion, D. (2008). Decompositions of a higher-order tensor in block terms - Part III: Alternating least squares algorithms. *SIAM journal on Matrix Analysis and Applications* **30** 1067-1083.

Derksen, H. (2013). Kruskal's uniqueness inequality is sharp. *Linear Algebra and its Applications* **438** 708–712.

Diggle, P. J. and Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association* **81** 230–233.

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

Fu, T. and Gao, X. Q. (2006). Simultaneous diagonalization with similarity transformation for non-defective matrices. *Proceedings of the IEEE ICA SSP 2006* **4** 1137–1140.

Gassiat, E., Cleynen, A. and Robin, S. (2013). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing*, forthcoming.

Gassiat, E. and Rousseau, J. (2014). Non parametric finite translation mixtures and extensions. *Bernoulli*, forthcoming.

Green, B. (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika* **16** 151–166.

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

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

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis. UCLA Working Papers in Phonetics 16:1–84.

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

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

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

IFERROUDJENE, R., ABED-MERAIM, K. and BELOUCHRANI, A. (2009). A new Jacobi-like method for joint diagonalization of arbitrary non-defective matrices. *Applied Mathematics and Computation* **211** 363–373.

IFERROUDJENE, R., ABED-MERAIM, K. and BELOUCHRANI, A. (2010). Joint diagonalization of non defective matrices using generalized Jacobi rotations. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on* 345-348.

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

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

KRUSKAL, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41** 281–293.

KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra and its Applications* **18** 95–138.

LEVINE, M., HUNTER, D. R. and CHAUVEAU, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416.

LUCIANI, X. and ALBERA, L. (2010). Joint eigenvalue decomposition using polar matrix factorization. In *Latent Variable Analysis and Signal Separation. Lecture Notes in Computer Sciences* **6365** 555–562. Springer.

LUCIANI, X. and ALBERA, L. (2014). Canonical polyadic decomposition based on joint eigenvalue decomposition. *Chemometrics and Intelligent Laboratory Systems* **132** 152 - 167.

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

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

NEWEY, W. K. and MCFADDEN, D. L. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, **4** 36 2111–2245. Elsevier.

PETRIE, T. (1969). Probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* **40** 97–115.

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

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics* **39** 1878–1915.

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

SIDIROPOULOS, N. D. and BRO, R. (2000). On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics* **14** 229–239.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14** 75–100.

SORENSEN, M., DE LATHAUWER, L., COMON, P., ICART, S. and DENEIRE, L. (2013). Canonical polyadic decomposition with a columnwise orthonormal factor matrix. *SIAM Journal on Matrix Analysis and Applications* **33** 1190–1213.

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

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.

APPENDIX A: TECHNICAL PROOFS

PROOF OF THEOREM 6. For generic $r\times r\kappa$ matrix $M=(M_1,M_2,\ldots,M_\kappa)$ denote the objective function as

$$L(Q,M)=\sum_{k=1}^{\kappa}\|\mathrm{off}(Q^{-1}M_kQ)\|_F^2.$$

Write $Q_{-i,-j}$ for the principal minors of Q. Because for any $Q\in\mathscr{Q}$ we have

$$[Q^{-1}]_{i,j}=(-1)^{i+j}\det Q_{-i,-j}$$

and $\det A$ is a polynomial function of A, the function $L(Q,M)$ is continuous in each of its arguments. Let

$$L_0(Q)=L(Q,C),\qquad L_n(Q)=L(Q,\widehat{C}).$$

Note that $\mathscr{Q}_0=\arg\inf_{Q\in\mathscr{Q}}L_0(Q)$ is the equivalence class containing all $Q\in\mathscr{Q}$ that are equal to Q_0 up to a permutation and direction of their columns.

