IDENTIFICATION OF MIXTURES OF DYNAMIC DISCRETE CHOICES

Ayden Higgins* University of Oxford

Koen Jochmans† Toulouse School of Economics

This version: January 19, 2023

Abstract

This paper provides new identification results for finite mixtures of Markov processes. Our arguments yield identification from knowledge of the cross-sectional distribution of three (or more) effective time-series observations under simple conditions. We explain how our approach and results are different from those in previous work by Kasahara and Shimotsu (2009) and Hu and Shum (2012). Most notably, outside information, such as monotonicity restrictions that link conditional distributions to latent types, is not needed.

JEL Classification: C14, C23, C51

Keywords: discrete choice, heterogeneity, Markov process, mixture, state dependence

Introduction

The analysis of dynamic discrete choices from short panel data is fundamental in applied work. Allowing for unobserved heterogeneity in such a setting is recognized to be important (Heckman 1981) but doing so in a flexible manner is known to be difficult. A leading paradigm is to presume that the population of agents is composed of a finite number of (latent) types, implying that the (marginal) distribution of the data takes the form of

*Address: Department of Economics, University of Oxford, 10 Manor Road, Oxford OX1 3UQ, United Kingdom. E-mail: ayden.higgins@economics.ox.ac.uk.

†Address: Toulouse School of Economics, 1 esplanade de l'Université, 31080 Toulouse, France. E-mail: koen.jochmans@tse-fr.eu.

Support from the European Research Council through grant ERC-2016-STG-715787, and from the French Government and the ANR under the Investissements d' Avenir program, grant ANR-17-EURE-0010 is gratefully acknowledged. An associate editor, three referees, Thierry Magnac and seminar participants at Brown, HKUST, Yale, and Montreal gave much appreciated feedback.

a finite mixture. Keane and Wolpin (1997), Eckstein and Wolpin (1999), Crawford and Schum (2005), Wang (2014), Nevo, Turner and Williams (2016), and Rossi (2017) are examples of papers that have taken this route.

In earlier work, Kasahara and Shimotsu (2009) have studied the identifiability of finite mixtures of first-order Markov processes. Their result is commonly invoked in the literature to claim identification; see, e.g., Arcidiacono and Miller (2011), Baum-Snow and Pavan (2012), Norets and Tang (2014), Chen (2017), Kalouptsidi, Scott and Souza-Rodrigues (2020), Kalouptsidi, Kitamura, Lima and Souza-Rodrigues (2021). The approach taken in Kasahara and Shimotsu (2009) closely follows work on (static) multivariate models with latent variables (in particular Anderson 1954 and Hall and Zhou 2003). Moreover, they exploit implications of the dynamic model to which the machinery for identification in the static case can be applied. These restrictions are, however, not sufficient to recover the type-specific distributions or the mixing distribution. This issue does not seem to be well appreciated and is a subtle consequence of the fact that the labelling of types is arbitrary, and can be changed without observable implications. We discuss this in more detail below. Hu and Shum (2012) follow a similar strategy to Kasahara and Shimotsu (2009), relying on restrictions inspired by the approach of Hu (2008). These restrictions again do not suffice to obtain identification. To resolve the issue, Hu and Shum (2012) supplement the model with outside information in the form of a set of monotonicity restrictions that link latent types to observable choices. Aside from practical issues that arise when taking such an approach to the data, monotonicity conditions may be difficult to justify or may simply not be available in many applications.

We develop a new identification argument that shows that the type-specific transition kernels, the type-specific distributions of the initial condition, and the mixing distribution are all recoverable from knowledge of the cross-sectional distribution of as little as four time-series observations under three simple conditions which ensure that the type-specific Markov processes are sufficiently different. Like Kasahara and Shimotsu (2009) and Hu and Shum (2012), we, too, exploit (different) multilinear restrictions that are reminiscent of those employed in the literature on multivariate mixtures (Bonhomme, Jochmans and

Robin 2016). However, these are only a subset of a larger set of restrictions implied by the Markovian structure of the model. While the subset of restrictions alone does not fully identify the unknown distributions, the full set of restrictions does. Identification is achieved without the need to impose additional structure, such as monotonicity restrictions. Our arguments also extend naturally to models with higher-order Markovian dependence. The general conclusion, then, is that mixtures of p-th order Markov processes can be identified from the cross-sectional distribution of $3 + p$ time-series observations.

The model is introduced in Section 1. Our assumptions and identification argument are presented in Section 2. A detailed comparison with the assumptions and approaches of Kasahara and Shimotsu (2009) and Hu and Shum (2012) is made in Section 3. We illustrate our assumptions in specific examples in Section 4. We show how our approach extends to models with higher-order Markovian dependence in Section 5. A short conclusion ends the paper. Appendices collect an auxiliary lemma and details on maximum-likelihood estimation via the EM algorithm.

1 Mixtures of dynamic discrete choices

Suppose that Z is a latent random variable that can take on q values, where q is a known integer. We normalize its support to the set of integers up to q, which is without loss of generality, and write $\mu_1, \ldots, \mu_q$ for its probability mass function. So, $\mu_z \coloneqq \mathbb{P}(Z = z) > 0$ for $1 \leq z \leq q$ and zero otherwise. Next let $\{X_t\}$ be a sequence of observable random variables that can take on r values. We presume that its support constitutes the set of integers up to r. This is merely for notational convenience in what is to follow; translation of the support to a general set is straightforward. Conditional on $Z = z$, the sequence $\{X_t\}$ follows a first-order Markov process. The process is initialized with a draw from the distribution

$$s_z(x) \coloneqq \mathbb{P}(X_1 = x | Z = z),$$

and subsequently evolves according to the time-homogeneous transition kernel

$$k_z(x, x') := \mathbb{P}(X_t = x' | X_{t-1} = x, Z = z).$$

This delivers a dynamic model of discrete choice with unobserved heterogeneity captured by a mixture over q latent types. The dynamic processes are allowed to be non-stationary in that the initial conditions are not assumed to have been drawn from the steady-state distribution.

Our goal is to recover the distribution of the latent types, $\mu_1, \ldots, \mu_q$, the distributions of the initial conditions, $s_1, \ldots, s_q$, and the transition kernels, $k_1, \ldots, k_q$, from knowledge of the joint distribution of X_1, X_2, X_3, X_4. Our arguments to follow can be generalized to the case where additional time-series observations are available and we discuss how to do so below. As latent types can be relabelled without any observable implications, identification here is to be understood as being up to an arbitrary re-ordering of types.

Before turning to the identification analysis it is useful to point out that restricting attention to univariate variables is without loss of generality. To see this, suppose that we are interested in a sequence of k-dimensional vectors $\boldsymbol{V}_t$ whose entries can take on, respectively, $r_1, \ldots, r_k$ values. Then we can always enumerate all values in the state space of $\boldsymbol{V}_t$ and define a scalar random variable X_t on this set of numbers that is a (known) one-to-one transformation of $\boldsymbol{V}_t$. This random variable can take on $r = r_1 \times \ldots \times r_k$ values. Identification of a mixture on X_t then implies identification of the corresponding mixture on $\boldsymbol{V}_t$. As will become apparent below, a larger r can only make the identification problem easier, and so observing more variables is helpful for identification. This connects to the discussion in Hall and Zhou (2003) and Hall, Neeman, Pakyari and Elmore (2005) on the identification power of multivariate mixtures. It might be the case that, for a partitioning of $\boldsymbol{V}_t$ into $\boldsymbol{Y}_t$ and $\boldsymbol{W}_t$, the ultimate goal is to recover the (type-specific) distribution of the initial condition and the transition kernel of $\boldsymbol{Y}_t$ conditional on $\boldsymbol{W}_t$. Here, $\boldsymbol{Y}_t$ is the outcome of interest while $\boldsymbol{W}_t$ plays the role of covariates. This corresponds to the point of view in Kasahara and Shimotsu (2009). Such conditional distributions are, of course, identified once the joint distribution is.

2 Identification

Assumptions. Our constructive identification approach employs three assumptions. We first introduce notation for probabilities that involve only observable variables. We will use the shorthands $p_{x_1,x_2} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2)$ and $p_{x_1,x_2,x_3} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ as well as $p_{x_1,x_2,x_3,x_4} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$. First, we collect all the bivariate probabilities into the two sets of r-vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r$ and $\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_r^*$, where we write

$$\boldsymbol{p}_x \coloneqq (p_{1,x}, \ldots, p_{r,x})^\top, \qquad \boldsymbol{p}_x^* \coloneqq (p_{x,1}, \ldots, p_{x,r})^\top.$$

Note that $\boldsymbol{p}_x \neq \boldsymbol{p}_x^*$, in general. Similarly, we construct $r \times r$ matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_r$ containing trivariate probabilities, letting

$$(\boldsymbol{P}_x)_{i,j} \coloneqq p_{j,x,i}.$$

Finally, we do the same for probabilities involving all four periods, by introducing the $r \times r$ matrices

$$(\boldsymbol{P}_{x,x'})_{i,j} \coloneqq p_{j,x,x',i},$$

with (x, x') ranging over all the r^2 possibilities. The probabilities involved here are all nonparametrically identified and these matrices may thus all be considered known for our purposes.

Our first assumption is a rank condition that is directly testable from the data.

