Bonhomme, Stéphane; Jochmans, Koen; Robin, Jean-Marc

Working Paper

Nonparametric spectral-based estimation of latent structures

cemmap working paper, No. CWP18/14

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Bonhomme, Stéphane; Jochmans, Koen; Robin, Jean-Marc (2014) : Nonparametric spectral-based estimation of latent structures, cemmap working paper, No. CWP18/14, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2014.1814

This Version is available at:
https://hdl.handle.net/10419/97371

cemmap
centre for microdata methods and practice

Nonparametric spectral-based estimation of latent structures

Stéphane Bonhomme
Koen Jochmans
Jean-Marc Robin

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP18/14

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

NONPARAMETRIC SPECTRAL-BASED ESTIMATION OF LATENT STRUCTURES

By Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin

CEMFI and University of Chicago, Sciences Po, and Sciences Po and UCL

First version: January 2013. This version: March 6, 2014

We present a constructive identification proof of p-linear decompositions of q-way arrays. The analysis is based on the joint spectral decomposition of a set of matrices. It has applications in the analysis of a variety of latent-structure models, such as q-variate mixtures of p distributions. As such, our results provide a constructive alternative to Allman, Matias and Rhodes [2009]. The identification argument suggests a joint approximate-diagonalization estimator that is easy to implement and whose asymptotic properties we derive. We illustrate the usefulness of our approach by applying it to nonparametrically estimate multivariate finite-mixture models and hidden Markov models.

1. Introduction. Longitudinal data are since long known to be a powerful tool to establish the nonparametric identification of latent structures. Early results on the identifiability of multivariate finite mixtures of Bernouilli distributions were derived by Green [1951] and Anderson [1954], and have been extended by Kasahara and Shimotsu [2009]. More recently, Hall and Zhou [2003] showed that mixtures of two arbitary distributions are generally identified as soon as three measurements are available, provided the component distributions are linearly independent and the outcomes satisfy a conditional-independence restriction. Allman, Matias and Rhodes [2009] have demonstrated that this result carries over to mixtures of more components. Their approach can be applied to a more general class of latent structures that feature some form of conditional independence, such as hidden Markov models with finite state spaces (see Petrie [1969] for seminal work on this) and to random-graph models. We note that, although the availability of two measurements can suffice in problems featuring additive structures, the work of Henry, Kitamura and Salanié [2013] shows that two measurements will only deliver set-identification of parameters in more general latent-structure models. Li and Vuong [1998], Bordes, Mottelet and Vandekerkhove [2006], and Gassiat and Rousseau [2013], among others, present results for additive models.

The work of Allman, Matias and Rhodes [2009] builds heavily on algebraic results for multiway arrays due to Kruskal [1976; 1977]. Although widely applicable, this approach is not constructive. Given identification, some authors have set out to develop methods to estimate latent structures. Benaglia, Chauveau and Hunter [2009] and Levine, Hunter and Chauveau [2011] have constructed EM-like algorithms for finite mixtures. Gassiat, Cleynen and Robin [2013] investigate the use of penalized-likelihood methods in hidden Markov models. However, because these approaches are not based directly on the identification argument, the estimator's asymptotic properties—that is, their consistency, convergence rates, and asymptotic distribution—are difficult to establish and are currently unknown. The results of Hall et al. [2005] on two-component mixtures suggest it should

AMS 2000 subject classifications: 15A18, 15A23, 15A69, 62G05, 62G20, 62H17, 62H30.
Keywords and phrases: estimation, joint diagonalization, latent structure, multilinear-equation system.

be possible to obtain estimators with conventional properties for such latent structures but, to the best of our knowledge, no such estimators have been proposed.

In this paper we first provide a constructive argument to identification of latent structures from longitudinal data. We next introduce a convenient estimator, discuss its implementation, and provide distribution theory. Our framework is the same as that of Allman, Matias and Rhodes [2009] and, as such, can be used to construct estimators for all structures mentioned above. In particular, we discuss the nonparametric estimation of multivariate finite mixtures of discrete and continuous measures, and of hidden Markov models as illustrations of our proposal. When the state spaces involved are finite, these estimators converge at the parametric rate. For absolutely-continuous measures we advocate the use of orthogonal-series estimators of densities to avoid tedious choices about discretizing their support. These estimators are shown to be consistent and to converge at the usual univariate rates. To evaluate the effectiveness of our approach we also provide some simulation evidence for each of the illustrations we consider.

Our approach works from a decomposition of multiway arrays and, as such, connects with the work of Kruskal [1976; 1977]. However, in contrast to Allman, Matias and Rhodes [2009], we go beyond appealing to his results to claim identification of the decomposition. Moreover, we show that a simple transformation of the array leads to a set of multilinear restrictions that identify the parameters of the latent structure at hand in a constructive manner. The restrictions in question take the form of a set of non-normal matrices that are jointly diagonalizable in the same basis. We show that this representation of the latent structure as the joint spectral-decomposition of an array can be cast into a least-squares problem for which an efficient computational approach exists. Our multilinear restrictions are similar to the factorization equations obtained by Anderson [1954], Hu [2008], and Kasahara and Shimotsu [2009], but are more general. In this way, our identification argument allows to reconcile their more ad hoc approaches to the generic but rather different and more abstract view taken by Allman, Matias and Rhodes [2009]. In addition, we construct a plug-in estimator based on the joint spectral decomposition and derive distribution theory. This theory, cast in the form of two theorems, can serve as a blueprint for deriving asymptotic theory for a large class of latent structures.

2. Identification via the joint spectral decomposition.

2.1. *Terminology and general notation.* We will be interested in linear combinations of multiway arrays and, in particular, of outer-product arrays. We borrow terminology and notational conventions from Kruskal [1976; 1977]. Let q be a finite integer. Let $\otimes$ denote the (tensor) outer product. A q-way array of dimension $\kappa_1 \times \kappa_2 \times \cdots \times \kappa_q$ is said to be an outer-product array, or a *q-ad*, if it factors as

$$\bigotimes_{i=1}^{q} x_i = x_1 \otimes x_2 \otimes \cdots \otimes x_q$$

for column vectors x_i of length κ_i. A one-ad is the vector x_1, a two-ad is the matrix $x_1 \otimes x_2$, and so on. We take the elements of the x_i to be real, although all arguments in this section continue to hold for complex values. A q-way array $\mathbb{X}$ is said to admit a *q-adic decomposition* if it can be written

as the sum of q-ads, i.e., if it factors as

$$\mathbb{X} = \sum_{j=1}^{p} \bigotimes_{i=1}^{q} x_{ij} \tag{2.1}$$

for a finite integer p. The *rank* of a q-way array is the smallest number of q-ads needed to obtain a q-adic decomposition. To avoid ambiguity when talking about identification later on, throughout, we will let p be the rank of $\mathbb{X}$. Then the p-linear q-adic decomposition of $\mathbb{X}$ in (2.1) is said to be *irreducible*. We note that another way to represent this q-adic decomposition, and that will be particularly useful for the analysis to follow, is as

$$\mathbb{X} = [X_1, X_2, \ldots, X_q], \tag{2.2}$$

where $X_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a $\kappa_i \times p$ matrix. Of course, in either case, the notation means that the q-way array $\mathbb{X}$ has $(e_1, e_2, \ldots, e_q)$th entry equal to $\mathbb{X}_{e_1,e_2,\ldots,e_q} = \sum_{j=1}^{p} \prod_{i=1}^{q} x_{ij}(e_i)$. Finally, any array can be seen as a collection of lower-dimensional arrays. Moreover, a q-way array has associated with it a set of *slabs*. The collection of slabs in the ith direction are the $(q-1)$-way arrays obtained on fixing the ith index. There are q such collections, one for each direction. For example, a matrix has two collections of slabs; its slabs in the first direction correspond to its rows, and its slabs in the second direction correspond to its columns. Similarly, in any direction, a three-way array can be seen as a collection of matrices.

The q-adic decomposition above is called *essentially unique* if the matrices X_i can be determined from knowledge of $\mathbb{X}$ up to re-arrangement and scale normalization of their respective columns. Permutational equivalence is a mostly trivial and inherently unresolvable ambiguity. In contrast, the indeterminacy of the scale of the columns of the X_i may be undesirable in some situations, and so a stronger concept than essential uniqueness is called for in such cases.

DEFINITION 1 (identification). *The X_i are identified from $\mathbb{X}$ if they can be uniquely recovered up to permutation of their columns.*

Any reference to either essential uniqueness or identification of q-adic decompositions made below implicitely assumes that p is known. We will get back to the identification of p in the discussion of our applications below.

2.2. *Essential uniqueness.* Essential uniqueness does not hold, in general, in the vector case and in the matrix case—that is, when either $q = 1$ or $q = 2$—unless additional constraints are imposed on the problem. To appreciate this point, note that both principal-component analysis and factor analysis, which are widely-used tools in data analysis, can be seen as methods that aim to recover two-adic compositions. In a linear factor model, a vector of κ observable outcomes, y, is related to a vector of $p \leq \kappa$ latent factors, f, and an error term u as

$$y = \Lambda f + u, \tag{2.3}$$

for a $\kappa \times p$ matrix of factor loadings. If f and u are orthogonal, $\operatorname{var} y = \Lambda (\operatorname{var} f) \Lambda^\top + \operatorname{var} u$. Suppose that $\operatorname{var} f$ is positive definite, so that it factors as $CC^\top$ for some matrix C; note that this

decomposition is not unique. Then, even if $\operatorname{var} u$ were known, the two-way array

$$\operatorname{var} y - \operatorname{var} u = [A, A] = AA^\top, \qquad A \equiv \Lambda C,$$

would not yield essential uniqueness of the matrix of factor loadings Λ. It is common to demand that the factors be orthogonal and have unit variance, that is, to set $\operatorname{var} f = \mathrm{I}_p$, in which case Λ can be recovered up to a rotation matrix; see Anderson and Rubin [1956].