Because $\|\widehat{C}-C\|_F=o_p(1)$ and $L(Q,M)$ is continuous in M, for all $Q\in\mathscr{Q}$,

$$L_n(Q)\xrightarrow{p}L_0(Q)$$

by the continuous-mapping theorem. Further, by the same argument, for all $Q,Q'\in\mathscr{Q}$,

$$|L_n(Q)-L_n(Q')|\leq O_p(1)\,\|Q-Q'\|_F,$$

because $L_n(Q)$ is polynomial in Q, and thus Lipschitz continuous. With $\mathscr{Q}$ compact it follows that $L_n(Q)$ is stochastically equicontinuous (see, for example, Newey and McFadden 1994, Lemma 2.9), and so

$$\sup_{Q\in\mathscr{Q}}|L_n(Q)-L_0(Q)|=o_p(1).$$

Then, for any open subset $\mathscr{O}$ of $\mathscr{Q}$ containing $\mathscr{Q}_0$, with complement $\mathscr{O}^c$, it holds that

$$L_0(\widehat{Q})<\inf_{Q\in\mathscr{O}^c}L_0(Q)$$

with probability approaching one. Hence, we have $\lim_{n\to\infty}\Pr(\widehat{Q}\in\mathscr{O})=1$ (Newey and McFadden 1994, Theorem 2.1). □

PROOF OF THEOREM 7. For the proof of Theorem 7 it is convenient to work with a different yet equivalent normalization on Q_0. More precisely, the set

$$\mathscr{Q} = \{Q : \det Q = 1, \|q_j\|_F = c \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m < \infty\}$$

is in a simple one-to-one correspondence with the set

$$\mathscr{Q}' = \{Q : \det Q = c^{-r}, \|q_j\|_F = 1 \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m < \infty\}.$$

The latter is easier to work with here because constraining columns to have unit norm is easier than requiring the determinant of the matrix to equal unity. In any case, the asymptotic distribution of $\widehat{Q}$ turns out not to depend on the choice of column norm.

We first derive the first-order conditions to the constrained minimization problem that defines $\widehat{Q}$. Given these, we can then proceed using standard arguments to derive the asymptotic distribution of the joint approximate diagonalizer.

Lagrangian and first-order conditions. It is useful to reformulate the joint approximate-diagonalization problem as

$$\min_{Q,R} \sum_{k=1}^{\kappa} \left\| \operatorname{off}(R\,\widehat{C}_k Q) \right\|_F^2, \qquad \text{s.t. } RQ = \mathrm{I}_r, \quad \|q_j\|_F = 1 \ \forall j.$$

With $[Q]_{i,j}$ and $[R]_{i,j}$ denoting the (i,j)th entries of matrices Q and R, respectively, the Lagrangian for this constrained minimization problem with respect to (Q, R) is

$$\begin{aligned} L(Q,R) = &\sum_{k=1}^{\kappa} \left\| \operatorname{off}(R\,\widehat{C}_k Q) \right\|_F^2 \\ &+ \sum_{i,j=1}^{r} \lambda_{ij} \left(\sum_{\ell=1}^{r} [R]_{i,\ell}[Q]_{\ell,j} - \delta_{ij} \right) + \sum_{j=1}^{r} \gamma_j (q_j' q_j - 1), \end{aligned}$$

for Lagrange multipliers $[\Lambda]_{i,j} = \lambda_{ij}$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_r)'$ associated with each of the constraints, and δ_{ij} denoting Kronecker's delta.

Application of the chain rule readily gives

$$\frac{\partial L(Q,R)}{\partial Q} = 2 \sum_{k=1}^{\kappa} \widehat{C}_k' R' \operatorname{off}(R\,\widehat{C}_k Q) + R'\Lambda + 2Q\Gamma, \tag{A.1}$$

$$\frac{\partial L(Q,R)}{\partial R} = 2 \sum_{k=1}^{\kappa} \operatorname{off}(R\,\widehat{C}_k Q)\, Q' \widehat{C}_k' + \Lambda Q', \tag{A.2}$$

with $\Gamma = \operatorname{diag}(\gamma)$. Substitute $R = Q^{-1}$ in (A.2) and solve $\frac{\partial L(Q,R)}{\partial R} = 0$ for Λ to get

$$\Lambda = -2\sum_{k=1}^{\kappa} \operatorname{off}(Q^{-1}\widehat{C}_k Q)\,(Q^{-1}\widehat{C}_k Q)'.$$

