Estimating Multivariate Latent-Structure Models

Stéphane Bonhomme, Koen Jochmans, Jean-Marc Robin

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

Revised on August 19, 2015

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 joint-diagonalization problem. Algorithms are available for computation and we derive distribution theory. We further develop asymptotic theory for orthogonal-series estimators of component densities in mixture models and emission densities in hidden Markov models.

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

*Supported by European Research Council grant ERC-2010-StG-0263107-ENMUH.

†Supported by Sciences Po's SAB grant 'Nonparametric estimation of finite mixtures'.

‡Supported by European Research Council grant ERC-2010-AdG-269693-WASP and by Economic and Social Research Council grant RES-589-28-0001 through the Centre for Microdata Methods and Practice.

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

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

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 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. However, their statistical properties—their consistency, convergence rates, and asymptotic distribution—are difficult to establish and are currently unknown.[1]

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

[1]There are results on inference in semi- and nonparametric finite-mixture models and hidden Markov models in several more restrictive settings. These include location models (Bordes, Mottelet and Vandekerkhove 2006; Hunter, Wang and Hettmansperger 2007; and Gassiat and Rousseau 2014), multivariate finite mixtures with identically distributed outcome variables (Hettmansperger and Thomas 2000; Bonhomme, Jochmans and Robin 2014), and two-component mixtures (Hall and Zhou 2003; Jochmans, Henry and Salanié 2014).

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 defined 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 find that our approach performs 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], Domanov and De Lathauwer [2013a;b; 2014b;a], Anandkumar et al. [2014], and Chiantini, Ottaviani and Vannieuwenhoven [2014; 2015]. Although our strategy has some similarity with this literature, both our conclusions and our simultaneous-diagonalization problem are different. Most importantly, our simultaneous-diagonalization formulation can deal with noise, making it useful as a tool for statistical inference.

In the context of multivariate finite mixtures of identically distributed variables, Kasahara and Shimotsu [2009] and Bonhomme, Jochmans and Robin [2014] also used (different) joint-diagonalization arguments to obtain nonparametric identification results. 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. 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 simulation experiments involving finite mixtures and a hidden Markov model. The supplementary

material (Bonhomme, Jochmans and Robin 2015) contains some additional results and discussion, as well as all technical proofs.

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 $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)'$ 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 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.

Suppose that the scalar random variable Y_i can take on a finite number κ_i of values. Let $\boldsymbol{p}_{ij} = (p_{ij1}, p_{ij2}, \ldots, p_{ij\kappa_i})'$ denote the probability distribution of Y_i given $Z = j$. Let $\otimes$ denote the outer (tensor) product. The joint probability distribution of $Y_1, Y_2, \ldots, Y_q$ given $Z = j$ then is the q-way table

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

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

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

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, $\{\boldsymbol{p}_{ij}\}$ 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 $\boldsymbol{p}_{ij}$. Moreover, for any collection of vector-valued transformations $y \mapsto \boldsymbol{\chi}_i(y)$ we have

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

provided the expectation exists. Of course, identification of linear functionals follows from identification of the component distributions, but (2.2) can be useful for the construction of estimators. To illustrate this 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. Let f_{ij} be the density of Y_i given $Z = j$. In this case, the q-variate finite-mixture model with r latent classes states that the joint density function of the outcomes $Y_1, Y_2, \ldots, Y_q$ factors as

$$\sum_{j=1}^{r} \pi_j \prod_{i=1}^{q} f_{ij}, \tag{2.3}$$

again for mixing proportions $\pi_1, \pi_2, \ldots, \pi_r$. This is an infinite-dimensional version of (2.1). Setting $\boldsymbol{\chi}_i$ in (2.2) to a set of indicators that partition the state space of Y_i yields a decomposition as in (2.1) for a discretized version of the mixture model. This approach has been used by Allman, Matias and Rhodes [2009] and Kasahara and Shimotsu [2014] in proving identification.

An alternative approach, which will prove convenient for the construction of density estimators, is as follows. Suppose that $(Y_1, Y_2, \ldots, Y_q)$ lives in the q-dimensional space $\mathscr{Y}^q \subseteq \mathscr{R}^q$. Let $L^2_\rho[\mathscr{Y}]$ be the space of functions that are square-integrable with respect to the weight function ρ on $\mathscr{Y}$, endowed with the inner product

$$\langle h_1, h_2 \rangle = \int_{\mathscr{Y}} h_1(y)\, h_2(y)\, \rho(y)\, \mathrm{d}y,$$

and the L^2_ρ-norm $\|h\|_2 = \sqrt{\langle h, h \rangle}$. Let $\{\varphi_k, k > 0\}$ be a class of functions that form a complete orthonormal basis for $L^2_\rho[\mathscr{Y}]$. When $\mathscr{Y}$ is compact, polynomials such as those belonging to the Jacobi class—e.g., Chebychev or Legendre polynomials—can serve this purpose. When $\mathscr{Y} = (-\infty, +\infty)$, Hermite polynomials are a natural choice.

Assume that $f_{ij} \in L^2_\rho[\mathscr{Y}]$. The projection of f_{ij} onto the subspace spanned by $\varphi_1, \varphi_2, \ldots, \varphi_\varkappa$ for any integer $\varkappa$ is

$$\mathrm{Proj}_\varkappa f_{ij} = \sum_{k=1}^{\varkappa} b_{ijk}\, \varphi_k,$$

where the

$$b_{ijk} = \langle \varphi_k, f_{ij} \rangle = E[\varphi_k(Y_i)\rho(Y_i)|Z = j]$$

are the (generalized) Fourier coefficients of f_{ij}. The projection converges to f_{ij} in L^2_ρ-norm, that is, $\|\mathrm{Proj}_\varkappa f_{ij} - f_{ij}\|_2 \to 0$ as $\varkappa \to \infty$. Such projections are commonly-used tools in the approximation of functions and underlie orthogonal-series estimators of densities.

The Fourier coefficients are not directly observable. For chosen integers $\kappa_1, \kappa_2, \ldots, \kappa_q$, define

$$\boldsymbol{b}_{ij} = E[\boldsymbol{\varphi}_{\kappa_i}(Y_i)\rho(Y_i)|Z = j],$$

where $\boldsymbol{\varphi}_{\kappa_i} = (\varphi_1, \varphi_2, \ldots, \varphi_{\kappa_i})'$, which are linear functionals of the f_{ij}. Then (2.2) yields

$$\mathbb{B} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} \boldsymbol{b}_{ij} \tag{2.4}$$

for $\mathbb{B} = E[\bigotimes_{i=1}^q \boldsymbol{\varphi}_{\kappa_i}(Y_i)\rho(Y_i)]$. 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 5 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. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)'$ be the stationary distribution of Z_i. Write $\boldsymbol{K}$ for the $r \times r$ matrix of transition probabilities; so $\boldsymbol{K}(j_1, j_2)$ is the probability of moving from state j_1 to state j_2. The observable scalar 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. This is a hidden Markov model with r latent states and q observable outcomes.