The situation is rather different for $q = 3$ and beyond, where essential uniqueness holds under weak conditions on the columns of the X_i. The seminal work of Kruskal [1977] provides sufficient conditions for essential uniqueness in three-way arrays. An extension of his result to arbitrary q can be found in Sidiropoulos and Bro [2000], who show that the q-adic decomposition of $\mathbb{X}$ above is essentially unique provided that

$$\sum_{i=1}^{q} \text{k-rank}\, X_i \geq 2p + (q-1), \tag{2.4}$$

where k-rank stands for Kruskal rank (Harshman and Lundy 1984). Recall that the k-rank of matrix X_i is the largest number k so that every collection of k columns are linearly independent. Thus, $\text{k-rank}\, X_i \leq \operatorname{rank} X_i \leq \min\{\kappa_i, p\}$ and, when all matrices involved have maximal k-rank, (2.4) becomes

$$\sum_{i=1}^{q} \min\{\kappa_i, p\} \geq 2p + (q-1). \tag{2.5}$$

Further, when the X_i are random, $\text{k-rank}\, X_i = \min\{\kappa_i, p\}$ holds generically. Clearly, this condition cannot be satisfied for one-way and two-way arrays. Also, it becomes less stringent as q increases. The large literature on independent component analysis and blind source separation is concerned with recovering Λ in models of the form in (2.3) when factors are assumed to be mutually independent and independent of the error term, and represents a prime example of the usefulness of Kruskal's work; see, e.g., Comon [1994].

In recent work, Allman, Matias and Rhodes [2009] showed that a variety of statistical problems can be represented as linear combinations of outer-product arrays in which the x_{ij} are non-negative and satisfy a scale constraints. As such they were able to succesfully apply Kruskal's result to establish new identification results for latent structures such as multivariate finite-mixture models, hidden Markov models, and random-graph models. Despite the considerable improvement on the existing literature on latent structures made by Allman, Matias and Rhodes [2009], direct application of Kruskal's method does not provide an estimator for these models, and as such leaves something to be desired.

Our chief aim below is to present a constructive approach to recovering q-adic decompositions, and equally to provide distribution theory for estimators of the decomposition based on it. We will focus on identification in the sense of Definition 1 rather than on essential uniqueness as the former concept is of considerably more importance in statistical and econometric application, such as the ones just mentioned. To achieve identification rather than mere essential uniqueness we will assume

that, besides $\mathbb{X}$ itself, its lower-dimensional subarrays, too, are directly observable. Moreover, we assume that $\mathbb{X}(\mathcal{Q})$ is observable for any index set that is a subset of $\{1, 2, \ldots, q\}$, where

$$\mathbb{X}(\mathcal{Q}) = \sum_{j=1}^{p} \bigotimes_{i \in \mathcal{Q}} x_{ij}.$$

As will become clear below, this assumption is rather natural in the applications that we have in mind. In the q-variate finite-mixture model of Hall and Zhou [2003], for example, $\mathbb{X}(\mathcal{Q})$ would simply be the contingency table of a subset of $|\mathcal{Q}|$ measurements, which is clearly observable for any $|\mathcal{Q}| \leq q$ as soon as $\mathbb{X}$ is. Here, the x_{ij} are probabilities and so, in this particular setting, the problem enforces a scale constraint on the columns of X_i. However, the availability of lower-dimensional subarrays will also allow us to tackle situations where such a constraint is absent.

2.3. *Identification.* To present our approach we may restrict attention to three-way arrays. This is without loss of generality, as any q-way array may always be unfolded into a three-way array by collapsing the X_i into matrices of the larger dimension $\prod_i \kappa_i \times p$ by consecutively taking their the columnwise Kronecker product (see, e.g., Sorensen et al. 2013, and below) and applying our techniques to the three-away array so obtained. Thus, our aim will be to recover the X_i making up the three-adic decomposition

$$\mathbb{X} = \sum_{j=1}^{p} x_{1j} \otimes x_{2j} \otimes x_{3j} = [X_1, X_2, X_3], \tag{2.6}$$

in the sense of Definition 1.

We impose the following condition.

Assumption 1 (rank). rank $X_i = p$ *for all* i.

This assumption is slightly stronger than required and can be relaxed; compare with (2.4). We maintain it here for reasons of parsimony.

Fix $i \in \{1, 2, 3\}$ and let $\mathcal{Q}_i \equiv \{i_1, i_2\}$ denote the remaining indices. We will recover each X_i from a transformation of the array $\mathbb{X}$ based on the matrix $\mathbb{X}(\mathcal{Q}_i) = X_{i_1} X_{i_2}^\top$, which we refer to as the *marginalization* of the array in the ith direction. Observe that $\mathbb{X}(\mathcal{Q}_i)$ is a $\kappa_{i_1} \times \kappa_{i_2}$ matrix whose rank is p by Assumption 1. Hence, there exist matrices U_i and V_i of dimension $p \times \kappa_{i_1}$ and $p \times \kappa_{i_2}$, respectively, having the property that

$$U_i \mathbb{X}(\mathcal{Q}_i) V_i^\top = (U_i X_{i_1}) (V_i X_{i_2})^\top = \mathrm{I}_p \,.$$

U_i and V_i can readily be constructed from the singular-value decomposition of $\mathbb{X}(\mathcal{Q}_i)$. Note that this decomposition implies that

$$(V_i X_{i_2})^\top = Q_i^{-1}, \qquad Q_i \equiv U_i X_{i_1}.$$

Now consider the collection of slabs of $\mathbb{X}$ in the ith direction. That is, all matrices obtained on fixing the ith index at one of its κ_i values. Their are κ_i such slabs. The outer-product structure of $\mathbb{X}$ implies that the kth such slab is the $\kappa_{i_1} \times \kappa_{i_2}$ matrix

$$S_{ik} = X_{i_1} D_{ik} X_{i_2}^\top,$$

where $D_{ik} \equiv \text{diag}_k\, X_i$ denotes the $p \times p$ diagonal matrix that has the kth row of matrix X_i on its diagonal. Moreover, pre- and postmultiplying the slabs with U_i and $V_i^\top$ yields the collection of $p \times p$ matrices

$$W_{ik} \equiv U_i S_{ik} V_i^\top = Q_i\, D_{ik}\, Q_i^{-1}, \qquad k = 1, 2, \ldots, \kappa_i. \tag{2.7}$$

This process of transforming the slabs of $\mathbb{X}$ in the ith direction by its marginalization in the ith direction we call the *whitening* of the array with respect to dimension i. To link the whitened array to Kruskal's results, note that they are the slabs in the third direction of a three-way array, say $\mathbb{W}_i$, that decomposes as

$$\mathbb{W}_i = [Q_i, Q_i^{-\top}, X_i]. \tag{2.8}$$

Because each matrix in this decomposition has rank p, Kruskal [1977] implies this decomposition to be essentially unique. Clearly, for given i, it is only required that k-rank $X_i \geq 2$, and not that rank $X_i = p$, but working under this assumption would destroy the symmetry in the argument. Moving on, recall that essential uniqueness implies that the we can unravel the decomposition to recover

$$Q_i P C_1, \qquad Q_i^{-\top} P C_2, \qquad X_i P C_3,$$

up to a permutation matrix P and diagonal matrices C_1, C_2, C_3 satisfying $C_1 C_2 C_3 = \mathrm{I}_p$. To establish identification of X_i we must, therefore, further show that the last such diagonal matrix, C_3, equals the identity matrix. Because from (2.8) it follows that any tri-adic decomposition of the array $\mathbb{W}_i$ into the form $[W_1, W_2, W_3]$ must satisfy the constraints $W_1^\top = W_2^{-1}$ and $W_1^{-1} = W_2^\top$, we have that $C_1 = C_2^{-1}$ must hold. Then $C_3 = \mathrm{I}_p$ follows from the fact that $C_1 C_2 C_3 = \mathrm{I}_p$.

As the above holds for all $i \in \{1, 2, 3\}$, we have demonstrated the following.

LEMMA 1 (identification). *Let Assumption 1 hold. Then X_i is identified from $\mathbb{W}_i$.*

Rather than stating this result as a theorem, we prefer to highlight an alternative representation in Theorem 1 below.

Equation (2.7) shows that the slabs of $\mathbb{W}_i$ are diagonalizable in the same basis. That is, their spectral decompositions share the same eigenvectors; the columns of Q_i. Furthermore, the associated eigenvalues, which do vary with the slabs, equal the columns of X_i. Indeed, recovering X_i is equivalent to recovering the collection of diagonal matrices $(D_{i1}, D_{i2}, \ldots, D_{i\kappa_i})$. We refer to the decomposition in (2.7) as the *joint spectral decomposition* of an array. The matrix Q_i is referred to as the *joint diagonalizer* of the array.

To state our main identification theorem, let $\|\cdot\|_F$ be the Frobenius norm and write Q_p for the set of $p \times p$ positive-definite matrices whose columns have unit Euclidean norm. In what follows, the normalization of the columns is inconsequential.

THEOREM 1 (joint spectral decomposition). *Let Assumption 1 hold. Then $(D_{i1}, D_{i2}, \ldots, D_{i\kappa_i})$ are identified as*

$$D_{ik} = Q_i^{-1} W_{ik} Q_i,$$

where

$$Q_i = \arg\min_{Q\in \mathrm{Q}_p} \sum_{k=1}^{\kappa_i} \left\| W_{ik} - Q D_k(Q) Q^{-1} \right\|_F^2, \qquad D_k(Q) \equiv \operatorname{diag}\left[Q^{-1} W_{ik}\, Q\right]$$

for all i.

Theorem 1 shows that identification of the X_i boils down to recovering a joint diagonalizer, which is characterized as the solution to a least-squares problem.

A large literature in the field of independent component analysis is dedicated to the construction of efficient algorithms for the simultaneous diagonalization of a set of matrices. In such applications—as in (2.3) above—the matrices to be diagonalized are typically normal, and so the joint diagonalizer is an orthonormal matrix. Orthonormality is convenient, as it drastically reduces the space of matrices to be searched over. Although it may be the case under additional restrictions, the applications we have in mind will typically not involve normal matrices; see our illustrations below. Fortunately, several recent contributions have set out to tackle the computational issues that arise in the general case when the joint diagonalizer is allowed to be non-orthogonal. Fu and Gao [2006] were the first to propose a simple algorithm for the joint diagonalization of a set of arbitrary non-defective matrices. More recently, Iferroudjene, Meraim and Belouchrani [2009] and Luciani and Albera [2010] have proposed computational refinements.

