Nonparametric estimation of non-exchangeable latent-variable models*

Stéphane Bonhomme
University of Chicago
sbonhomme@uchicago.edu

Koen Jochmans
Sciences Po Paris
koen.jochmans@sciencespo.fr

Jean-Marc Robin†
Sciences Po Paris and University College London
jeanmarc.robin@sciencespo.fr.

January 27, 2016

Abstract

We propose a two-step method to nonparametrically estimate multivariate models in which the observed outcomes are independent conditional on a discrete latent variable. Applications include microeconometric models with unobserved types of agents, regime-switching models, and models with misclassification error. In the first step, we estimate weights that transform moments of the marginal distribution of the data into moments of the conditional distribution of the data for given values of the latent variable. In the second step, these conditional moments are estimated as weighted sample averages. We illustrate the method by estimating a model of wages with unobserved heterogeneity on PSID data.

JEL codes: C14, C33, C38, J31
Keywords: Latent variable models, unobserved heterogeneity, finite mixtures, hidden Markov models, nonparametric estimation, panel data, wage dynamics

*We are grateful to the guest editors and two anonymous referees for comments that helped improve the paper. Robin acknowledges financial support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and from the European Research Council (ERC) grant ERC-2010-AdG-269693-WASP.

†Mail address for correspondence: Sciences Po, Département d'économie, 28 rue des Saints Pères, 75007 Paris, France.

1 Introduction

Latent variable models (LVMs) are of central interest in empirical microeconomics, where unobserved heterogeneity, censoring, and measurement error in variables are common; see Hu (2015) for a recent review of the literature. In many economic applications the latent variables are discrete. Examples are models with discrete covariates and misclassification errors (Mahajan, 2006; Hu, 2008), models of individual earnings dynamics (Keane and Wolpin, 1997; Geweke and Keane, 2000), structural discrete choice models (Kasahara and Shimotsu, 2009), or classification errors in dynamic discrete choice models (Keane and Sauer, 2009). LVMs are also useful in empirical macroeconomics, for example the switching regime model of Hamilton (1989) and state space models more generally.

Hall and Zhou (2003), Allman et al. (2009), and others provide powerful nonparametric identification results for finite mixture models and related LVMs based on the availability of short panel data. A review of this literature is given in Chauveau et al. (2014). These results cover in particular the class of models that we focus on in this paper: finite mixtures of conditionally-independent measurements, with possibly different distributions (i.e. *non-exchangeable* measurements). Hidden Markov models (HMM, or regime-switching models) are particular members of the class of finite mixtures where, rather than remaining fixed, the latent variable follows a Markov chain. Allman et al. (2009) show that for these models three measurements are generically sufficient for identification. One of them can have coarse support, such as a binary variable. Although identification is now well understood, nonparametric estimation is still a subject of active research.

In this paper, we develop a two-step procedure for estimating conditional expectations of general functions of observed measurements given unobserved types, without imposing parametric restrictions on the underlying distributions. We build on and extend the results derived in Bonhomme et al. (2016) (first submitted in 2013; BJR1 hereafter) and Bonhomme et al. (2015) (first submitted in 2014; BJR2 hereafter). In the first step, weights are estimated that operate like the individual posterior probabilities of unobserved types

calculated in the E-step of the EM algorithm.[1] The second step is analogous to the M-step: conditional moments given unobserved types are estimated as weighted sample averages. However, unlike in the EM algorithm, only one iteration suffices to deliver a consistent estimator. This method exploits the multilinear structure of the problem for fast estimation of the weights,[2] and readily lends itself to asymptotic analysis.

BJR1 focus on finite mixtures of iid distributions. BJR2 consider the non-exchangeable case, including HMMs. BJR2 use orthogonal polynomials for density estimation and show how the Fourier coefficients can be obtained using techniques related to, yet different from, those used in BJR1. This allows one to estimate conditional moments given latent types, but only after estimating the entire conditional distribution. The current paper shows how BJR1 and BJR2 can be adapted in order to estimate conditional moments of continuous outcomes given the unobserved types without first estimating the entire conditional distributions in the non-exchangeable case. Our method works under the identification restrictions of Allman et al. (2009): three measurements are necessary, two measurements have at least as many points of support as the number of latent types, while the third measurement may have a coarser support (such as binary). In addition, we show how to estimate the conditional densities of outcomes and the state transition probabilities in nonstationary hidden Markov models, using four periods of panel data.

The key difference between the exchangeable and non-exchangeable cases lies in the way the estimation weights are constructed. In models with identically distributed outcomes, the identifying restrictions take the form of a simultaneous diagonalization problem for a set of *symmetric* matrices. With non-exchangeable outcomes, a set of general, *non-symmetric* matrices are now simultaneously diagonalizable in the same basis. The joint diagonalization algorithm that we use in this paper takes advantage of recent developments in the signal

[1] See Benaglia et al. (2009) and Levine et al. (2011) for applications of the EM algorithms to the nonparametric estimation of finite mixtures.

[2] This method may be called a "spectral" method because it is based on eigenvalue and singular value decompositions. Related techniques may be found the signal processing literature, see Comon and Jutten (2010) and Cichocki et al. (2015) for recent surveys.

processing literature, and it is numerically fast and stable. In contrast, our experience with GMM and polynomial restrictions is that standard nonlinear solvers may not work well when the number of parameters to estimate becomes large. Our approach allows for a larger number of potential applications than BJR1, while preserving the computational simplicity of their method.

Our work contributes to a growing literature using spectral methods. Notably, Song et al. (2013) develop an estimation procedure related to the one in BJR1. Their method applies to both the "symmetric view" case (exchangeable) and the "multi-view" case (non-exchangeable), thanks to a symmetrization technique due to Anandkumar et al. (2012) that allows transforming the non-symmetric identifying matrices into symmetric ones. For this method to work, all three measurements must have as many points of support as the number of types. Symmetrization techniques are also used by Anandkumar et al. (2014) and De Castro et al. (2015). Lastly, Anandkumar et al. (2012) and Hsu et al. (2012) also propose spectral algorithms for finite mixture models and hidden Markov models for discrete, non-exchangeable measurements which are related to the transformation algorithm that we use in BJR2 and in this paper.

Relative to these references, our original contribution is as follows. None of these alternative methods use a joint diagonalization algorithm. Jointly enforcing model restrictions as we do may help improve the precision of the estimates compared to methods based on a single diagonalization. Also, from BJR1 and BJR2 it follows that nonparametric density estimation based on joint diagonalization leads to optimal convergence rates. Lastly, we provide a complete identification and estimation procedure for the case where only three measurements are available, one of them with possibly coarse support (Propositions 1 and 2). We also discuss identification and estimation of hidden Markov models in the non stationary case (Proposition 3).

An attractive feature of our approach is that it allows for a simple treatment of continuous outcomes. In particular, kernel estimators of component densities can be obtained by reweighting, and the bandwidths can be chosen using standard techniques such as cross-

validation. Our estimator being a weighted mean, with weights being functions of a finite-dimensional parameter, asymptotic theory is standard, in contrast with iterated algorithms such as EM, for which no asymptotic theory has yet been proposed. At the same time, relative to full information methods, method of moments methods such as the one we advocate in this paper may be less efficient. The relative asymptotic efficiency of the different approaches is currently unknown.[3]

As an empirical illustration, we use our method to document the structure and evolution of wage distributions in the US. As documented by a large literature, allowing for unobserved heterogeneity is particularly important in this context. For example, augmenting canonical models of earnings by allowing for type heterogeneity, Geweke and Keane (2000) and Gu and Koenker (2014) found that heterogeneity is quantitatively important for explaining and forecasting earnings trajectories. The models estimated by these authors are parametric, and thus restrict the channels through which type heterogeneity is allowed to affect earnings. To assess the impact of unobserved factors on the entire wage distributions, we fit a nonparametric model with time-invariant unobserved heterogeneity to PSID data spanning a period of two decades.

The outline of the paper is as follows. In Section 2 we present the latent variable models and describe a number of examples. In Section 3 we introduce our two-step estimation strategy and report simulation evidence on its performance. Section 4 discusses a number of extensions, including applying the framework to models with time-varying unobserved types. In Section 5 we apply our method to PSID data.

[3]Such efficiency calculations are difficult because of the lack of asymptotic theory for EM-based estimators. Even though one may expect full-information approaches to be more efficient asymptotically, an important issue with the EM approach is the lack of data-driven, component-specific bandwidth. See, e.g., Chauveau et al. (2014) for more on this.

2 Framework and examples

2.1 Finite mixtures

Let $(Y_1, ..., Y_M)$ be a random vector of observed outcome variables with joint cumulative distribution function (cdf) $F(y_1, ..., y_M)$. Let $X \in \{1, ..., K\}$ be a discrete latent random variable with K points of support.

Assumption 1 (Finite Mixture)**.** Y_1,..., Y_M *are mutually independent conditional on* X*.*

Under Assumption 1,

$$F(y_1, ..., y_M) = \sum_{k=1}^{K} \pi_k \, F_{k1}(y_1) \cdots F_{kM}(y_M), \tag{2.1}$$

where $\pi_k = \Pr(X = k)$, and F_{km} denotes the conditional cdf of outcome Y_m given $X = k$. Our goal is to construct estimators of the conditional distributions F_{km} and moments thereof, as well as of the probabilities $(\pi_1, \ldots, \pi_K)$, from a random sample on $(Y_1, ..., Y_M)$ drawn from the model in (2.1), without imposing functional-form restrictions on the distributions F_{km}.

Conditions that ensure identification are now well known (see, e.g., Allman et al. 2009). We will assume that the number of components, K, is known,[4] that the number of outcome variables, M, is at least equal to three, and that certain rank conditions to be detailed below are satisfied. When $M = 3$, these rank conditions require that at least two measurements have at least K points of support. The third measurement is not restricted beyond the fact that it has at least two points of support (as in Hu, 2015, for example). When $M > 3$, these support requirements can be relaxed further.

We now review several applications of these models in economics.

[4]Identification when K is unknown is difficult. Moreover, in the nonparametric context, there may be multiple K for which a decomposition as in (2.1) can be obtained. Kasahara and Shimotsu (2014) show that a lower bound on K is identified under weak conditions.

Example 1 (Unobserved heterogeneity and wage dynamics)**.** Consider a panel data model for individual log wages measured over M periods, $Y_1, ..., Y_M$. Suppose that individuals can be clustered into different groups indexed by $X \in \{1, ..., K\}$, which correspond to different types of unobserved ability. Under Assumption 1, wages are conditionally independent over time given ability type. This model encompasses the simple additive one-factor model estimated by Gottschalk and Moffitt (1994), with an individual time-invariant fixed-effect and a transitory, serially-independent shock.