Next, substitute this value for Λ and $R = Q^{-1}$ in (A.1) and premultiply with Q' to get

$$\sum_{k=1}^{\kappa} (Q^{-1}\widehat{C}_k Q)' \operatorname{off}(Q^{-1}\widehat{C}_k Q) - \operatorname{off}(Q^{-1}\widehat{C}_k Q)(Q^{-1}\widehat{C}_k Q)' + Q'Q\Gamma = 0.$$

Force the columns of Q to have unit Euclidean norm, so that $\operatorname{diag}(Q'Q) = \mathrm{I}_r$, to see that

$$\Gamma = -\operatorname{diag}\left(\sum_{k=1}^{\kappa} (Q^{-1}\widehat{C}_k Q)' \operatorname{off}(Q^{-1}\widehat{C}_k Q) - \operatorname{off}(Q^{-1}\widehat{C}_k Q)(Q^{-1}\widehat{C}_k Q)'\right),$$

as $\Gamma = \operatorname{diag}(Q'Q\Gamma)$ because Γ is diagonal. Then the first-order condition for Q of our constrained minimization problem is obtained on plugging this expression back in to (A.1). To write it compactly, let

$$\Delta(M) = M' \operatorname{off}(M) - \operatorname{off}(M)M'$$

for any matrix M and

$$S(Q, M) = \sum_{k=1}^{\kappa} (Q')^{-1}\Delta(Q^{-1}M_k Q) - Q \operatorname{diag}(\Delta(Q^{-1}M_k Q))$$

for any $M = (M_1, M_2, \ldots, M_\kappa)$. Then

$$S(Q, \widehat{C}) = 0$$

is the score equation defining $\widehat{Q}$.

Expansion of first-order conditions. With $S(Q, M)$ polynomial in each of its arguments, an expansion around Q_0 and C gives

$$\frac{dS(Q_0, M)}{dM}\bigg|_{M=C} \operatorname{vec}(\widehat{M} - M) + \frac{dS(Q, C)}{dQ}\bigg|_{Q=Q_0} \operatorname{vec}(\widehat{Q} - Q_0) = o_p(n^{-1/2}), \tag{A.3}$$

where

$$\frac{dS(Q, M)}{dM} = \frac{\partial \operatorname{vec} S(Q, M)}{\partial \operatorname{vec}(M)'}, \qquad \frac{dS(Q, M)}{dQ} = \frac{\partial \operatorname{vec} S(Q, M)}{\partial \operatorname{vec}(Q)'}.$$

To derive the asymptotic distribution of $\widehat{Q}$ we need to calculate both these derivatives, and evaluate at true values.

Start with the derivative with respect to Q. First observe that

$$\frac{dQ^{-1}MQ}{dQ} = \left(\mathrm{I}_r \otimes Q^{-1}M\right) - \left((Q^{-1}MQ)' \otimes Q^{-1}\right). \tag{A.4}$$

Furthermore, $\operatorname{vec}(\operatorname{off} M) = \operatorname{vec}(M - \operatorname{diag} M) = (\mathrm{I}_{r^2} - \mathrm{S}_r)\operatorname{vec}(M)$ and, by an application of the chain rule,

$$\begin{aligned}\frac{d\Delta(Q^{-1}MQ)}{dQ} &= \left\{\operatorname{off}(Q^{-1}MQ)' \ominus \operatorname{off}(Q^{-1}MQ)\right\} \frac{dQ^{-1}MQ}{dQ} \\ &\quad - \left\{(Q^{-1}MQ) \ominus (Q^{-1}MQ)'\right\} \{\mathrm{I}_{r^2} - \mathrm{S}_r\} \frac{dQ^{-1}MQ}{dQ}.\end{aligned} \tag{A.5}$$