We advocate the use of the algorithm of Iferroudjene, Meraim and Belouchrani [2009]. Their method is a Jacobi-like routine that, in contrast to the other procedures, minimizes the criterion put forth in Theorem 1. Working from a least-squares formulation is useful for our purposes as it readily suggests the construction of an extremum estimator based on a sample version of the criterion, and facilitates the derivation of distribution theory. As such, Theorem 1 is constructive. We present the estimator and derive its asymptotic properties in the next section.

2.4. *Impact of* q. Before turning to estimation we first discuss how the decomposition

$$\mathbb{X} = [X_1, X_2, \ldots, X_q]$$

can be recovered via our approach under conditions that weaken as q increases. Let $\odot$ denote the Khatri-Rao product, or columnwise Kronecker product. For example, $X_2 \odot X_1$ is the $\kappa_1\kappa_2 \times p$ matrix obtained on stacking the matrices $X_1 D_{2k}$. For a moment, set $q = 4$. The slabs of $\mathbb{X}$ in the fourth direction are three-way arrays whose slabs in the third direction are again $\kappa_1 \times \kappa_2$ matrices. The (k_1, k_2)th such matrix equals $X_1 D_{3k_1} D_{4k_2} X_2^\top$, where k_1 and k_2 range over the sets $\{1, 2, \ldots, \kappa_3\}$ and $\{1, 2, \ldots, \kappa_4\}$, respectively. For each k_2, stack the resulting collection of κ_3 matrices beneath each other. Then is gives the collection

$$(X_3 \odot X_1) D_{4k_2} X_2^\top, \quad k_2 = 1, 2, \ldots, \kappa_4,$$

which are the slabs in the third direction of the array $[(X_3 \odot X_1), X_2, X_4]$, to which Theorem 1 can be applied. More generally, to recover X_i, $\mathbb{X}$ can always be re-arranged into a three-way array with q-adic decomposition

$$[\odot_{i_1\in\mathcal{Q}_1} X_{i_1}, \odot_{i_2\in\mathcal{Q}_2} X_{i_2}, X_i], \tag{2.9}$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are index sets that partition $\{1, 2, \ldots, q\}$. For our results to be applicable to this array it suffices that the matrices in (2.9) satisfy the rank condition in Assumption 1. As $\odot_{i\in\mathcal{Q}} X_i$ is of dimension $\prod_{i\in\mathcal{Q}} \kappa_i \times p$, this requirement clearly becomes easier to fulfill as the number of matrices to be recovered, q, increases. It is interesting to compare this finding with Theorem 4 of Allman, Matias and Rhodes [2009], which states that the X_i are essentially unique if there exists a tri-partition $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$ of $\{1, 2, \ldots, q\}$ such that

$$\text{k-rank}\{\odot_{i\in\mathcal{Q}_1} X_i\} + \text{k-rank}\{\odot_{i\in\mathcal{Q}_2} X_i\} + \text{k-rank}\{\odot_{i\in\mathcal{Q}_3} X_i\} \geq 2(p+1)$$

is satisfied. Observe that both these conditions are weaker than (2.4). These conclusions are a generalization of the observation by Hall and Zhou [2003] that, in the finite-mixture context, the usual curse of dimensionality does not arise and, rather, works in reverse. Further, it is useful to note that the whitening process implies that the matrices to be diagonalized are of dimension $p \times p$, regardless of q and κ_i. Thus, even in large-dimensional problems, the computational complexity of the diagonalization problem is limited.

3. Estimation by joint approximate diagonalization. Consider the general case in which a set of estimable non-defective matrices $(W_1, W_2, \ldots, W_\kappa)$ can be jointly diagonalized by a finite $p \times p$ matrix Q_0 whose determinant is strictly positive, in the sense that $\det Q_0 > \varepsilon$ for some $\varepsilon > 0$. Let $(D_1, D_2, \ldots, D_\kappa)$ denote diagonal matrices with the corresponding eigenvalues on the diagonal. Rather than the W_k, we have at our disposal only estimators $\widehat{W}_k$, say, constructed from a random sample of size n. A natural estimator of Q_0, then, is the minimizer of the sample analog of the criterion stated in Theorem 1. Moreover, we take

$$\widehat{Q} \equiv \arg\min_{Q\in \mathrm{Q}_p^\varepsilon} \sum_{k=1}^{\kappa} \big\| \widehat{W}_k - Q\widehat{D}_k(Q)Q^{-1} \big\|_F^2, \qquad \widehat{D}_k(Q) \equiv \operatorname{diag}\big[Q^{-1}\widehat{W}_i\, Q\big], \tag{3.1}$$

to be our estimator, where $\mathrm{Q}_p^\varepsilon \equiv \{Q \in \mathrm{Q}_p : \det Q > \varepsilon\}$ is the parameter space over which the minimization takes place. An estimator of D_k follows as

$$\widehat{D}_k \equiv \widehat{D}_k(\widehat{Q}) = \operatorname{diag}\big[\widehat{Q}^{-1}\widehat{W}_k\, \widehat{Q}\big]. \tag{3.2}$$

Sampling noise in the $\widehat{W}_k$ prevents them from sharing the same set of eigenvectors. Indeed, in general, there does not exist a Q such that $Q^{-1}\widehat{W}_k\, Q$ will be exactly diagonal for all k. 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 $\widehat{Q}$ the *joint approximate-diagonalizer* of the set of $\widehat{W}_k$. The algorithm of Iferroudjene, Meraim and Belouchrani [2009] can be applied to (3.1) to compute $\widehat{Q}$, from which the $\widehat{D}_k$ follow in turn.

Distribution theory for the joint approximate-diagonalizer is not available, and so we provide it here. We write J_p for a $p\times p$ matrix of ones and let $E_k \equiv \big[\big(\mathrm{I}_p \otimes D_k\big) - \big(D_k \otimes \mathrm{I}_p\big)\big](\mathrm{I}_p \otimes Q_0^{-1})$. Introduce

$$G_k \equiv E_k^\top \operatorname{diag}[\operatorname{vec}(\mathrm{J}_p - \mathrm{I}_p)]\,\big(Q_0^\top \otimes \mathrm{I}_p\big)(\mathrm{I}_p \otimes Q_0^{-1}), \qquad H \equiv \sum_{k=1}^{\kappa} E_k^\top \operatorname{diag}[\operatorname{vec}(\mathrm{J}_p - \mathrm{I}_p)]\, E_k,$$

and construct $G \equiv (G_1, G_2, \ldots, G_\kappa)$, $\widehat{W} \equiv (\widehat{W}_1, \widehat{W}_2, \ldots, \widehat{W}_\kappa)$, and $W \equiv (W_1, W_2, \ldots, W_\kappa)$. The following theorem contains an asymptotically-linear representation of $\widehat{Q}$.

THEOREM 2 (eigenvectors). *Let* $\|\widehat{W} - W\|_F = O_P(1/\sqrt{n})$ *and suppose that* Q_0 *is an interior element of* O_p^ε. *Then* $\|\widehat{Q} - Q_0\|_F = O_P(1/\sqrt{n})$ *and*

$$\sqrt{n}\,\mathrm{vec}(\widehat{Q} - Q_0) = -H^{-1}G\,\sqrt{n}\,\mathrm{vec}(\widehat{W} - W) + o_P(1)$$

as $n \to \infty$.

If, further, $\sqrt{n}\,\mathrm{vec}(\widehat{W} - W) \overset{A}{\sim} \mathcal{N}(0, V)$ for some asymptotic covariance matrix V, then Theorem 2 implies that

$$\sqrt{n}\,\mathrm{vec}(\widehat{Q} - Q_0) \overset{A}{\sim} \mathcal{N}\big(0, H^{-1}GVG^\top H^{-1}\big).$$

Given consistency of $\mathrm{vec}\,\widehat{Q}$, a consistent plug-in estimator of the asymptotic variance can easily be constructed provided we have at our disposal a matrix $\widehat{V}$ that satisfies $\|\widehat{V} - V\|_F = o_P(1)$.

With Theorem 2 at hand, asymptotic results for the eigenvalues are immediate.

THEOREM 3 (eigenvalues). *Let the conditions of Theorem 2 hold. Then* $\|\widehat{D}_k - D_k\|_F = o_P(1/\sqrt{n})$ *and*

$$\sqrt{n}\,\mathrm{vec}(\widehat{D}_k - D_k) = \mathrm{diag}[\mathrm{vec}\,\mathrm{I}_p]\big\{E_k^\top \sqrt{n}\,\mathrm{vec}(\widehat{Q} - Q_0) + \big(Q_0^\top \otimes Q_0^{-\top}\big)\sqrt{n}\,\mathrm{vec}(\widehat{W}_k - W_k)\big\} + o_P(1)$$

as $n \to \infty$.

In our context, the proof to Theorem 1 shows that the input matrices are estimates of the form

$$\widehat{W}_k = \widehat{U}\widehat{S}_k\widehat{V}^\top,$$

where $\widehat{U}$ and $\widehat{V}$ are estimators of $U \equiv \Lambda^{-1/2}A$ and $V \equiv \Lambda^{-1/2}B$ for a diagonal matrix Λ and orthonormal matrices A, B such that $A\Lambda B^\top$ is a singular-value decomposition of a marginalization matrix M. If a $\sqrt{n}$-consistent and asymptotically-normal estimator of M is available, then its left and right singular vectors and its singular values are $\sqrt{n}$-consistent and asymptotically-normal estimators of A, B, and Λ, respectively, if all singular values of M are simple (see, e.g., Eaton and Tyler [1991]; Magnus [1985]; Bura and Pfeiffer [2008]). The latter condition holds generically. Hence, the estimator of the input matrices will be asymptotically-linear if the estimator of the slabs is. In all the applications we consider in the next section, this will be the case.

