IDENTIFICATION OF MIXTURES OF DYNAMIC DISCRETE CHOICES

Ayden Higgins*
Faculty of Economics
University of Cambridge

Koen Jochmans†
Toulouse School of Economics
University of Toulouse Capitole

November 9, 2021

Abstract

This paper provides new identification results for finite mixtures of Markov processes. Our arguments are constructive and show that identification can be achieved from knowledge of the cross-sectional distribution of three (or more) effective time-series observations under simple conditions. Our approach is contrasted with the ones taken in prior work by Kasahara and Shimotsu (2009) and Hu and Shum (2012). Most notably, monotonicity restrictions that link conditional distributions to latent types are not needed. Maximum likelihood is considered for the purpose of estimation and inference. Implementation via the EM algorithm is straightforward. Its performance is evaluated in a simulation exercise.

JEL Classification: C14, C23 C51

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

*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: amh239@cam.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 n° 715787 (MiMo), and from the French Government and the ANR under the Investissements d' Avenir program, grant ANR-17-EURE-0010 is gratefully acknowledged.

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 a finite mixture. Keane and Wolpin (1997), Crawford and Schum (2005), Aguirregabiria and Mira (2007), and Arcidiacono and Miller (2011) are examples of papers that have taken this route in different settings.

Kasahara and Shimotsu (2009) and Hu and Shum (2012) have studied (nonparametric) identification of mixtures of first-order Markov processes. Their arguments closely follow work on (static) multivariate models with latent variables (in particular Anderson 1954 and Hu 2008). Moreover, they both exploit (different) 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 fully recover the type-specific distributions or the mixing distribution. This under-identification 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). To achieve identification, Hu and Shum (2012) supplement these restrictions with outside information in the form of a set of monotonicity restrictions that link latent types to observable choices.

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 two simple conditions that 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. However, we show that these are 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.

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 explore maximum likelihood for estimation and inference in Section 4. As usual with mixture models, the EM algorithm is attractive for implementation. Furthermore, in the current setting, both the E-step as the M-step are available in closed form. Our arguments extend naturally to models with higher-order Markovian dependence, and we show how in Section 5. The general conclusion, then, is that, under suitable conditions, mixtures of p-th order Markov processes are identified from the cross-sectional distribution of $3 + p$ time-series observations.

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.

We let $\{X_t\}$ be a sequence of observable random variables that can take on r values. For notational convenience we presume that its support constitutes the set of integers up to r; translation 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

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

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 nonparametrically 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.

2 Identification

Assumptions. A constructive identification approach can be devised under two simple conditions. Before these are stated we introduce notation for probabilities that involve only observable variables. We define the notational shorthands $p_{x_1,x_2} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2)$ and $p_{x_1,x_2,x_3} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ for bivariate and trivariate probabilities, and $p_{x_1,x_2,x_3,x_4} \coloneqq \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ for four-variate probabilities. We collect the bivariate probabilities in the two sets of r-vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r$ and $\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_r^*$, where

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

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 Assumption 1 implies 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. The need for this is intuitive. On the other hand, the rank condition allows for the presence of, for example, absorbant states, i.e., that, for some x and z, $k_z(x, x) = 1$, and so $k_z(x, x') = 0$ for all $x' \neq x$. It also does not require that state dependence necessarily be present in order to go through. Indeed, we may have that $k_z(x', x) = k_z(x'', x)$ for all pairs (x', x'') and for all x.

To state our second assumption we let

$$f_{z,x,m}(x_1, \ldots, x_m) \coloneqq k_z(x, x_1)\, k_z(x_1, x_2) \cdots k_z(x_{m-1}, x_m) k_z(x_m, x),$$

which is the probability of arriving in State x via the walk $x_1, \ldots, x_m$ when starting at State x itself, conditional on $Z = z$. Stack the probabilities over the different possible walks $x_1, \ldots, x_m$ in the vector $\boldsymbol{f}_{z,x,m}$ and collect these distributions over latent types in the matrix $\boldsymbol{F}_{x,m} \coloneqq (\boldsymbol{f}_{1,x,m}, \ldots, \boldsymbol{f}_{q,x,m})$.

Assumption 2. *For each x there exists an integer o such that the columns of the matrix*

$$\boldsymbol{F}_x \coloneqq (\boldsymbol{F}_{x,m_1}^\top, \ldots, \boldsymbol{F}_{x,m_o}^\top)^\top$$

are all distinct.

This requirement is quite weak. A simple sufficient condition for Assumption 2 is that there exists an m for which the distributions $\boldsymbol{f}_{z,x,m}$ and $\boldsymbol{f}_{z',x,m}$ are different for all pairs (z, z').

