LEARNING MARKOV PROCESSES WITH LATENT VARIABLES FROM LONGITUDINAL DATA

Ayden Higgins[*]
University of Oxford

Koen Jochmans[†]
University of Toulouse Capitole

This version: May 9, 2023

Abstract

We present a constructive proof of identification of the parameters of a bivariate Markov chain when only one of the two random variables is observable. This setup generalizes the hidden Markov model in various useful directions, allowing for state dependence in the observables and allowing the transition kernel of the hidden Markov chain to depend on past observables. We give conditions under which the transition kernel and the distribution of the initial condition are uniquely recoverable (up to an arbitrary permutation of the latent states) from the cross-sectional joint distribution of four (or more) time-series observations.

JEL Classification: C14, C23

Keywords: finite mixture, Markov process, regime switching, state dependence.

[*] 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.

Financial support from the European Research Council, 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.

1 Introduction

Let $Y_0, Y_1, \ldots, Y_T$ be a bivariate random process on a finite state space. Once the random variable Y_0 has been drawn from an initial distribution, the sequence $Y_1, \ldots, Y_T$ evolves according to a time-homogenous Markov chain. Partition Y_t as (X_t, Z_t). The random variables $X_0, X_1, \ldots, X_T$ can take on r values and are observable. The random variables $Z_0, Z_1, \ldots, Z_T$ can take on q values and are latent. We complete the model with the assumption that the transition probability

$$\mathbb{P}\{X_t = x_t, Z_t = z_t | X_{t-1} = x_{t-1}, Z_{t-1} = z_{t-1}\}$$

factors as

$$\mathbb{P}\{X_t = x_t | X_{t-1} = x_{t-1}, Z_t = z_t\} \times \mathbb{P}\{Z_t = z_t | X_{t-1} = x_{t-1}, Z_{t-1} = z_{t-1}\}. \tag{1.1}$$

This is a redundancy statement on further lags of the latent variable and is intuitive. The state space of the X_t is taken to be the set of positive integers up to r and the state space of the Z_t is normalized to the set of positive integers up to q. The former restriction is imposed for notational convenience—translation to a general set is immediate—while the latter normalization, given that the process in question is unobserved, is without loss of generality. Also, given that the state spaces are finite, our focus on scalar random variables is innocuous.

Our aim in this paper is to recover the distribution of the initial condition Y_0 and the (time invariant) transition probabilities from Y_{t-1} to Y_t from the distribution of only $X_0, X_1, \ldots, X_T$. We show that, subject to conditions spelled out below, this is possible as soon as three transitions are observed, i.e., $T \geq 3$. Identification is to be understood as being up to an arbitrary permutation of the state space of the Z_t. Indeed, as these random variables are unobserved, their support can be relabelled without any observable implications. Such invariance is standard in models that feature latent variables and is harmless for our purposes.

The problem that we study has applications in, for example, economics (e.g., Miller 1984), speech recognition and natural-language modelling (e.g., Rabiner 1989), molecular biology (e.g., Krogh, Brown, Mian, Sjölander and Haussler 1994), and graphical modelling (e.g., Bishop 2006). Our setup encompasses several specifications that each have received considerable attention in the literature. The first such model is the hidden Markov model (see Cappé, Moulines and Rydén 2005, Allman, Matias and Rhodes 2009, Gassiat, Cleynen and Robin 2016, and Bonhomme, Jochmans and Robin 2016a). In this case, $X_0, X_1, \ldots, X_T$ is assumed to be a stationary sequence whose components are independent conditional on $Z_0, Z_1, \ldots, Z_T$, the distribution of X_t given $Z_0, Z_1, \ldots, Z_T$ depends only on Z_t, and the sequence $Z_0, Z_1, \ldots, Z_T$ is a (stationary) Markov chain. In such a setting, (X_0, Z_0) is assumed to be a draw from the steady-state distribution and the transition probability in (1.1) is taken to further simplify to

$$\mathbb{P}\{X_t = x_t | Z_t = z_t\} \times \mathbb{P}\{Z_t = z_t | Z_{t-1} = z_{t-1}\}.$$

Our framework is more general in the following three ways. First, it does not require full stationarity. Second, it allows for the observable variables to be dependent even after conditioning on the latent variables, as X_t is Markovian conditional on Z_t. Third, it allows the transition probabilities in the hidden chain to depend on past realizations of the observable chain. That is, the evolution of the unobservable chain depends on the observable variables. This is important, and Pouzo, Psaradakis and Sola (2022) discuss the usefulness of this generalization and provide many examples.