Example 2 (Misclassification error)**.** Suppose we wish to explore the relationship between an outcome Y_1 and a discrete covariate X, but one only observes an error-laden version of X, say Y_2. Assume that a second measurement Y_3 of X is available, and that Y_1, Y_2 and Y_3 are mutually independent given X. Then Assumption 1 holds, and the methods of this paper can be applied. In this example, the conditional independence requirement is an assumption of conditional ignorability, which is conventional in the literature on measurement error. Note that, while in this application it is natural to assume that Y_2 and X have the same (discrete) support, our setup allows the second measure Y_3 to possibly have a coarser support.

LVMs have been used in a number of other economic applications. Studies in empirical industrial organization, for example, make intensive use of dynamic discrete choice models with unobserved type heterogeneity (Kasahara and Shimotsu, 2009). In the analysis of games with finitely many equilibria, treating the realized equilibrium as a latent variable may lead to a similar LVM structure as the one we study here (Bajari et al., 2011; Hahn and Moon, 2010).

2.2 Regime-switching models

Consider now a panel model where the latent state is time-varying, $(X_1, ..., X_M)$. In a model of earnings dynamics, X_m could denote the latent skills of a worker evolving over

time as a result of health shocks or job training, for example. We restrict the dynamics of X_m to be first-order Markov, and we make the following assumption.

Assumption 2 (Hidden Markov Models)**.** *For all $m > 1$,*

1. *Y_m is independent of $Y_{m-1}, ..., Y_1$ and $X_{m-1}, ..., X_1$ given X_m;*

2. *X_m is independent of $Y_{m-1}, ..., Y_1$ and $X_{m-2}, ..., X_1$ given X_{m-1}.*

Under Assumption 2 the model has a hidden Markov structure. Note that the present setup differs from stationary hidden-Markov models popular in the time-series literature (e.g. Gassiat et al., 2013; Gassiat and Rousseau, 2013). There, asymptotics are done for M diverging. Here, in contrast, we consider a panel data setup with fixed M, and we do not assume stationarity. The conditional distribution of Y_m given X_m may depend on m, as well as the transition probability from state X_{m-1} to state X_m.

In principle we could define a vector-valued latent variable $X = (X_1, ..., X_M)$ and treat the model with time-varying latent states as a standard finite mixture model in (2.1), with X being the latent variable. However, doing so would lead to a mixture with a potentially very large number of components, as the cardinality of the state space of X grows rapidly with M. This may be problematic in practice, as nonparametric identification requires restricting the number of latent types.

The Markovian assumption significantly reduces the dimensionality of the unobserved states. To see why this is so, consider the case $M = 3$, and note that by Assumption 2, we have

$$(Y_3, X_3) \perp\!\!\!\perp (Y_2, Y_1, X_1) \mid X_2 \quad \text{and} \quad Y_2 \perp\!\!\!\perp (Y_1, X_1) \mid X_2,$$

where $\perp\!\!\!\perp$ denotes statistical independence. Hence (Y_1, X_1), Y_2 and (Y_3, X_3) are mutually independent given X_2. It follows that Y_1, Y_2, and Y_3 are independent given X_2. This, therefore, implies that Assumption 1 is satisfied for $X = X_2$. We will show in Section 4 that the techniques developed for finite mixtures can also be applied to models with time-varying unobserved states.

3 Two-step estimation

Now consider the model in (2.1) and set $M = 3$, and denote the three scalar measurements as Y_1, Y_2, Y_3. The theory to follow can be extended to accommodate more than 3 measurements (see the next section), and the results can easily be adapted to deal with vector-valued measurements. As a notational shorthand, we write $\mathbb{E}_k W = \mathbb{E}(W|X = k)$ for the conditional expectation of any random variable W.

In this section we show how to consistently estimate linear functionals of the form $\mathbb{E}_k\varphi(Y_m)$ for any measurable univariate function φ. Particular cases of interest are power functions, $\varphi(u) = u^p$, which deliver conditional moments of outcomes. Also, setting $\varphi_y(u) = \mathbf{1}\{u \leq y\}$ gives $\mathbb{E}_k\varphi_y(Y_m) = F_{km}(y)$, the conditional cdf. Finally, if $\varphi_y(u) = h^{-1}\kappa(h^{-1}(u - y))$, then $\mathbb{E}_k\varphi_y(Y_m)$ is the conditional density of $Y_m + h\varepsilon$ at point y, where ε is a random error with density κ. This delivers a kernel density estimator of the density function of F_{km} that is particularly easy to implement.

3.1 Identification

Let $\psi_1, ..., \psi_J$ be a set of $J \geq K$ univariate functions, and let $\Psi = (\psi_1, ..., \psi_J)'$. In addition, we define the following $J \times J$ matrix,[5]

$$A = \mathbb{E}\left[\Psi(Y_1)\Psi(Y_2)'\right] = \sum_{k=1}^{K} \pi_k \, \mathbb{E}_k\Psi(Y_1) \, \mathbb{E}_k\Psi(Y_2)'. \tag{3.2}$$

Identification rests on the following restriction on the matrix A and the number of types K.

Assumption 3. *A has rank K.*

Assumption 3 is satisfied provided both $\mathbb{E}_1\Psi(Y_1), \ldots, \mathbb{E}_K\Psi(Y_1)$ and $\mathbb{E}_1\Psi(Y_2), \ldots, \mathbb{E}_K\Psi(Y_2)$ are linearly independent, and $\pi_k > 0$ for all k.

[5] Alternatively, one could use different functions ψ_j, and a different J, for each measurement Y_1, Y_2. Here we focus on the case where A is a square matrix in order to keep the notation simple.

Under Assumption 3, the singular value decomposition (SVD) of A is $A = USV'$, where U and V are $J \times K$ matrices with orthogonal and unitary columns, and S is a $K \times K$ diagonal and non singular matrix with non-negative elements. The matrix A allows to construct two *whitening* matrices, $W_1 = S^{-\frac{1}{2}}U'$ and $W_2 = S^{-\frac{1}{2}}V'$, such that the matrices

$$B(\varphi) = W_1 \mathbb{E}\left[\Psi(Y_1)\Psi(Y_2)'\varphi(Y_3)\right] W_2' \tag{3.3}$$

have their eigenvalues equal to the unknown conditional moments $\mathbb{E}_k\varphi(Y_3)$. More precisely, we show in Appendix A.1 the following proposition.

Proposition 1. *Let Assumptions 1 and 3 hold. The set of matrices $B(\varphi)$, for all univariate functions φ, can be jointly diagonalized in the same basis, and the conditional moments $\mathbb{E}_k\varphi(Y_3)$ are their eigenvalues. That is, there exists a non singular $K \times K$ matrix Q such that, for all φ,*

$$Q^{-1}B(\varphi)Q = D_3(\varphi), \tag{3.4}$$

for $D_3(\varphi) = \text{diag}\left(\mathbb{E}_1\varphi(Y_3), ..., \mathbb{E}_K\varphi(Y_3)\right)$. The matrix Q is unique up to column swapping and rescaling provided for all $k \neq k'$ there exists φ such that $\mathbb{E}_k\varphi(Y_3) \neq \mathbb{E}_{k'}\varphi(Y_3)$.

Let $\tau_k(Y_1, Y_2)$ denote the k-th diagonal element of the random matrix whose expectation is $B(\varphi)$, i.e.

$$\tau_k(Y_1, Y_2) = e_k'Q^{-1}W_1\Psi(Y_1)\Psi(Y_2)'W_2'Qe_k, \tag{3.5}$$

where e_k is the kth column of the $K \times K$ identity matrix. Proposition 1 implies that, for any univariate function φ, the functionals

$$\mathbb{E}_k\varphi(Y_3) = \mathbb{E}\left[\tau_k(Y_1, Y_2)\varphi(Y_3)\right], \quad k = 1, ..., K, \tag{3.6}$$

are identified up to relabeling the types. The weights τ_k thus transform moments of the distribution of Y_3 into moments of the type-k distributions.

It is interesting to compare the weights $\tau_k(Y_1, Y_2)$ with the posterior probabilities

$$p_k(Y_1, Y_2, Y_3) = \frac{\pi_k f_{k1}(Y_1) f_{k2}(Y_2) f_{k3}(Y_3)}{\sum_{\ell=1}^K \pi_\ell f_{\ell 1}(Y_1) f_{\ell 2}(Y_2) f_{\ell 3}(Y_3)},$$

where f_{km} denotes the conditional probability density (or mass) function of Y_m given $X = k$. The ratios of posterior to prior probabilities, p_k/π_k, also transform functionals of the distribution of Y_3 into functionals of the type-k distributions. Specifically,

$$\mathbb{E}_k\varphi(Y_3) = \mathbb{E}\left[\frac{p_k(Y_1, Y_2, Y_3)}{\pi_k}\varphi(Y_3)\right].$$

However, the posterior probabilities p_k depend on the conditional densities f_{km}, which are unknown and need first to be nonparametrically estimated, whereas the weights τ_k depend only on the matrices W_1, W_2, Q.

Proposition 1 shows that the type-specific distributions of Y_3 are nonparametrically identified up to relabeling. This result is closely related to Theorem 1 of BJR2 and Lemma 3.2 of Anandkumar et al. (2012). A noteworthy feature of Proposition 1 is that it provides a set of joint restrictions on the matrix Q, for all functions φ. We will enforce these joint restrictions in estimation. In addition, the restrictions involve moments of the form $\mathbb{E}_k\varphi(Y_3)$. This will be useful to construct simple empirical counterparts of those moments that converge at the parametric rate.

In many situations, Proposition 1 will be enough to identify moments $\mathbb{E}_k\varphi(Y_m)$, for all $m = 1, 2, 3$. It suffices to apply Proposition 1 three times redefining $A = \mathbb{E}\left[\Psi(Y_{m_1})\Psi(Y_{m_2})'\right]$, for all couples $(m_1, m_2) \in \{(1, 2), (1, 3), (2, 3)\}$. Each choice of A delivers a different Q, with a possibly different labeling of the unobserved types.[6]

However, Proposition 1 cannot directly be applied for identifying $\mathbb{E}_k\varphi(Y_m)$, $m \in \{1, 2\}$ when, say, Y_3 is a binary variable and $\mathbb{E}\left[\Psi(Y_1)\Psi(Y_3)'\right]$ does not satisfy the rank condition of Assumption 3. The next result shows that the type-specific distributions of Y_1 and Y_2, as well as the type proportions, are also identified for the same choice of matrix A, and up to the same labeling of types as in Proposition 1.