Assumption 1. *For each x the $r \times r$ matrix $\boldsymbol{P}_x$ has rank q.*

A necessary condition for this assumption to hold is that $r \geq q$. The decomposition in Equation (2.1) below shows that the main implication of Assumption 1 is that, for each x, the conditional distributions of X_t given $X_{t-1} = x$ and $Z = z$ (seen as a function of z) are linearly independent.

To state our remaining assumptions we let

$$k_z^m(x; x_1, \ldots, x_m; x') \coloneqq k_z(x, x_1) \left(\bigtimes_{i=1}^{m-1} k_z(x_i, x_{i+1}) \right) k_z(x_m, x'),$$

be the probability of the walk from x to x' via the m intermediate stops $x_1, \ldots, x_m$ through the Markov chain of type z. To cover walks that go directly from x to x', that is, without passing through any intermediate points, we adopt the convention that $k_z^0(x; x') = k_z(x, x')$. So, m can be any non-negative integer.

Our second assumption involves walks from x back to x. For each triplet (z, x, m), let $\boldsymbol{f}_{z,x,m}$ be the vector that contains the probabilities $k_z^m(x; x_1, \ldots, x_m; x)$ across all possible sequences $x_1, \ldots, x_m$. Next collect these vectors across the q different types in the matrix

$$\boldsymbol{F}_{x,m} \coloneqq (\boldsymbol{f}_{1,x,m}, \ldots, \boldsymbol{f}_{q,x,m}).$$

For example, $\boldsymbol{F}_{x,0} = (k_1(x, x), \ldots, k_q(x, x))$. For $m > 0$ $\boldsymbol{F}_{x,m}$ has as many rows as there are unique sequences $x_1, \ldots, x_m$.

Assumption 2. *For each x there exists an integer $\bar{m}$ such that the columns of the matrix*

$$\boldsymbol{F}_x \coloneqq (\boldsymbol{F}_{x,0}^\top, \ldots, \boldsymbol{F}_{x,\bar{m}}^\top)^\top$$

are all distinct.

This requirement is quite weak. A simple sufficient (but substantially too strong) condition for Assumption 2 to go through is that, for each x, there exists a walk from x going back to x that occurs with a different probability for each type.

Our last assumption concerns a point x_0 and walks between it and all other points x.

Assumption 3. *There exists a value x_0 such that for each $x \neq x_0$ there exists a finite non-negative integer m such that*

$$(i)\ k_z^m(x_0; x_1, \ldots, x_m; x) \quad \textit{or} \quad (ii)\ k_z^m(x; x_1, \ldots, x_m; x_0)$$

is non-zero for some sequence $x_1, \ldots, x_m$ for each z. The sequence $x_1, \ldots, x_m$ and its length m may be different for the different x.

Assumption 3 requires that we can connect the value x_0 to all other values $x \neq x_0$ by some walk for all types. Only a single such x_0 is required. The developments to follow will make it apparent that evaluating whether any particular sequence of values satisfies this assumption can be done by testing whether a collection of $q \times q$ matrices have maximal rank.

Identification. We now proceed to establish identification. The proof can be broken down into four logical steps. In the first step we show that the model gives rise to a set of multilinear restrictions. In this format it becomes apparent that model parameters are identified if we are able to recover a collection of matrices up to a common permutation of their columns. In the second step we show that a subset of these multilinear restrictions can be used to recover each of these matrices up to an arbitrary permutation of their columns. In the third step we use the remaining multilinear restrictions to enforce an ordering on the columns that is common across matrices. In the fourth and final step we back out the parameters of our model.

1) Multilinear restrictions. We begin by constructing, for each x, the $r \times q$ matrices $\boldsymbol{K}_x$ and $\boldsymbol{L}_x$ as

$$(\boldsymbol{K}_x)_{x',z} \coloneqq k_z(x, x'), \qquad (\boldsymbol{L}_x)_{x',z} \coloneqq \mu_z \, s_z(x') \, k_z(x', x).$$

Next, we appeal to the Markovian structure of our model to see that, for each x, the factorization

$$\boldsymbol{P}_x = \boldsymbol{K}_x \boldsymbol{L}_x^\top \tag{2.1}$$

holds. Assumption 1 states that each $r \times r$ matrix $\boldsymbol{P}_x$ has rank equal to q. Hence, it has the singular-value decomposition

$$\boldsymbol{P}_x = \boldsymbol{U}_x \boldsymbol{E}_x \boldsymbol{V}_x^\top,$$

for unitary $r \times q$ matrices of, respectively, left and right singular vectors, $\boldsymbol{U}_x$ and $\boldsymbol{V}_x$, and $q \times q$ diagonal matrices $\boldsymbol{E}_x$ of singular values. It then follows that, if we use the shorthands $\boldsymbol{A}_x \coloneqq \boldsymbol{E}_x^{-1/2} \boldsymbol{U}_x^\top$ and $\boldsymbol{B}_x \coloneqq \boldsymbol{E}_x^{-1/2} \boldsymbol{V}_x^\top$,

$$\boldsymbol{A}_x \boldsymbol{P}_x \boldsymbol{B}_x^\top = \boldsymbol{I}_q, \tag{2.2}$$

with $\boldsymbol{I}_q$ being the $q \times q$ identity matrix. Now introduce the $q \times q$ matrix $\boldsymbol{Q}_x \coloneqq \boldsymbol{A}_x \boldsymbol{K}_x$. Combining Equation (2.1) with Equation (2.2) reveals that

$$\boldsymbol{I}_q = \boldsymbol{A}_x \boldsymbol{P}_x \boldsymbol{B}_x^\top = (\boldsymbol{A}_x \boldsymbol{K}_x)(\boldsymbol{B}_x \boldsymbol{L}_x)^\top = \boldsymbol{Q}_x \boldsymbol{Q}_x^{-1},$$

and so $\boldsymbol{Q}_x^{-\top} = \boldsymbol{B}_x \boldsymbol{L}_x$ must hold. Here, and later we use the superscript $-\top$ as a notational shorthand for the inverse of a matrix transpose, i.e., $\boldsymbol{Q}_x^{-\top} = (\boldsymbol{Q}_x^\top)^{-1}$.

We now turn to the distribution of all four observable variables. Notice that, in the same way as before,

$$\boldsymbol{P}_{x,x'} = \boldsymbol{K}_{x'} \boldsymbol{D}_{x,x'} \boldsymbol{L}_x^\top,$$

where $\boldsymbol{D}_{x,x'} \coloneqq \operatorname{diag}(k_1(x,x'), \ldots, k_q(x,x'))$ collects the transition probabilities from state x to x' for each of the different types z. Hence,

$$\boldsymbol{C}_{x,x'} \coloneqq \boldsymbol{A}_{x'} \boldsymbol{P}_{x,x'} \boldsymbol{B}_x^\top = \boldsymbol{Q}_{x'} \boldsymbol{D}_{x,x'} \boldsymbol{Q}_x^{-1} \tag{2.3}$$

for all (x, x'). If we can recover the matrices $\boldsymbol{Q}_x$ for all x we can invert the above system for each pair (x, x') to identify the matrix $\boldsymbol{D}_{x,x'}$ and, consequently the transition kernel k_z for each z. Identification of the remaining parameters of the model then follows readily. We thus next set out to recover $\boldsymbol{Q}_x$ for all x. Note that we can at best hope to recover these matrices up to a common permutation of their columns, as Equation (2.3) is invariant to such a permutation. Also note that it does not suffice to recover $\boldsymbol{Q}_x$ and $\boldsymbol{Q}_{x'}$ up to a different permutation of their columns, as this is insufficient to be able to solve Equation (2.3) for the transition kernels.

2) Identification of $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ up to arbitrary permutation matrices. A first implication of Equation (2.3) is that $\boldsymbol{C}_{x,x} = \boldsymbol{Q}_x \boldsymbol{D}_{x,x} \boldsymbol{Q}_x^{-1}$, so that $\boldsymbol{Q}_x$ is a matrix of eigenvectors. Furthermore, we also have that, for each x and all x_1,

$$\boldsymbol{C}_{x_1,x} \boldsymbol{C}_{x,x_1} = (\boldsymbol{Q}_x \boldsymbol{D}_{x_1,x} \boldsymbol{Q}_{x_1}^{-1}) (\boldsymbol{Q}_{x_1} \boldsymbol{D}_{x,x_1} \boldsymbol{Q}_x^{-1}) = \boldsymbol{Q}_x (\boldsymbol{D}_{x,x_1} \boldsymbol{D}_{x_1,x}) \boldsymbol{Q}_x^{-1}$$

and, more generally, that for each x and for any sequence of m values $x_1, \ldots, x_m$,

$$\boldsymbol{C}_{x_m,x} \boldsymbol{C}_{x_{m-1},x_m} \cdots \boldsymbol{C}_{x_1,x_2} \boldsymbol{C}_{x,x_1} = \boldsymbol{Q}_x (\boldsymbol{D}_{x,x_1} \boldsymbol{D}_{x_1,x_2} \cdots \boldsymbol{D}_{x_{m-1},x_m} \boldsymbol{D}_{x_m,x}) \boldsymbol{Q}_x^{-1}.$$