Suppose that Y_i is discrete and that its state space contains κ points of support. Write $\boldsymbol{p}_j$ for the probability vector of Y_i given $Z_i = j$, that is, the emission distributions. Let $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r)$ be the $\kappa \times r$ matrix of emission distributions and write $\boldsymbol{\Pi} = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_r)$. The Markovian assumption implies that Y_i and Z_{i-1} are independent given Z_i. Hence, the columns of the matrix

$$\boldsymbol{B} = \boldsymbol{P}\boldsymbol{K}' = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r)$$

contain the probability distributions of Y_i for given values of Z_{i-1}. Likewise, Y_i and Z_{i+1} are independent given Z_i, and so the matrix

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{K}\boldsymbol{\Pi}^{-1} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_r)$$

gives the distributions of Y_i for given values of Z_{i+1}. Finally, Y_{i-1}, Y_i, and Y_{i+1} are independent given Z_i. Thus, with $q = 3$ measurements, the hidden Markov model implies that the contingency table of (Y_1, Y_2, Y_3) factors as

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \, (\boldsymbol{a}_j \otimes \boldsymbol{p}_j \otimes \boldsymbol{b}_j). \tag{2.5}$$

A detailed derivation is provided in the supplementary material; also see Gassiat, Cleynen and Robin [2013, Theorem 2.1] and Allman, Matias and Rhodes [2009, Section 6.1] for alternative derivations. When $q > 3$ we may bin several outcomes together and proceed as before, by using the unfolding argument in Subsection 3.1.

Equation (2.5) shows that appropriate conditioning allows viewing the hidden Markov model as a finite-mixture model, thus casting it into the framework of finite mixtures with conditionally-independent (although not identically-distributed) outcomes as in (2.1). Here, the parameters of interest are the emission distributions $\{\boldsymbol{p}_j\}_{j=1}^{r}$ and the stationary distribution of the Markov chain $\boldsymbol{\pi}$, and also the matrix of transition probabilities $\boldsymbol{K}$.

When the Y_i are continuously distributed, (2.5) becomes a mixture as in (2.3), and we may again work with projections of the densities onto an orthogonal basis.

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

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

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

for some integer r, non-zero weights $\pi_1, \pi_2, \ldots, \pi_r$, and vectors $\boldsymbol{x}_{ij} \in \mathscr{R}^{\kappa_i \times 1}$.

Our interest lies in nonparametrically recovering $\{\boldsymbol{x}_{ij}\}$ and $\{\pi_j\}$ from knowledge of $\mathbb{X}$ and r. Clearly, these parameters are not unique, in general. For example, a permutation of the $\boldsymbol{x}_{ij}$ and π_j leaves $\mathbb{X}$ unaffected, and a common scaling of the $\boldsymbol{x}_{ij}$ combined with an inverse scaling of the π_j, too, does not change the q-way array. However, the work of Kruskal [1976; 1977], Sidiropoulos and Bro [2000], Jiang and Sidiropoulos [2004] and Domanov and De Lathauwer [2013a;b], among others, gives simple sufficient conditions for uniqueness of the decomposition up to these two indeterminacies. These conditions cannot be satisfied when $q < 3$.

While permutational equivalence of possible decompositions of $\mathbb{X}$ is an inherently unresolvable ambiguity, indeterminacy of the scale of the vectors $\boldsymbol{x}_{ij}$ is undesirable in many situations. Indeed, in arrays of the general form in (2.2), recovering the scale of the $\boldsymbol{x}_{ij}$ and the constants π_j is fundamental. In some cases natural scale restrictions may be present. Indeed, in (2.1) the $\boldsymbol{x}_{ij}$ 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] derived 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 that does not require knowledge of the scale of the $\boldsymbol{x}_{ij}$. 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

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

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. For example, in the mixture model in (2.1), lower-dimensional submodels are just 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. In more general situations, 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. Indeed, in contrast to when the array is a contingency table, here, there is no natural scale constraint on the $\boldsymbol{x}_{ij}$. So, summing the array in one direction does not yield an array that decomposes as in (3.2). Nonetheless, expectations concerning any subset of the random variables can still be computed in (2.2) and so submodels as defined in (3.2) are observable. In the supplementary material we adapt our main identification result (Theorem 1 below) to settings where submodels are not available and marginalizations are used instead.

Note that, throughout, we take r in (3.1) to be known. This ensures $\{\boldsymbol{x}_{ij}\}$ and $\{\pi_j\}$ to be unambiguously defined. For a 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. For example, in the multivariate mixture model in Section 2.1, r is the number of fitted mixture components and the rank is the smallest number of components that would allow us to write the joint distribution of the variables as a mixture that satisfies the required conditional-independence restriction as in (2.1). The rank need not be equal to r. Moreover, besides the factorization of $\mathbb{P}$ in terms of $\pi_1, \pi_2, \ldots, \pi_r$ and $\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}, \ldots, \boldsymbol{p}_{ir}\}$ in (2.1), there may exist a different set of, say, r' weights $\pi'_1, \pi'_2, \ldots, \pi'_{r'}$ and distributions $\{\boldsymbol{p}'_{i1}, \boldsymbol{p}'_{i2}, \ldots, \boldsymbol{p}'_{ir'}\}$ that also yield a representation of $\mathbb{P}$ as a mixture. Identifying the number of components is a difficult issue. Recent work by Kasahara and Shimotsu [2014] shows that a simple lower bound on the number of components is nonparametrically identified (and estimable).

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. To write this compactly, let $\odot$ be the Khatri-Rao product. Then, for vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$,

$$\bigodot_{i=1}^{q} \boldsymbol{a}_i = \boldsymbol{a}_1 \odot \boldsymbol{a}_2 \odot \cdots \odot \boldsymbol{a}_q$$

is the vector containing all interactions between the elements of the $\boldsymbol{a}_i$. The end result of iterated unfolding towards direction i, say, is a three-way array of the form

$$\sum_{j=1}^{r} \pi_j \left(\bigodot_{i_1 \in \mathcal{Q}_1} \boldsymbol{x}_{i_1 j} \otimes \boldsymbol{x}_{ij} \otimes \bigodot_{i_2 \in \mathcal{Q}_2} \boldsymbol{x}_{i_2 j} \right),$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are two index sets that partition $\{1, 2, \ldots, q\} \backslash \{i\}$. We will illustrate this in the context of density estimation in Section 5.

3.2. *Identification via simultaneous diagonalization.* We thus focus on a three-way array $\mathbb{X}$ of dimension $\kappa_1 \times \kappa_2 \times \kappa_3$ that factors as a tri-ad, that is,

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

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

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

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

Our first theorem concerns identification of the $\boldsymbol{X}_i$ as the eigenvalues of a set of matrices and is the cornerstone of our argument. The proof of this result is constructive and will be the basis for our estimator in Section 4 below.