Proposition 2. *Given Q from Proposition 1, for all univariate functions φ and $k =$*

[6]Theorem 2 of BJR2 shows how to recover a common labeling of the types across the different measurements.

$1, ..., K$,

$$\mathbb{E}_k\varphi(Y_1) = \frac{e_k' Q' W_2 \mathbb{E}\left[\Psi(Y_2)\varphi(Y_1)\right]}{e_k' Q' W_2 \mathbb{E}\Psi(Y_2)}, \tag{3.7}$$

$$\mathbb{E}_k\varphi(Y_2) = \frac{e_k' Q^{-1} W_1 \mathbb{E}\left[\Psi(Y_1)\varphi(Y_2)\right]}{e_k' Q^{-1} W_1 \mathbb{E}\Psi(Y_1)}. \tag{3.8}$$

Furthermore, the type-k proportion satisfies

$$\pi_k = e_k' Q^{-1} W_1 \mathbb{E}\Psi(Y_1) \cdot e_k' Q' W_2 \mathbb{E}\Psi(Y_2). \tag{3.9}$$

Equations (3.7), (3.8), and (3.9) hold irrespective of the choice of observationally-equivalent eigenvector matrix Q. Moments $\mathbb{E}_k\varphi(Y_1)$, $\mathbb{E}_k\varphi(Y_2)$, and proportions π_k are thus identified up to the labeling chosen for $\mathbb{E}_k\varphi(Y_3)$, but they are not subject to the scale indeterminacy of the matrix Q.

3.2 Estimation

Propositions 1 and 2 suggest a two-step estimation strategy. In the first step, the matrix Q is estimated by approximately jointly diagonalizing empirical counterparts of matrices $B(\psi_1), ..., B(\psi_J)$. The weights τ_k in (3.5) can then be estimated. In the second step, any functional of the type-specific distributions associated with a given measurement can be estimated as a simple weighted average. We now detail the two estimation steps. We work with an iid sample (Y_{i1}, Y_{i2}, Y_{i3}), $i = 1, .., N$.

Step 1: Weights

Let us first estimate the matrices $B(\varphi)$ in Proposition 1 by

$$\widehat{B}(\varphi) = \widehat{W}_1 \widehat{\mathbb{E}}\left[\Psi(Y_1)\Psi(Y_2)'\varphi(Y_3)\right] \widehat{W}_2', \tag{3.10}$$

where $\widehat{\mathbb{E}}(Z) = \frac{1}{N}\sum_{i=1}^N Z_i$, and $\widehat{W}_1 = \widehat{S}^{-\frac{1}{2}}\widehat{U}'$ and $\widehat{W}_2 = \widehat{S}^{-\frac{1}{2}}\widehat{V}'$, with $(\widehat{U}, \widehat{S}, \widehat{V})$ coming from the SVD of $\widehat{A} = \widehat{\mathbb{E}}\left(\Psi(Y_1)\Psi(Y_2)'\right)$.

Proposition 1 implies that Q is the matrix of joint eigenvectors of all matrices $B(\varphi)$. As in BJR2, we estimate Q by approximate joint diagonalization of the matrices $\widehat{B}(\psi_j)$, $j = 1, ..., J$, i.e.

$$\widehat{Q} = \arg\min_{Q \in \mathcal{Q}} \sum_{j=1}^{J} \text{off}\left(Q^{-1}\widehat{B}(\psi_j)Q\right), \tag{3.11}$$

where $\text{off}(A) = \sum_{k=1}^{K}\sum_{\ell \neq k} a_{k\ell}^2$ denotes the sum of squared off-diagonal coefficients of a square matrix $A = [a_{k\ell}]$, and the set $\mathcal{Q}$ of $K \times K$ matrices enforces a scaling constraint; in practice we normalize $\det Q = 1$.

The objective function in (3.11) can be minimized using the algorithms of Iferroudjene et al. (2009, 2010) or Luciani and Albera (2010).[7] These algorithms allow for fast computation of the matrix $\widehat{Q}$.

Finally, we construct the weight functions,

$$\widehat{\omega}_{1k}(y_1) = e_k'\widehat{Q}^{-1}\widehat{W}_1\Psi(y_1), \quad \widehat{\omega}_{2k}(y_2) = e_k'\widehat{Q}'\widehat{W}_2\Psi(y_2), \quad k = 1, ..., K.$$

The product $\widehat{\tau}_k(y_1, y_2) = \widehat{\omega}_{1k}(y_1)\widehat{\omega}_{2k}(y_2)$ is an estimate of $\tau_k(y_1, y_2)$ in (3.5).

Remark. Note that Algorithm 4 in Anandkumar et al. (2012, 2015) allows to transform the problem of diagonalizing the non-symmetric matrices $\widehat{B}(\psi_j)$ in the same basis into the joint diagonalization of a set of symmetric matrices. Hence, an alternative approach would be to use the algorithm of Cardoso and Souloumiac (1993), which is a well-known algorithm used in Independent Component Analysis, and which we used in BJR1. However, as we show in Appendix A.1.3, this symmetrization algorithm delivers matrices of the form $C_3\Omega C_3'$ and $C_3\Omega D_3(\psi_j)C_3'$, and identification requires the matrix C_3 to be of full column rank K. As already emphasized, this is not likely to hold if the third measurement Y_3 has coarse support.[8]

[7] In the Monte Carlo and the application we use the Matlab code that Xavier Luciani and Laurent Albera kindly provided to us.

[8] The symmetrization algorithm (without joint diagonalization) was used by Song et al. (2013) and De Castro et al. (2015) for estimating component densities.

Step 2: Averaging

Let φ be a univariate function. Let $\theta_{km} = \mathbb{E}_k\varphi(Y_m)$, for all $(k,m) \in \{1,...,K\} \times \{1,2,3\}$. For all k, we can estimate the functionals θ_{k1}, θ_{k2}, and θ_{k3} as weighted averages

$$\widehat{\theta}_{k1} = \frac{\widehat{\mathbb{E}}\left[\widehat{\omega}_{2k}(Y_2)\varphi(Y_1)\right]}{\widehat{\mathbb{E}}\widehat{\omega}_{2k}(Y_2)}, \quad \widehat{\theta}_{k2} = \frac{\widehat{\mathbb{E}}\left[\widehat{\omega}_{1k}(Y_1)\varphi(Y_2)\right]}{\widehat{\mathbb{E}}\widehat{\omega}_{1k}(Y_1)}, \quad \widehat{\theta}_{k3} = \widehat{\mathbb{E}}\left[\widehat{\omega}_{1k}(Y_1)\widehat{\omega}_{2k}(Y_2)\varphi(Y_3)\right], \tag{3.12}$$

and type proportions as

$$\widehat{\pi}_k = \widehat{\mathbb{E}}\left[\widehat{\omega}_{1k}(Y_1)\right]\widehat{\mathbb{E}}\left[\widehat{\omega}_{2k}(Y_2)\right]. \tag{3.13}$$

Note that (3.13) does not guarantee that the type proportions be non negative and sum up to one. In practice, these constraints can be imposed *ex post*, by projecting the vector $(\widehat{\pi}_1,...,\widehat{\pi}_K)$ on the K-dimensional simplex. Similarly, estimates of cdfs may be re-arranged in order to be non-decreasing (as in Chernozhukov et al., 2009), and the density estimates below can be guaranteed to be non negative by using for example the procedure of Gajek (1986).

Given that conditional moments of outcomes given the unobserved types take the form of simple weighted averages with pre-estimated weights, one can readily show that they are root-N consistent and asymptotically normal under standard conditions. In Appendix A.2 we derive the form of the influence function of the estimator of $\theta_{k3} = \mathbb{E}_k\varphi(Y_3)$ given by (3.12) as an example, using results from BJR1 and BJR2. The estimator is root-N consistent under the following additional assumptions: 1) $\mathbb{E}[\psi_j^2(Y_m)]$ is finite for all $j = 1,...,J$ and $m = 1,2$; 2) $\mathbb{E}[\varphi^2(Y_3)]$ is finite; and 3) all eigenvalues of matrix A are simple. The asymptotic distributions of conditional moments of other measurements and type proportions can be derived similarly.

3.3 Simulations

3.3.1 Experiment 1: continuous outcomes

We illustrate the performance of our estimators by means of two Monte Carlo experiments. The first is taken from Levine et al. (2011). This allows a comparison of our results with the parametric EM estimator, the nonparametric EM estimator, and the estimator in BJR1. The design is as follows. Three measurements are drawn from a mixture model with two latent types. The distribution of each measurement is a bivariate mixture of normals with means zero and three, respectively, and unit variances. Moreover,

$$F_{1m}(y) = \Phi(y), \qquad F_{2m}(y) = \Phi(y-3),$$

for all m, and we will provide results for the different mixing proportions $\pi_1 \in \{.2, .4, .6, .8\}$ This is a symmetric design, but our estimator does not use this information. We will estimate the mean (μ_{km}) and standard deviation (σ_{km}) of each component using the formulae in (3.12). The results we report below are for samples of size $N = 500$ and were obtained over $1,000$ Monte Carlo simulations.

We implemented our procedures for Ψ set to the leading J orthonormalized Hermite polynomials. We report results for $J \in \{5, 10\}$ to evaluate the impact of J on the results. To estimate the joint diagonalizer Q, we set $\varphi_j = \psi_j$ for all j.

Table 1 contains the mean and the standard deviation (in italics) of our estimators of μ_{km} and σ_{km} for each k, m. Biases are generally moderate. However, standard deviations can be quite large. In particular, standard deviations tend to be larger for the first two outcomes when π_1 is closer to zero or one. Inspection of (3.7) and (3.8) suggests that, as estimates corresponding to these outcomes are ratios of two components, they may be poorly estimated when the denominator is close to zero.[9] The estimator for the third outcome is much more stable. We also see that the estimates tend to be more precise when

[9] An interesting possibility, which we do not study in this paper, would be to add a regularization term to the denominator, chosen as a decreasing function of the sample size.

J is 10 instead of 5. However, even in that case there is a loss of efficiency compared to EM estimators and the method of BJR1 tailored to the exchangeable case, as may be seen when comparing Table 1 to Table 1 in BJR1.