4. Finite-mixture models. The identification of finite mixtures from longitudinal data has recently received quite some attention. Assume there are q measurements $(y_1, y_2, \ldots, y_q)$ and p latent classes. The cornerstone model assumes that the measurements are independent within a given group (extensions to dynamics are possible, see Kasahara and Shimotsu [2009] and also below). Let π_j be the marginal probability of belonging to group j. Then the distribution of $(y_1, y_2, \ldots, y_q)$ factors as

$$\mu(y_1, y_2, \ldots, y_q) = \sum_{j=1}^{p} \pi_j \prod_{i=1}^{q} \mu_{ij}(y_i), \tag{4.1}$$

where μ_{ij} is the conditional univariate distribution function of measurement i when it belongs to group j. Hettmansperger and Thomas [2000] and Hall and Zhou [2003] showed that, in a bivariate

mixture, both the component distributions and the mixing proportions are identified when at least three measurements are available. For the general case, Hall et al. [2005] showed there exists a $\overline{q}(p)$ so that component distributions and the mixing proportions are identified if $q > \overline{q}(p)$. The link between finite mixtures and three-way arrays was made by Allman, Matias and Rhodes [2009] who, using Kruskal's results mentioned above, established generic identifiability for any p when the μ_{ij} are linearly independent and $q \geq 3$. In the context of finite mixtures, the failure of Kruskal's rank condition when $q = 2$ is the underlying cause for the partial-identification results obtained by Hall and Zhou [2003] and Henry, Kitamura and Salanié [2013]. Point identification can be restored by imposing additional restrictions; Bordes, Mottelet and Vandekerkhove [2006] and Hunter, Wang and Hettmansperger [2007] consider location models under a symmetry condition on the noise distribution, and Henry, Jochmans and Salanié [2013] achieve identification under a tail restriction on the component distributions in more general models.

In spite of these important results, there has been little work on the nonparametric estimation of mixtures from longitudinal data. Hall and Zhou [2003] and Hall et al. [2005] discuss estimation in the two-component case but their procedures do not naturally extend to the general case. Recently, Benaglia, Chauveau and Hunter [2009] and Levine, Hunter and Chauveau [2011] have developed an algorithm akin to the conventional EM algorithm of Dempster, Laird and Rubin [1977] to estimate both the mixing proportions, π_j, and the component distributions, μ_{ij}. However, as mentioned by the authors, it seems rather difficult to establish the estimators' statistical properties, such as its consistency and convergence rates.

Below we illustrate how our joint-diagonalization approach yields estimators of finite mixtures with conventional asymptotic properties.

4.1. *Mixtures of discrete measures.* Suppose first that the y_i are discrete random variables with κ_i points of support. Let p_{ij} be the vector that collects the κ_i point probabilities of measurement i conditional on belonging to group j. Conditional independence implies that the contingency table of any collection $\mathcal{Q}$ of $|\mathcal{Q}|$ measurements is

$$\mathbb{P}(\mathcal{Q}) = \sum_{j=1}^{p} \pi_j \bigotimes_{i \in \mathcal{Q}} p_{ij}.$$

Apart from the presence of π_j, which can be seen as scale factors, this representation directly fits our framework. Nonetheless, the component distributions p_{ij} and the mixing proportions π_j will be separately identified because, essentially, the mixing proportions do not vary with $\mathcal{Q}$. Moreover, the contingency table of all the measurements, $\mathbb{P}$, admits the q-adic decomposition

$$\mathbb{P} = [P_1\Pi, P_2, \ldots, P_q] = [P_1, P_2\Pi, \ldots, P_q] = \cdots = [P_1, P_2, \ldots, P_q\Pi],$$

where $P_i \equiv (p_{i1}, p_{i2}, \ldots, p_{ip})$ and $\Pi \equiv \operatorname{diag} \pi$ for $\pi \equiv (\pi_1, \pi_2, \ldots, \pi_p)^\top$. All these representations are equivalent. However, because π equally shows up in all lower-dimensional subarrays, whitening the array with respect to its marginalization in any direction will always absorb the mixing proportions in the joint diagonalizer.

Impose the following conditions.

ASSUMPTION 2 (rank). *The μ_{ij} are linearly independent for each i, $\pi_j > 0$ for each j, and $q \geq 3$.*

Assumption 2 is identical to the conditions in Theorem 8 of Allman, Matias and Rhodes [2009]. Linear independence of the p_{ij} implies that all the P_i have maximal column rank. Ruling out the possibility that $\pi_j = 0$ is natural and, together with the rank condition on the P_i, ensures that the mixture is irreducible, so that the components are unambiguously defined.

Showing Proposition 1 can be done by direct application of Theorem 1.

PROPOSITION 1 (identification). *Let Assumption 2 hold. Then P_i is identified from a joint-spectral decomposition for each i.*

Assumption 2 requires $\kappa_i \geq p$ for all i. This effectively demands the y_i to have more support points than there are latent groups. This ensures identification as soon as $q = 3$. In line with the discussion above, Proposition 1 can be shown to hold under weaker conditions on the columns of P_i if q is larger than this minimum value.

Turning to estimation, a natural way to proceed is to replace $\mathbb{P}(\mathcal{Q})$ by its sample counterpart and to apply the techniques introduced above. Moreover, for any $\mathcal{Q}$, an estimator of $\mathbb{P}(\mathcal{Q})$ is simply the empirical frequency table of the data. When the support of the μ_{ij} is large, a smoothed approach may be preferable to avoid the issue of empty cells; see Aitchison and Aitken [1976]. In any case, such an estimator converges at the parametric rate and is asymptotically normal. Theorem 3 then delivers consistent and asymptotically-linear point estimators of the p_{ij}.

A direct consequence of Proposition 1 is that the mixing proportions, too, are identified. Moreover,

$$\pi = P_i^+ \, \mathbb{P}(\{i\}), \qquad P_i^+ \equiv \left(P_i^\top P_i\right)^{-1} P_i^\top,$$

for each i because the univariate mixture for y_i has associated with it the lower-dimensional subarray $\mathbb{P}(\{i\}) = P_i \pi$ and P_i has maximal column rank.

COROLLARY 1 (mixing proportions). *Let Assumption 2 hold. Then π is identified as the solution to a least-squares fit of $\mathbb{P}(\{i\})$ on P_i.*

Given $\sqrt{n}$-consistent estimators of P_i and $\mathbb{P}(\{i\})$, a plug-in version of π, too, will be $\sqrt{n}$-consistent. Its asymptotic distribution can be deduced by an application of the delta method.

So far we always assumed the number of components to be known. Assumption 2, however, implies the following.

COROLLARY 2 (number of components). *Let Assumption 2 hold. Then p is identified as the rank of $\mathbb{P}(\{i_1, i_2\})$ for any $i_1, i_2 \in \{1, 2, \ldots, q\}$.*

This result relates directly to a recent contribution of Kasahara and Shimotsu [2014]. Moreover, a consistent estimator of p is easily formed via a sequential-testing procedure based on infering the rank of an empirical analog of the $\mathbb{P}(\{i_1, i_2\})$. In empirical applications, this procedure can give guidance on the number of components. An issue that we have not addressed in our asymptotics is accounting for the resulting estimation uncertainty p.

4.2. *Mixtures of continuous measures.* Return to (4.1) but, now, suppose that the function μ_{ij} are absolutely continuous. Let f_{ij} be the associated densities. Then

$$f(y_1, y_2, \ldots, y_q) = \sum_{j=1}^{p} \pi_j \prod_{i=1}^{q} f_{ij}(y_i). \tag{4.2}$$

The identification results from the previous subsection can be generalized by discretizing the support of the variables. Such an approach would be in line with Allman, Matias and Rhodes [2009] and with Kasahara and Shimotsu [2014] but is not attractive for the construction of estimators. An alternative approach that has the advantage of yielding nice estimators is to consider a discretization in the frequency domain, as was recently done for a special case of (4.2) in Bonhomme, Jochmans and Robin [2013].

Suppose that the f_{ij} are supported on the compact interval $[-1, 1]$. The translation to generic compact intervals is straightforward. Let $\{\phi_i, i > 0\}$ be a set of polynomials that form a complete orthonormal system with respect to a weight function ρ on $[-1, 1]$. For example, polynomials such as those belonging to the Jacobi class—e.g., Chebychev polynomials 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 $\varphi_{\kappa_i} \equiv (\phi_1, \phi_2, \ldots, \phi_{\kappa_i})^\top$ is

$$\mathrm{Proj}_{\kappa_i} f_{ij} \equiv \varphi_{\kappa_i}^\top \gamma_{ij}, \qquad \gamma_{ij} \equiv \int_{-1}^{1} \varphi_{\kappa_i}(u)\rho(u) f_{ij}(u)\,\mathrm{d}u, \tag{4.3}$$

for any integer κ_i. The vector γ_{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 $\|f\|_2 \equiv \big(\int_{-1}^{1} f(u)^2 \rho(u)\,\mathrm{d}u\big)^{1/2}$. 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). Now, convergence in L^2_ρ-norm of two functions implies that they share the same Fourier coefficients and can differ only on a set of measure zero. Hence, recovering the function f_{ij} is equivalent to recovering its Fourier coefficients.

Now, γ_{ij} is the expectation of $\varphi_{\kappa_i}\rho$ against the density f_{ij}, which is not directly observable. However, in the same way, we can define the projection of $f(y_{i_1}, y_{i_2}, \ldots, y_{i_{|\mathcal{Q}|}})$ for any index set $\mathcal{Q}$. Its Fourier coefficients equal the array

$$\mathbb{G}(\mathcal{Q}) \equiv \int_{-1}^{1}\int_{-1}^{1}\cdots\int_{-1}^{1} \bigotimes_{i\in\mathcal{Q}} \varphi_{\kappa_i}(u_i)\rho(u_i)\, f(u_{i_1}, u_{i_2}, \ldots, u_{i_{|\mathcal{Q}|}})\ \mathrm{d}(u_{i_1}, u_{i_2}, \ldots, u_{i_{|\mathcal{Q}|}}), \tag{4.4}$$

and is nonparametrically identified. From (4.2), (4.3), and (4.4) it follows that

$$\mathbb{G}(\mathcal{Q}) = \sum_{j=1}^{p} \pi_j \bigotimes_{i\in\mathcal{Q}} \gamma_{ij},$$

which fits our framework. Note that, here, the problem does not impose a natural scale constraints on the entries of $\mathbb{G}$. Nonetheless, lower-dimensional subarrays are available as the Fourier coefficients of the joint densities of a subset of the measurements.