THEOREM 1 (Columns of $\boldsymbol{X}_i$). *If $\boldsymbol{X}_{i_1}$ and $\boldsymbol{X}_{i_2}$ both have full column rank and $\mathbb{X}_{\{i_1,i_2\}}$ is observable, then $\boldsymbol{X}_{i_3}$ 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 $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{\kappa_3}$ denote these matrices for $i = 3$. So, the matrix $\boldsymbol{A}_k$ is obtained from $\mathbb{X}$ by fixing its third index to the value k, that is, $\boldsymbol{A}_k = \mathbb{X}(:, :, k)$, using obvious array-indexing notation. Also, let $\boldsymbol{A}_0 = \mathbb{X}_{\{1,2\}}$. Note that all of $\boldsymbol{A}_0$ and $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{\kappa_3}$ are observable matrices of dimension $\kappa_1 \times \kappa_2$.

The lower-dimensional submodel $\boldsymbol{A}_0$ has the structure

$$\boldsymbol{A}_0 = \boldsymbol{X}_1 \boldsymbol{\Pi} \boldsymbol{X}_2'. \tag{3.3}$$

Because the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ both have rank r and because all π_j are non-zero by definition, the matrix $\boldsymbol{A}_0$, too, has rank r. Therefore, it has a singular-value decomposition

$$\boldsymbol{A}_0 = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$$

for unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ of dimension $\kappa_1 \times r$ and $\kappa_2 \times r$, respectively, and a non-singular $r \times r$ diagonal matrix $\boldsymbol{S}$. Now construct $\boldsymbol{W}_1 = \boldsymbol{S}^{-1/2}\boldsymbol{U}'$ and $\boldsymbol{W}_2 = \boldsymbol{S}^{-1/2}\boldsymbol{V}'$. Then,

$$\boldsymbol{W}_1 \boldsymbol{A}_0 \boldsymbol{W}_2' = (\boldsymbol{W}_1 \boldsymbol{X}_1 \boldsymbol{\Pi}^{1/2})(\boldsymbol{W}_2 \boldsymbol{X}_2 \boldsymbol{\Pi}^{1/2})' = \boldsymbol{Q}\boldsymbol{Q}^{-1} = \boldsymbol{I}_r,$$

where $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix and $\boldsymbol{Q} = \boldsymbol{W}_1 \boldsymbol{X}_1 \boldsymbol{\Pi}^{1/2}$.

Moving on, each of $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{\kappa_3}$ has the form

$$\boldsymbol{A}_k = \boldsymbol{X}_1 \boldsymbol{\Pi} \boldsymbol{D}_k \boldsymbol{X}_2', \qquad \boldsymbol{D}_k = \operatorname{diag}_k \boldsymbol{X}_3,$$

where $\operatorname{diag}_k \boldsymbol{X}$ denotes the diagonal matrix whose diagonal equals the kth row of matrix $\boldsymbol{X}$. Applying the same transformation to $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{\kappa_3}$

yields the collection of $r \times r$ matrices

$$\boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2' = \boldsymbol{Q} \boldsymbol{D}_k \boldsymbol{Q}^{-1}. \tag{3.4}$$

So, the matrices $\{\boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2'\}$ are diagonalizable in the same basis, namely, the columns of matrix $\boldsymbol{Q}$. The associated eigenvalues $\{\boldsymbol{D}_k\}$ equal the columns of the matrix $\boldsymbol{X}_3$. 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 $\boldsymbol{W}_1 \boldsymbol{A}_{k_1} \boldsymbol{W}_2'$ and $\boldsymbol{W}_1 \boldsymbol{A}_{k_2} \boldsymbol{W}_2'$ 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 $\boldsymbol{X}_3$ 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 $\boldsymbol{X}_i$ and the weights on the diagonal of $\boldsymbol{\Pi}$. This is so because the matrix $\boldsymbol{\Pi}$ equally shows up in the lower-dimensional submodels, and so transforming $\boldsymbol{A}_k$ to $\boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2'$ absorbs the weights into the joint diagonalizer $\boldsymbol{Q}$ in (3.4).

Also note that the dimension of the matrices in (3.4) is $r \times r$, independent of the size of the original matrices $\boldsymbol{X}_i$. On the other hand, larger matrices $\boldsymbol{X}_i$ could be beneficial for identification, as it becomes easier for them to satisfy the requirement of full column rank.

The full-rank condition that underlies Theorem 1 has a simple testable implication. Indeed, by (3.3), it implies that the matrix $\boldsymbol{A}_0$ has rank r. As this matrix is observable, so is its rank and, hence, our key identifying assumption is refutable. In applications this can be done using any of a number of available rank tests. We refer to Kasahara and Shimotsu [2014] and Bonhomme, Jochmans and Robin [2014] for practical details on the implementation of such procedures.

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 $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ 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 $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ are all identified up to a common permutation of their columns.*

THEOREM 3 (Weights). *If $\boldsymbol{X}_i$ is identified up to a permutation of its columns and has full column rank, and if $\mathbb{X}_{\{i\}}$ is observable, then $\boldsymbol{\pi}$ is identified up to the same permutation.*

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

$$\mathbb{X}_{\{i\}} = \boldsymbol{X}_i \boldsymbol{\pi}.$$

Given $\boldsymbol{X}_i$, the one-dimensional submodel yields linear restrictions on the weight vector $\boldsymbol{\pi}$. Moreover, if $\boldsymbol{X}_i$ is known and has maximal column rank, these equations can be solved for $\boldsymbol{\pi}$, giving

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

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

In the supplement we apply Theorems 1–3 to the finite-mixture model and the hidden Markov model of Section 2 to obtain constructive proofs of identification.

4. Estimation by joint approximate diagonalization. The proof of 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 joint matrix diagonalization has recently received considerable attention in the field of independent component analysis, and 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.

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 $\boldsymbol{x}_{ij}$ by solving a sample version of the joint diagonalization problem in (3.4), 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 based on the proof of Theorem 2 (see the supplementary material).