Identification. 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.

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').

A first implication of Equation (2.3) is that

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

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

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

That is, $\boldsymbol{Q}_x$ is a joint diagonalizer of a set of matrices. Notice that the z-th diagonal entry of $\boldsymbol{D}_{x_1,x} \boldsymbol{D}_{x_2,x_1} \cdots \boldsymbol{D}_{x,x_m}$ is $f_{z,x,m}(x_m, \ldots, x_1)$. Moreover, the eigenvalues of the set of matrices $\boldsymbol{C}_{x_1,x} \boldsymbol{C}_{x_2,x_1} \cdots \boldsymbol{C}_{x,x_m}$ (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 $\boldsymbol{C}_{x'_1,x} \boldsymbol{C}_{x'_2,x'_1} \cdots \boldsymbol{C}_{x,x'_{m'}}$ (as a function of $x'_1, \ldots, x'_{m'}$) for any different walk length m'. Take a set of o such walk lengths, $m_1, \ldots, m_o$. This delivers a joint diagonalization problem whose eigenvalues are the rows of the matrix $\boldsymbol{F}_x$. By Assumption 2 there exists an o 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 the permutation of the entries, from the fact that

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

where the first equality uses the model structure and the second follows from the result above, and $\boldsymbol{\iota}_q$ denotes the q-vector of ones. Indeed, re-arrangement of this expression reveals that

$$(\boldsymbol{\Omega}_x \boldsymbol{\Delta}_x)^\top \boldsymbol{\iota}_q = \tilde{\boldsymbol{Q}}_x^\top \boldsymbol{B}_x \boldsymbol{p}_x.$$

A second implication of Equation (2.3) is that, for each (x, x'),

$$\tilde{\boldsymbol{Q}}_{x'}^{-1} \boldsymbol{C}_{x,x'} \tilde{\boldsymbol{Q}}_x = \boldsymbol{\Delta}_{x'}^{-1} \boldsymbol{\Omega}_{x'}^{-1} \boldsymbol{D}_{x,x'} \boldsymbol{\Omega}_x \boldsymbol{\Delta}_x =: \tilde{\boldsymbol{D}}_{x,x'}$$

is diagonal. Re-arranging this equation yields

$$(\boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}_{x'})\,\tilde{\boldsymbol{D}}_{x,x'} = \boldsymbol{\Delta}_x^{-1}(\boldsymbol{\Omega}_{x'}^{-1}\boldsymbol{D}_{x,x'}\boldsymbol{\Omega}_x)\boldsymbol{\Delta}_x.$$

It is easy to see that the matrix on the right-hand side must be diagonal; a proof is provided in Lemma A.1 in the Appendix. It is also easy to see that, because the product of two permutation matrices itself is a permutation matix, the matrix on the left-hand side can be diagonal if and only if

$$\boldsymbol{\Delta}_x^{-1}\boldsymbol{\Delta}_{x'} = \boldsymbol{I}_q;$$

see Lemma A.2 for a proof. This implies that $\boldsymbol{\Delta}_x = \boldsymbol{\Delta}_{x'}$. As this holds for all (x, x') we have that $\boldsymbol{\Delta}_x = \boldsymbol{\Delta}$ for some $\boldsymbol{\Delta}$, independent of the value x.

We have just shown that the r matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ are identified up to a common permutation of their columns. From Equation (2.3) we may then directly recover the transition kernels as

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

up to the same permutation.

Because the diagonal of $\boldsymbol{D}_{x,x'}$ constitutes the x'-th row of matrix $\boldsymbol{K}_x$, we can construct the matrices $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_r$, with their columns arranged in the same common order. Now, Assumption 1 implies that each $\boldsymbol{K}_x$ has maximal column rank. The Markovian structure, in turn, implies that $\boldsymbol{p}_x^* = \boldsymbol{K}_x\boldsymbol{\lambda}_x$, where $\boldsymbol{\lambda}_x \coloneqq (s_1(x)\,\mu_1, \ldots, s_q(x)\,\mu_q)^\top$. Consequently, calculating

$$\boldsymbol{\lambda}_x = (\boldsymbol{K}_x^\top\boldsymbol{K}_x)^{-1}\boldsymbol{K}_x^\top\boldsymbol{p}_x^*$$

for each x gives the joint distribution of types and initial conditions, again up to the same common permutation. Collecting $\boldsymbol{\Lambda} \coloneqq (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)^\top$ and letting $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_q)^\top$ we recover the distribution of latent types as

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^\top\boldsymbol{\iota}_r.$$

Finally, with $\boldsymbol{S} \coloneqq (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_q)$ for $\boldsymbol{s}_z \coloneqq (s_z(1), \ldots, s_z(r))^\top$, we can write $\boldsymbol{\Lambda} = \boldsymbol{S}\,\mathrm{diag}(\boldsymbol{\mu})$ and so

$$\boldsymbol{S} = \boldsymbol{\Lambda}\,\mathrm{diag}(\boldsymbol{\mu})^{-1}$$

yields the type-specific distributions of the initial condition.