Let $\Gamma_i \equiv (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{ip})$.

PROPOSITION 2 (identification). *Let Assumption 2 hold. Then* $\operatorname{rank}\Gamma_i = p$ *for sufficiently large* κ_i.

Proposition 2 implies that Proposition 1 and Corollaries 1 and 2 carry over to the case of finite mixtures of continuous distributions. For the remainder of this subsection, we assume that the κ_i are chosen sufficiently large so that $\operatorname{rank}\Gamma_i = p$ holds.

To construct an estimator of the f_{ij}, let $\{y_{m1}, y_{m2}, \ldots, y_{mq}\}_{m=1}^n$ denote a random sample of size n. The entries of the array $\mathbb{G}(\mathcal{Q})$ can be estimated as sample averages over the orthogonal polynomials, weighted against ρ. Moreover, a sample analog is

$$\widehat{\mathbb{G}}(\mathcal{Q}) \equiv n^{-1} \sum_{m=1}^{n} \bigotimes_{i \in \mathcal{Q}} \varphi_{\kappa_i}(y_{mi})\rho(y_{mi}).$$

Joint approximate diagonalization then yields estimators of the γ_{ij}, say $\widehat{\gamma}_{ij}$, which can be used to construct the orthogonal-series estimator

$$\widehat{f}_{ij} \equiv \varphi_{\kappa_i}^{\top} \widehat{\gamma}_{ij} \tag{4.5}$$

of f_{ij} for each i, j.

Let $\|\cdot\|_\infty$ be the supremum norm. Some weak regularity conditions on the basis functions are collected in Assumption 3.

ASSUMPTION 3 (regularity). *The sequence* $\{\phi_i, i \geq 0\}$ *is dominated by a function* ψ, *which is continuous on* $(-1, 1)$ *and positive almost everywhere on* $[-1, 1]$. *Both* $\psi\rho$ *and* $\psi^2\rho$ *are integrable. There exists a sequence of constants* $\{\zeta_\kappa, \kappa > 0\}$ *so that* $\|\sqrt{\varphi_\kappa^{\top}\varphi_\kappa}\|_\infty \leq \zeta_\kappa$.

These conditions are satisfied for the class of Jacobi polynomials, for example.

To present convergence rates for our orthogonal-series estimator we require an assumption about the smoothness of the density that is being estimated.

ASSUMPTION 4 (smoothness). *The* f_{ij} *are continuous and the* $(\psi\rho)^2 f_{ij}$ *are integrable. There exists a constant* $\beta \geq 1$ *such that* $\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\|_\infty = O(\kappa_i^{-\beta})$.

Convergence in L^2_ρ-norm implies that $\lim_{\kappa_i \to \infty} \|\gamma_{ij}\|_F$ is finite. The coefficient β is a measure of how fast the Fourier coefficients shrink. In general, β is larger the smoother the underlying function that is being approximated.

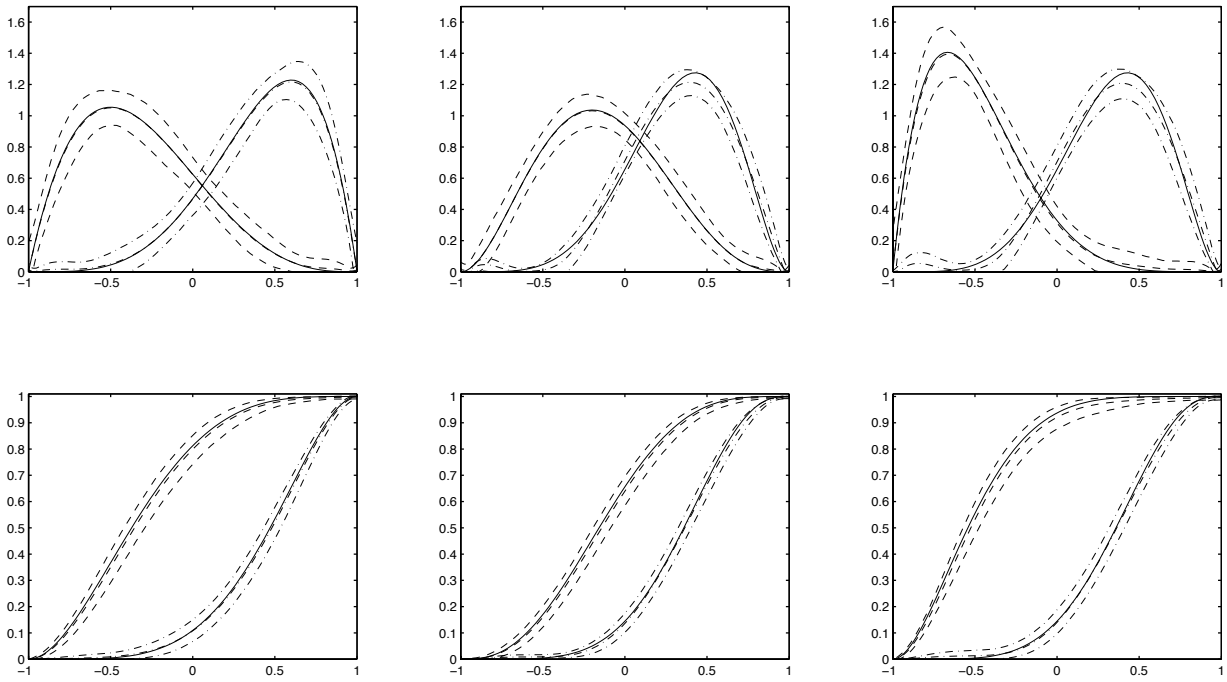
Under these conditions we obtain integrated squared-error and uniform convergence rates.

PROPOSITION 3 (convergence rates). *Let Assumptions 2–4 hold. Then*

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

for all i, j.

FIG 1. *Component densities and distributions*



Proposition 3 shows that there is no penality in terms of convergence rate for not observing group membership of the observations. Under slightly stronger tail conditions on the f_{ij} we can equally derive an asymptotically-linear representation of the density estimator at a fixed point and present a pointwise asymptotic-normality result by suitably adapting the arguments in Bonhomme, Jochmans and Robin [2013]. Such results are useful to construct $\sqrt{n}$-consistent estimators of functionals of f_{ij} or semiparametric two-step estimators of Euclidean parameters.

Assumptions 3 and 4 imply that the Fourier coefficients are estimated at the parametric rate. Thus,

$$\widehat{\pi} \equiv \widehat{\Gamma}_i^{+}\widehat{\mathbb{G}}(\{i\}),$$

with $\widehat{\Gamma}_i \equiv (\widehat{\gamma}_{i1}, \widehat{\gamma}_{i2}, \ldots, \widehat{\gamma}_{ip})$, is a $\sqrt{n}$-consistent and asymptotically-normal estimator of the mixing proportions for any i.

An interesting yet unresolved issue is the selection of the number of terms in the series expansion of the f_{ij}. Because we recover the γ_{ij} from a multiway array that involves all f_{ij} and must satisfy a rank condition, the problem appears far more complicated than in standard orthogonal-series estimation for which some selection rules exist (see, e.g., Diggle and Hall 1986).

To illustrate the performance of the orthogonal-series estimator we applied it to simulated data from a three-variate two-component mixture of generalized beta distributions on the interval $[-1, 1]$. The solid lines in the upper plots of Figure 1 represent these densities. The solid lines in the lower plots are the corresponding distribution functions. The data was drawn from a mixture of these

TABLE 1
Mixing proportions

	$n = 500$				$n = 1000$			
	mean		std		mean		std	
	π_1	π_2	π_1	π_2	π_1	π_2	π_1	π_2
$i = 1$	.5133	.4794	.0257	.0260	.5090	.4869	.0186	.0186
$i = 2$	.5130	.4854	.0300	.0301	.5092	.4895	.0204	.0205
$i = 3$	.4978	.4948	.0319	.0320	.4980	.4989	.0231	.0229

distributions with $\pi_1 = \pi_2 = .5$, with the right-skewed distribution is labelled as the first component. We estimated the densities by means of our joint approximate-diagonalization estimator using the leading five Chebychev polynomials of the first kind as basis functions on data of size $n = 500$, and then applied the correction of Gajek [1986] to the resulting estimator to obtain bona fide estimates. We next used Clenshaw and Curtis [1960] quadrature to construct an estimator of the 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. Table 1 provides the mean and standard deviation of the estimated mixing proportions over the Monte Carlo simulations for $n = 500$ and with $n = 1000$. In the table, π_1 denotes the proportion associated with the right-skewed density while π_2 denotes the proportion corresponding to the left-skewed density. Again, the point estimates are broadly correctly centered. Further, the standard deviation decreases roughly with a factor $1/\sqrt{n}$ as n is doubled, confirming that the mixing proportions are estimated at the parametric rate.