4.1. *Estimator.* Consider a generic situation in which a set of κ $r \times r$ matrices $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_\kappa$ can be jointly diagonalized by an $r \times r$ invertible matrix $\boldsymbol{Q}_0$, that is,

$$\boldsymbol{C}_k = \boldsymbol{Q}_0 \boldsymbol{D}_k \boldsymbol{Q}_0^{-1}, \tag{4.1}$$

for diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_\kappa$. Knowledge of the joint eigenvectors implies knowledge of the eigenvalues, as

$$\boldsymbol{D}_k = \boldsymbol{Q}_0^{-1} \boldsymbol{C}_k \boldsymbol{Q}_0. \tag{4.2}$$

The matrix $\boldsymbol{Q}_0$ is not unique. Moreover, let $\operatorname{off} \boldsymbol{Q} = \boldsymbol{Q} - \operatorname{diag} \boldsymbol{Q}$ and let $\|\boldsymbol{Q}\|_F = \sqrt{\operatorname{trace}(\boldsymbol{Q}'\boldsymbol{Q})}$ denote the Frobenius norm. Then any solution to the least-squares problem

$$\min_{\boldsymbol{Q}} \sum_{k=1}^{\kappa} \|\operatorname{off}(\boldsymbol{Q}^{-1} \boldsymbol{C}_k \boldsymbol{Q})\|_F^2 \tag{4.3}$$

is a joint diagonalizer in the sense of (4.1). Each of these delivers the same set of eigenvalues in (4.2) (up to a joint permutation).

The statistical problem of interest in this section is to perform inference on the $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_\kappa$ when we only observe noisy versions of the input matrices $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_\kappa$, say $\widehat{\boldsymbol{C}}_1, \widehat{\boldsymbol{C}}_2, \ldots, \widehat{\boldsymbol{C}}_\kappa$. The sampling noise in the $\widehat{\boldsymbol{C}}_k$ prevents them from sharing the same set of eigenvectors. Indeed, in general, there does not exist a $\boldsymbol{Q}$ such that $\boldsymbol{Q}^{-1} \widehat{\boldsymbol{C}}_k \boldsymbol{Q}$ will be exactly diagonal for all k. For this, the least-squares formulation in (4.2)–(4.3) is important as it readily suggests using, say $\widehat{\boldsymbol{Q}}$, any solution to

$$\min_{\boldsymbol{Q} \in \mathscr{Q}} \sum_{k=1}^{\kappa} \|\operatorname{off}(\boldsymbol{Q}^{-1} \widehat{\boldsymbol{C}}_k \boldsymbol{Q})\|_F^2, \tag{4.4}$$

where $\mathscr{Q}$ is an appropriately-specified space of matrices to search over; see below. The estimator $\widehat{\boldsymbol{Q}}$ 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 $\widehat{\boldsymbol{Q}}$ the joint approximate-diagonalizer of $\widehat{\boldsymbol{C}}_1, \widehat{\boldsymbol{C}}_2, \ldots, \widehat{\boldsymbol{C}}_\kappa$. An estimator of the $\boldsymbol{D}_k$ (up to a joint permutation of their eigenvalues) then is

$$\widehat{\boldsymbol{D}}_k = \operatorname{diag}(\widehat{\boldsymbol{Q}}^{-1} \widehat{\boldsymbol{C}}_k \widehat{\boldsymbol{Q}}). \tag{4.5}$$

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.

4.2. *Asymptotic theory.* For our problem to be well-defined we assume that the matrix of joint eigenvectors is bounded. In (4.4), we may therefore restrict attention to the set of $r \times r$ matrices $\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_r)$ defined as

$$\mathscr{Q} = \{\boldsymbol{Q} : \det \boldsymbol{Q} = 1, \|\boldsymbol{q}_j\|_F = c \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m\}$$

for some $m \in (0, \infty)$. 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 (4.4). Let $\boldsymbol{Q}_*$ be any solution to (4.3) on $\mathscr{Q}$ and let $\mathscr{Q}_0 \subset \mathscr{Q}$ be the set of all matrices $\boldsymbol{Q}_*\boldsymbol{\Delta}\boldsymbol{\Theta}$ for permutation matrices $\boldsymbol{\Delta}$ and diagonal matrices $\boldsymbol{\Theta}$ whose diagonal entries are equal to 1 and -1 and have $\det \boldsymbol{\Theta} = 1$. Then $\mathscr{Q}_0$ is the set of solutions to (4.3) on $\mathscr{Q}$.

Construct the $r \times r\kappa$ matrix $\boldsymbol{C} = (\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_\kappa)$ by concatenation and define $\widehat{\boldsymbol{C}}$ similarly.

THEOREM 4 (Consistency). *If the set $\mathscr{Q}_0$ belongs to the interior of $\mathscr{Q}$, $\widehat{\boldsymbol{C}} = \boldsymbol{C} + o_p(1)$, and $\widehat{\boldsymbol{Q}} \in \mathscr{Q}$ satisfies*

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

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

Each $\boldsymbol{Q} \in \mathscr{Q}_0$ has associated with it a permutation matrix $\boldsymbol{\Delta}$ and a diagonal matrix $\boldsymbol{\Theta}$ as just defined so that $\boldsymbol{Q} = \boldsymbol{Q}_*\boldsymbol{\Delta}\boldsymbol{\Theta}$. Theorem 4 states that (up to a subsequence) we have that $\widehat{\boldsymbol{Q}} \xrightarrow{p} \boldsymbol{Q}_*\boldsymbol{\Delta}_0\boldsymbol{\Theta}_0$ for well-defined $\boldsymbol{\Delta}_0$ and $\boldsymbol{\Theta}_0$. We may then set $\boldsymbol{Q}_0 = \boldsymbol{Q}_*\boldsymbol{\Delta}_0\boldsymbol{\Theta}_0$ in (4.1). It then equally follows that

$$\widehat{\boldsymbol{D}}_k \xrightarrow{p} \boldsymbol{D}_k = \boldsymbol{\Delta}_0'\boldsymbol{D}_k^*\boldsymbol{\Delta}_0,$$

where $\boldsymbol{D}_k$ is as in (4.2) and $\boldsymbol{D}_k^* = \boldsymbol{Q}_*^{-1}\boldsymbol{C}_k\boldsymbol{Q}_*$, both of which are equal up to a permutation. Thus, the consistency of the eigenvalues (up to a joint permutation) follows from the consistency of the estimator of the input matrices $\boldsymbol{C}$.

To provide distribution theory, let

$$\boldsymbol{D}_{k_1} \ominus \boldsymbol{D}_{k_2} = (\boldsymbol{D}_{k_1} \otimes \boldsymbol{I}_{\dim \boldsymbol{D}_{k_2}}) - (\boldsymbol{I}_{\dim \boldsymbol{D}_{k_1}} \otimes \boldsymbol{D}_{k_2})$$

denote the Kronecker difference between the square matrices $\boldsymbol{D}_{k_1}$ and $\boldsymbol{D}_{k_2}$. Construct the $r^2 \times r^2\kappa$ matrix

$$\boldsymbol{T} = ((\boldsymbol{D}_1 \ominus \boldsymbol{D}_1), (\boldsymbol{D}_2 \ominus \boldsymbol{D}_2), \ldots, (\boldsymbol{D}_\kappa \ominus \boldsymbol{D}_\kappa))$$

by concatenation and let

$$\boldsymbol{G} = (\boldsymbol{I}_r \otimes \boldsymbol{Q}_0) \left(\sum_{k=1}^{\kappa} (\boldsymbol{D}_k \ominus \boldsymbol{D}_k)^2 \right)^+ \boldsymbol{T} \, (\boldsymbol{I}_\kappa \otimes \boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1}),$$

where $\boldsymbol{Q}^+$ is the Moore-Penrose pseudo inverse of $\boldsymbol{Q}$. Theorem 5 contains distribution theory for our estimator of the matrix of joint eigenvectors $\widehat{\boldsymbol{Q}}$ in (4.4).