We have shown the following result.

Theorem 1. *Let Assumptions 1 and 2 hold. Then the distributions of the initial condition, s_z, the transition kernels, k_z, and the type probabilities, μ_z, may all be nonparametrically identified, up to a common permutation of the latent types, from the distribution of four consecutive observations.*

Having access to longer time series allows to weaken Assumption 1. Say that 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 $\boldsymbol{P}_x$ the joint distribution of $X_1, \ldots, X_{T-1}$ at $X_{\lfloor T/2 \rfloor} = x$ and $\boldsymbol{P}_{x,x'}$ 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. Consequently, we can accommodate up to $r^{(T-2)/2}$ latent types when T is even and $r^{(T-3)/2}$ latent types when T is odd. Assumption 2 remains unchanged.

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 was 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

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

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_6 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 as 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.

If Conditions (i)–(iii) hold for all x Proposition 6 of Kasahara and Shimotsu (2009) can be applied to each of them. Identification here is up to an arbitrary ordering of the latent types, however, and separate application of 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 $\boldsymbol{\Delta}_1\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\Delta}_r\boldsymbol{\lambda}_r$, 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 outcomes, the latent types, too, may follow a first-order Markov process. This, of course, nests our setup. On the other hand, their approach requires that $r = q$, i.e., that the outcomes can not take on more values as there are latent types, which is restrictive and not imposed here.

Hu and Shum (2012) recover the unknown probabilities from the distribution of only four outcomes, as do we. To do so they impose our Assumption 1 together with 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 joint-diagonalization system. 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 of the solution to their joint-diagonalization problem. As such, it plays the same role as does our Assumption 2. However, our Assumption 2 is arguably weaker in that we only require that there exist (collections of) walks along the type-specific Markov chains that occur with different probability for the different types.

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

$\boldsymbol{\Delta}_1\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\Delta}_r\boldsymbol{\lambda}_r$ 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 completing the identification analysis. Our Theorem 1 shows that a common (albeit arbitrary) ordering is identified from the data. Therefore, monotonicity restrictions linking types to outcomes can be dispensed with.[1]

4 Estimation

The proof of Theorem 1 is constructive. The key to the construction of an estimator based on it is a routine to (approximately) solve 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).

[1] Hu and Shum (2012, Theorem 1 and Corollary 1) also treat the case where the transition kernels are time dependent. They show that, if five time-series observations are available, their arguments can be used to establish identification of the transition kernels relating time period three to time period four. It turns out that, under their conditions, our techniques can be modified to yield identification of the transition kernel relating time period two to time period three from only four time-series observations. This, then, equally identifies the type-specific marginal distributions of the outcome in those two periods, as well as the distribution of latent types, which does not change with time. Again, this result does not require monotonicity restrictions.

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). Initializing the algorithm at differents sets of starting values can be used to ensure that the global maximizer is found. We defer to McLachlan and Krishnan (2008) for additional discussion and references.