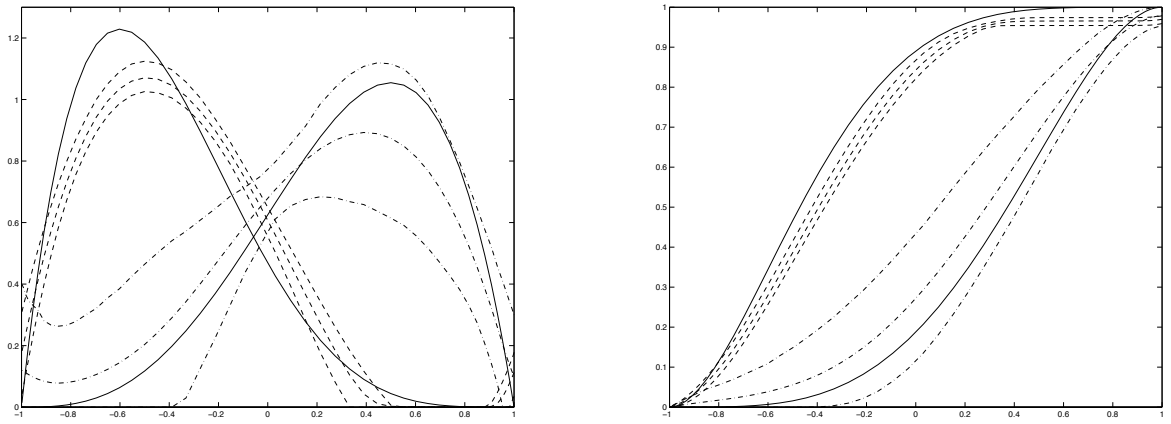
5. Hidden Markov models. As a final example, we discuss hidden Markov models. Such a model can be seen as a stationary finite-mixture model in which group membership changes according to a Markov switching process. Thus, compared to our mixture from above, here, we allow for (latent) Markovian dynamics. To set up the model, let $\{s_t, t > 0\}$ be a strictly stationary stochastic process that can take on values in $\{1, 2, \ldots, p\}$ with probability $\{\pi_1, \pi_2, \ldots, \pi_p\}$, and whose dependency structure is first-order Markov. Let K denote the $p \times p$ transition matrix of the chain. Then,

$$K(e_1, e_2) = \Pr[s_t = e_2 | s_{t-1} = e_1].$$

Rather than s_t we observe outcome variables $\{y_t, t > 0\}$. These outcome variables are taken to be jointly-independent conditional on $\{s_t, t > 0\}$, and the distribution of y_t depends on $\{s_t, t > 0\}$ only through the current state, s_t. Let μ_j denote the conditional distribution of y_t given $s_t = j$. The μ_j are called the emission distributions and, together with the vector of marginal probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_p)^\top$ and the transition matrix K, are the parameters of interest in the hidden Markov model. The recent monograph by Cappé, Moulines and Rydén [2005] provides an extensive overview of the literature and many illustrations.

Because of the dynamics in the latent states s_t, it may not be immediately obvious that a hidden Markov model fits our setup. Nonetheless, as noted by Allman, Matias and Rhodes [2009] and more recently by Gassiat, Cleynen and Robin [2013], a close inspection of the problem allows to fit the hidden Markov model in the framework of three-way arrays, and therefore makes it amendable to our approach.

FIG 2. *Emission densities and distributions*



Sufficient conditions for identification are collected in Assumption 5.

ASSUMPTION 5 (rank). *The μ_j are linearly independent,* $\operatorname{rank} K = p$, *and* $q \geq 3$.

These conditions should by now be familiar. They yield identification both when the state space of y_t is finite and when it is a continuous interval.

Consider the case where y_t can take on values in the finite set $\{v_1, v_2, \ldots, v_\kappa\}$. For $i \in \{1, 2, \ldots, q\}$, consider $\kappa \times p$ matrices P_i whose entries are

$$P_i(e_1, e_2) = \Pr[y_i = v_{e_1} | s_2 = e_2].$$

Note that the conditioning argument involves s_2 regardless of i. This asymmetry is a consequence of the Markovian dependence in the latent states. Then it is easily shown that the q-way contingency table of $(y_1, y_2, \ldots, y_q)$ decomposes as

$$\mathbb{P} = [P_1 \Pi, P_2, \ldots, P_q],$$

where, again, $\Pi = \operatorname{diag} \pi$. The presence of π with P_1 is due to the initial state being drawn from the stationary distribution of the Markov chain. Note that the columns of the matrix P_2 are the emission distributions. Hence, given the similarity with the mixture model from the previous section, identification of the emission distributions is immediate. Further, and again similar to before, the vector of marginal mixing probabilities, π, can be recovered as

$$\pi = P_2^+ \, \mathbb{P}(\{1\}),$$

where P_2^+ is the Moore-Penrose pseudoinverse of P_2. Finally, the transition matrix can be recovered from a slightly more convoluted identity involving $\mathbb{P}(\{1, 2\})$. Note that

$$\Pr[y_1 = a, y_2 = b] = \sum_{e_1=1}^{p} \sum_{e_2=1}^{p} \Pr[y_2 = b | s_2 = e_2] \Pr[y_1 = a | s_1 = e_1] \Pr[s_2 = e_2 | s_1 = e_1] \Pr[s_1 = e_1],$$

TABLE 2
Hidden Markov model

parameter	value	mean	std
$\Pr[s_t = 1]$	.7591	.7255	.0755
$\Pr[s_t = 0]$	.2409	.2554	.0786
$K(0,0)$	.5000	.5731	.3056
$K(0,1)$	.5000	.3913	.3494
$K(1,0)$	.1587	.1352	.0587
$K(1,1)$	.8413	.8500	.0608

and so $\mathbb{P}(\{1,2\}) = P_2 \Pi K P_2^\top$ holds by stationarity. Hence, pre- and postmultiplying by P_2^+ yields ΠK from which

$$K = \Pi^{-1} P_2^+ \mathbb{P}(\{1,2\})(P_2^+)^\top$$

follows because all objects on the right-hand side have already been shown to be identified. When the μ_j are absolutely continuous and the associated density functions f_j, say, are compactly supported and square integrable, we can again work via a discretization in the frequency domain. In this case, everything continues to go through after replacing the P_i by Γ_i, using obvious notation. We have thus established the following result.

PROPOSITION 4 (identification). *Let Assumption 5 hold. Then μ_j, π, and K are all identified.*

For both the discrete- and the continuous-outcome case, asymptotic results for estimators of the hidden Markov model can be established in the same way as before. In particular, under regularity conditions, plug-in estimators based on the identification approach just laid out are consistent and converge at the conventional univariate rates.

Rather than going into more detail on this here, we provide the results of a small Monte Carlo experiment in which we generated s_t via the stationary probit model

$$s_t = 1\{s_{t-1} \geq \varepsilon_t\}, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

and subsequently drew y_t from a left-skewed beta distribution on $[-1,1]$ when $s_t = 0$ and from a right-skewed beta distribution when $s_t = 1$. The solid lines in Figure 2 present the emission densities (left plot) and distributions (right plot). The stationary distribution of the Markov chain and its transition matrix are given in Table 2. The data generating process is such that it is more likely to deliver observations from the right-skewed regime. Indeed, the event $s_t = 1$ occurs with probability roughly equal to .75 in equilibrium, and it is characterized by positive state dependence, that is, $K(1,1) > K(1,0)$. We implemented our procedures in the same way as in the previous section, using Chebychev polynomials to estimate the emission densities and Clenshaw-Curtis quadrature to recover estimates of the emission distributions. Figure 2 has the same layout as Figure 1. It again shows our algorithm to be effective in recovering the underlying densities and distributions. As anticipated, the left-skewed density is estimated less precisely than is the right-skewed density. Table 2 shows a similar pattern for the estimators of π and K. The point estimates are broadly correct, on average. Also, the bias and standard deviation are both smaller for transition from the highly-probable state, as could have been expected.

Acknowledgements. We thank participants to the Cowles Foundation summer conference in econometrics at Yale in June 2013, the ESRC Econometric Study Group in Bristol in July 2013, and the Asian meetings of the Econometric Society in Singapore in August 2013. We are grateful to Xiaohong Chen and Marc Henry for comments, and to Laurent Albera and Xavier Luciani for sharing the code for their algorithm in Luciani and Albera [2010] with us.

APPENDIX

PROOF OF THEOREM 1. By Lemma 1 and (2.7), Q_i and $(D_{ik}, D_{i2}, \ldots, D_{i\kappa})$ uniquely solve

$$\min_{Q\in \mathrm{Q}_p,\, \{D_k\}\in \mathrm{D}_p} \sum_{k=1}^{\kappa} \big\| W_{ik} - Q D_k Q^{-1} \big\|_F^2, \tag{A.1}$$

where D_k denotes the set of $p \times p$ diagonal matrices. For given Q, the solution for D_{ik} is easily seen to be $D_k(Q)$. Profiling out the diagonal matrices from (A.1) yields the objective function for the joint diagonalizer stated in Theorem 1. □

PROOF OF THEOREM 2. To see that $\|\widehat{Q} - Q_0\|_F = o_P(1)$, note that (i) the parameter space Q_p^ε is a compact set; (ii) the sample objective function converges to the population criterion stated in Theorem 1 uniformly on Q_p^ε because of consistency of the first-stage estimators and because the objective function is Lipschitz continuous; (iii) the population criterion is uniquely minimized at Q_0 by the identification result in Theorem 1; and (iv) the population criterion is continuous in Q because the Frobenius norm is continuous. Theorem 2.1 in Newey and McFadden [1994] then yields consistency.

To derive asymptotic theory for the joint approximate-diagonalization estimator we start by setting up the Lagrangian for the constrained optimization problem in (3.1). To do so it is useful to reformulate the problem as

$$\min_{(Q,R)} \sum_{k=1}^{\kappa} \big\| \widehat{O}_k(Q,R) \big\|_F^2 \quad \text{s.t. } QR^\top = \mathrm{I}_p,\ Q^\top R = \mathrm{I}_p, \qquad \widehat{O}_k(Q,R) \equiv \operatorname{off}[R^\top \widehat{W}_k Q],$$

where $\operatorname{off} A \equiv A - \operatorname{diag} A$ and we have used the fact that the Frobenius norm is invariant under rotations. Denote the (i,j)th entries of Q and R as q_{ij} and r_{ij}, respectively, and let $\delta_{ij} \equiv 1\{i=j\}$ denote Kronecker's delta. Then the Lagrangian is

$$\widehat{L}(Q,R) \equiv \sum_{k=1}^{\kappa} \big\| \widehat{O}_k(Q,R) \big\|_F^2 + \sum_{i,j=1}^{p} \lambda_{ij} \left\{ \sum_{\ell=1}^{p} r_{\ell i} q_{\ell j} - \delta_{ij} \right\} + \sum_{i,j=1}^{p} \gamma_{ij} \left\{ \sum_{\ell=1}^{p} q_{i\ell} r_{j\ell} - \delta_{ij} \right\},$$