THEOREM 5 (Asymptotic distribution). *If* $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F = O_p(n^{-1/2})$, *then*

$$\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}_0) = \boldsymbol{G}\,\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{C}} - \boldsymbol{C}) + o_p(1)$$

as $n \to \infty$.

If, further, $\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{C}} - \boldsymbol{C}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$ for some covariance matrix $\boldsymbol{V}$, Theorem 5 implies that

$$\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}\,\boldsymbol{V}\boldsymbol{G}')$$

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 $\boldsymbol{C}_k = \boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2'$, where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ follow from a singular-value decomposition of $\boldsymbol{A}_0$. An estimator of $\boldsymbol{C}_k$ can thus be constructed using a sample analog of $\boldsymbol{A}_0$ to estimate $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, together with a sample analog of $\boldsymbol{A}_k$. If the estimators of $\boldsymbol{A}_0$ and $\boldsymbol{A}_k$ are $\sqrt{n}$-consistent and asymptotically normal and all non-zero singular values of $\boldsymbol{A}_0$ are simple, then $\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{C}} - \boldsymbol{C}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$ holds. A detailed derivation of $\boldsymbol{V}$ 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 $\widehat{\boldsymbol{D}} = (\widehat{\boldsymbol{D}}_1, \widehat{\boldsymbol{D}}_2, \ldots, \widehat{\boldsymbol{D}}_\kappa)$, our estimator of the eigenvalues $\boldsymbol{D} = (\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_\kappa)$. To state it, let $\boldsymbol{S}_r = \operatorname{diag}(\operatorname{vec} \boldsymbol{I}_r)$ be an $r^2 \times r^2$ selection matrix; note that $\boldsymbol{S}_r \operatorname{vec} \boldsymbol{Q} = \operatorname{vec}(\operatorname{diag} \boldsymbol{Q})$. Let

$$\boldsymbol{H} = (\boldsymbol{I}_\kappa \otimes \boldsymbol{S}_r)\ (\boldsymbol{I}_\kappa \otimes \boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1}).$$

Theorem 6 follows.

THEOREM 6 (Asymptotic distribution). *If* $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F = O_p(n^{-1/2})$, *then*

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

as $n \to \infty$.

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

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

as $n \to \infty$.

5. Application to density estimation. With discrete outcomes, both the finite-mixture model in (2.1) and the hidden Markov model in (2.5) are finite dimensional. Further, the 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 on the asymptotic behavior of the eigenvalues from the previous section (i.e., Theorem 6) can directly be applied to deduce the large-sample behavior of the parameter estimates.

With continuous outcomes, as in (2.3), the main parameters of the model are density functions. Such an infinite-dimensional problem is not directly covered by the arguments from the previous section. Nonetheless, we will show that Theorem 5 can be used to obtain density estimators with standard asymptotic properties.

5.1. *Estimator.* We provide convergence rates and distribution theory for series estimators based on (2.4). By the results of Subsection 2.3, this also covers the estimation of emission densities in a hidden Markov model with continuous outcome variables. Recall from above that the projections

$$\operatorname{Proj}_{\kappa_i} f_{ij} = \boldsymbol{\varphi}_{\kappa_i}' \boldsymbol{b}_{ij}$$

yield the multilinear restrictions

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

where $\boldsymbol{\varphi}_{\kappa_i}$ is the vector containing the κ_i leading polynomials from the orthogonal system $\{\varphi_k, k>0\}$. As we will show, for fixed $\kappa_1, \kappa_2, \ldots, \kappa_q$, the array $\mathbb{B}$ provides sufficient information for nonparametric identification of Fourier coefficients through the associated joint diagonalizer. Moreover, in the asymptotic analysis, $\kappa_1, \kappa_2, \ldots, \kappa_q$ are all held fixed.

For the purpose of this section we may fix attention to a given index i. By unfolding $\mathbb{B}$ towards direction i we obtain the (equivalent) three-way array

$$\mathbb{B}_i = E\big[\boldsymbol{\phi}^{\mathcal{Q}_1} \otimes \boldsymbol{\phi}^{\mathcal{Q}_2} \otimes \boldsymbol{\varphi}_{\kappa_i}(Y_i)\rho(Y_i)\big],$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ partition the index set $\{1,2,\ldots,q\}\backslash\{i\}$ (see Section 3) and we have introduced the notational shorthand

$$\boldsymbol{\phi}^{\mathcal{Q}} = \bigodot_{i' \in \mathcal{Q}} \boldsymbol{\varphi}_{\kappa_{i'}}(Y_{i'})\rho(Y_{i'}).$$

The array $\mathbb{B}_i$ can be analyzed using our diagonalization approach. Following the notation from the proof of Theorem 1, the two-dimensional submodel associated with $\mathbb{B}_i$ is the matrix

$$\boldsymbol{A}_0 = E\big[\boldsymbol{\phi}^{\mathcal{Q}_1} \otimes \boldsymbol{\phi}^{\mathcal{Q}_2}\big],$$

while the array $\mathbb{B}_i$ itself consists of the first κ_i matrices of the set $\{\boldsymbol{A}_k, k>0\}$, where

$$\boldsymbol{A}_k = E\big[\big(\boldsymbol{\phi}^{\mathcal{Q}_1} \otimes \boldsymbol{\phi}^{\mathcal{Q}_2}\big)\, \varphi_k(Y_i)\rho(Y_i)\big].$$

All these matrices are of dimension $\prod_{i_1 \in \mathcal{Q}_1} \kappa_{i_1} \times \prod_{i_2 \in \mathcal{Q}_2} \kappa_{i_2}$. A singular-value decomposition of $\boldsymbol{A}_0$ provides matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ so that the κ_i matrices $\boldsymbol{W}_1\boldsymbol{A}_k\boldsymbol{W}_2'$ are jointly diagonalizable by, say, $\boldsymbol{Q}$. From the proof of Theorem 1, the matrix $\boldsymbol{Q}$ is unique (up to the usual normalizations on the sign and norm of its columns and a joint permutation of the columns, as discussed before) as soon as the conditions in Theorem 1 are satisfied.