That is, $\boldsymbol{Q}_x$ is a joint diagonalizer of a set of matrices. Notice that the z-th diagonal entry of $\boldsymbol{D}_{x,x_1} \boldsymbol{D}_{x_1,x_2} \cdots \boldsymbol{D}_{x_{m-1},x_m} \boldsymbol{D}_{x_m,x}$ is $k_z^m(x; x_1, \ldots, x_m; x)$. Moreover, the eigenvalues of the set of matrices $\boldsymbol{C}_{x_m,x} \boldsymbol{C}_{x_{m-1},x_m} \cdots \boldsymbol{C}_{x_1,x_2} \boldsymbol{C}_{x,x_1}$ (as a function of $x_1, \ldots, x_m$) are the rows of the matrix $\boldsymbol{F}_{x,m}$. Further, because the joint diagonalizer is independent of m, the

same $\boldsymbol{Q}_x$ equally diagonalizes the matrices $\boldsymbol{C}_{x'_{m'},x}\boldsymbol{C}_{x'_{m'-1},x'_{m'}}\cdots\boldsymbol{C}_{x'_1,x'_2}\boldsymbol{C}_{x,x'_1}$ (as a function of $x'_1,\ldots,x'_{m'}$) for any different walk length m'. Take the set of all walk lengths from zero to $\bar{m}$. This delivers a joint diagonalization problem whose eigenvalues are the rows of the matrix $\boldsymbol{F}_x$. By Assumption 2 there exists an $\bar{m}$ for which the columns of $\boldsymbol{F}_x$ are all distinct. It then follows from Theorem 6.1 of De Lathauwer, De Moor and Vandewalle (2004) that the matrix $\boldsymbol{Q}_x$ is unique up to the scale and ordering of its columns. That is, a joint diagonalization problem identifies the matrix $\tilde{\boldsymbol{Q}}_x := \boldsymbol{Q}_x\boldsymbol{\Omega}_x\boldsymbol{\Delta}_x$, where $\boldsymbol{\Omega}_x$ is a diagonal scaling matrix and $\boldsymbol{\Delta}_x$ is a permutation matrix.

The diagonal matrix $\boldsymbol{\Omega}_x$ can be recovered, up to permutation of the entries on its diagonal, from the observation that

$$\boldsymbol{u}_x := \boldsymbol{B}_x\,\boldsymbol{p}_x = \boldsymbol{B}_x\boldsymbol{L}_x\,\boldsymbol{\iota}_q = \boldsymbol{Q}_x^{-\top}\boldsymbol{\iota}_q,$$

where the first transition uses the model structure, the second follows from the definition of $\boldsymbol{Q}_x$, and $\boldsymbol{\iota}_q$ denotes the q-vector of ones. Moreover,

$$\tilde{\boldsymbol{Q}}_x^{\top}\boldsymbol{u}_x = \boldsymbol{\Delta}_x^{-1}\boldsymbol{\Omega}_x\boldsymbol{\iota}_q = \boldsymbol{\Delta}_x^{-1}\boldsymbol{\Omega}_x\boldsymbol{\Delta}_x\boldsymbol{\iota}_q,$$

using that each row of any permutation matrix sums to unity. It is easy to see that the matrix $\boldsymbol{\Delta}_x^{-1}\boldsymbol{\Omega}_x\boldsymbol{\Delta}_x$ on the right-hand side is diagonal; a proof is provided in Lemma A.1 in the Appendix. We, therefore, indeed recover

$$\tilde{\boldsymbol{\Omega}}_x := \boldsymbol{\Delta}_x^{-1}\boldsymbol{\Omega}_x\boldsymbol{\Delta}_x$$

for all x.

3) Identification of the joint eigenvectors up to a common permutation. Moving on, take a particular x and let the value x_0 and the sequence $x_1,\ldots,x_m$ be such that Condition (i) in Assumption 3 is satisfied for this x; working with Condition (ii) instead of Condition (i) is analogous and is, therefore, not dealt with further. Using the same argument as before it is easy to see that a second implication of Equation (2.3) is that

$$\tilde{\boldsymbol{D}}_{x_0,x} := \tilde{\boldsymbol{Q}}_x^{-1}(\boldsymbol{C}_{x_m,x}\boldsymbol{C}_{x_{m-1},x_m}\cdots\boldsymbol{C}_{x_0,x_1})\,\tilde{\boldsymbol{Q}}_{x_0} = \boldsymbol{\Delta}_x^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0},$$

where the matrices $\dot{\boldsymbol{D}}_{x_0,x} \coloneqq \boldsymbol{\Omega}_{x_0}(\boldsymbol{D}_{x_0,x_1}\boldsymbol{D}_{x_1,x_2}\cdots\boldsymbol{D}_{x_m,x})\,\boldsymbol{\Omega}_x^{-1}$ are diagonal. We can write

$$\tilde{\boldsymbol{D}}_{x_0,x} = \boldsymbol{\Delta}_x^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0} = \boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}_{x_0}(\boldsymbol{\Delta}_{x_0}^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0}). \tag{2.4}$$

Note that the matrix $\boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}_{x_0}$ is a product of permutation matrices and, hence, is itself a permutation matrix. Therefore, $\tilde{\boldsymbol{D}}_{x_0,x}$ is equal to $\boldsymbol{\Delta}_{x_0}^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0}$ up to ordering of the rows. The latter matrix is diagonal. The diagonal entries of $\boldsymbol{\Delta}_{x_0}^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0}$ are thus identified by the column sums of the matrix $\tilde{\boldsymbol{D}}_{x_0,x}$. From Equation (2.4), coupled with Assumption 3, it follows that

$$\boldsymbol{H}_{x,x_0} \coloneqq \boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}_{x_0} = \tilde{\boldsymbol{D}}_{x_0,x}(\boldsymbol{\Delta}_{x_0}^{-1}\dot{\boldsymbol{D}}_{x_0,x}\boldsymbol{\Delta}_{x_0})^{-1}$$

is identified for all x. With these matrices in hand we may re-arrange the diagonal entries of the scaling matrices in a common order, as

$$\boldsymbol{H}_{x,x_0}^{-1}\tilde{\boldsymbol{\Omega}}_x\,\boldsymbol{H}_{x,x_0} = \boldsymbol{\Delta}_{x_0}^{-1}\boldsymbol{\Omega}_x\boldsymbol{\Delta}_{x_0} =: \bar{\boldsymbol{\Omega}}_x,$$

and subsequently recover

$$\bar{\boldsymbol{Q}}_x \coloneqq \tilde{\boldsymbol{Q}}_x\boldsymbol{H}_{x,x_0}\bar{\boldsymbol{\Omega}}_x^{-1} = \boldsymbol{Q}_x\boldsymbol{\Delta}_{x_0}.$$

Now, because Assumption 3 is satisfied for each $x \neq x_0$ we can repeat the same argument for each $x \neq x_0$. Thus, we have identified the matrices of joint eigenvectors $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ up to a common permutation of their columns. We can now recover the parameters of our mixture model, up to the permutation $\boldsymbol{\Delta}_{x_0}$.

4) Identification of the model parameters. First, from Equation (2.3) we recover the transition kernels as

$$\bar{\boldsymbol{Q}}_{x'}^{-1}\boldsymbol{C}_{x,x'}\bar{\boldsymbol{Q}}_x = \boldsymbol{\Delta}_{x_0}^{-1}\boldsymbol{D}_{x,x'}\boldsymbol{\Delta}_{x_0} =: \bar{\boldsymbol{D}}_{x,x'}.$$

Because the diagonal of $\boldsymbol{D}_{x,x'}$ constitutes the x'-th row of matrix $\boldsymbol{K}_x$, knowledge of $\bar{\boldsymbol{D}}_{x,x'}$ for all (x, x') allows us to construct the matrix $\bar{\boldsymbol{K}}_x \coloneqq \boldsymbol{K}_x\boldsymbol{\Delta}_{x_0}$ for each x. Next, the Markov structure of the model implies that

$$\boldsymbol{p}_x^* = \boldsymbol{K}_x\boldsymbol{\lambda}_x,$$

where $\boldsymbol{\lambda}_x \coloneqq (s_1(x)\,\mu_1, \ldots, s_q(x)\,\mu_q)^\top$. By consequence of Assumption 1, each $\boldsymbol{K}_x$ has maximal column rank. Letting a $+$ superscript on a matrix indicate its left inverse, we

may thus calculate

$$\bar{\boldsymbol{K}}_x^+ \boldsymbol{p}_x^* = \boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{K}_x^+ \boldsymbol{p}_x^* = \boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\lambda}_x =: \bar{\boldsymbol{\lambda}}_x,$$

for each x. Collecting $\bar{\boldsymbol{\Lambda}} := (\bar{\boldsymbol{\lambda}}_1, \ldots, \bar{\boldsymbol{\lambda}}_r)^\top$ yields the joint distribution of types and initial conditions. Indeed, with $\boldsymbol{\Lambda} := (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)^\top$, we see that $\bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{\Delta}_{x_0}$. Finally, letting $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_q)^\top$ and $\boldsymbol{S} := (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_q)$, where $\boldsymbol{s}_z := (s_z(1), \ldots, s_z(r))^\top$ is the distribution of the initial condition for type z, we first recover the distribution of latent types by averaging-out the initial condition, as in

$$\bar{\boldsymbol{\Lambda}}^\top \boldsymbol{\iota}_r = \boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\Lambda}^\top \boldsymbol{\iota}_r = \boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\mu} =: \bar{\boldsymbol{\mu}},$$

and then recover the type-specific distributions of the initial condition as

$$\bar{\boldsymbol{\Lambda}} \operatorname{diag}(\bar{\boldsymbol{\mu}})^{-1} = \boldsymbol{\Lambda}\boldsymbol{\Delta}_{x_0} \boldsymbol{\Delta}_{x_0}^{-1} \operatorname{diag}(\boldsymbol{\mu})^{-1} \boldsymbol{\Delta}_{x_0} = \boldsymbol{\Lambda} \operatorname{diag}(\boldsymbol{\mu})^{-1} \boldsymbol{\Delta}_{x_0} = \boldsymbol{S}\boldsymbol{\Delta}_{x_0} =: \bar{\boldsymbol{S}},$$

which uses the fact that $\boldsymbol{\Lambda} = \boldsymbol{S} \operatorname{diag}(\boldsymbol{\mu})$ by definition and $\operatorname{diag}(\bar{\boldsymbol{\mu}}) = \boldsymbol{\Delta}_{x_0}^{-1} \operatorname{diag}(\boldsymbol{\mu}) \boldsymbol{\Delta}_{x_0}$.