3.3.2 Experiment 2: coarse support

The second design we consider is a modification of the first which allows us to evaluate our procedure when one of the measurements has a coarse support. To do so we generate the first two outcomes as before, but now restrict the third outcome to have a probability mass function supported only on the set $\{0, 1, 2\}$, with mass functions

$$\Pr(Y_3 = v|X = 0) = \begin{cases} .50 & \text{if } v = 0 \\ .34 & \text{if } v = 1 \\ .16 & \text{if } v = 2 \end{cases}, \qquad \Pr(Y_3 = v|X = 1) = \begin{cases} .16 & \text{if } v = 0 \\ .68 & \text{if } v = 1 \\ .16 & \text{if } v = 2 \end{cases}.$$

In this case, $\mu_{13} = \mathbb{E}_1 Y_3 = .6587$ and $\mu_{23} = \mathbb{E}_2 Y_3 = 1$, and the corresponding standard deviations are $\sigma_{13} = .7363$ and $\sigma_{23} = .5633$, respectively. The rest of the design and implementation are the same as in the first experiment.

The simulation results are collected in Table 2. As in the first experiment, we see that while biases are moderate some of the standard deviations are large, particularly for the first two outcomes when π_1 is closer to zero or one and $J = 5$. The results when $J = 10$ are more encouraging. Developing a data-driven choice of J is an interesting question for future work.

4 Extensions

4.1 Additional measurements

If $M > 3$ measurements are available, the above results can easily be adapted. Suppose for example that one has 4 measurements $Y_1, ..., Y_4$. In order to estimate $\mathbb{E}_k \varphi(Y_4)$ one can use

$$\widehat{A} = \widehat{\mathbb{E}} \left[\Psi_2(Y_1, Y_2) \Psi(Y_3)' \right],$$

Table 1: Simulation results for Experiment 1

π_1	μ_{11}	μ_{21}	μ_{12}	μ_{22}	μ_{13}	μ_{23}	σ_{11}	σ_{21}	σ_{12}	σ_{22}	σ_{13}	σ_{23}
						$J = 5$						
.2	-0.010	2.993	-0.020	2.998	0.008	2.989	0.876	1.003	0.873	0.994	0.977	0.997
	0.284	*0.080*	*0.238*	*0.082*	*0.124*	*0.091*	*0.454*	*0.091*	*0.421*	*0.097*	*0.104*	*0.083*
.4	-0.002	2.994	0.001	2.992	-0.003	2.966	0.953	0.992	0.957	0.983	0.988	1.001
	0.150	*0.108*	*0.155*	*0.110*	*0.088*	*0.121*	*0.255*	*0.163*	*0.253*	*0.169*	*0.068*	*0.132*
.6	-0.001	2.990	0.004	2.994	-0.003	2.902	0.987	0.952	0.965	0.947	0.986	1.014
	0.102	*0.164*	*0.143*	*0.163*	*0.069*	*0.174*	*0.140*	*0.323*	*0.231*	*0.321*	*0.052*	*0.229*
.8	0.015	3.283	0.022	3.080	-0.002	2.247	1.019	0.885	1.002	0.862	0.992	1.082
	0.095	*11.757*	*0.154*	*1.986*	*0.064*	*0.735*	*0.149*	*0.664*	*0.243*	*0.675*	*0.063*	*0.486*
						$J = 10$						
.2	0.020	2.993	0.000	2.995	0.010	2.976	0.969	1.004	0.921	1.000	0.980	0.996
	0.203	*0.061*	*0.166*	*0.075*	*0.130*	*0.063*	*0.340*	*0.057*	*0.310*	*0.077*	*0.131*	*0.045*
.4	0.009	2.998	0.005	2.996	0.004	2.967	1.000	1.002	0.985	0.997	0.988	0.996
	0.106	*0.079*	*0.099*	*0.078*	*0.085*	*0.073*	*0.130*	*0.078*	*0.128*	*0.086*	*0.069*	*0.051*
.6	0.000	2.988	0.007	2.990	0.001	2.947	0.996	1.001	1.001	0.992	0.989	0.996
	0.074	*0.107*	*0.078*	*0.096*	*0.068*	*0.098*	*0.080*	*0.131*	*0.084*	*0.123*	*0.050*	*0.071*
.8	0.002	2.977	0.003	2.965	0.000	2.795	1.002	0.952	1.002	0.957	0.990	1.027
	0.062	*0.284*	*0.101*	*0.220*	*0.058*	*0.257*	*0.085*	*0.368*	*0.082*	*0.342*	*0.042*	*0.160*

Table 2: Simulation results for Experiment 2

π_1	μ_{11}	μ_{21}	μ_{12}	μ_{22}	μ_{13}	μ_{23}	σ_{11}	σ_{21}	σ_{12}	σ_{22}	σ_{13}	σ_{23}
						$J = 5$						
.2	-0.043	2.983	-0.037	2.994	0.654	0.977	0.863	0.978	0.857	0.976	0.731	0.550
	0.533	*0.377*	*0.442*	*0.268*	*0.063*	*0.073*	*0.574*	*0.169*	*0.549*	*0.175*	*0.035*	*0.052*
.4	-0.015	2.997	-0.010	3.001	0.654	0.988	0.892	0.961	0.896	0.942	0.728	0.556
	0.247	*0.160*	*0.251*	*0.160*	*0.076*	*0.049*	*0.412*	*0.275*	*0.411*	*0.292*	*0.040*	*0.036*
.6	-0.009	3.024	0.005	3.013	0.646	0.991	0.931	0.867	0.940	0.876	0.719	0.558
	0.184	*0.277*	*0.218*	*0.272*	*0.095*	*0.039*	*0.309*	*0.457*	*0.358*	*0.453*	*0.050*	*0.030*
.8	0.012	3.154	0.012	3.324	0.539	0.994	0.991	0.861	0.960	0.759	0.625	0.561
	0.157	*3.127*	*0.236*	*3.756*	*0.274*	*0.039*	*0.255*	*0.716*	*0.367*	*0.737*	*0.180*	*0.030*
						$J = 10$						
.2	-0.025	2.998	-0.016	2.990	0.654	0.942	0.848	0.998	0.856	0.995	0.732	0.548
	0.362	*0.088*	*0.339*	*0.247*	*0.047*	*0.074*	*0.516*	*0.110*	*0.498*	*0.123*	*0.026*	*0.050*
.4	0.011	2.993	-0.001	3.003	0.655	0.977	0.974	0.993	0.944	0.974	0.732	0.556
	0.163	*0.111*	*0.161*	*0.107*	*0.052*	*0.051*	*0.262*	*0.153*	*0.277*	*0.154*	*0.027*	*0.035*
.6	0.008	2.996	0.004	3.001	0.648	0.985	0.992	0.973	0.984	0.952	0.728	0.559
	0.110	*0.149*	*0.109*	*0.152*	*0.064*	*0.040*	*0.156*	*0.235*	*0.159*	*0.256*	*0.034*	*0.029*
.8	0.007	2.997	0.008	3.021	0.626	0.988	1.001	0.892	0.993	0.848	0.706	0.559
	0.085	*0.534*	*0.107*	*0.373*	*0.116*	*0.033*	*0.116*	*0.501*	*0.154*	*0.498*	*0.070*	*0.026*

where

$$\Psi_2(y_1, y_2) = \Psi(y_1) \otimes \Psi(y_2)$$

is a vector of interactions $\psi_{j_1}(y_1)\psi_{j_2}(y_2)$, and estimate Q as a joint diagonalizer of matrices

$$\widehat{B}(\psi_j) = \widehat{W}_1 \widehat{\mathbb{E}} \left[\Psi_2(Y_1, Y_2) \Psi(Y_3)' \psi_j(Y_4) \right] \widehat{W}_2'.$$

Letting

$$\widehat{\omega}_{12k}(y_1, y_2) = e_k' \widehat{Q}^{-1} \widehat{W}_1 \Psi_2(y_1, y_2), \quad \widehat{\omega}_{3k}(y_3) = e_k' \widehat{Q}' \widehat{W}_2 \Psi(y_3),$$

we can estimate $\theta_{k12} = \mathbb{E}_k \varphi(Y_1, Y_2)$, $\theta_{k3} = \mathbb{E}_k \varphi(Y_3)$, $\theta_{k4} = \mathbb{E}_k \varphi(Y_4)$, and π_k, respectively, as

$$\widehat{\theta}_{k12} = \frac{\widehat{\mathbb{E}} \left[\widehat{\omega}_{3k}(Y_3) \varphi(Y_1, Y_2) \right]}{\widehat{\mathbb{E}} \widehat{\omega}_{3k}(Y_3)}, \quad \widehat{\theta}_{k3} = \frac{\widehat{\mathbb{E}} \left[\widehat{\omega}_{12k}(Y_1, Y_2) \varphi(Y_3) \right]}{\widehat{\mathbb{E}} \widehat{\omega}_{12k}(Y_1, Y_2)}, \quad \widehat{\theta}_{k4} = \widehat{\mathbb{E}} \left[\widehat{\omega}_{12k}(Y_1, Y_2) \widehat{\omega}_{3k}(Y_3) \varphi(Y_4) \right],$$

$$\text{and } \widehat{\pi}_k = \widehat{\mathbb{E}} \left[\widehat{\omega}_{12k}(Y_1, Y_2) \right] \widehat{\mathbb{E}} \left[\widehat{\omega}_{3k}(Y_3) \right].$$

Everything works as before because (Y_1, Y_2), Y_3, and Y_4 are independent given X.

There are many possibilities to combine the restrictions implied by the model in estimation. Characterizing semi-parametric efficient estimators in this context is a very interesting question, which exceeds the scope of this paper.

4.2 Density estimation

In models with continuous measurements, one can construct kernel density estimators of type-specific densities as well. Consider as an example the conditional density f_{k3} of Y_3 given $X = k$. Let κ be a kernel function and $h > 0$ be a bandwidth parameter. Let us define

$$\widehat{f}_{k3}(y) = \widehat{\mathbb{E}} \left[\widehat{\tau}_k(Y_1, Y_2) \frac{1}{h} \kappa \left(\frac{Y_3 - y}{h} \right) \right]. \tag{4.14}$$

Under conditions similar to the ones in Proposition 2 in BJR1, this density estimator is $\sqrt{Nh}$-consistent for $f_{k3}(y)$ and asymptotically normal. In addition, $\widehat{f}_{k3}(y)$ is (pointwise) asymptotically equivalent to the infeasible estimator obtained upon replacing $\widehat{\tau}_k(Y_1, Y_2)$ in (4.14) by its population counterpart $\tau_k(Y_1, Y_2)$ given by (3.5). For density estimation,

an appealing feature of our approach is that bandwidths may be chosen using data-driven methods such as cross-validation. See BJR1 for details.