Given $\boldsymbol{Q}$, we can compute

$$\boldsymbol{Q}^{-1}(\boldsymbol{W}_1\boldsymbol{A}_k\boldsymbol{W}_2')\,\boldsymbol{Q} = \operatorname{diag}(b_{i1k}, b_{i2k}, \ldots, b_{irk}),$$

where, recall, $b_{ijk} = E[\varphi_k(Y_i)\rho(Y_i)|Z = j]$ for any integer k (including those k that exceed κ_i). Equivalently, the kth Fourier coefficient of f_{ij} can be written as

$$b_{ijk} = \boldsymbol{e}_j' \left(\boldsymbol{Q}^{-1}(\boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2')\, \boldsymbol{Q}\right) \boldsymbol{e}_j, \tag{5.1}$$

where $\boldsymbol{e}_j$ is the $r \times 1$ selection vector whose jth entry is equal to one and its other entries are all equal to zero.

Our orthogonal-series estimator of f_{ij} is based on sample analogs of the b_{ijk} in (5.1). We estimate the array $\mathbb{B}$ as

$$\widehat{\mathbb{B}} = n^{-1} \sum_{m=1}^{n} \bigotimes_{i=1}^{q} \boldsymbol{\varphi}_{\kappa_i}(Y_{im})\rho(Y_{im}),$$

where $\{Y_{1m}, Y_{2m}, \ldots, Y_{qm}\}_{m=1}^n$ is a size-n sample drawn at random from the mixture model. From this we estimate b_{ijk} for any k as

$$\hat{b}_{ijk} = \boldsymbol{e}_j' \left(\widehat{\boldsymbol{Q}}^{-1}(\widehat{\boldsymbol{W}}_1 \widehat{\boldsymbol{A}}_k \widehat{\boldsymbol{W}}_2')\, \widehat{\boldsymbol{Q}}\right) \boldsymbol{e}_j = n^{-1} \sum_{m=1}^{n} \boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j \; \varphi_k(Y_{im})\rho(Y_{im}),$$

using obvious notation to denote sample counterparts in the first expression and introducing the matrix

$$\widehat{\boldsymbol{\Omega}}_m = \widehat{\boldsymbol{Q}}^{-1}(\widehat{\boldsymbol{W}}_1 \big(\boldsymbol{\phi}_m^{\mathcal{Q}_1} \otimes \boldsymbol{\phi}_m^{\mathcal{Q}_2}\big)\, \widehat{\boldsymbol{W}}_2')\, \widehat{\boldsymbol{Q}}$$

in the second expression; here, we let $\boldsymbol{\phi}_m^{\mathcal{Q}} = \bigodot_{i' \in \mathcal{Q}} \boldsymbol{\varphi}_{\kappa_{i'}}(Y_{i'm})\rho(Y_{i'm})$. The associated orthogonal-series estimator of $f_{ij}(y)$ for some chosen integer $\varkappa$ is

$$\begin{aligned} \hat{f}_{ij}(y) &= \sum_{k=1}^{\varkappa} \hat{b}_{ijk}\, \varphi_k(y) \\ &= n^{-1} \sum_{m=1}^{n} \boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j \; \sum_{k=1}^{\varkappa} \varphi_k(Y_{im})\varphi_k(y)\rho(Y_{im}). \end{aligned} \tag{5.2}$$

Note that, in the absence of $\boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j$, this expression collapses to a standard series estimator of the marginal density of Y_i. Hence, the term $\boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j$ can be understood as a weight that transforms this estimator into one of the conditional density of Y_i given $Z = j$. Equation (5.2) generalizes the kernel estimator of Bonhomme, Jochmans and Robin [2014]. The term $\boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j$

plays the same role as the posterior classification probability (normalized to sum up to one across observations) in the EM algorithm as well as in its nonparametric version (Levine, Hunter and Chauveau 2011, Equations (15)–(17)). A computational advantage here is that the series estimator is available in closed form once $\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j$ has been computed while EM requires iterative computation of density estimates and classification probabilities until convergence.

A natural way of choosing the number of series terms in (5.2) would be by minimizing the squared L^2_ρ-loss,

$$\|\hat{f}_{ij} - f_{ij}\|_2^2,$$

as a function of $\varkappa$. In the supplement we show that an empirical counterpart of this criterion (up to terms that do not involve $\varkappa$) is

$$\sum_{k=1}^{\varkappa} \hat{b}_{ijk}^2 - \frac{2n^{-1}}{n-1}\sum_{m=1}^{n}\sum_{o\neq m} \boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j\,\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_o\boldsymbol{e}_j \sum_{k=1}^{\varkappa}\varphi_k(Y_{io})\varphi_k(Y_{im})\rho(Y_{io})\rho(Y_{im}).$$

Apart from the weight functions, this is the usual cross-validation objective for orthogonal-series estimators (Hall 1987).

Before turning to the statistical properties of $\hat{f}_{ij}$ we note that, although we maintain a hard thresholding procedure in (5.2), our approach can equally be combined with other popular smoothing policies that shrink the impact of higher-order Fourier coefficients; see Efromovich [1999, Chapter 3] for a discussion on such policies.

5.2. *Asymptotic theory.* Under mild conditions, the series estimator in (5.2) exhibits standard large-sample behavior. The precise conditions depend on the choice of orthogonal system, i.e., $\{\varphi_k, k > 0\}$. We give two sets of conditions that cover the most popular choices.

When the component densities are supported on compact intervals we can restrict attention to $[-1, 1]$ without loss of generality; translation to generic compact sets is straightforward. In this case we will allow for polynomial systems that satisfy the following general requirements. Here and later we let $\|\cdot\|_\infty$ denote the supremum norm.

A.1 The sequence $\{\varphi_k, k > 0\}$ is dominated by a function ψ, which is continuous on $(-1, 1)$ and positive almost everywhere on $[-1, 1]$. ρ,

$\psi\rho$, and $\psi^2\rho$ are integrable, and there exists a sequence of constants $\{\zeta_\varkappa, \varkappa > 0\}$ so that $\|\sqrt{\boldsymbol{\varphi}'_\varkappa \boldsymbol{\varphi}_\varkappa}\|_\infty \leq \zeta_\varkappa$.

These conditions are rather weak. They are satisfied for the popular class of Jacobi polynomials, for example, which includes Chebychev polynomials of the first kind, Chebychev polynomials of the second kind, and Legendre polynomials.

In this case we will need the following regularity from the component densities.

A.2 The $(\psi\rho)^4 f_{ij}$ are integrable.

The weaker requirement that the $(\psi\rho)^2 f_{ij}$ are integrable will suffice to obtain the convergence rates in Theorem 7 below, but **A.2** will be needed to obtain the pointwise asymptotic-normality result in Theorem 8.

When the component densities are supported on the whole real line, we will take $\{\varphi_k, k > 0\}$ to be the orthonormalized system of Hermite functions.