Therefore, combining (A.4) and (A.5), and using that $D_k = Q_0^{-1}C_kQ_0$ and $\operatorname{off}(D_k) = 0$, we have

$$\begin{aligned}\left.\frac{d\Delta(Q^{-1}C_kQ)}{dQ}\right|_{Q=Q_0} &= (D_k \ominus D_k)\,(\mathrm{I}_{r^2} - \mathrm{S}_r)\,(D_k \ominus D_k)\left(\mathrm{I}_r \otimes Q_0^{-1}\right) \\ &= (D_k \ominus D_k)^2\left(\mathrm{I}_r \otimes Q_0^{-1}\right)\end{aligned}$$

for all k, where the last transition follows from the fact that $\mathrm{S}_r(D_k \ominus D_k) = 0$ because S_r selects only the $\{(ir+(i+1), ir+(i+1))\}_{i=0}^{r-1}$ entries of the $r^2 \times r^2$ matrix $D_k \ominus D_k$, and these are equal to zero. Then

$$\left.\frac{dS(Q,C)}{dQ}\right|_{Q=Q_0} = \left(\mathrm{I}_r \otimes Q_0^{-1}\right)' \left\{\sum_{k=1}^{\kappa} (D_k \ominus D_k)^2\right\} \left(\mathrm{I}_r \otimes Q_0^{-1}\right) \tag{A.6}$$

follows readily.

Now turn to the derivative with respect to M. Proceeding in the same way as before, now using that

$$\frac{dQ^{-1}MQ}{dM} = Q' \otimes Q^{-1},$$

we obtain

$$\left.\frac{d\Delta(Q_0^{-1}MQ_0)}{dM}\right|_{M=C_k} = -(D_k \ominus D_k)\,(Q_0' \otimes Q_0^{-1}),$$

which yields

$$\left.\frac{dS(Q_0,M)}{dM_k}\right|_{M_k=C_k} = -(\mathrm{I}_r \otimes Q_0^{-1})'(D_k \ominus D_k)(Q_0' \otimes \mathrm{I}_r)(\mathrm{I}_r \otimes Q_0^{-1})$$

for each k, and so concatenating these matrices finally gives

$$\left.\frac{dS(Q_0, M)}{dM}\right|_{M=C} = -(\mathrm{I}_r \otimes Q_0^{-1})' \, T \, (\mathrm{I}_\kappa \otimes Q_0' \otimes Q_0^{-1}). \tag{A.7}$$

Combining (A.3) with (A.6) and (A.7) then yields

$$\operatorname{vec}(\widehat{Q} - Q_0) = G \operatorname{vec}(\widehat{C} - C) + o_p(n^{-1/2}),$$

with matrix G as defined in the main text. This completes the proof of the theorem. □

Proof of Theorem 8. Because we have $\|\widehat{C}_k - C_k\| = O_p(n^{-1/2})$ and $\|\widehat{Q} - Q_0\| = O_p(n^{-1/2})$, a linearization of

$$\widehat{D}_k - D_k = \operatorname{diag}(\widehat{Q}^{-1}\widehat{C}_k\widehat{Q} - Q_0^{-1}C_kQ_0)$$

up to $o_p(n^{-1/2})$ will yield the result. Moreover, the term inside the diagonal operator equals

$$(\widehat{Q} - Q_0)^{-1}C_kQ_0 + Q_0^{-1}(\widehat{C}_k - C_k)Q_0 + Q_0^{-1}C_k(\widehat{Q} - Q_0) + o_p(n^{-1/2}).$$

Because matrix inversion is a continuous transformation, the delta method can further be applied to yield

$$\operatorname{vec}\left((\widehat{Q} - Q_0)^{-1}C_kQ_0\right) = -\left(D_k \otimes Q_0^{-1}\right) \operatorname{vec}\left(\widehat{Q} - Q_0\right) + o_p(n^{-1/2}).$$