We have shown the following result.

Theorem 1. *Let Assumptions 1–3 hold. Then the distributions of the initial condition, s_z, the transition kernels, k_z, and the type probabilities, μ_z, may all be identified, up to a common permutation of the latent types, from the distribution of four consecutive observations.*

Remark. Having access to longer time series allows to weaken Assumption 1. Say we have access to the joint distribution of $X_1, \ldots, X_T$. Let $\lfloor \cdot \rfloor$ denote the floor function. Redefine the matrix $\boldsymbol{K}_x$ to let its z-th column be the (vectorized) distribution of $X_{\lfloor T/2 \rfloor + 1}, \ldots, X_{T-1}$ given $X_{\lfloor T/2 \rfloor} = x$ and $Z = z$. A typical entry of this matrix in column z has the multiplicative structure

$$k_z(x, x_1) \bigtimes_{i=1}^{\lfloor (T-1)/2 \rfloor - 1} k_z(x_i, x_{i+1})$$

and depends only on the number of time periods involved. Conformably redefine the matrix $\boldsymbol{L}_x$ so that its z-th column reflects the joint distribution of $X_1, \ldots, X_{\lfloor T/2 \rfloor}$ and Z at $X_{\lfloor T/2 \rfloor} = x$ and $Z = z$. Then we can mimic the proof of Theorem 1 with $\boldsymbol{P}_x$ being the

joint distribution of $X_1, \ldots, X_{T-1}$ at $X_{\lfloor T/2 \rfloor} = x$ and $\boldsymbol{P}_{x,x'}$ being the joint distribution of $X_1, \ldots, X_T$ at $X_{\lfloor T/2 \rfloor} = x$ and $X_{\lfloor T/2 \rfloor + 1} = x'$, both arranged as two-way tables. Indeed, we again have that

$$\boldsymbol{P}_x = \boldsymbol{K}_x \boldsymbol{L}_x^\top, \qquad \boldsymbol{P}_{x,x'} = \boldsymbol{K}_{x'} \boldsymbol{D}_{x,x'} \boldsymbol{L}_x^\top.$$

Assumption 1 now involves matrices that are of dimension $r^{(T-2)/2} \times r^{(T-2)/2}$ when T is even and of dimension $r^{(T-1)/2} \times r^{(T-3)/2}$ when T is odd. Assumptions 2 and 3 require no modification.

The above discussion yields the following corollary

Corollary 1. *The maximum number of types that can be accommodated in our framework equals $r^{(T-2)/2}$ when T is even and $r^{(T-3)/2}$ when T is odd.*

It is also useful to note that under Assumption 1 the number of types is identified as the rank of $\boldsymbol{P}_x$.

3 Comparison to prior work

Kasahara and Shimotsu (2009). Identification of mixtures of dynamic discrete choices has previously been considered by Kasahara and Shimotsu (2009). Their Proposition 6 provides an identification result for the matrix of transition probabilities $\boldsymbol{K}_x$ and the vector of joint probabilities $\boldsymbol{\lambda}_x$ for a *fixed* value x from the joint distribution of six outcomes. The conditions under which this result is obtained are (in our notation) that (i) the vector $\boldsymbol{\lambda}_x$ only has positive entries; (ii) there exists a collection of points $x_1, \ldots, x_{q-1}$ such that the $q \times q$ matrix $\boldsymbol{M}_x$ with

$$(\boldsymbol{M}_x)_{z,i} := \begin{cases} 1 & \text{if } i = 1 \\ k_z(x, x_{i-1})\, k_z(x_{i-1}, x) & \text{if } i > 1 \end{cases}$$

is invertible; and (iii) for some x', $k_z(x, x') > 0$ for all z and $k_z(x, x') \neq k_{z'}(x, x')$ for all $z' \neq z$.

The approach of Kasahara and Shimotsu (2009) is built around the observation that the joint distribution of X_2, X_4, X_6, conditional on the fact that X_1, X_3, X_5 all take on the

value x, factors as a static tri-variate mixture. This argument works around the Markovian dependence, whereas ours exploits it. It also makes clear why they require six time-series observations as opposed to our four.

The difference between the approach of Kasahara and Shimotsu (2009) and ours makes a precise comparison between the requirements underlying them difficult. Still, in the argument of Kasahara and Shimotsu (2009), Conditions (i) and (ii) play a similar role to does our Assumption 1, although we do not require Condition (i) and our techniques avoid the need to work with only a subset of the support points to ensure that the resulting matrix is square. Condition (iii), in turn, is used by Kasahara and Shimotsu (2009) to ensure uniqueness of an eigendecomposition. As such it fulfills the role of our Assumption 2 in their context. Condition (iii) is too strong for that purpose, however. Indeed, a look at their proof shows that their result continues to go through under the weaker requirement that the columns of $\boldsymbol{K}_x$ are all distinct. This follows from an application of Theorem 6.1 of De Lathauwer, De Moor and Vandewalle (2004) to their set of multilinear restrictions. Kasahara and Shimotsu (2009) have no analog of our Assumption 3 as their argument is for a given value x.

If Conditions (i)–(iii) hold for all x Proposition 6 of Kasahara and Shimotsu (2009) can be applied to each of them (see, e.g., the discussion in Kasahara and Shimotsu (2009, Remark 5(iii))). Identification here is up to an arbitrary ordering of the latent types, however, and separate application of their Proposition 6 does not ensure that the same ordering of latent types is recovered in all of the cases. Hence, this argument only identifies $\boldsymbol{K}_1\boldsymbol{\Delta}_1, \ldots, \boldsymbol{K}_r\boldsymbol{\Delta}_r$ and $\boldsymbol{\Delta}_1^{-1}\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\Delta}_r^{-1}\boldsymbol{\lambda}_r$, where $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_r$ are arbitrary permutation matrices. This does not suffice to reconstruct the transition kernels, nor does it lead to identification of the distributions of the initial condition or the distribution of the latent types.

Hu and Shum (2012). In related work, Hu and Shum (2012) entertain a framework where, in addition to the observable variables, the latent types themselves, too, may follow a first-order Markov process. This nests our specification. On the other hand, their approach

requires that $r = q$, i.e., that the observable variables cannot take on more values than there are latent types. This is a substantial restriction that, together with Assumption 1, implies that the matrices $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_r$ are invertible, which is crucial in the argument of Hu and Shum (2012). Their technique could be applied after binning the r support points into q bins, but this would not yield identification of the model on the original data. The requirement that $r = q$ is not imposed here. Moreover, our derivations highlight the identifying power of having $r > q$.

Hu and Shum (2012) recover the unknown probabilities from the distribution of only four outcomes, as do we. To do so they impose, along with Assumption 1, the requirement that, for each x, there exists an x' and a pair $(x_1, x_2) \neq (x', x)$ such that $k_z(x', x)$, $k_z(x_1, x)$, $k_z(x', x_2)$, and $k_z(x_1, x_2)$ are all strictly positive for all z, and that, in addition, it holds that

$$\frac{k_z(x', x)\, k_z(x_1, x_2)}{k_z(x', x_2)\, k_z(x_1, x)} \neq \frac{k_{z'}(x', x)\, k_{z'}(x_1, x_2)}{k_{z'}(x', x_2)\, k_{z'}(x_1, x)}$$

for all $z \neq z'$. The first of these two conditions is used to set up a matrix-diagonalization problem. It states that, for every x, there exist two states, x' and x_1, from which x can be reached by all types, and that there exists another state, x_2, which is equally reachable from these starting points by all types. Our results here reveal that such restrictions are unnecessary to achieve identification in our setup. The second condition further requires the transition probabilities along these states to be sufficiently different for different latent types. This condition is used by Hu and Shum (2012) to ensure uniqueness (up to scale and permutation) of the eigenvectors in their diagonalization problem. As such, it plays a similar role as our Assumption 2. However, our Assumption 2 is arguably weaker in that it only requires there to exist (collections of) walks along the type-specific Markov chains that occur with different probability for the different types. Moreover, we use different multilinear restrictions and exploit many of them through a joint diagonalization system, which demands weaker restrictions on the associated eigenvalues.