4.3 Regime-switching models

We now consider panel data models with time-varying latent variables. In these models, multiple measurements may be particularly useful because they can allow to identify and estimate the transition probabilities of the latent states X_t, $t \in \{1, ..., T\}$. We show in this section that one can nonparametrically identify and estimate $\Pr(X_2 = k)$ and $\mathbb{E}_k\varphi(Y_t) = \mathbb{E}\left[\varphi(Y_t)|X_t = k\right]$ for $t = 2, ..., T-1$, and $\Pr(X_t|X_{t-1})$ for $t = 3, ..., T-1$. The first and last transitions cannot be recovered nonparametrically without further assumptions.

4.3.1 Three measurements

Consider first the case of three measurements (Y_1, Y_2, Y_3). Under Assumption 2, (Y_1, Y_2, Y_3) are independent given X_2. It follows that one can apply the results obtained above with

$$A = \mathbb{E}\left[\Psi(Y_1)\Psi(Y_3)'\right], \quad B(\varphi) = W_1\mathbb{E}\left[\Psi(Y_1)\Psi(Y_3)'\varphi(Y_2)\right]W_2'.$$

Assuming that A has maximal rank and that $\Pr(X_2 = k) = \pi_{k2} > 0$ for all k, these matrices identify $\mathbb{E}\left[\varphi(Y_2)|X_2 = k\right] = \mathbb{E}_k\varphi(Y_2)$ and π_{k2} for all k, and also $\mathbb{E}\left[\varphi(Y_1)|X_2 = k\right]$ and $\mathbb{E}\left[\varphi(Y_3)|X_2 = k\right]$. Yet, it is not possible in general to identify the conditional moments $\mathbb{E}_k\varphi(Y_1)$ and $\mathbb{E}_k\varphi(Y_3)$ or the probabilities $\Pr(X_1 = k, X_2 = \ell)$ and $\Pr(X_2 = k, X_3 = \ell)$.

In the stationary case, the conditional distributions and transition probabilities remain constant over time, and both $\Pr(X_t = k|X_{t-1} = \ell)$ and all $\mathbb{E}_k\varphi(Y_t)$ may be identified based on three measurements (see BJR2). We now show how a fourth measurement allows to identify $\Pr(X_2 = k, X_3 = \ell)$ in the general, non stationary case.

4.3.2 Four measurements

Let the matrix used for whitening now be

$$A = \mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\right].$$

Moreover, let Π denote the $K \times K$ matrix whose (k,ℓ)-element is $\Pr(X_2 = k, X_3 = \ell)$.

Assumption 4. *A has rank K and Π is non singular.*

Let us denote the SVD of A as $A = USV'$, and let $W_1 = S^{-\frac{1}{2}}U'$ and $W_2 = S^{-\frac{1}{2}}V'$. Let

$$B_2(\varphi) = W_1\mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\varphi(Y_2)\right]W_2', \quad B_3(\varphi) = W_1\mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\varphi(Y_3)\right]W_2',$$

and let

$$D_2(\varphi) = \operatorname{diag}\left(\mathbb{E}_1\varphi(Y_2), ..., \mathbb{E}_K\varphi(Y_2)\right), \quad D_3(\varphi) = \operatorname{diag}\left(\mathbb{E}_1\varphi(Y_3), ..., \mathbb{E}_K\varphi(Y_3)\right),$$

for $\mathbb{E}_k\varphi(Y_t) = \mathbb{E}\left[\varphi(Y_t)|X_t = k\right]$.

The following result shows that the joint distribution of (Y_2, X_2, Y_3, X_3) is nonparametrically identified.

Proposition 3. *Let Assumptions 2 and 4 hold. Let Q and R be two non-singular $K \times K$ matrices solutions to the simultaneous diagonalization problems,*

$$Q^{-1}B_2(\varphi)Q = D_2(\varphi), \qquad R^{-1}B_3(\varphi)R = D_3(\varphi),$$

for all univariate functions φ. Q and R are unique up to rescaling and permutation of their columns provided for all $k \neq k'$ there exists φ and φ' such that $\mathbb{E}_k\varphi(Y_2) \neq \mathbb{E}_{k'}\varphi(Y_2)$ and $\mathbb{E}_k\varphi'(Y_3) \neq \mathbb{E}_{k'}\varphi'(Y_3)$. Conditional moments of Y_2 and Y_3, $\mathbb{E}_k\varphi(Y_2)$ and $\mathbb{E}_k\varphi(Y_3)$, are identified as the eigenvalues. Moreover, the probability matrix of (X_2, X_3) is given, up to permutation of its rows and columns, by

$$\Pi = \operatorname{diag}\left(Q'W_2\mathbb{E}\Psi(Y_4)\right) \times (Q^{-1}R) \times \operatorname{diag}\left(R^{-1}W_1\mathbb{E}\Psi(Y_1)\right).$$

Proposition 3 allows to construct estimators $\widehat{Q}$ and $\widehat{R}$ by solving two approximate joint diagonalization problems. An estimator of Π is then given by

$$\widehat{\Pi} = \text{diag}\left(\widehat{Q}'\widehat{W}_2\widehat{\mathbb{E}}\Psi(Y_4)\right) \times (\widehat{Q}^{-1}\widehat{R}) \times \text{diag}\left(\widehat{R}^{-1}\widehat{W}_1\widehat{\mathbb{E}}\Psi(Y_1)\right).$$

Conditional moments $\mathbb{E}_k\varphi(Y_2)$ and $\mathbb{E}_k\varphi(Y_3)$ can then be estimated as simple weighted averages, as above. The asymptotic distributions of all these quantities can be derived using essentially the same arguments as in the case of time-invariant heterogeneity detailed in Appendix A.2.

5 Illustration on wage distributions

A simple representation of individual log wages is

$$Y_{it} = X_i + \eta_{it}, \tag{5.15}$$

where Y_{it} may be log wages or residuals from a standard Mincer equation, X_i is a worker effect, and η_{it} is an idiosyncratic white noise process. In a classic paper, Gottschalk and Moffitt (1994) estimate model (5.15) on log earnings residuals, and contrast US earnings inequality in the 1970s with earnings inequality in the 1980s. Model (5.15) has been extended in various directions, replacing the worker effect by a random walk with individual-specific drift or initial condition, or replacing the white noise by a more general ARMA process, see for example Moffitt and Gottschalk (2012). In this section, we take a nonparametric approach and show how finite mixtures can be used to document the structure and evolution of wage inequality in the U.S.

From the PSID 1969–1998 we construct a set of non-overlapping three-period balanced subpanels.[10] In each subpanel, we compute log hourly wages Y_{it}. Taking instead residuals

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

from a pooled regression of log wages on a set of time dummies, years of schooling, and a second-degree polynomial in experience gave similar results.

We first estimate conditional means and variances of log wages given the unobserved worker types (Figures 1 and 2). Througout, we use the estimator $\widehat{\theta}_{k3}$ as defined in (3.12). We focus on a small number of types, $K = 3$, for ease of exposition. In this way, one can think of the latent X as an indicator for low, intermediate, and high values of unobserved ability, for example. We label latent groups by decreasing order of the conditional means.

The first two groups have rather stable log wage means, which increase after 1990. The last group's mean steadily decreases throughout the whole period. All groups show increasing dispersion over time, accelerating after 1990. The standard deviations of groups 1 and 3 show similar trends, and their levels are higher than the standard deviation of group 2. These differences confirms the usefulness of allowing for type-specific differences in distributions, beyond differences in means.

Figure 3 shows how the total variance of log wages decomposes into within-group (WG) and between-group (BG) components. The BG-component clearly takes the bigger share (about 75%).

We then estimate the conditional densities for each subpanel using the weighted kernel density estimator in equation (4.14). The densities were estimated using our weighted kernel density estimator with bandwidth set by cross-validation. Figure 4 contains the estimated conditional densities for a selection of subpanels. All component densities are estimated unimodal and rather symmetric. These nonparametric results could be useful to guide the choice of parametric specifications of wage distributions.

Appendix

A.1 Proofs

A.1.1 Proof of Proposition 1

Define the $J \times K$ matrices

$$C_m = [\mathbb{E}_1\Psi(Y_m), ..., \mathbb{E}_K\Psi(Y_m)], \quad m \in \{1, 2, 3\},$$

and the $K \times K$ diagonal matrix $\Omega = \text{diag}(\pi_1, ..., \pi_K)$. By Assumption 1 (conditional independence) we have

$$A_{12} \equiv \mathbb{E}[\Psi(Y_1)\Psi(Y_2)'] = \sum_{k=1}^{K} \pi_k \, \mathbb{E}_k\Psi(Y_1) \, \mathbb{E}_k\Psi(Y_2)' = C_1\Omega C_2', \tag{A.1}$$

and, for any scalar function φ,

$$A_{123}(\varphi) \equiv \mathbb{E}[\Psi(Y_1)\Psi(Y_2)'\varphi(Y_3)] = C_1\Omega D_3(\varphi)C_2', \tag{A.2}$$

where we have denoted $D_3(\varphi) = \text{diag}(\mathbb{E}_1\varphi(Y_3), ..., \mathbb{E}_K\varphi(Y_3))$.

Next, write the singular value decomposition (SVD) of A_{12} as

$$A_{12} = USV',$$

where U and V are $J \times K$, with orthogonal columns, and S is $K \times K$ diagonal. All these matrices have rank K by Assumption 3. Let $W_1 = S^{-\frac{1}{2}}U'$ and $W_2 = S^{-\frac{1}{2}}V'$, and let

$$Q = W_1C_1\Omega, \tag{A.3}$$

which is also non-singular by Assumption 3. Equation (A.1) then implies that

$$W_1C_1\Omega C_2'W_2' = W_1A_{12}W_2' = I_K,$$

where I_K is the identity matrix of size K. Hence

$$C_2'W_2' = Q^{-1}. \tag{A.4}$$

It thus follows from (A.2) that

$$\begin{aligned} Q^{-1}W_1\mathbb{E}\left[\Psi(Y_1)\Psi(Y_2)'\varphi(Y_3)\right]W_2'Q &= Q^{-1}W_1C_1\Omega D_3(\varphi)C_2'W_2'Q \\ &= D_3(\varphi), \end{aligned}$$

which is equation (3.4) of Proposition 1. The matrices

$$B(\varphi) = W_1\mathbb{E}\left[\Psi(Y_1)\Psi(Y_2)'\varphi(Y_3)\right]W_2'$$

can thus be diagonalized in the same basis, and the moments $\mathbb{E}_k\varphi(Y_3)$ are their eigenvalues.

Lastly, by Theorem 6.1 in De Lathauwer et al. (2004) the matrix Q of joint eigenvectors is unique up to scaling and permutation of its columns.