The remaining right-hand side terms are already linear in the estimators $\widehat{Q}$ and $\widehat{C}_k$. Then, using that $Q_0^{-1}C_k = D_kQ_0^{-1}$, $\operatorname{vec}(\widehat{Q}^{-1}\widehat{C}_k\,\widehat{Q} - Q_0^{-1}C_kQ_0)$ equals

$$(Q_0' \otimes Q_0^{-1}) \operatorname{vec}(\widehat{C}_k - C_k) - (D_k \ominus D_k)(\mathrm{I}_r \otimes Q_0^{-1}) \operatorname{vec}(\widehat{Q} - Q_0) + o_p(n^{-1/2}).$$

Now, $\operatorname{vec}(\widehat{D}_k - D_k) = \mathrm{S}_r \operatorname{vec}(\widehat{Q}^{-1}\widehat{C}_k\,\widehat{Q} - Q_0^{-1}C_kQ_0)$, implying asymptotic linearity of the estimated eigenvalues for each k. Further, concatenating the influence functions gives

$$\operatorname{vec}(\widehat{D} - D) = \left(\mathrm{I}_\kappa \otimes (\mathrm{S}_r(Q_0' \otimes Q_0^{-1}))\right) \operatorname{vec}(\widehat{C} - C) + o_p(n^{-1/2}).$$

This proves the theorem because

$$\mathrm{I}_\kappa \otimes (\mathrm{S}_r(Q_0' \otimes Q_0^{-1})) = (\mathrm{I}_\kappa \otimes \mathrm{S}_r)\,(\mathrm{I}_\kappa \otimes Q_0' \otimes Q_0^{-1}) = H,$$

as claimed. □

PROOF OF THEOREM 9. For the proof it suffices to consider the case with $q = 3$. Without loss of generality we fix $i = 3$ throughout. As in the proof to Theorem 1,

$$A_0 = B_1 \Pi B_2', \qquad A_k = B_1 \Pi D_k B_2', \qquad D_k = \operatorname{diag}_k B_3.$$

The Fourier coefficients are then estimated by solving the sample version of

$$C_k = W_1 A_k W_2' = Q D_k Q^{-1}.$$

The proof consists of two steps. We first derive integrated squared-error and uniform convergence rates for the infeasible estimator that assumes that the matrices Q and W_1, W_2 are observable without noise. That is, for the estimator

$$\widetilde{f}_{ij} = \varphi'_{\kappa_i} \widetilde{b}_{ij},$$

where the $\widetilde{b}_{ij}$ are constructed from $\widetilde{D}_k = \operatorname{diag}[(Q^{-1}W_1)\, \widehat{A}_k\, (W_2'Q)]$. We then show that the additional noise in the feasible estimator

$$\widehat{f}_{ij} = \varphi'_{\kappa_i} \widehat{b}_{ij},$$

that is, the one that uses $\widehat{D}_k = \operatorname{diag}[(\widehat{Q}^{-1}\widehat{W}_1)\, \widehat{A}_k\, (\widehat{W}_2'\widehat{Q})]$, is asymptotically negligible.

We begin by showing that $\|\widetilde{b}_{ij} - b_{ij}\|_F = O_p(\sqrt{\kappa_i/n})$. The convergence rates for $\widetilde{f}_{ij}$ will then follow easily. Write $a_{k_1k_2k}$ for the (k_1, k_2)th entry of A_k and let $\widehat{a}_{k_1k_2k}$ be its estimator. Note that

$$\widehat{a}_{k_1k_2k} = \frac{1}{n} \sum_{m=1}^{n} \phi_{k_1}(Y_{1m}) \rho(Y_{1m}) \phi_{k_2}(Y_{2m}) \rho(Y_{2m}) \phi_k(Y_{3m}) \rho(Y_{3m})$$