Under these conditions, Hu and Shum (2012, Lemma 3 and Corollary 2) establish an analog of Kasahara and Shimotsu (2009, Proposition 6), recovering $\boldsymbol{K}_1\boldsymbol{\Delta}_1, \ldots, \boldsymbol{K}_r\boldsymbol{\Delta}_r$ and $\boldsymbol{\Delta}_1^{-1}\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\Delta}_r^{-1}\boldsymbol{\lambda}_r$ for unknown permutation matrices $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_r$. To be able to proceed

further, they additionally assume that, for each x, there exists a functional, say the mean,

$$\delta_{x,z} \coloneqq \sum_{x'=1}^{r} k_z(x,x')\, x',$$

for which it is known that $\delta_{x,1} < \cdots < \delta_{x,q}$. This assigns empirical content to the types. Moreover, it allows to recover the matrices $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_r$ with their columns arranged in a common order, thereby resolving the remaining ambiguity and establishing identification. Our Theorem 1 shows that, under Assumption 3, a common (albeit arbitrary) ordering is identified from the data. Therefore, monotonicity restrictions that link types to outcomes can be dispensed with.

4 Illustrations

Dynamic binary choice. The most basic situation that fits our setup has both binary observables and binary type heterogeneity. Given adding-up constraints, the model depends on the success probabilities $\omega_z \coloneqq k_z(1,2)$, $\varpi_z \coloneqq k_z(2,2)$, and $\pi_z = s_z(1)$ for $z \in \{1,2\}$, and the single mixing proportion μ_1.

With four times periods we have

$$\boldsymbol{P}_1 = \begin{pmatrix} (1-\omega_1) & (1-\omega_2) \\ \omega_1 & \omega_2 \end{pmatrix} \begin{pmatrix} \mu_1 & 0 \\ 0 & 1-\mu_1 \end{pmatrix} \begin{pmatrix} \pi_1\,(1-\omega_1) & \pi_2\,(1-\omega_2) \\ (1-\pi_1)\,(1-\varpi_1) & (1-\pi_2)\,(1-\varpi_2) \end{pmatrix}^{\top}$$

and

$$\boldsymbol{P}_2 = \begin{pmatrix} (1-\varpi_1) & (1-\varpi_2) \\ \varpi_1 & \varpi_2 \end{pmatrix} \begin{pmatrix} \mu_1 & 0 \\ 0 & 1-\mu_1 \end{pmatrix} \begin{pmatrix} \pi_1\,\omega_1 & \pi_2\,\omega_2 \\ (1-\pi_1)\,\varpi_1 & (1-\pi_2)\,\varpi_2 \end{pmatrix}^{\top}.$$

Assumption 1 requires these matrices to have full rank. Equivalently, all matrices on the right-hand side need to be invertible. Given that $0 < \mu_1 < 1$, $\boldsymbol{P}_1$ is invertible if and only if

$$\omega_1 \neq \omega_2, \qquad \pi_1\,(1-\omega_1)\,(1-\pi_2)\,(1-\varpi_2) \neq \pi_2\,(1-\omega_2)\,(1-\pi_1)\,(1-\varpi_1) \tag{4.5}$$

Similarly, $\boldsymbol{P}_2$ is invertible if and only if

$$\varpi_1 \neq \varpi_2, \qquad \pi_1\,\omega_1\,(1-\pi_2)\,\varpi_2 \neq \pi_2\,\omega_2\,(1-\pi_1)\,\varpi_1 \tag{4.6}$$

The two conditions $\omega_1 \neq \omega_2$ and $\varpi_1 \neq \varpi_2$ simply state that the transition probabilities of the Markov chain differ across types. The other two conditions further involve the initial conditions π_1 and π_2. They state that the ratio of the type-specific probabilities of the event $(X_1 = 1, X_2 = x)$ to the event $(X_1 = 2, X_2 = x)$ must be different across the two types, for $x \in \{1, 2\}$. They can equally be interpreted as restrictions on the initial conditions. Rearrangement yields the equivalent statements that

$$\frac{1-\pi_1/1-\pi_2}{\pi_1/\pi_2} \neq \frac{1-\omega_1/1-\omega_2}{1-\varpi_1/1-\varpi_2}, \qquad \frac{1-\pi_1/1-\pi_2}{\pi_1/\pi_2} \neq \frac{\varpi_2/\varpi_1}{\omega_2/\omega_1},$$

where we presume for a moment that the probabilities showing up in denominators are non-zero.

Assumption 2 is implied by Assumption 1. Indeed, setting $\bar{m} = 0$, Assumption 2 requires that direct walks from 1 to 1 and from 2 to 2 occur with different probabilities across types. This means that we require that $(1 - \omega_1) \neq (1 - \omega_2)$ and that $\varpi_1 \neq \varpi_2$. These conditions already appeared in (4.5) and (4.6), respectively, where they were shown to validate Assumption 1.

For Assumption 3 we need that either $\omega_z > 0$ or $\varpi_z < 1$ for both types. One implication is that Assumption 3 allows for a state to be absorbent for one type; if it were absorbant for both types Assumption 1 would fail.

In this example $r = q$ and so the approach of Hu and Shum (2012) could equally be followed. In addition to Assumption 1 they require, in place of Assumption 2, that either

$$\frac{\omega_z}{(1-\omega_z)}\frac{(1-\varpi_z)}{\varpi_z}, \qquad \text{or} \qquad \frac{(1-\omega_z)}{\omega_z}\frac{\varpi_z}{(1-\varpi_z)},$$

is known to be well defined for both z and is different for different z. Next, they need a monotonicity condition. Here, this condition boils down to imposing that $\omega_1 < \omega_2$ (where the direction of the inequality is without loss of generality because types are arbitrary) together with either $\varpi_1 < \varpi_2$ or $\varpi_1 > \varpi_2$. The choice between the two amounts to taking a stand on the difference in magnitude of state dependence across types. To illustrate consider a standard random-coefficient model (e.g., Browning and Carro 2007, 2014) with parametrization

$$\omega_z = F(\alpha_z), \qquad \varpi_z = F(\alpha_z + \rho_z),$$

for some specified strictly-increasing continuous distribution function F and type-specific parameters (α_z, ρ_z). Here, Assumption 1 demands that $\alpha_1 \neq \alpha_2$ and $\alpha_1 + \rho_1 \neq \alpha_2 + \rho_2$. With $\alpha_1 < \alpha_2$ the monotonicity requirement then further requires a choice to be made as to whether $\rho_1 > \rho_2 + (\alpha_2 - \alpha_1)$ or $\rho_1 < \rho_2 + (\alpha_2 - \alpha_1)$. With the transition probabilities in hand identification follows from

$$\alpha_z = F^{-1}(\omega_z), \qquad \rho_z = F^{-1}(\varpi_z) - F^{-1}(\omega_z).$$

Without monotonicity Hu and Shum's (2012) argument is not guaranteed to yield the ω_z and the ϖ_z up to a common ordering, and so the ρ_z parameters cannot be point identified.

Numerical example. Next consider a numerical example with $r = 3$ and $q = 2$, and transition kernels

k_1	1	2	3
1	1	0	0
2	$7/10$	$1/10$	$2/10$
3	$2/10$	$6/10$	$2/10$

,

k_2	1	2	3
1	$3/10$	$2/10$	$5/10$
2	0	1	0
3	$6/10$	$2/10$	$2/10$

.

Here both types have a terminal state (equal to their type number). Nonetheless, the transition probabilities are linearly independent, so $\boldsymbol{K}_x$ has maximal rank for all x. Further, a simple sufficient condition for $\boldsymbol{L}_x$ to have maximal rank here is that $s_2(1) > 0$ and that $s_1(2) > 0$. Then Assumption 1 is satisfied.

For Assumption 2, reading off the diagonal entries of the transition matrices we see that

$$\boldsymbol{F}_{1,0} = \begin{pmatrix} 1 & 3/10 \end{pmatrix}, \qquad \boldsymbol{F}_{2,0} = \begin{pmatrix} 1/10 & 1 \end{pmatrix}, \qquad \boldsymbol{F}_{3,0} = \begin{pmatrix} 2/10 & 2/10 \end{pmatrix}.$$

The last of these row vectors does not have distinct elements. Hence, although Assumption 2 holds for $x \in \{1, 2\}$ with $\bar{m} = 0$ it does not for $x = 3$. However, looking at walks over one intermediate point yields

$$\boldsymbol{F}_{1,1} = \begin{pmatrix} 1 & 9/100 \\ 0 & 0 \\ 0 & 30/100 \end{pmatrix}, \qquad \boldsymbol{F}_{2,1} = \begin{pmatrix} 0 & 0 \\ 1/100 & 1 \\ 12/100 & 0 \end{pmatrix}, \qquad \boldsymbol{F}_{3,1} = \begin{pmatrix} 0 & 30/100 \\ 12/100 & 0 \\ 4/100 & 4/100 \end{pmatrix}.$$

These matrices all have distinct columns. Hence, Assumption 2 is satisfied for all x with $\bar{m} = 1$.

Assumption 3 is satisfied because the final row of both k_1 and k_2 only has non-zero entries. Therefore, we can walk from $x_0 = 3$ to each $x \in \{1, 2, 3\} \backslash \{x_0\}$ for both types z by taking one step through the Markov chain. This is more than enough to validate Assumption 3.