Remark. Note that

$$\mathbb{E}\Psi(Y_1) = C_1\Omega e,$$

denoting as e the $K \times 1$ vector of ones. Hence, $\widetilde{Q} = Q\Delta^{-1}$, for some invertible diagonal matrix $\Delta = \text{diag}(\delta), \delta \in \mathbb{R}^{K\times 1}$, is identified up to permutation of its columns. Now,

$$W_1\mathbb{E}\Psi(Y_1) = \widetilde{Q}\Delta e = \widetilde{Q}\delta,$$

so $\delta = \widetilde{Q}^{-1}W_1\mathbb{E}\Psi(Y_1)$, from which it follows that Q is identified up to permutation of its columns.

A.1.2 Proof of Proposition 2

Let φ be an $\mathbb{R}$-valued, univariate function. We have, by Assumption 1,

$$\begin{aligned} \mathbb{E}\left[\Psi(Y_2)\varphi(Y_1)\right] &= C_2\Omega v_1(\varphi), \\ \mathbb{E}\left[\Psi(Y_1)\varphi(Y_2)\right] &= C_1\Omega v_2(\varphi), \end{aligned}$$

where $v_m(\varphi) = (\mathbb{E}_1\varphi(Y_m), ..., \mathbb{E}_K\varphi(Y_m))', m = 1, 2$. Let Q be one solution to the simultaneous diagonalization problem in Proposition 1. Then, by equations (A.3) and (A.4), there

exists $\lambda_k \neq 0, k = 1, ..., K$, and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_K)$ such that, up to columns permutation,

$$Q = W_1 C_1 \Omega \Lambda^{-1}, \quad Q^{-1} = \Lambda C_2' W_2'.$$

Hence,

$$W_2 \mathbb{E}\left[\Psi(Y_2)\varphi(Y_1)\right] = (Q^{-1})' \Lambda^{-1} \Omega v_1(\varphi), \tag{A.5}$$

$$W_1 \mathbb{E}\left[\Psi(Y_1)\varphi(Y_2)\right] = Q \Lambda v_2(\varphi). \tag{A.6}$$

Taking $\varphi = 1$ we obtain

$$\lambda_k = e_k' Q^{-1} W_1 \mathbb{E}\Psi(Y_1), \tag{A.7}$$

$$\pi_k = \lambda_k e_k' Q' W_2 \mathbb{E}\Psi(Y_2). \tag{A.8}$$

Note that $\pi_k \neq 0$ for all k by Assumption 3. It follows that, for any φ,

$$v_1(\varphi) = \Omega^{-1} \Lambda Q' W_2 \mathbb{E}\left[\Psi(Y_2)\varphi(Y_1)\right], \tag{A.9}$$

$$v_2(\varphi) = \Lambda^{-1} Q^{-1} W_1 \mathbb{E}\left[\Psi(Y_1)\varphi(Y_2)\right]. \tag{A.10}$$

Combining this with (A.7) and (A.8) yields (3.7) and (3.8).

A.1.3 A symmetrization result by Anandkumar et al. (2012)

Define $A_{ij} = \mathbb{E}\left[\Psi(Y_i)\Psi(Y_j)'\right] = C_i \Omega C_j'$ for all $i \neq j \in \{1, 2, 3\}$. Let $A_{12} = USV'$ be the SVD of matrix A_{12}, with $S \in \mathbb{R}^{K \times K}$ a non singular diagonal matrix. Define

$$\widetilde{A}_{12} = U' A_{12} V, \quad \widetilde{A}_{13} = U' A_{13}, \quad \widetilde{A}_{32} = A_{32} V.$$

Note that $\widetilde{A}_{12} = U' C_1 \Omega C_2' V = S$ is invertible. It follows that matrices $U' C_1$ and $C_2' V$ are invertible as Ω has non zero diagonal entries. Then,

$$\widetilde{A}_{13}' (\widetilde{A}_{12}')^{-1} \widetilde{A}_{32}' = C_3 \Omega C_1' U \left[(C_1' U)^{-1} \Omega^{-1} (V' C_2)^{-1}\right] V' C_2 \Omega C_3' = C_3 \Omega C_3'.$$

Moreover, define $\widetilde{A}_{123}(\varphi) = U'A_{123}(\varphi)V$. Then

$$\widetilde{A}_{32}\widetilde{A}_{12}^{-1}\widetilde{A}_{123}(\varphi)\widetilde{A}_{12}^{-1}\widetilde{A}_{13} = C_3\Omega C_2'V\left[(C_2'V)^{-1}\Omega^{-1}(U'C_1)^{-1}\right] \times \left[U'C_1\Omega D_3(\varphi)C_2'V\right] \times \left[(C_2'V)^{-1}\Omega^{-1}(U'C_1)^{-1}\right]U'C_1\Omega C_3' = C_3\Omega D_3(\varphi)C_3'.$$

It follows that the methods of BJR1 directly apply under the additional restriction that C_3 has rank K. However, as pointed out in the text, this condition may be unlikely when Y_3 has coarse support.

A.1.4 Proof of Proposition 3

Define the $J \times K$ matrices

$$C_1 = \left(\mathbb{E}\left[\Psi(Y_1) \,|\, X_2 = 1\right], ..., \mathbb{E}\left[\Psi(Y_1) \,|\, X_2 = K\right]\right),$$
$$C_4 = \left(\mathbb{E}\left[\Psi(Y_4) \,|\, X_3 = 1\right], ..., \mathbb{E}\left[\Psi(Y_4) \,|\, X_3 = K\right]\right).$$

By Assumption 2 we have

$$A = \mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\right] = \sum_{k=1}^{K}\sum_{\ell=1}^{K}\Pr(X_2 = k, X_3 = \ell)\mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)' \,|\, X_2 = k, X_3 = \ell\right]$$
$$= \sum_{k=1}^{K}\sum_{\ell=1}^{K}\Pr(X_2 = k, X_3 = \ell)\mathbb{E}\left[\Psi(Y_1) \,|\, X_2 = k\right]\mathbb{E}\left[\Psi(Y_4)' \,|\, X_3 = \ell\right],$$

making use of the fact that, under Assumption (2),

$$f(Y_1, Y_2, Y_3, Y_4|X_2, X_3) = f(Y_1|X_2)f(Y_2|X_2)f(Y_3|X_3)f(Y_4|X_3),$$

where $f(Y|Z)$ denotes the density of Y conditional on Z for any Y, Z.

Hence

$$A = C_1\Pi C_4'. \tag{A.11}$$

It is also straightforward to verify that

$$\mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\varphi(Y_2)\right] = C_1D_2(\varphi)\Pi C_4',$$
$$\mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\varphi(Y_3)\right] = C_1\Pi D_3(\varphi)C_4',$$

for $D_t(\varphi) = \text{diag}\left(\mathbb{E}_1\varphi(Y_t), ..., \mathbb{E}_K\varphi(Y_t)\right)$, with $\mathbb{E}_k\varphi(Y_t) = \mathbb{E}\left[\varphi(Y_t)|X_t = k\right]$.

Using the SVD of $A\,(= USV')$, and defining W_1 and W_2 as in the text, let

$$Q = W_1 C_1, \tag{A.12}$$

which is non-singular by Assumption 4. From (A.11) we get

$$W_1 C_1 \Pi C_4' W_2' = I_K.$$

Hence

$$\Pi C_4' W_2' = Q^{-1}. \tag{A.13}$$

Moreover,

$$\begin{aligned} Q^{-1} B_2(\varphi) Q &= D_2(\varphi), \\ Q^{-1} B_3(\varphi) Q &= \Pi D_3(\varphi) \Pi^{-1}, \end{aligned}$$

where $B_t(\varphi) = W_1 \mathbb{E}\left[\Psi(Y_1)\Psi(Y_4)'\varphi(Y_t)\right] W_2'$. Hence, similarly as in Proposition 1, $\mathbb{E}_k\varphi(Y_2)$ and $\mathbb{E}_k\varphi(Y_3)$ follow as the eigenvalues of two simultaneous diagonalization problems. The matrices of common eigenvectors, Q and $Q\Pi$, are therefore also unique up to rescaling and permutation of their columns.

This implies that, for two $K \times K$ non-singular diagonal matrices Λ and Δ, and up to relabeling of their columns, we have

$$Q = W_1 C_1 \Lambda, \quad R = W_1 C_1 \Pi \Delta,$$

where Q and R are any solutions to

$$Q^{-1} B_2(\varphi) Q = D_2(\varphi), \quad R^{-1} B_3(\varphi) R = D_3(\varphi),$$

for all φ.

Now, note that, by Assumption 2, and denoting as e the $K \times 1$ vector of ones,

$$\mathbb{E}\Psi(Y_1) = C_1 \Pi e,$$

so

$$W_1 \mathbb{E}\Psi(Y_1) = R\Delta^{-1} e,$$

from which it follows that

$$\Delta^{-1} = \operatorname{diag}\left(R^{-1} W_1 \mathbb{E}\Psi(Y_1)\right).$$

Likewise,

$$\mathbb{E}\Psi(Y_4) = C_4 \Pi' e,$$

so

$$W_2 \mathbb{E}\Psi(Y_4) = (Q')^{-1} \Lambda e,$$

from which it follows that

$$\Lambda = \operatorname{diag}\left(Q' W_2 \mathbb{E}\Psi(Y_4)\right).$$

Combining results, we finally obtain

$$\Pi = \operatorname{diag}\left(Q' W_2 \mathbb{E}\Psi(Y_4)\right) \times (Q^{-1} R) \times \operatorname{diag}\left(R^{-1} W_1 \mathbb{E}\Psi(Y_1)\right).$$

A.2 Asymptotic theory

The parameter of interest is

$$\theta = \mathbb{E}_k \varphi(Y_3) = \mathbb{E}[\tau_k(Y_1, Y_2)\, \varphi(Y_3)]$$

for fixed k. The estimator is

$$\hat{\theta} = \widehat{\mathbb{E}}\left[\hat{\tau}_k(Y_1, Y_2)\, \varphi(Y_3)\right],$$

with the weight functions $\hat{\tau}_k(Y_1, Y_2) = \widehat{\omega}_{1k}(Y_1)\widehat{\omega}_{2k}(Y_2)$.

To present the asymptotic distribution of $\hat{\theta}$, note that it is a plug-in version of the infeasible estimator

$$\tilde{\theta} = \widehat{\mathbb{E}}\left[\tau_k(Y_1, Y_2)\, \varphi(Y_3)\right],$$

that is, the estimator that would be used if the weights were known. This estimator is a simple sample average, and so the central limit theorem can be directly applied to show that $\sqrt{N}(\tilde{\theta} - \theta)$ is asymptotically normal. It remains only to quantify the impact of estimating the weights. Thus, we need to derive the asymptotic behavior of $\sqrt{N}(\hat{\theta} - \tilde{\theta})$. This requires quantifying the impact of (i) the whitening step, and (ii) the joint approximate diagonalization step. We turn to each of these next.