B.1 The sequence $\{\varphi_k, k > 0\}$ has members

$$\varphi_k(y) = 2^{-(k-1)/2}\,((k-1)!)^{-1/2}\pi^{-1/4}\,e^{-y^2/2}\,h_{k-1}(y),$$

where $\{h_k, k \geq 0\}$ is the system of the Hermite polynomials, in which case $\|\sqrt{\boldsymbol{\varphi}'_\varkappa \boldsymbol{\varphi}_\varkappa}\|_\infty \leq \zeta_\varkappa$ for $\zeta_\varkappa \propto \sqrt{\varkappa}$.

We will also impose the following regularity and smoothness conditions.

C.1 The f_{ij} are continuous.

C.2 $\|\mathrm{Proj}_\varkappa f_{ij} - f_{ij}\|_\infty = O(\varkappa^{-\beta})$ for some constant $\beta \geq 1$.

C.3 The singular values of $\boldsymbol{A}_0$ are all simple.

Convergence in L^2_ρ-norm implies that $\lim_{\varkappa\to\infty} \sum_{k=1}^{\varkappa} b_{ijk}^2$ is finite, and so that the Fourier coefficient associated with φ_k shrinks to zero as $k \to \infty$. The constant 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. Simplicity of the singular values of $\boldsymbol{A}_0$ holds generically and is used here to ensure that the matrices $\boldsymbol{W}_1, \boldsymbol{W}_2$ are continuous transformations of $\boldsymbol{A}_0$. This is a technical requirement used to derive the convergence rates of their plug-in estimators.

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

THEOREM 7 (Convergence rates). *Let either* ***A.1***–***A.2*** *and* ***C.1***–***C.3*** *or* ***B.1*** *and* ***C.1***–***C.3*** *hold. Then*

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

for all i, j.

The rates in Theorem 7 equal the conventional univariate rates of series estimators; see, e.g., Newey [1997]. Thus, the fact that Z is latent does not affect the convergence speed of the density estimates.

To present distribution theory for the orthogonal-series estimator at a fixed point y let

$$\hat{\sigma}_{ij}(y) = \sqrt{n^{-1}\sum_{m=1}^{n}\left(\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j\sum_{k=1}^{\varkappa}\varphi_k(Y_{im})\varphi_k(y)\rho(Y_{im}) - \hat{f}_{ij}(y)\right)^2},$$

which is a sample standard deviation, and denote $f_i = \sum_{j=1}^r \pi_j f_{ij}$ in the following theorem.

THEOREM 8 (Asymptotic distribution). *Suppose that* $n, \varkappa \to \infty$ *so that* $\varkappa^2/n \to 0$ *and* $n\varkappa^{-2\beta} \to 0$. *Then*

$$\frac{\hat{f}_{ij}(y) - f_{ij}(y)}{\hat{\sigma}_{ij}(y)/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

for each $y \in \mathscr{Y}$ *that lies in an interval on which* f_i *is of bounded variation.*

Under **A.1**–**A.2** $\hat{\sigma}_{ij}(y)$ grows like $\|\boldsymbol{\varphi}_\varkappa(y)\|_F$, and this depends on the polynomial system used. Because **A.1** states that $\|\sqrt{\boldsymbol{\varphi}_\varkappa'\boldsymbol{\varphi}_\varkappa}\|_\infty = O(\zeta_\varkappa)$, a weak bound on the convergence rate that holds for all y is $O_p(\zeta_\varkappa/\sqrt{n})$. With Legendre polynomials, for example, the orthogonal-series estimator has a variance of order $\varkappa/n$, which is the same as that of an estimator based on a random sample from f_{ij} (Hall 1987). Likewise, under **B.1** we have that $\hat{\sigma}_{ij}(y)$ grows like $\varkappa^{1/4}$ and so the variance of the estimator is of the order $\sqrt{\varkappa}/n$. This is again the standard convergence rate for conventional Hermite series estimators (Liebsher 1990).

6. Monte Carlo illustrations. We evaluated the performance of the orthogonal-series estimator via simulation. We report root mean integrated squared error (RMISE) calculations for designs taken from Levine, Hunter and Chauveau [2011]. This allows us to compare our estimator to the EM-like approaches proposed in the literature. We also investigate the accuracy of the pointwise asymptotic approximation of the density estimator in Theorem 8 in a Monte Carlo experiment based on a hidden Markov model. Throughout this section we use Hermite polynomials as basis functions, set $\kappa_i = 10$ for all i, and use the cross-validation technique introduced above to select the number of series terms. Joint approximate diagonalization was done using the algorithm of Luciani and Albera [2010; 2014]. We also computed the estimator using the algorithms of Fu and Gao [2006] and Iferroudjene, Abed-Meraim and Belouchrani [2009; 2010] and found very similar results to the ones reported below

6.1. *RMISE comparisons.* We evaluate the RMISE of the estimator $\hat{f}_{ij}$,

$$\sqrt{E\|\hat{f}_{ij} - f_{ij}\|_2^2},$$

as approximated by 500 Monte Carlo replications. The first set of designs involves mixtures of normals, where

$$f_{ij}(y) = \phi\left(y - \mu_{ij}\right).$$

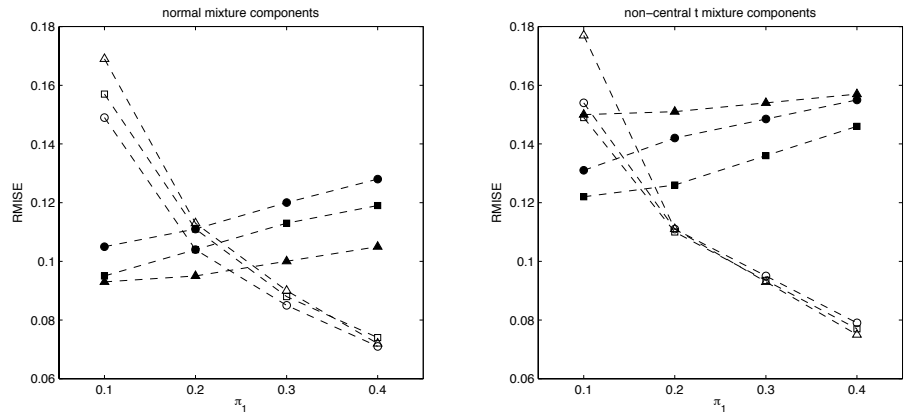
The second set of designs deals with mixtures of central and non-central t-distributions, that is,

$$f_{ij}(y) = t_{10}(y; \mu_{ij}),$$

where we let $t_d(y; \mu)$ denote a t-distribution with d degrees of freedom and non-centrality parameter μ. We set $q = 3$, $r = 2$, so the data is drawn from a three-variate two-component mixture. The parameters of the component densities are set to $(\mu_{11}, \mu_{21}, \mu_{31}) = (0, 0, 0)$ for the first component and $(\mu_{12}, \mu_{22}, \mu_{32}) = (3, 4, 5)$ for the second component. We consider various choices for the mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2)'$.