Binary choice with a state variable. Our results are relevant for identification of structural dynamic discrete-choice models with unobserved heterogeneity. Suppose that we have binary variables $Y_t \in \{0, 1\}$ and $W_t \in \{0, 1\}$. Then we can define the new variable

$$X_t = \begin{cases} 1 & \text{if } (Y_t, W_t) = (0, 0) \\ 2 & \text{if } (Y_t, W_t) = (1, 0) \\ 3 & \text{if } (Y_t, W_t) = (0, 1) \\ 4 & \text{if } (Y_t, W_t) = (1, 1) \end{cases},$$

and apply our results to the distribution of X_1, X_2, X_3, X_4. Say that Y_t is the choice variable and that W_t is the state variable. A common assumption on the transition probability $\mathbb{P}(Y_t = y_t, W_t = w_t | Y_{t-1} = y_{t-1}, W_{t-1} = w_{t-1}, Z = z)$ in empirical work is that it factors as

$$\mathbb{P}(Y_t = y_t | W_t = w_t, Z = z) \times \mathbb{P}(W_t = w_t | Y_{t-1} = y_{t-1}, W_{t-1} = w_{t-1}, Z = z).$$

We can parametrise the transition kernel in terms of the conditional success probabilities

$$\varpi_z(w) \coloneqq \mathbb{P}(Y_t = 1 | W_t = w, Z = z), \quad \omega_z(y, w) \coloneqq \mathbb{P}(W_t = 1 | Y_{t-1} = y, W_{t-1} = w, Z = z).$$

Then the transition from (y, w) to (y', w') for type z takes the form

$$(\varpi_z(w')^{y'} (1 - \varpi_z(w'))^{1-y'}) \times (\omega_z(y, w)^{w'} (1 - \omega_z(y, w))^{1-w'}).$$

We discuss our assumptions in the context of this model next, maintaining binary type heterogeneity.

Assumption 1 demands the 4×2 matrices $\boldsymbol{K}_x$ and $\boldsymbol{L}_x$ to have maximal column rank for all $x \in \{1, 2, 3, 4\}$. As an example, the first of these matrices, $\boldsymbol{K}_1$, contains the type-specific

conditional transition probabilities when starting at $(y, w) = (0, 0)$. The matrix is equal to

$$\boldsymbol{K}_1 = \begin{pmatrix} (1-\omega_1(0,0))\,(1-\varpi_1(0)) & (1-\omega_2(0,0))\,(1-\varpi_2(0)) \\ (1-\omega_1(0,0))\,\varpi_1(0) & (1-\omega_2(0,0))\,\varpi_2(0) \\ \omega_1(0,0)\,(1-\varpi_1(1)) & \omega_2(0,0)\,(1-\varpi_2(1)) \\ \omega_1(0,0)\,\varpi_1(1) & \omega_2(0,0)\,\varpi_2(1) \end{pmatrix}.$$

Clearly, the presence of two binary variables makes the rank condition easier to satisfy than in the first example, where we had a single binary variable. For example, if $\omega_1(0,0) = 1$ and $\omega_2(0,0) = 0$, then both types can be perfectly separated and no restriction on $\varpi_1(1)$ and $\varpi_2(0)$ is needed to ensure that $\boldsymbol{K}_1$ has maximal column rank. If, on the other hand, $\omega_1(0,0) = 1$ and $\omega_2(0,0) = 1$ would hold, the first two rows of $\boldsymbol{K}_1$ would have only zeros and we would effectively fall back to the previous example; here we would require that $\varpi_1(1) \neq \varpi_2(1)$.

Many other cases are possible. A case of particular interest is obtained by imposing the additional model restriction (Magnac and Thesmar 2002) that $\omega_z(y, w) = \omega(y, w)$ for all z; that is, that type heterogeneity only affects the choice variable, and not the state variable. We can factor

$$\boldsymbol{K}_1 = \begin{pmatrix} (1-\omega(0,0)) & 0 & 0 & 0 \\ 0 & (1-\omega(0,0)) & 0 & 0 \\ 0 & 0 & \omega(0,0) & 0 \\ 0 & 0 & 0 & \omega(0,0) \end{pmatrix} \begin{pmatrix} (1-\varpi_1(0)) & (1-\varpi_2(0)) \\ \varpi_1(0) & \varpi_2(0) \\ (1-\varpi_1(1)) & (1-\varpi_2(1)) \\ \varpi_1(1) & \varpi_2(1) \end{pmatrix}.$$

Provided that $0 < \omega(0,0) < 1$, elementary row operations reveal that the rank condition is satisfied if $\varpi_1(0) \neq \varpi_2(0)$ or $\varpi_1(1) \neq \varpi_2(1)$. The matrices $\boldsymbol{K}_2, \boldsymbol{K}_3, \boldsymbol{K}_4$ have the same structure. As the matrix $\boldsymbol{K}_x$ depends on x only through the diagonal matrix on the right-hand side of the above equation, it suffices to replace $\omega(0,0)$ by the relevant $\omega(y, w)$. Further note that, if the state variable does not depend on latent type, its transition matrix is nonparametrically identified without any restrictions. Hence, in this case, the $\omega(y, w)$ can effectively be considered as known. The remainder of Assumption 1 involves a rank condition on $\boldsymbol{L}_x$, and, as before, this can again be interpreted as a restriction on the initial conditions.

We next verify Assumption 2 in this example. For $\bar{m} = 0$ the assumption requires that

$$\begin{aligned} &(1 - \varpi_z(0))\,(1 - \omega_z(0,0)), \qquad && \varpi_z(0)\,(1 - \omega_z(1,0)), \\ &(1 - \varpi_z(1)) \quad \omega_z(0,1), && \varpi_z(1) \quad \omega_z(1,1), \end{aligned}$$

all vary with z. The presence of the state variable again provides an additional source from which such variation may arise. Under the assumption that $\omega_z(y,w) = \omega(y,w)$ for all z, thus shutting down this additional channel, and considering the case where $0 < \omega(y,w) < 1$ the requirement is equivalent to demanding that

$$\varpi_1(0) \neq \varpi_2(0), \qquad \text{and} \qquad \varpi_1(1) \neq \varpi_2(1).$$

Assumption 1 required only one of these two inequalities to hold. However, we can consider larger values for $\bar{m}$. With $\bar{m} = 1$, for example, we equally consider probabilities involving walks with one intermediate stop. For $x = 1$, for example, we can consider the probabilities of the walks $1, x', 1$ for $x' \in \{1,2,3,4\}$. The probabilities of the first two walks are equal to

$$(1 - \varpi_z(0))^2\,(1 - \omega(0,0))^2,$$

and

$$\varpi_z(0)\,(1 - \omega(0,0))\,(1 - \varpi_z(0))\,(1 - \omega(1,0)),$$

respectively. They differ across z if and only if $\varpi_1(0) \neq \varpi_2(0)$. This permits $\varpi_1(1) = \varpi_2(1)$. Similarly, the probabilities of the walks over the two remaining stops are

$$(1 - \varpi_z(1))\,\omega(0,0)\,(1 - \varpi_z(0))\,(1 - \omega(0,1))$$

and

$$\varpi_z(1)\,\omega(0,0)\,(1 - \varpi_z(0))\,(1 - \omega(1,1)).$$

They vary with type if either $(1 - \varpi_1(1))\,(1 - \varpi_1(0)) \neq (1 - \varpi_2(1))\,(1 - \varpi_2(0))$ or if $\varpi_1(1)(1 - \varpi_1(0)) \neq \varpi_2(1)(1 - \varpi_2(0))$. These conditions, in turn, permit $\varpi_1(0) = \varpi_2(0)$. So, again, Assumption 2 is easily satisfied given Assumption 1.

Assumption 3, finally, again involves probabilities of walks and, in particular, requires that their exists an $x_0 \in \{1,2,3,4\}$ from which we can travel to $x \in \{1,2,3,4\}\backslash\{x_0\}$ for

both types z. If we maintain that the evolution of the state variable does not depend on z and the relevant probabilities are non-zero, the probability of such walks are positive if $0 < \varpi_z(0) < 1$ and $0 < \varpi_z(1) < 1$, that is, if the choice variable has full support for each value of the state variable and for each of the latent types.

An alternative to Assumption 3 would be to impose monotonicity restrictions. In the current model, such restrictions would need to be based on functionals of the variable X_t, and so would involve the joint distribution of the choice variable and the state variable. This may be challenging. Even if possible, they may imply sign and magnitude restrictions that may not be desirable to impose. In addition, the formulation of any such restrictions is not invariant to how the auxiliary variable X_t is defined. As, here, this variable is an artificial construction, there are many possible ways to do this.

5 Higher-order Markov dependence

Our argument can be extended to models with higher-order dynamics. To see how this can be done, take a model with second-order Markov dependence. The transition kernel is now

$$k_z(x, x', x'') \coloneqq \mathbb{P}(X_t = x''|X_{t-1} = x', X_{t-2} = x, Z = z).$$

For each pair (x, x'), collect the type-specific distributions in the $r \times q$ matrix $\boldsymbol{K}_{x,x'}$ and, similarly, construct the $r \times q$ matrix $\boldsymbol{L}_{x,x'}$ as

$$(\boldsymbol{L}_{x,x'})_{x'',z} \coloneqq \mathbb{P}(X_3 = x', X_2 = x, X_1 = x'', Z = z).$$

Then

$$\boldsymbol{P}_{x,x'} = \boldsymbol{K}_{x,x'}\boldsymbol{L}_{x,x'}^{\top}.$$

If we have access to the joint distribution of five observations we can define the collection of matrices

$$(\boldsymbol{P}_{x,x',x''})_{i,j} \coloneqq p_{j,x,x',x'',i}$$

in complete analogy to before. We see that

$$\boldsymbol{P}_{x,x',x''} = \boldsymbol{K}_{x',x''}\boldsymbol{D}_{x,x',x''}\boldsymbol{L}_{x,x'}^{\top}$$

where, now, $\boldsymbol{D}_{x,x',x''}$ is the $q \times q$ diagonal matrix that contains the $k_z(x, x', x'')$. The factorizations in the above equations are of the same form as those obtained in Section 2, and the arguments followed there can be modified to apply here.