Whitening. Recall that the whitening is done using a plug-in estimator of the singular-value decomposition of the matrix

$$A = \mathbb{E}[\Psi(Y_1)\,\Psi(Y_2)'] = USV' = U_K S_K V_K',$$

where we now let S_K be the $K \times K$ block of S containing the non-zero singular values, and let U_K and V_K denote the associated left and right singular vectors. We denote as U, S and V the $J \times J$ matrices that contain U_K, V_K and S_K, respectively. Note that this notation differs from the one used in the main text. The whitening matrices

$$W_1 = S_K^{-\frac{1}{2}} U_K', \qquad W_2 = S_K^{-\frac{1}{2}} V_K',$$

are then estimated using the singular-value decomposition of

$$\widehat{A} = \widehat{\mathbb{E}}\left[\Psi(Y_1)\,\Psi(Y_2)'\right],$$

which is the empirical counterpart of A.

Let $\overset{\text{col}}{\otimes}$ and $\overset{\text{row}}{\otimes}$ be the columnwise and rowwise Kronecker product, respectively, and let $\ominus$ be the "Kronecker difference".[11] Define

$$\begin{aligned} J_{W_1} &= -(U \otimes I)\,(S^2 \ominus S_K^2)^{+}(U' \otimes W_1) - \frac{1}{4}(W_1' \overset{\text{col}}{\otimes} I)\,S_K^{-1}\,(W_1 \overset{\text{row}}{\otimes} W_1) \\ J_{W_2'} &= (I \otimes V)\,(S_K^2 \ominus S^2)^{+}(W_2 \otimes V') - \frac{1}{4}(I \overset{\text{col}}{\otimes} W_2')\,S_K^{-1}\,(W_2 \overset{\text{row}}{\otimes} W_2), \end{aligned}$$

[11] That is, $A \ominus B = A \otimes I_{\dim B} - I_{\dim A} \otimes B$.

where I denotes the identity matrix of conformable dimension and A^+ is the Moore-Penrose pseudo inverse of matrix A. In the following result we assume that the non-zero singular values of A are simple. This allows us to avoid issues related to asymptotic distributions depending on the multiplicity of singular values in a complicated way; see Eaton and Tyler (1991).

Lemma 1. *Assume that $\mathbb{E}[\psi_j^2(Y_m)]$ is finite for all $j = 1, ..., J$ and $m = 1, 2$, and suppose that all non-zero singular values of A are simple. Then*

$$\begin{aligned}\sqrt{N}\text{vec}(\widehat{W}_1 - W_1) &= J_{W_1}\sqrt{N}\text{vec}(\widehat{A}\widehat{A}' - AA') + o_p(1),\\ \sqrt{N}\text{vec}(\widehat{W}_2' - W_2') &= J_{W_2'}\sqrt{N}\text{vec}\,(\widehat{A}'\widehat{A} - A'A) + o_p(1),\end{aligned}$$

and are asymptotically normal.

Proof. The results can be proved by adapting the proof of Lemmas S.1 and S.2 in BJR1 to the eigendecompositions $AA' = US^2U'$ and $A'A = VS^2V'$. The condition $\mathbb{E}[\psi_j^2(Y_m)] < \infty$ allows to apply the Lindeberg-Lévy CLT to $\sqrt{N}\text{vec}(\widehat{A} - A)$. □

Note that under the conditions of Lemma 1 we have

$$\begin{aligned}\text{vec}(\widehat{A}\widehat{A}' - AA') &= (A \otimes I)\,\text{vec}(\widehat{A} - A) + (I \otimes A)\,\text{vec}(\widehat{A}' - A') + o_p(N^{-1/2}),\\ \text{vec}(\widehat{A}'\widehat{A} - A'A) &= (I \otimes A)'\,\text{vec}(\widehat{A} - A) + (A \otimes I)'\,\text{vec}(\widehat{A}' - A') + o_p(N^{-1/2}).\end{aligned}$$

Diagonalization. Introduce the shorthand

$$\overline{B}_j = \mathbb{E}[\Psi(Y_1)\Psi(Y_2)'\psi_j(Y_3)],$$

and write the whitened matrices compactly as

$$B_j = B(\psi_j) = W_1\overline{B}_jW_2'.$$

We estimate Q by the joint approximate diagonalizer of the sample counterparts of the B_j,

$$\widehat{B}_j = \widehat{W}_1\widehat{\overline{B}}_j\widehat{W}_2'.$$

Let vert denote the vertical concatenation operator, for example $B = \text{vert}[B_1, B_2, \ldots, B_J]$ and $\widehat{B} = \text{vert}[\widehat{B}_1, \widehat{B}_2, \ldots, \widehat{B}_J]$, and similarly let horz denote the horizontal concatenation operator. Introduce the matrix

$$H = (I \otimes Q) \left(\sum_{j=1}^{J} (D_j \ominus D_j)^2 \right)^{+} \text{horz}[D_1 \ominus D_1, \ldots, D_J \ominus D_J] \, (I \otimes Q' \otimes Q^{-1}).$$

Lemma 2. *Assume that $\mathbb{E}[\psi_j^2(Y_m)]$ is finite for all $j = 1, ..., J$ and $m = 1, 2$, and suppose that all non-zero singular values of A are simple. Then*

$$\sqrt{N}\text{vec}(\widehat{Q} - Q) \quad = \quad H \sqrt{N} \, \text{vec}(\widehat{B} - B) + o_p(1),$$

and is asymptotically normal.

Proof. Follows directly from Theorem 5 in BJR2. □

Under the conditions of Lemma 2,

$$\begin{aligned} \text{vec}(\widehat{B} - B) = \; & \text{vert}[W_2 \overline{B}_1' \otimes I, \ldots, W_2 \overline{B}_J' \otimes I] \, \text{vec}(\widehat{W}_1 - W_1) \\ & + \; \text{vert}[I \otimes W_1 \overline{B}_1, \ldots, I \otimes W_1 \overline{B}_J] \, \text{vec}(\widehat{W}_2 - W_2) \\ & + (I \otimes W_2 \otimes W_1) \, \text{vec}(\widehat{\overline{B}} - \overline{B}) + o_p(N^{-1/2}), \end{aligned}$$

where $\overline{B} = \text{vert}[\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_J]$ and $\widehat{\overline{B}} = \text{vert}[\widehat{\overline{B}}_1, \widehat{\overline{B}}_2, \ldots, \widehat{\overline{B}}_J]$.

Feasible estimator. With Lemmas 1 and 2 in hand, a standard argument (as in the proof of Theorem 2 in BJR1 gives

$$\hat{\theta} - \theta = \widehat{\mathbb{E}} \left[\tau_k(Y_1, Y_2) \varphi(Y_3) - \theta \right] + \left(\nu_{2k} (e_k' \otimes I) Z_1 + \nu_{1k} (I \otimes e_k) Z_2 \right) + o_p(N^{-1/2}),$$

where the second right-hand side term represents the contribution to the influence function of the estimation noise in the weights. It features the terms

$$\nu_{1k} = e_k' Q^{-1} W_1 B(\varphi), \qquad \nu_{2k} = e_k' Q' W_2 B(\varphi)',$$

and the random variables

$$\begin{aligned} Z_1 &= (I \otimes Q^{-1})\text{vec}(\widehat{W}_1 - W_1) - (W_1' \otimes I)(Q' \otimes Q)^{-1}\text{vec}(\widehat{Q} - Q) \\ Z_2 &= (Q' \otimes I)\text{vec}(\widehat{W}_2' - W_2') + (I \otimes W_2')\text{vec}(\widehat{Q} - Q), \end{aligned}$$

where expressions for $\text{vec}(\widehat{W}_1 - W_1)$, $\text{vec}(\widehat{W}_2' - W_2')$, and $\text{vec}(\widehat{Q} - Q)$ are given above.

It follows that $\hat{\theta}$ is asymptotically normal provided that the variance of $\varphi(Y_3)$ exists. It also follows that its asymptotic variance can be readily characterized.

References

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

Anandkumar, A., D. Foster, D. Hsu, S. Kakade, and Y.-K. Liu (2015). A spectral algorithm for latent dirichlet allocation. *Algorithmica 72*(1), 193–214.

Anandkumar, A., D. P. Foster, D. Hsu, S. M. Kakade, and Y. Liu (2012). A spectral algorithm for latent dirichlet allocation. *CoRR abs/1204.6703*.

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

Anandkumar, A., D. Hsu, and S. M. Kakade (2012). A method of moments for mixture models and hidden markov models. *JMLR Workshop and Conference Proceedings, COLT 23(33)*, 1–34.

Bajari, P., J. Hahn, H. Hong, and G. Ridder (2011). A note on semiparametric estimation of finite mixtures of discrete choice models with application to game theoretic models. *International Economic Review 52*(3), 807–824.

Benaglia, T., T. Chauveau, and D. R. Hunter (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., K. Jochmans, and J.-M. Robin (2015). Estimating multivariate latent-structure models. *Forthcoming Annals of Statistics*.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016). Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 211–229.

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

Chauveau, D., D. R. Hunter, and M. Levine (2014). Semi-parametric estimation for conditional independence multivariate mixture models. Technical Report No 14-02, Department of Statistics, Purdue University.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*, 559–575.

Cichocki, A., D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. Phan (2015, March). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *Signal Processing Magazine, IEEE 32*(2), 145–163.

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

De Castro, Y., E. Gassiat, and C. Lacour (2015, January). Minimax adaptive estimation of non-parametric hidden markov models. *ArXiv e-prints*.

De Lathauwer, L., B. D. Moor, and J. Vandewalle (2004). Computation of the canonical decomposition by means of a simultaneous generalized Shur decomposition. *SIAM journal of matrix analysis and applications 26*, 295–327.

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

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

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

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

Geweke, J. and M. Keane (2000). An empirical analysis of earnings dynamics among men in the PSID: 1968-1989. *Journal of Econometrics 96*(2), 293–356.

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

Gu, J. and R. Koenker (2014). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. Technical report, University of Illinois, Department of Economics.

Hahn, J. and H. R. Moon (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory 26*, 863–881.

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

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica 57*(2), 357–84.

Hsu, D., S. M. Kakade, and T. Zhang (2012). A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences 78*, 1460–1480.