is an unbiased estimator of $a_{k_1k_2k}$. Hence, for any k,

$$\begin{aligned} E\big\|\widehat{A}_k - A_k\big\|_F^2 &= \sum_{k_1=1}^{\kappa_1} \sum_{k_2=1}^{\kappa_2} E\big[(\widehat{a}_{k_1k_2k} - a_{k_1k_2k})^2\big] \\ &= \sum_{k_1=1}^{\kappa_1} \sum_{k_2=1}^{\kappa_2} \frac{E\big[\phi_{k_1}(Y_1)^2 \rho(Y_1)^2 \phi_{k_2}(Y_2)^2 \rho(Y_2)^2 \phi_k(Y_3)^2 \rho(Y_3)^2\big] - a_{k_1k_2k}^2}{n} \\ &\le \sum_{k_1=1}^{\kappa_1} \sum_{k_2=1}^{\kappa_2} \frac{\prod_{i'=1}^{q} \sum_{j=1}^{r} \pi_j \left(\int_{-1}^{1} \psi(y)^2 \rho(y)^2 f_{i'j}(y)\, \mathrm{d}y \right) - a_{k_1k_2k}^2}{n}. \end{aligned}$$

As the $\psi^2\rho^2 f_{i'j}$ are integrable and the Fourier coefficients $a_{k_1k_2k_3}$ are square summable, we have that $E\|\widehat{A}_k - A_k\|_F^2 = O(1/n)$ uniformly in k. Hence, $\sum_{k=1}^{\kappa_i}\|\widehat{A}_k - A_k\|_F^2 = O_p(\kappa_i/n)$ follows from Markov's inequality, and also

$$\begin{aligned}\|\widetilde{b}_{ij} - b_{ij}\|_F^2 &\leq \sum_{k=1}^{\kappa_i} \big\|\widetilde{D}_k - D_k\big\|_F^2 \\ &\leq \big\|Q^{-1}W_1 \otimes Q'W_2\big\|_F^2 \sum_{k=1}^{\kappa_i} \big\|\widehat{A}_k - A_k\big\|_F^2 = O_p(\kappa_i/n)\end{aligned}$$

follows by the Cauchy-Schwarz inequality. This establishes the rate result on the Fourier coefficients sought for. Now turn to the convergence rates for $\widetilde{f}_{ij}$. By orthonormality of the ϕ_k,

$$\begin{aligned}\big\|\widetilde{f}_{ij} - f_{ij}\big\|_2^2 &= \big\|\widetilde{f}_{ij} - \mathrm{Proj}_{\kappa_i} f_{ij}\big\|_2^2 + \big\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\big\|_2^2 \\ &= \|\widetilde{b}_{ij} - b_{ij}\|_F^2 + \big\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\big\|_2^2.\end{aligned}$$

The first right-hand side term is known to be $O_p(\kappa_i/n)$ from above. For the second right-hand side term, by Assumption 2,

$$\big\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\big\|_2^2 \leq \int_{-1}^{1} \big\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\big\|_\infty^2 \,\rho(y)\,\mathrm{d}y = O(\kappa_i^{-2\beta})$$

because ρ is integrable. This establishes the integrated squared-error rate for $\widetilde{f}_{ij}$. To obtain the uniform convergence rate, use the triangle inequality to see that

$$\big\|\widetilde{f}_{ij} - f_{ij}\big\|_\infty \leq \big\|\widetilde{f}_{ij} - \mathrm{Proj}_{\kappa_i} f_{ij}\big\|_\infty + \big\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\big\|_\infty.$$

By the Cauchy-Schwarz inequality in the first step and by the uniform bound on the norm of the basis functions and the convergence rate of $\|\widetilde{b}_{ij} - b_{ij}\|_F$ in the second, the first right-hand side term satisfies

$$\big\|\widetilde{f}_{ij} - \mathrm{Proj}_{\kappa_i} f_{ij}\big\|_\infty \leq \big\|\sqrt{\varphi'_{\kappa_i}\varphi_{\kappa_i}}\big\|_\infty \,\|\widetilde{b}_{ij} - b_{ij}\|_F = O\big(\zeta_{\kappa_i}\big)\; O_p\big(\sqrt{\kappa_i/n}\big).$$

By Assumption 2, $\|\mathrm{Proj}_{\kappa_i} f_{ij} - f_{ij}\|_\infty = O(\kappa_i{}^{-\beta})$. This yields the uniform convergence rate.