The general conclusion, then, is that, under suitable modifications of Assumptions 1 to 3, identification of a mixture of Markov processes is possible from the cross-sectional distribution of as little as three effective time-series observations. If dependence is present up to order p, we need $3 + p$ observations.

Conclusion

We have derived a constructive identification result for finite mixtures of dynamic discrete choices. Our method of proof differs from Kasahara and Shimotsu (2009) and Hu and Shum (2012), who rely on arguments from the literature on static mixture models, and is able to deliver full identification without the need to impose monotonicity restrictions. The chief observation behind it is that, while the model implies a collection of multilinear restrictions akin to those used in the analysis of multivariate mixtures, these are only a small subset of the restrictions that arise from the dynamics in the model. This subset of restrictions, in isolation, does not yield identification. The full set of restrictions, however, does.

Our arguments yield identification from three effective time-series observations. Results of Hall and Zhou (2003) and Henry, Kitamura and Salanié (2014) (in a different context) suggest that (point) identification from shorter panels is unlikely to be possible, in general, without imposing additional restrictions. An example of such additional restrictions is given in Gupta, Kumar and Vassilvitskii (2016), where a specific approach to identification of first-order Markov processes from two effective time periods is considered. A necessary (but not sufficient) requirement for their approach to go through is that (in addition to Assumption 1) we have that $r \geq 2q$.

Appendix A

Lemma A.1. *Let $\boldsymbol{P}$ be a permutation matrix and let $\boldsymbol{D}$ be a diagonal matrix. Then $\boldsymbol{P}^{-1}\boldsymbol{D}\boldsymbol{P}$ is a diagonal matrix.*

Proof. We show that $\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ is diagonal. Because $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{P}^{-1} = \boldsymbol{P}^{\top}$ and so $\boldsymbol{P}^{-1}\boldsymbol{D}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, from which the result follows. Because $\boldsymbol{P}$ is a permutation matrix each of its rows and columns contains a single one; the other entries are all zero. Let $\sigma(i)$ be the mapping which yields the column that contains the one in the i-th row and let σ^{-1} be the inverse mapping. Then

$$(\boldsymbol{P}\boldsymbol{D})_{i,j} = \sum_k (\boldsymbol{P})_{i,k}(\boldsymbol{D})_{k,j} = (\boldsymbol{P})_{i,j}(\boldsymbol{D})_{j,j} = \begin{cases} (\boldsymbol{D})_{\sigma(i),\sigma(i)} & \text{if } j = \sigma(i) \\ 0 & \text{otherwise} \end{cases},$$

where the first equality follows by definition, the second from the fact that $\boldsymbol{D}$ is diagonal, and the third from the fact that $\boldsymbol{P}$ is a permutation matrix. Next, using this result yields

$$(\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1})_{i,j} = (\boldsymbol{P}\boldsymbol{D})_{i,\sigma(i)}\,(\boldsymbol{P})_{j,\sigma(i)} = \begin{cases} (\boldsymbol{D})_{\sigma(i),\sigma(i)} & \text{if } j = i \\ 0 & \text{otherwise} \end{cases},$$

so that, indeed, $\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ is a diagonal matrix. □

Appendix B

The proof of Theorem 1 is constructive. The key to constructing an estimator based on it is a routine that (approximately) solves the set of equations in (2.3) based on estimators of the matrices on the left-hand side. Such a problem is related to, but different from, joint approximate diagonalization. An algorithm for doing so is provided in a companion paper (Higgins and Jochmans 2021). Alternatively, maximum likelihood estimation is feasible in our context. As it is efficient and yields estimated distributions that are easily ensured to satisfy non-negativity and adding-up constraints it carries our preference. A natural way to proceed with implementation is via the EM algorithm (Dempster, Laird and Rubin 1977).

Likelihood. Let $\boldsymbol{X} \coloneqq (X_1, \ldots, X_T)$ be a random sequence drawn from the mixture model and let $\boldsymbol{x} \coloneqq (x_1, \ldots, x_T)$ be a particular realization of this sequence. The probability mass function of $\boldsymbol{X}$ at $\boldsymbol{x}$ takes the form

$$\sum_{z=1}^{q} \mu_z \, \ell_z(\boldsymbol{x}; \boldsymbol{\vartheta}_z),$$

where

$$\ell_z(\boldsymbol{x}; \boldsymbol{\vartheta}_z) \coloneqq \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | Z = z) = \bigtimes_{x=1}^{r} s_z(x)^{\{x_1 = x\}} \bigtimes_{x'=1}^{r} k_z(x, x')^{n_{x,x'}(\boldsymbol{x})}.$$

Here, the $r + r^2$ vector $\boldsymbol{\vartheta}_z$ collects the steady-state distribution s_z and the transition matrix k_z, we use $\{\cdot\}$ to denote the indicator function, and write $n_{x,x'}(\boldsymbol{x})$ for the number of transitions from x to x' that appear in $\boldsymbol{x}$.

The log-likelihood function for a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is

$$\sum_{i=1}^{n} \log \left(\sum_{z=1}^{q} \mu_z \, \ell_z(\boldsymbol{X}_i; \boldsymbol{\vartheta}_z) \right).$$

Let $Z_1, \ldots, Z_n$ denote the (latent) types. The complete-data log-likelihood function equals

$$L_n(\boldsymbol{\Theta}) \coloneqq \sum_{i=1}^{n} \sum_{z=1}^{q} \{Z_i = z\} \left(\log \mu_z + \log \ell_z(\boldsymbol{X}_i; \boldsymbol{\vartheta}_z) \right),$$

where $\boldsymbol{\Theta}$ collects all $\mu_1, \ldots, \mu_q$ and $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_q$. The EM algorithm iterates on $L_n(\boldsymbol{\Theta})$ and, in our case, is guaranteed to deliver a local maximizer of the log-likelihood (Wu 1983). We defer to McLachlan and Krishnan (2008) for additional discussion and references on the EM algorithm in a mixture context.

EM iteration. An iteration starting at $\hat{\boldsymbol{\Theta}}$ proceeds as follows. In the E-step we compute the expectation of $L_n(\boldsymbol{\Theta})$ given the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ under the distribution induced by $\hat{\boldsymbol{\Theta}}$. This yields the criterion

$$\mathbb{E}_{\hat{\boldsymbol{\Theta}}}(L_n(\boldsymbol{\Theta}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \sum_{i=1}^{n} \sum_{z=1}^{q} \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}}) \left(\log \mu_z + \log \ell_z(\boldsymbol{X}_i; \boldsymbol{\vartheta}_z) \right),$$

where

$$\omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}}) \coloneqq \frac{\hat{\mu}_z \, \ell_z(\boldsymbol{X}_i; \hat{\boldsymbol{\vartheta}}_z)}{\sum_{z'} \hat{\mu}_{z'} \, \ell_{z'}(\boldsymbol{X}_i; \hat{\boldsymbol{\vartheta}}_{z'})}$$

is the posterior probability that $Z_i = z$. In the M-step we maximize the criterion with respect to $\boldsymbol{\Theta}$ to get $\hat{\hat{\boldsymbol{\Theta}}}$, say. Inspection of $\mathbb{E}_{\hat{\boldsymbol{\Theta}}}(L_n(\boldsymbol{\Theta})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ reveals that $\hat{\hat{\boldsymbol{\Theta}}}$ can be written in closed form. With the solution forced to consist of valid probability distributions we find

$$\hat{\hat{\mu}}_z = \frac{\sum_{i=1}^n \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}})}{n}$$

and

$$\hat{\hat{s}}_z(x) = \frac{\sum_{i=1}^n \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}})\{X_{i,1} = x\}}{\sum_{i=1}^n \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}})}, \quad \hat{\hat{k}}_z(x, x') = \frac{\sum_{i=1}^n \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}})\, n_{x,x'}(\boldsymbol{X}_i)}{\sum_{x''=1}^r \sum_{i=1}^n \omega_z(\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}})\, n_{x,x''}(\boldsymbol{X}_i)}.$$

We subsequently replace $\hat{\boldsymbol{\Theta}}$ by $\hat{\hat{\boldsymbol{\Theta}}}$ and start a new iteration. The procedure is repeated until convergence.

Simulations. We provide results from a Monte Carlo experiment on a two-component mixture of binary decisions. The type-specific transition kernels, k_1 and k_2, are specified as

k_1	1	2
1	$^2/_{10}$	$^8/_{10}$
2	$^7/_{10}$	$^3/_{10}$

k_2	1	2
1	$^8/_{10}$	$^2/_{10}$
2	$^3/_{10}$	$^7/_{10}$

and we mix the two types with $\mu_1 = {}^4/_{10}$ and $\mu_2 = 1 - \mu_1 = {}^6/_{10}$. The type-specific Markov chains are initialized with a draw from their steady-state distributions. In each of 10,000 Monte Carlo replications we estimate the model by maximum likelihood, using the EM algorithm (initiated at a range of different starting values with a terminal condition on the improvement of the likelihood), and estimate the information as the outer-product of the score vector, evaluated at the maximizer.