where $\{\lambda_{ij}\}$ and $\{\gamma_{ij}\}$ are sets of Lagrange multipliers. Let $\circ$ denote the Hadamard product. To compute the vector of first derivatives with respect to $\operatorname{vec} Q$ we use the fact that $\|A\|_F^2 = \operatorname{trace} AA^\top$ and that

$$\operatorname{vec}[\operatorname{off} A] = \operatorname{vec}[(\mathrm{J}_p - \mathrm{I}_p) \circ A] = \operatorname{diag}[\operatorname{vec}(\mathrm{J}_p - \mathrm{I}_p)] \operatorname{vec} A$$

for any $p \times p$ matrix, together with elementary properties of the trace operator and its derivative (see, e.g., Magnus and Neudecker 2007, Chapter 9). By an application of the chain rule, it is easily

seen that

$$\frac{\partial \widehat{L}(Q,R)}{\partial \operatorname{vec} Q} = 2\sum_{k=1}^{\kappa} \left(\mathrm{I}_p \otimes \widehat{W}_k^\top R\right) \operatorname{vec} \widehat{O}_k(Q,R) + \left(\mathrm{I}_p \otimes R\right) \operatorname{vec} \Lambda + \left(R^\top \otimes \mathrm{I}_p\right) \operatorname{vec} \Gamma, \tag{A.2}$$

where Λ and Γ are $p \times p$ matrices that collect the Lagrange multipliers $\{\lambda_{ij}\}$ and $\{\gamma_{ij}\}$, respectively. To compute the first-derivative vector of the Lagrangian with respect to R we further rely on the $p \times p$ commutation matrix (e.g., Magnus and Neudecker 2007, pp. 46–48), K_p. The commutation matrix is defined through the equality $\operatorname{vec} A^\top = \mathrm{K}_p \operatorname{vec} A$, where A is any $p \times p$ matrix, and satisfies $\mathrm{K}_p = \mathrm{K}_p^\top = \mathrm{K}_p^{-1}$. We obtain

$$\frac{\partial \widehat{L}(Q,R)}{\partial \operatorname{vec} R} = 2\sum_{k=1}^{\kappa} \mathrm{K}_{p^2} \left(\widehat{W}_k Q \otimes \mathrm{I}_p\right) \operatorname{vec} \widehat{O}_k(Q,R) + \mathrm{K}_{p^2} \left(Q \otimes \mathrm{I}_p\right) \operatorname{vec} \Lambda + \mathrm{K}_{p^2} \left(\mathrm{I}_p \otimes Q^\top\right) \operatorname{vec} \Gamma.$$

Solve $\partial \widehat{L}(Q,R)/\partial \operatorname{vec} R = 0$ for $\operatorname{vec} \Lambda$ and enforce $RQ^\top = \mathrm{I}_p$ to see that

$$\operatorname{vec} \Lambda = \left(Q^{-1} \otimes \mathrm{I}_p\right) \operatorname{vec} \Gamma - 2\sum_{k=1}^{\kappa} \left(Q^{-1} \widehat{W}_k Q \otimes R\right) \operatorname{vec} \widehat{O}_k(Q,R).$$

Substitute this result back into (A.2) to dispense with the Lagrange multipliers. Enforce the constraints by replacing $R^\top$ by Q^{-1} to arrive at

$$\widehat{M}(Q) \equiv \frac{\partial \widehat{L}(Q,Q^{-\top})}{\partial \operatorname{vec} Q} = 2\sum_{k=1}^{\kappa} \left[\left(\mathrm{I}_p \otimes Q^{-1}\widehat{W}_k\right)^\top - \left(Q^{-1}\widehat{W}_k Q \otimes Q^{-\top}\right)\right] \operatorname{vec} \widehat{O}_k(Q,Q^{-\top}). \tag{A.3}$$

This function is the first derivative of the concentrated problem in Theorem 2 defining $\widehat{Q}$, and $\widehat{Q}$ solves

$$\widehat{M}(Q) = 0,$$

the associated score equation. $\widehat{M}(Q)$ is a plug-in version of the derivative of the population objective function in Theorem 1, $M(Q)$, say. Note that

$$M(Q) = 2\sum_{k=1}^{\kappa} \left[\left(\mathrm{I}_p \otimes Q^{-1}W_k\right)^\top - \left(Q^{-1}W_k Q \otimes Q^{-\top}\right)\right] \operatorname{vec} O_k(Q), \qquad O_k(Q) \equiv \operatorname{off}[Q^{-1}W_k Q],$$

which differs from $\widehat{M}(Q)$ only in its dependence on W_k rather than on $\widehat{W}_k$. Introduce the matrices of second derivatives

$$M_Q(Q) \equiv \frac{\partial M(Q)}{\partial \operatorname{vec}[Q]^\top}, \qquad M_{W_k}(Q) \equiv \frac{\partial M(Q)}{\partial \operatorname{vec}[W_k]^\top}.$$

A Taylor expansion of $\widehat{M}(\widehat{Q}) = 0$ around $\operatorname{vec} Q_0$ and the $\operatorname{vec} W_k$, together with the $\sqrt{n}$-consistency of their respective estimators, gives

$$\widehat{M}(\widehat{Q}) = 0 = M(Q_0) + \sum_{k=1}^{\kappa} M_{W_k}(Q_0) \left(\operatorname{vec} \widehat{W}_k - \operatorname{vec} W_k\right) + M_Q(Q_0) \left(\operatorname{vec} \widehat{Q} - \operatorname{vec} Q_0\right) + o_P(1/\sqrt{n}),$$

from which, in turn, follows the asymptotically-linear representation

$$\sqrt{n}\operatorname{vec}(\widehat{Q}-Q_0) = -M_Q(Q_0)^{-1}\sum_{k=1}^{\kappa} M_{W_k}(Q_0)\,\sqrt{n}\operatorname{vec}(\widehat{W}_k - W_k) + o_P(1/\sqrt{n}).$$

To establish Theorem 2 it remains only to verify that (i) $\frac{1}{2}M_Q(Q_0) = H$ and (ii) $\frac{1}{2}M_{W_k}(Q_0) = G_k$. For (i), note that

$$\begin{aligned}\frac{1}{2}M_Q(Q) = &\left\{\left(\mathrm{I}_p \otimes (Q^{-1}W_k)^\top\right) - \left(Q^{-1}W_kQ \otimes Q^{-\top}\right)\right\}\frac{\partial \operatorname{vec}[O_k(Q)]}{\partial \operatorname{vec}[Q]^\top}\\ &+ \operatorname{vec}[O_k(Q)^\top \otimes \mathrm{I}_p]\left\{\frac{\partial \operatorname{vec}[\mathrm{I}_p \otimes (Q^{-1}W_k)^\top]}{\partial \operatorname{vec}[Q]^\top} - \frac{\partial \operatorname{vec}[Q^{-1}W_kQ \otimes Q^{-\top}]}{\partial \operatorname{vec}[Q]^\top}\right\},\end{aligned}$$

and that the second right-hand side term vanishes when $Q = Q_0$. A calculation then establishes that

$$\frac{\partial \operatorname{vec} O_k(Q)}{\partial \operatorname{vec}[Q]^\top} = \operatorname{diag}[\operatorname{vec}(\mathrm{J}_p - \mathrm{I}_p)]\left[\left(\mathrm{I}_p \otimes Q^{-1}W_k\right) - \left((Q^{-1}W_kQ)^\top \otimes Q^{-1}\right)\right],$$

from which (i) follows after simplifying the resulting expression using that $Q_0^{-1}W_k = D_kQ_0^{-1}$. Also, because

$$\frac{\partial \operatorname{vec} O_k(Q)}{\partial \operatorname{vec}[W_k]^\top} = \operatorname{diag}[\operatorname{vec}(\mathrm{J}_p - \mathrm{I}_p)]\left(Q^\top \otimes Q^{-1}\right),$$

(ii) follows in the same way. □

Proof of Theorem 3. Because both $\|\widehat{W}_k - W_k\| = O_P(1/\sqrt{n})$ and $\|\widehat{Q} - Q_0\| = O_P(1/\sqrt{n})$ hold,

$$\widehat{Q}^{-1}\widehat{W}_k\widehat{Q} - Q_0^{-1}W_kQ_0 = (\widehat{Q}-Q_0)^{-1}W_kQ_0 + Q_0^{-1}(\widehat{W}_k - W_k)Q_0 + Q_0^{-1}W_k(\widehat{Q}-Q_0) + o_P(1/\sqrt{n})$$

follows from a linearization. For the first right-hand side term, because matrix inversion is a continuous transformation, the delta method can further be applied to yield

$$\operatorname{vec}\left((\widehat{Q}-Q_0)^{-1}W_kQ_0\right) = -\left(Q_0^\top W_k^\top \otimes \mathrm{I}_p\right)\left(Q_0^{-\top} \otimes Q_0^{-1}\right)\operatorname{vec}\left(\widehat{Q}-Q_0\right) = -\left(D_k \otimes Q_0^{-1}\right)\operatorname{vec}\left(\widehat{Q}-Q_0\right).$$

The remaining right-hand side terms are already linear in the estimators $\widehat{Q}$ and $\widehat{W}_k$. As $\operatorname{diag} A = \mathrm{I}_p \circ A$ for any compatible matrix A,

$$\operatorname{vec}(\widehat{D}_k - D_k) = \operatorname{diag}[\operatorname{vec}\mathrm{I}_p]\left\{E_k^\top \operatorname{vec}\left(\widehat{Q}-Q_0\right) + \left(Q_0^\top \otimes Q_0^{-\top}\right)\operatorname{vec}(\widehat{W}_k - W_k)\right\} + o_P(1/\sqrt{n}),$$

as claimed. □

Proof of Proposition 3. It suffices to consider the case with $q = 3$. Fix i throughout. Denote the slabs of $\mathbb{G}$ in the ith direction as S_{ik} and let $\widehat{S}_{ik}$ be their estimators. The proof consists of two steps. We first derive integrated squared-error and uniform convergence rates for the infeasible estimator that assumes the whitening can be performed without statistical noise. We then show that the additional noise in the feasible estimator is asymptotically negligible.