To extend the results to the feasible density estimator we first show that the presence of estimation noise in Q and (W_1, W_2) implies that

$$\|\widehat{b}_{ij} - \widetilde{b}_{ij}\|_F = O_p(n^{-1/2}) + O_p(\sqrt{\kappa_i}/n). \tag{A.8}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\|\widehat{b}_{ij} - \widetilde{b}_{ij}\|_F^2 &\leq \sum_{k=1}^{\kappa_i} \|\widehat{D}_k - \widetilde{D}_k\|_F^2 \\ &\leq \|\widehat{Q}^{-1}\widehat{W}_1 \otimes \widehat{Q}'\widehat{W}_2 - Q^{-1}W_1 \otimes Q'W_2\|_F^2 \sum_{k=1}^{\kappa_i} \|\widehat{A}_k\|_F^2.\end{aligned}$$

Because both $\widehat{Q}$ and $(\widehat{W}_1, \widehat{W}_2)$ are $\sqrt{n}$-consistent,

$$\|\widehat{Q}^{-1}\widehat{W}_1 \otimes \widehat{Q}'\widehat{W}_2 - Q^{-1}W_1 \otimes Q'W_2\|_F^2 = O_p(1/n).$$

Also, from above, we have that

$$\sum_{k=1}^{\kappa_i} \|\widehat{A}_k\|_F^2 \leq 2\sum_{k=1}^{\kappa_i} \|A_k\|_F^2 + 2\sum_{k=1}^{\kappa_i} \|\widehat{A}_k - A_k\|_F^2 = O(1) + O_p(\kappa_i/n).$$

Together, these results imply (A.8). Next,

$$\|\widehat{f}_{ij} - f_{ij}\|_2^2 \leq 2\|\widehat{b}_{ij} - \widetilde{b}_{ij}\|_F^2 + 2\|\widetilde{f}_{ij} - f_{ij}\|_2^2.$$

From above, the first right-hand side term is $O_p(1/n) + O_p(\kappa_i/n^2)$ while the second right-hand side term is $O_p(\kappa_i/n + \kappa_i^{-2\beta})$. Therefore, the difference between $\widehat{b}_{ij}$ and $\widetilde{b}_{ij}$ has an asymptotically-negligible impact on the density estimator, and

$$\|\widehat{f}_{ij} - f_{ij}\|_2^2 = O_p(\kappa_i/n + \kappa_i^{-2\beta}).$$

For the uniform convergence, similarly, the triangle inequality gives the bound

$$\|\widehat{f}_{ij} - f_{ij}\|_\infty \leq \|\widehat{f}_{ij} - \widetilde{f}_{ij}\|_\infty + \|\widetilde{f}_{ij} - f_{ij}\|_\infty.$$

Again,

$$\|\widehat{f}_{ij} - \widetilde{f}_{ij}\|_\infty \leq \|\sqrt{\varphi'_{\kappa_i}\varphi_{\kappa_i}}\|_\infty \|\widehat{b}_{ij} - \widetilde{b}_{ij}\|_F = O_p(\zeta_{\kappa_i}/\sqrt{n}) + O_p(\zeta_{\kappa_i}\sqrt{\kappa_i}/n),$$

which is of a smaller stochastic order than is $\|\widetilde{f}_{ij} - f_{ij}\|_\infty$. This concludes the proof. □

University of Chicago
1126 E. 59th Street
Chicago, IL 60637
U.S.A.
E-mail: sbonhomme@uchicago.edu

Sciences Po
28 rue des Saints Pères
75007 Paris
France
E-mail: koen.jochmans@sciencespo.fr

Sciences Po
28 rue des Saints Pères
75007 Paris
France
E-mail: jeanmarc.robin@sciencespo.fr