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

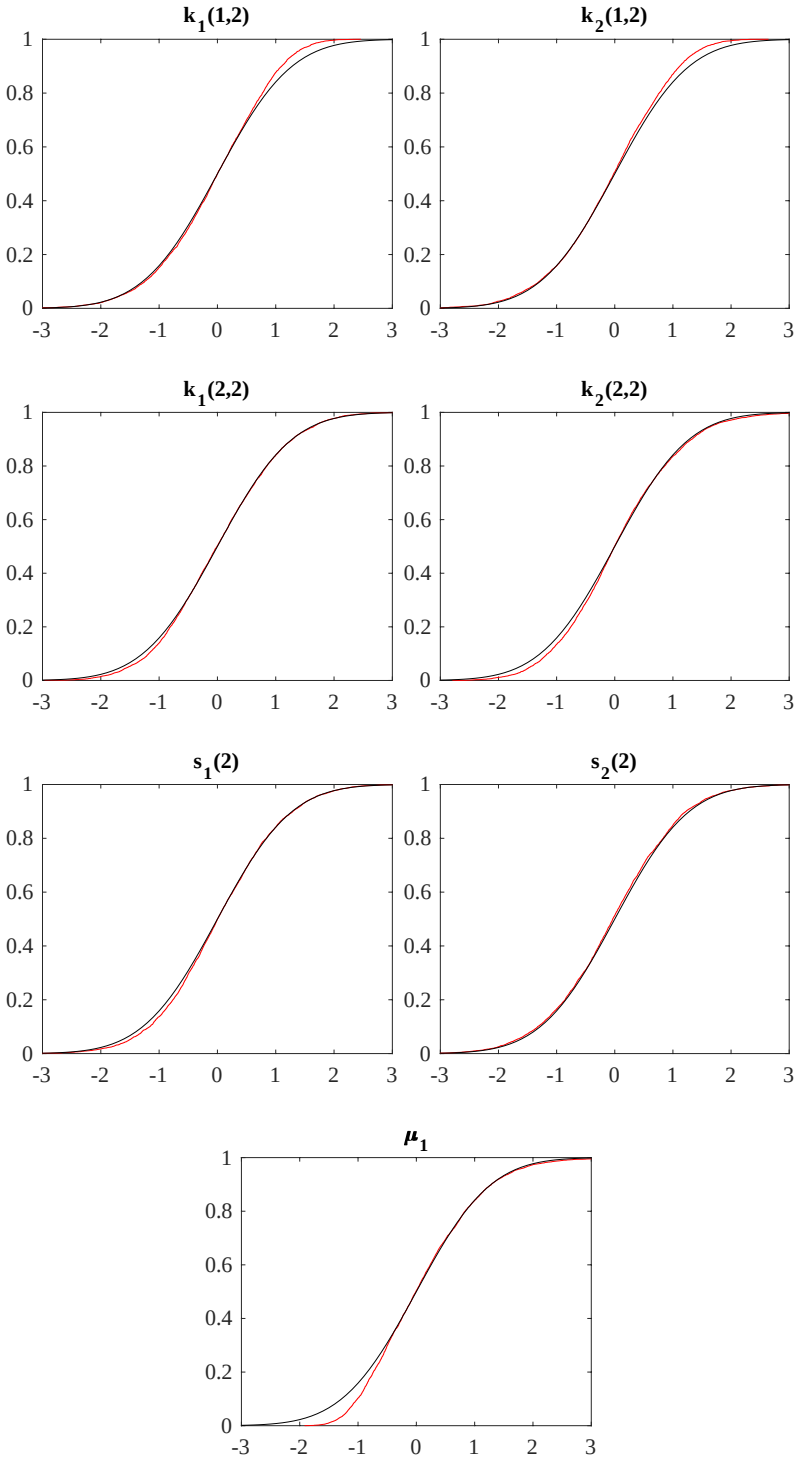
k_1	1	2
1	$2/10$	$8/10$
2	$7/10$	$3/10$

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.

The plots in Figure 1 contain the empirical cumulative distribution functions (in red) of the Studentized point estimators for the different parameters, obtained from four-wave panel data sets consisting of 500 observations. Each plot also provides the benchmark

Figure 1: Studentized empirical distributions



standard normal distribution. 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. Inspection of the simulation output revealed that these stem from estimation error in the variance. The various parameter estimators are all virtually unbiased. The normal approximation quickly improves as either n grows or as longer panels are considered. We omit additional numerical output.

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 now is

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

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 and 2, 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. In contrast, an extension of the approach underlying the result in Proposition 6 of Kasahara and Shimotsu (2009) would require $3\,(p + 1)$ observations.

Conclusion

We have derived a constructive nonparametric 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 identification 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

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 that 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. □

Lemma A.2. *Let $\boldsymbol{P}$ be a permutation matrix and let $\boldsymbol{D}$ be a diagonal matrix. Then $\boldsymbol{P}\boldsymbol{D}$ is a diagonal matrix if and only if $\boldsymbol{P}$ is the identity matrix.*

Proof. The 'if' part of the lemma is immediate. For the 'only if' part, we note that, in the same way as in the proof of Lemma A.1,

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

Thus the (potentially) non-zero entries of $\boldsymbol{P}\boldsymbol{D}$ are at the co-ordinates $(i, \sigma(i))$. These are the co-ordinates of the ones in the permutation matrix $\boldsymbol{P}$. Hence, only if $\boldsymbol{P}$ is diagonal will $\boldsymbol{P}\boldsymbol{D}$ be diagonal. The result then follows from the fact that $\boldsymbol{P}$ is a permutation matrix and so necessarily equal to the identity matrix when diagonal. □

References

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica 75*, 1–54.

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.

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.

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

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 of 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.

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.

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

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