Tables B.1 and B.2 provide the mean, median, standard deviation, average standard error, and interquartile range of the point estimator (over the Monte Carlo replications) together with the empirical size of a two-sided t-test with a theoretical size of 5%. Table B.1 concerns the case where we observe four outcomes for each of 500 observations (so the minimum of the three transition needed for our results to apply). Table B.2 reports results

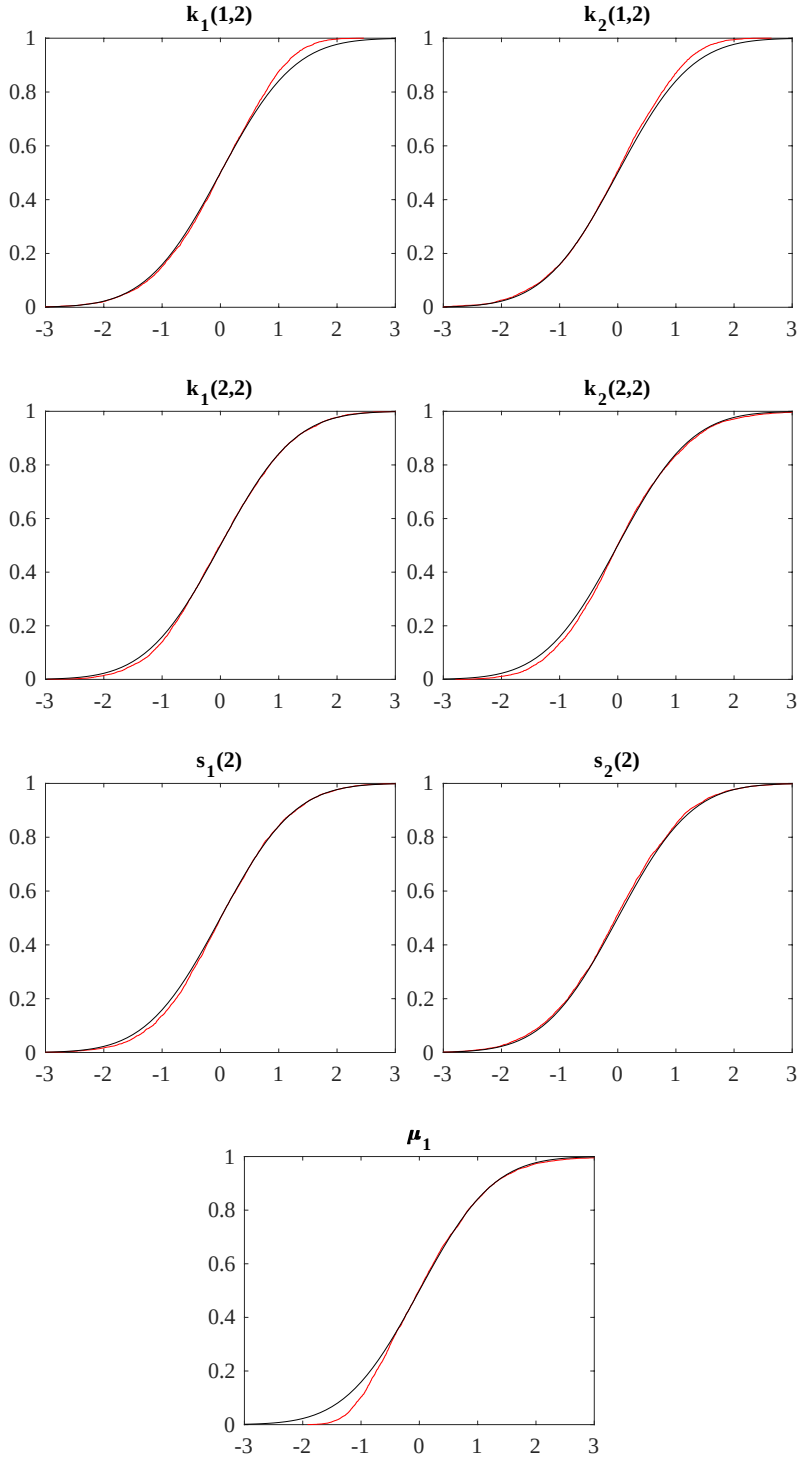
Table B.1: Descriptive statistics of simulation results for $T = 4$

	$k_1(1,2)$	$k_1(2,2)$	$s_1(2)$	$k_2(0,1)$	$k_2(1,1)$	$s_2(2)$	μ_1
value	0.800	0.300	0.533	0.200	0.700	0.400	0.400
mean	0.794	0.305	0.533	0.194	0.705	0.398	0.414
median	0.797	0.303	0.533	0.197	0.703	0.399	0.402
std dev	0.058	0.048	0.046	0.035	0.042	0.036	0.068
std error	0.067	0.051	0.046	0.039	0.045	0.036	0.079
iqr	0.078	0.065	0.061	0.046	0.054	0.048	0.096
size	0.037	0.049	0.047	0.025	0.028	0.042	0.025

Table B.2: Descriptive statistics of simulation results for $T = 5$

	$k_1(1,2)$	$k_1(2,2)$	$s_1(2)$	$k_2(0,1)$	$k_2(1,1)$	$s_2(2)$	μ_1
value	0.800	0.300	0.533	0.200	0.700	0.400	0.400
mean	0.800	0.301	0.534	0.199	0.701	0.399	0.404
median	0.801	0.301	0.534	0.199	0.701	0.400	0.401
std dev	0.040	0.035	0.042	0.024	0.029	0.033	0.046
std error	0.041	0.036	0.042	0.024	0.030	0.033	0.047
iqr	0.055	0.049	0.056	0.032	0.039	0.044	0.064
size	0.053	0.050	0.052	0.043	0.047	0.050	0.040

Figure B.1: Studentized empirical distributions (red) together with the standard-normal reference distribution (black) for $T = 4$



for four transitions. The results show good performance of the estimator with bias being negligible relative to the standard deviation. Inference is slightly conservative in Table B.1 due to the standard error being slightly upward biased for some of the parameters. In Table B.2 this underrejection is virtually eliminated. The plots in Figure B.1 contain the empirical cumulative distribution functions (in red) of the Studentized point estimators for the different parameters associated with the simulations for the four-wave data. Each plot also provides the standard-normal distribution as a benchmark (in black). Overall, the normal approximation performs well. Some deviations can be observed in the upper (lower) tail of the transition probabilities for type 1 (type 2) and in the lower tail of the distribution of the proportion of type 1 individuals.

References

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

Arcidiacono, P. and R. A. Miller (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica 79*, 1823–1867.

Baum-Snow, N. and R. Pavan (2012). Understanding the city size wage gap. *Review of Economic Studies 79*, 88–127.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016). Estimating multivariate latent-structure models. *Annals of Statistics 44*, 540–563.

Browning, M. and J. M. Carro (2007). Heterogeneity and microeconometrics modeling. In R. W. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances In Economics and Econometrics*, Volume III, Chapter 3, pp. 47–74. Cambridge University Press.

Browning, M. and J. M. Carro (2014). Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics 178*, 805–823.

Chen, L.-Y. (2017). Identification of discrete choice dynamic programming models with nonparametric distribution of unobservables. *Econometric Theory 33*, 551–577.

Crawford, G. S. and M. Schum (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica 73*, 1137–1173.

De Lathauwer, L., B. De 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.

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

Eckstein, Z. and K. Wolpin (1999). Why youths drop out of high school: The impact of preferences, opportunities, and abilities. *Econometrica 67*, 1295–1340.

Gupta, R., R. Kumar, and S. Vassilvitskii (2016). On mixtures of Markov chains. Thirtieth Conference on Neural Information Processing Systems, Barcelona.

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

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

Heckman, J. J. (1981). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, pp. 91–139.

Henry, M., Y. Kitamura, and B. Salanié (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics 5*, 123–144.

Higgins, A. and K. Jochmans (2021). Joint approximate asymmetric diagonalization by non-orthogonal matrices. Mimeo.

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

Hu, Y. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics 171*, 32–44.

Kalouptsidi, M., Y. Kitamura, L. Lima, and E. Souza-Rodrigues (2021). Counterfactual analysis for structural dynamic discrete choice models. Mimeo.

Kalouptsidi, M., P. T. Scott, and E. Souza-Rodrigues (2020). Linear IV regression estimators for structural dynamic discrete choice models. Forthcoming in *Journal of Econometrics*.

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

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*, 473–522.

Magnac, T. and D. Thesmar (2002). Identifying dynamic discrete decision processes. *Econometrica 70*, 801–816.

McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions*. Wiley.

Nevo, A., J. L. Turner, and J. W. Williams (2016). Usage-based pricing and demand for residential broadband. *Econometrica 84*, 411–443.

Norets, A. and X. Tang (2014). Semiparametric inference in dynamic binary choice models. *Review of Economic Studies 81*, 1229–1262.

Rossi, F. (2017). Lower price or higher reward? Measuring the effect of consumers' preferences on reward programs. *Management Science 64*, 4451–4470.

Wang, Y. (2014). Dynamic implications of subjective expectations: Evidence from adult smokers. *American Economic Journal: Applied Economics 6*, 1–37.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*, 95–103.