The second model nested in our setup is obtained by complementing the restrictions of the hidden Markov model with the condition that the latent variable does not change over time, i.e., that $\mathbb{P}\{Z_t = Z_0\} = 1$ for all $t \geq 1$. In this case, the joint distribution of $X_0, X_1, \ldots, X_T$ factors as a multivariate mixture model (see, e.g., Hall and Zhou 2003) with identically distributed measurements (as in Bonhomme, Jochmans and Robin 2016b and Vandermeulen and Scott 2020). The latent variable Z_0 governs which mixture component the sequence $X_0, X_1, \ldots, X_T$ is drawn from.

Finally, a dynamic version of the mixture model is a third special case. Here, after drawing (X_0, Z_0) from an initial distribution, the Z_t are held fixed at Z_0 while the X_t follow a Markov chain whose transition kernel depends on Z_0. Here, the distribution of X_t

given Z_0 can depend on t as the initial condition (X_0, Z_0) need not be a draw from the steady-state distribution of the process (presuming that such a distribution exists). This type of structure was initially considered by Kasahara and Shimotsu (2009) and, more recently, by Higgins and Jochmans (2021).

Our focus in this paper is on identification. Given the discreteness of the variables involved estimation can be done by maximum likelihood using some version of the EM algorithm, which is well understood. Ailliot and Pène (2015) provide a discussion and many references on this.

The question of identification in our model cannot be answered by relying on existing methods for hidden Markov chains or for multivariate mixtures. While our arguments bare some similarity with the approaches taken in Bonhomme, Jochmans and Robin (2016a) and Higgins and Jochmans (2021), the fact that the observed and unobserved variables are allowed to be jointly Markovian makes the key restrictions used there inapplicable here. As far as we are aware, the only other work that has taken up the problem of identification here is Hu and Shum (2012). However, their argument only applies to the case where the state spaces of the observable and the latent Markov chain are the same, i.e., $r = q$. Additionally, they assume that there exists a functional of the distribution of X_t given $X_{t-1} = x, Z_t = z$ that is known to be strictly monotonic in z for all x. This allows them to circumvent indeterminacies due to a permutational invariance in their argument and obtain identification. The approach that we take in this paper does not require this type of assumption. This is important because monotonicity restrictions of this type can be hard to justify in practice. They also pose obvious difficulties when trying to enforce them during estimation.

The plan of the paper is as follows. Section 2 contains our identification argument and its proof for the case $T = 3$. Section 3 provides a discussion on the assumptions that underly our result and on how the proof adjusts in the case $T > 3$. It also contrasts our approach with that of Hu and Shum (2012), highlighting where the restriction that $r = q$ and their monotonicity requirement come into play. A short conclusion section ends the paper.

2 Identification

We first show how identification can be achieved from knowledge of the joint distribution of four observations, X_0, X_1, X_2, X_3. Below we discuss how our argument generalizes to the case where longer time series are available.

Consider the $r \times r$ matrix

$$(\boldsymbol{P}_x)_{i,j} \coloneqq \mathbb{P}\{X_0 = j, X_1 = x, X_2 = i\}$$

for each x. Our model implies that X_0 and X_2 are independent conditional on (X_1, Z_1). Therefore,

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

where we introduce $r \times q$ matrices

$$(\boldsymbol{A}_x)_{i,z} \coloneqq \mathbb{P}\{X_t = i | X_{t-1} = x, Z_{t-1} = z\}, \qquad (\boldsymbol{B}_x)_{i,z} \coloneqq \mathbb{P}\{X_0 = i, X_1 = x, Z_1 = z\}.$$

Impose the following assumption.