Figure 1 plots the RMISE as a function of the mixing proportion π_1 for samples of size $n = 500$. The results for the first and second component for each outcome variable are labelled consecutively as $\circ, \square, \triangle$ and as $\bullet$ $\blacksquare, \blacktriangle$, respectively.

FIG 1. *RMISE of the orthogonal-series density estimator*



The patterns of the RMISE are comparable to those for the EM-like estimators in Levine, Hunter and Chauveau [2011, Figure 1], although the magnitudes are larger here. The latter observation agrees with the intuition that joint estimation of classification probabilities and component densities (as in EM) should be more efficient than sequential estimation (as here). However, a precise comparison between the methods is complicated by the fact that the EM approaches are kernel based while we work with orthogonal series, and because the tuning parameters (the bandwidths for EM and the number of series terms here) were selected in a different manner.

Our least-squares estimator of the mixing proportions was also evaluated in these designs and was found to perform well. The Monte Carlo results are provided in the supplementary material.

6.2. *Inference in a hidden Markov model.* We next consider inference in a hidden Markov model with $r = 2$ latent states and $q = 3$ outcome variables. The latent Markov chain has transition matrix and stationary distribution equal to

$$\boldsymbol{K} = \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix}, \qquad \boldsymbol{\pi} = \begin{pmatrix} .5 \\ .5 \end{pmatrix},$$

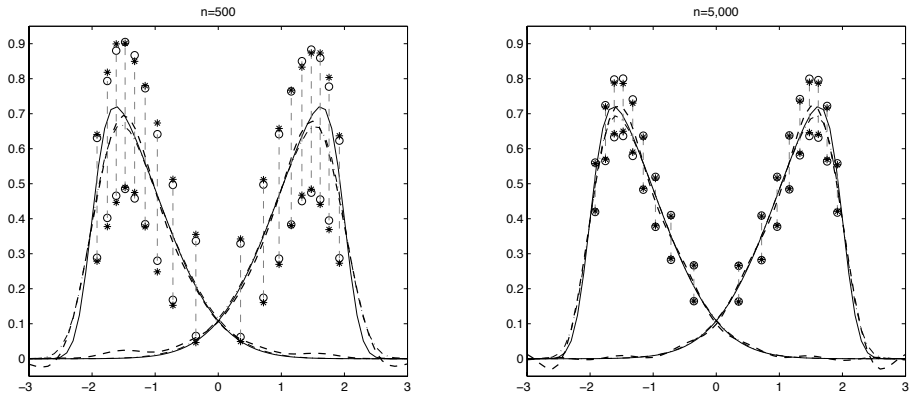
respectively. The emission densities f_1 and f_2 are skew-normal densities (Azzalini 1985),

$$f_j(y) = 2\phi(y - \mu_j)\,\Phi(\alpha_j(y - \mu_j)),$$

with $\mu_1 = -2$, $\alpha_1 = 5$ and $\mu_2 = -\mu_1$, $\alpha_2 = -\alpha_1$. The sign of the skewness parameters α_1, α_2 implies that f_1 is skewed to the right while f_2 is skewed to the left.

In each of 500 Monte Carlo replications, we estimated the two emission densities f_1 and f_2 using our orthogonal-series estimator and constructed 95% confidence intervals at the percentiles of f_1 and f_2. We present results for $n = 500$ (left plot) and $n = 5,000$ (right plot) graphically in Figure 2. Results for additional sample sizes are available in the supplementary material.

FIG 2. *Emission densities in the hidden Markov model*



Each plot in Figure 2 contains the true functions f_1 and f_2 (solid lines), and the mean (across the Monte Carlo replications) of our orthogonal-series estimator (dashed lines) as well as of an infeasible kernel-density estimator (dashed-dotted lines) computed from the subsample of observations that are in the respective latent state (see the supplementary material for more detail.) The plots show that, even in small samples, our estimator essentially co-incides with the infeasible estimator, on average.

Figure 2 also contains average 95% confidence intervals ($-\circ$), based on the pointwise distributional result in Theorem 8, for the emission densities at their respective percentiles. To assess the adequacy of our asymptotic approximation, the plots in the figure also provide 95% confidence intervals at the percentiles constructed using the empirical standard deviation of the

point estimates across the Monte Carlo replications ($-*$). Figure 2 shows that our estimated standard error captures well the small-sample variability of the orthogonal-series estimator.

Supplement. The supplement to this paper (Bonhomme, Jochmans and Robin 2015) contains additional details and discussion, omitted proofs, and additional simulation results.

Acknowledgements. We thank the Editor (Runze Li), an Associate Editor, three referees, Xiaohong Chen, Ignat Domanov, Marc Henry, and Nick Vannieuwenhoven for comments. We are grateful to Laurent Albera and Xavier Luciani for sharing the code for their diagonalization algorithm in Luciani and Albera [2010; 2014] with us. Early versions of this paper circulated as 'Nonparametric spectral-based estimation of latent structures'.

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.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12** 171–178.

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.

Bonhomme, S., Jochmans, K. and Robin, J. M. (2015). Supplement to "Estimating multivariate latent-structure models".

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.

Chiantini, L., Ottaviani, G. and Vannieuwenhoven, N. (2014). An algorithm for generic and low-rank specific identifiability of complex tensors. *SIAM Journal on Matrix Analysis and Applications* **35** 1265–1287.

Chiantini, L., Ottaviani, G. and Vannieuwenhoven, N. (2015). On generic identifiability of symmetric tensors of subgeneric rank. Mimeo.

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.

Domanov, I. and De Lathauwer, L. (2013a). On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part I: Basic results and uniqueness of one factor matrix. *SIAM Journal on Matrix Analysis and Applications* **34** 855–875.

Domanov, I. and De Lathauwer, L. (2013b). On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part II: Uniqueness of the overall decomposition. *SIAM Journal on Matrix Analysis and Applications* **34** 876–903.

Domanov, I. and De Lathauwer, L. (2014a). Canonical polyadic decomposition of third-order tensors: Reduction to generalized eigenvalue decomposition. *SIAM Journal on Matrix Analysis and Applications* **35** 636–660.

Domanov, I. and De Lathauwer, L. (2014b). Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL. Mimeo.

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. (1987). Cross-validation and the smoothing of orthogonal series density estimators. *Journal of Multivariate Analysis* **21** 181–206.

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.

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.

Jiang, T. and Sidiropoulos, N. D. (2004). Kruskal's permutation lemma and the

identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints. *IEEE Transactions on Signal Processing* **52** 2625–2636.

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

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.

Liebsher, E. (1990). Hermite series estimators for probability densities. *Metrica* **37** 321–343.

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.

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

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

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.

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

Sciences Po
Department of Economics
28 rue des Saints Pères
75007 Paris
France
E-mail: koen.jochmans@sciencespo.fr
(corresponding author)

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