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

Hu, Y. (2015). Microeconomic models with latent variables: Applications of measurement error models in empirical industrial organization and labor economics. Technical report, Cemmap Working Papers, CWP03/15.

Iferroudjene, R., K. A. Meraim, and A. Belouchrani (2009). A new jacobi-like method for joint diagonalization of arbitrary non-defective matrices. *Applied Mathematics and Computation 211*, 363–373.

Iferroudjene, R., K. A. Meraim, and A. Belouchrani (2010). Joint diagonalization of non defective matrices using generalized jacobi rotations. In *10th International Conference on Information Sciences, Signal Processing and their Applications, ISSPA 2010, Kuala Lumpur, Malaysia, 10-13 May, 2010*, pp. 345–348.

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

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

Keane, M. P. and R. M. Sauer (2009). Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior. *Econometrica 77*(3), 975–991.

Keane, M. P. and K. I. Wolpin (1997). The Career Decisions of Young Men. *Journal of Political Economy 105*(3), 473–522.

Levine, M., D. R. Hunter, and D. Chauveau (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika 98*, 403–416.

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

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica 74*(3), 631–665.

Moffitt, R. A. and P. Gottschalk (2012). Trends in the transitory variance of male earnings: Methods and evidence. *Journal of Human Resources 47*(1), 204–236.

Song, L., A. Anandkumar, B. Dai, and B. Xie (2013). Nonparametric estimation of multi-view latent variable models. *CoRR abs/1311.3287*.

Figure 1: Means

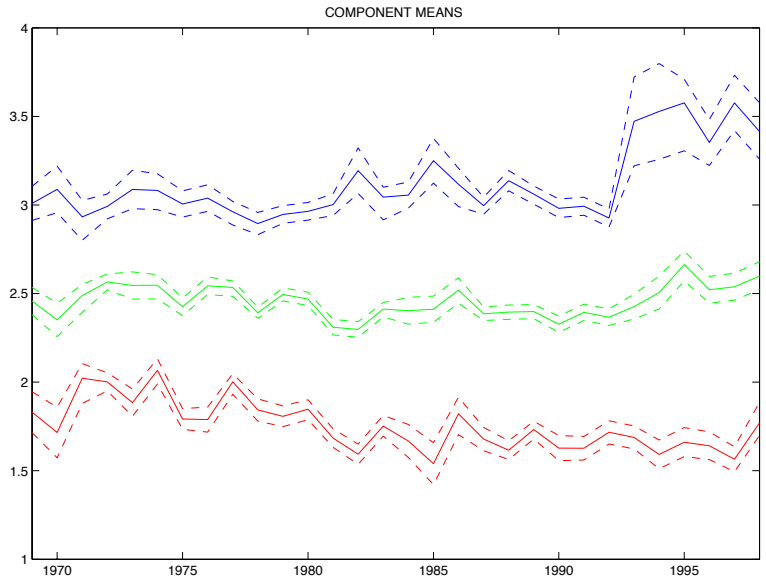


Figure 2: Standard deviations

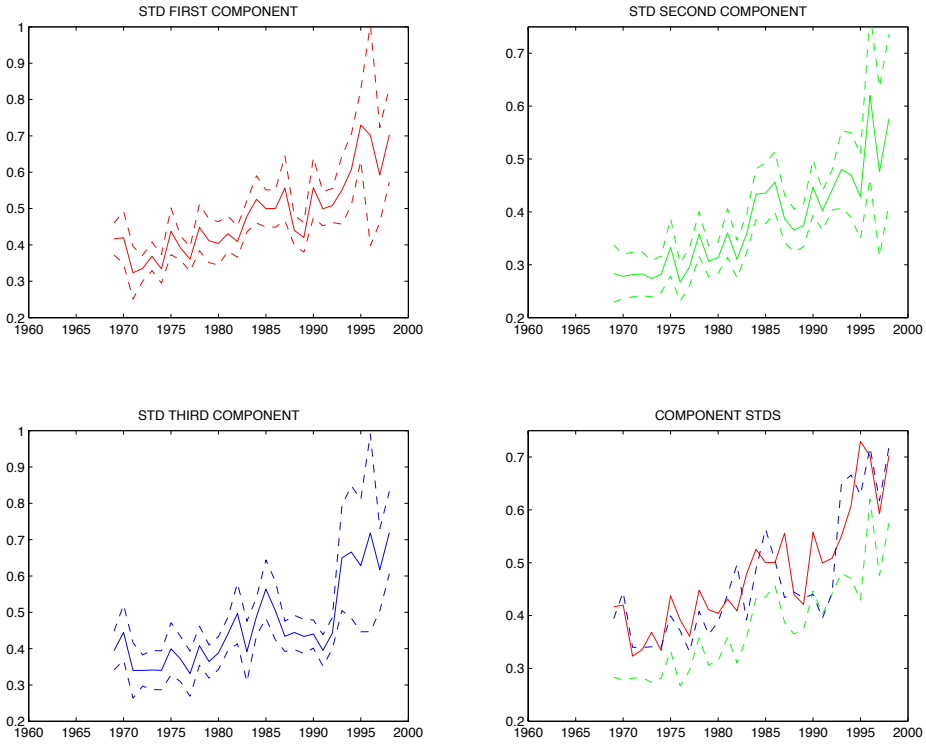


Figure 3: Within-between variance decompositions

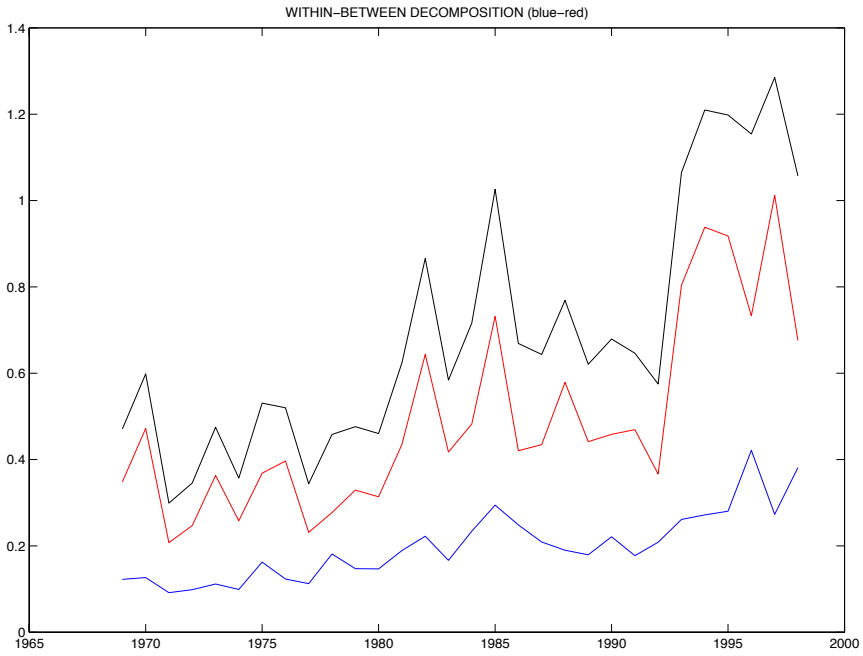
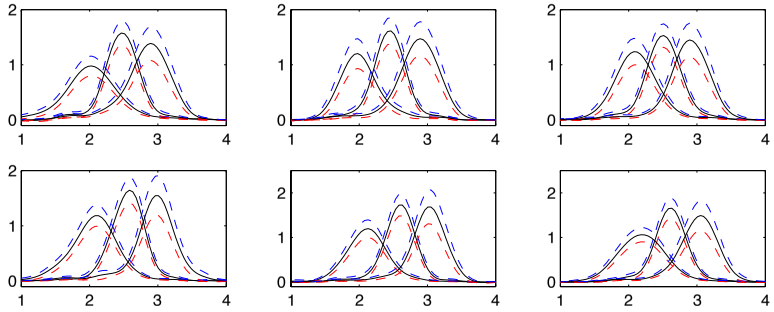
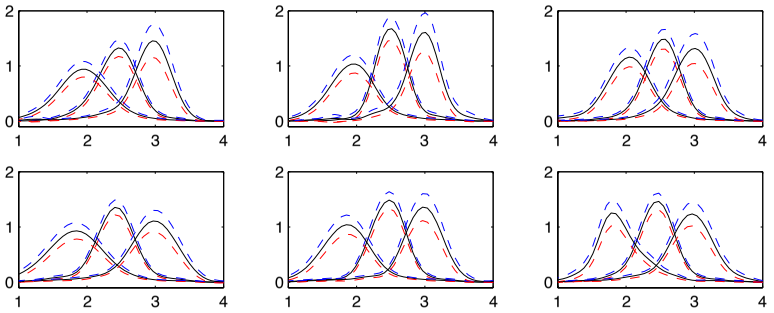


Figure 4: Component Densities

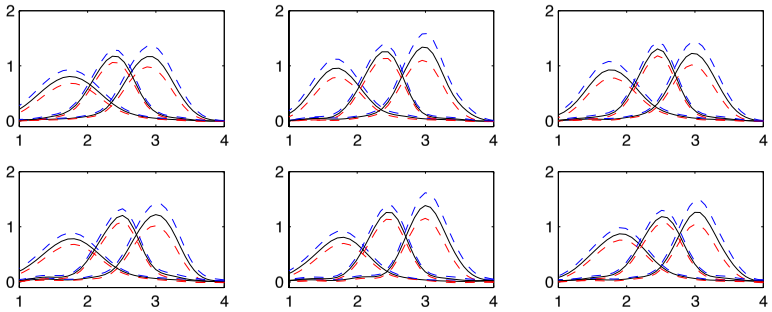
(a) 1969-1974



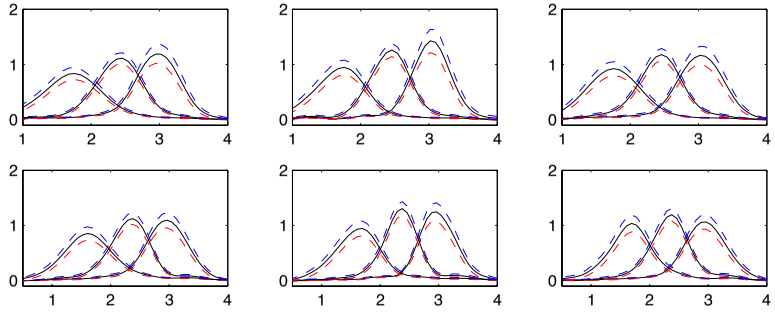
(b) 1975-1980



(c) 1981-1986



(d) 1987-1992



(e) 1993-1998