The infeasible estimator is given by

$$\widetilde{f}_{ij} \equiv \varphi_{\kappa_i}^\top \widetilde{\gamma}_{ij},$$

where the coefficient vector $\widetilde{\gamma}_{ij}$ is constructed from $\widetilde{D}_{ik} \equiv \operatorname{diag}[(Q_i^{-1}U_i)\,\widehat{S}_{ik}\,(V_i^\top Q_i)]$, where U_i and V_i are constructed from the singular-value decomposition of $\mathbb{G}(\{i_1, i_2\})$. The feasible estimator, in contrast, equals

$$\widehat{f}_{ij} = \varphi_{\kappa_i}^\top \widehat{\gamma}_{ij},$$

where the coefficient vector $\widehat{\gamma}_{ij}$ is constructed from $\widehat{D}_{ik} = \operatorname{diag}[(\widehat{Q}_i^{-1}\widehat{U}_i)\,\widehat{S}_{ik}\,(\widehat{V}_i^\top \widehat{Q}_i)]$, using obvious notation.

We begin by showing that $\|\widetilde{\gamma}_{ij} - \gamma_{ij}\|_F = O_P(\sqrt{\kappa_i/n})$. The convergence rates for $\widetilde{f}_{ij}$ will then follow easily. Write $s_{k_1k_2k}$ for the (k_1, k_2)th entry of S_{ik} and let $\widehat{s}_{k_1k_2k}$ be its estimator. First observe that, for any k,

$$\begin{aligned}
\mathbb{E}\left\|\widehat{S}_{ik} - S_{ik}\right\|_F^2 &= \sum_{k_1=1}^{\kappa_{i_1}} \sum_{k_2=1}^{\kappa_{i_2}} \mathbb{E}\left[(\widehat{s}_{k_1k_2k} - s_{k_1k_2k})^2\right] \\
&= \sum_{k_1=1}^{\kappa_{i_1}} \sum_{k_2=1}^{\kappa_{i_2}} \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=1}^{n} \phi_{k_1}(y_{mi_1})\rho(y_{mi_1})\phi_{k_2}(y_{mi_2})\rho(y_{mi_2})\phi_k(y_{mi})\rho(y_{mi}) - s_{k_1k_2k}\right)^2\right] \\
&= \sum_{k_1=1}^{\kappa_{i_1}} \sum_{k_2=1}^{\kappa_{i_2}} \frac{\mathbb{E}\left[\phi_{k_1}(y_{i_1})^2\rho(y_{i_1})^2\phi_{k_2}(y_{i_2})^2\rho(y_{i_2})^2\phi_k(y_i)^2\rho(y_i)^2\right] - s_{k_1k_2k}^2}{n} \\
&\le \sum_{k_1=1}^{\kappa_{i_1}} \sum_{k_2=1}^{\kappa_{i_2}} \frac{\sum_{j=1}^{p} \pi_j \left(\int_{-1}^{1} \psi(u)^2\rho(u)^2 f_{ij}(u)\,\mathrm{d}u\right) - s_{k_1k_2k}^2}{n}.
\end{aligned}$$

As the $\psi^2\rho^2 f_{ij}$ are integrable and the Fourier coefficients $s_{k_1k_2k_3}$ are square summable, we have that $\mathbb{E}\left\|\widehat{S}_{ik} - S_{ik}\right\|_F^2 = O(1/n)$ uniformly in k. Hence, $\sum_{k=1}^{\kappa_i} \mathbb{E}\left\|\widehat{S}_{ik} - S_{ik}\right\|_F^2 = O_P(\kappa_i/n)$ follows from Markov's inequality, and

$$\|\widetilde{\gamma}_{ij} - \gamma_{ij}\|_F^2 \le \sum_{k=1}^{\kappa_i} \left\|\widetilde{D}_{ik} - D_{ik}\right\|_F^2 \le \left\|Q_i^{-1}U_i \otimes Q_i^\top V_i\right\|_F^2 \sum_{k=1}^{\kappa_i} \left\|\widehat{S}_{ik} - S_{ik}\right\|_F^2 = O_P(\kappa_i/n)$$

follows by the Cauchy-Schwarz inequality. This establishes the rate result on the Fourier coefficients sought for. Now turn to the convergence rates. By orthonormality of the ϕ_i,

$$\left\|\widetilde{f}_{ij} - f_{ij}\right\|_2^2 = \left\|\widetilde{f}_{ij} - \operatorname{Proj}_{\kappa_i} f_{ij}\right\|_2^2 + \left\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\right\|_2^2 = \|\widetilde{\gamma}_{ij} - \gamma_{ij}\|_F^2 + \left\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\right\|_2^2.$$

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 4,

$$\left\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\right\|_2^2 \le \int_{-1}^{1} \left\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\right\|_\infty^2 \rho(u)\,\mathrm{d}u = O(\kappa_i^{-2\beta})$$

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

$$\left\|\widetilde{f}_{ij} - f_{ij}\right\|_\infty \le \left\|\widetilde{f}_{ij} - \operatorname{Proj}_{\kappa_i} f_{ij}\right\|_\infty + \left\|\operatorname{Proj}_{\kappa_i} f_{ij} - f_{ij}\right\|_\infty.$$

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 $\|\widetilde{\gamma}_{ij}-\gamma_{ij}\|_F$ in the second, the first right-hand side term satisfies

$$\big\|\widetilde{f}_{ij}-\mathrm{Proj}_{\kappa_i}\,f_{ij}\big\|_\infty \le \big\|\sqrt{\varphi_{\kappa_i}^\top\varphi_{\kappa_i}}\big\|_\infty\,\|\widetilde{\gamma}_{ij}-\gamma_{ij}\| = O\big(\zeta_{\kappa_i}\big)\;O_P\big(\sqrt{\kappa_i/n}\big).$$

By Assumption 4, $\big\|\,\mathrm{Proj}_{\kappa_i}\,f_{ij}-f_{ij}\big\|_\infty = O({\kappa_i}^{-\beta})$. 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_i and (U_i,V_i) implies that

$$\|\widehat{\gamma}_{ij}-\widetilde{\gamma}_{ij}\|_F = O_P(1/\sqrt{n})+O_P(\sqrt{\kappa_i}/n). \tag{A.4}$$

By the Cauchy-Schwarz inequality,

$$\big\|\widehat{\gamma}_{ij}-\widetilde{\gamma}_{ij}\big\|_F^2 \le \sum_{k=1}^{\kappa_i}\big\|\widehat{D}_{ik}-\widetilde{D}_{ik}\big\|_F^2 \le \big\|\widehat{Q}_i^{-1}\widehat{U}_i\otimes\widehat{Q}_i^\top\widehat{V}_i-Q_i^{-1}U_i\otimes Q_i^\top V_i\big\|_F^2\;\sum_{k=1}^{\kappa_i}\big\|\widehat{S}_{ik}\big\|_F^2.$$

Because both Q_i and (U_i,V_i) are $\sqrt{n}$-consistent, $\big\|\widehat{Q}_i^{-1}\widehat{U}_i\otimes\widehat{Q}_i^\top\widehat{V}_i-Q_i^{-1}U_i\otimes Q_i^\top V_i\big\|_F^2 = O_P(1/n)$. Also, from above, we have that

$$\sum_{k=1}^{\kappa_i}\big\|\widehat{S}_{ik}\big\|_F^2 \le 2\sum_{k=1}^{\kappa_i}\big\|S_{ik}\big\|_F^2+2\sum_{k=1}^{\kappa_i}\big\|\widehat{S}_{ik}-S_{ik}\big\|_F^2 = O(1)+O_P(\kappa_i/n).$$

Together, these results imply (A.4). Then

$$\|\widehat{f}_{ij}-f_{ij}\|_2^2 \le 2\|\widehat{\gamma}_{ij}-\widetilde{\gamma}_{ij}\|^2+2\|\widetilde{f}_{ij}-f_{ij}\|_2^2.$$

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 $\widehat{\gamma}_{ij}$ and $\widetilde{\gamma}_{ij}$ 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

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

And, again,

$$\|\widehat{f}_{ij}-\widetilde{f}_{ij}\|_\infty \le \big\|\sqrt{\varphi_{\kappa_i}^\top\varphi_{\kappa_i}}\big\|_\infty\,\big\|\widehat{\gamma}_{ij}-\widetilde{\gamma}_{ij}\big\| = O_P(\zeta_{\kappa_i}/\sqrt{n})+O_P(\zeta_{\kappa_i}\sqrt{\kappa_i}/n),$$

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

REFERENCES

Aitchison, J. and Aitken, G. C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420.

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.

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

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5** 111–150.

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. (2013). Nonparametric estimation of finite mixtures. Discussion Paper No 2013-09, Department of Economics, Sciences Po.

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

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

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

Clenshaw, C. W. and Curtis, A. R. (1960). A method for numerical integration on an automatic computer. *Numerical Mathematics* **2** 197–205.

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing* **36** 287–314.

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

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.

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

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.

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

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

Gassiat, E. and Rousseau, J. (2013). Non parametric finite translation mixtures with dependent regime. Mimeo.

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. and Lundy, M. E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In *Research Methods for Multimode Data Analysis* (H. G. Law, C. W. S. Jr., J. Hattie and R. P. McDonald, eds.) 122–215. Praeger.

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

Henry, M., Kitamura, Y. and Salanié, B. (2013). Partial identification of finite mixtures in econometric models. *Quantitative Economics*, to appear.

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.

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

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

Iferroudjene, R., Meraim, K. A. and Belouchrani, A. (2009). A new Jacobi-like method for joint diagonalization

of arbitrary non-defective matrices. *Applied Mathematics and Computation* **211** 363–373.

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.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of measurement error models using multiple indicators. *Journal of Multivariate Analysis* **65** 139–165.

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.

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

Magnus, J. R. and Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley Series in Probability and Statistics*. John Wiley & Sons.

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.

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

Sorensen, M., De Lathauwer, L., Comon, P., Icart, S. and Deneire, L. (2013). Canonical polyadic decomposition with a columnwise orthonormal factor matrix. *SIAM journal of matrix analysis and applications* **33** 1190–1213.

CEMFI and University of Chicago
1126 E. 59th Street
Chicago, IL 60637
U.S.A.
E-mail: bonhomme@cemfi.es

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

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