Assumption 1. *$\boldsymbol{P}_x$ has rank q for all x.*

Assumption 1 implies that there exist $q \times r$ matrices $\boldsymbol{U}_x$ and $\boldsymbol{V}_x$ such that

$$\boldsymbol{U}_x \boldsymbol{P}_x \boldsymbol{V}_x^\top = \boldsymbol{I}_q, \tag{2.3}$$

where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. The matrices $\boldsymbol{U}_x$ and $\boldsymbol{V}_x$ may be constructed from a singular-value decomposition of $\boldsymbol{P}_x$. Combining (2.2) with (2.3) and invoking Assumption 1 shows that

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

where we let $\boldsymbol{Q}_x \coloneqq \boldsymbol{U}_x \boldsymbol{A}_x$ and note that $(\boldsymbol{V}_x \boldsymbol{B}_x)^\top = \boldsymbol{Q}_x^{-1}$ follows as a consequence of the decomposition.

Next consider the $r \times r$ matrix

$$(\boldsymbol{P}_{x_1,x_2})_{i,j} \coloneqq \mathbb{P}\{X_0 = j, X_1 = x_1, X_2 = x_2, X_3 = i\}$$

for each pair (x_1, x_2). Similarly to (2.2), the Markov structure of our model implies the factorization

$$\boldsymbol{P}_{x_1,x_2} = \boldsymbol{A}_{x_2}\boldsymbol{K}_{x_1,x_2}\boldsymbol{B}_{x_1}^\top,$$

where the $q \times q$ matrix

$$(\boldsymbol{K}_{x_1,x_2})_{z_2,z_1} \coloneqq \mathbb{P}\{X_t = x_2, Z_t = z_2 | X_{t-1} = x_1, Z_{t-1} = z_1\},$$

contains the transition probabilities of our Markov chain at given values of the observable variables. Thus, the model implies the set of multilinear restrictions

$$\boldsymbol{M}_{x_1,x_2} \coloneqq \boldsymbol{U}_{x_2}\boldsymbol{P}_{x_1,x_2}\boldsymbol{V}_{x_1}^\top = (\boldsymbol{U}_{x_2}\boldsymbol{A}_{x_2})\boldsymbol{K}_{x_1,x_2}(\boldsymbol{V}_{x_1}\boldsymbol{B}_{x_1})^\top = \boldsymbol{Q}_{x_2}\boldsymbol{K}_{x_1,x_2}\boldsymbol{Q}_{x_1}^{-1}$$

for all pairs (x_1, x_2). The transition probabilities of the Markov chain can be recovered from these restrictions if we can learn the matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ up to a common permutation of their columns. We now turn to conditions under which this can be achieved. Once the transition probabilities are known the distribution of the initial condition $Y_0 = (X_0, Z_0)$ can be recovered with little additional work.

Observe that the redundancy condition in (1.1) gives the factorization

$$\boldsymbol{K}_{x_1,x_2} = \boldsymbol{D}_{x_1,x_2}\boldsymbol{T}_{x_1}$$

for $q \times q$ matrix

$$(\boldsymbol{T}_x)_{z_2,z_1} \coloneqq \mathbb{P}\{Z_t = z_2 | X_{t-1} = x, Z_{t-1} = z_1\}$$

and $q \times q$ diagonal matrix

$$(\boldsymbol{D}_{x_1,x_2})_{z,z} \coloneqq \mathbb{P}\{X_t = x_2 | X_{t-1} = x_1, Z_t = z\}.$$

If, for some (x_1, x_2), the matrix $\boldsymbol{K}_{x_1,x_2}$ is also invertible, then

$$\boldsymbol{G}_{x_2',x_2}^{x_1} \coloneqq \boldsymbol{K}_{x_1,x_2'}\boldsymbol{K}_{x_1,x_2}^{-1} = \boldsymbol{D}_{x_1,x_2'}\boldsymbol{D}_{x_1,x_2}^{-1}$$

is a diagonal matrix for any x_2'. Hence,

$$\boldsymbol{M}_{x_1,x_2'}\boldsymbol{M}_{x_1,x_2}^{-1} = \boldsymbol{Q}_{x_2'}\boldsymbol{G}_{x_2',x_2}^{x_1}\boldsymbol{Q}_{x_2}^{-1} \tag{2.4}$$

for any x_2'. Now, if there is another pair $(x_1', x_2') \neq (x_1, x_2)$ such that matrix $\boldsymbol{K}_{x_1',x_2'}$ is invertible,

$$\boldsymbol{M}_{x_1',x_2}\boldsymbol{M}_{x_1',x_2'}^{-1} = \boldsymbol{Q}_{x_2}\boldsymbol{G}_{x_2,x_2'}^{x_1'}\boldsymbol{Q}_{x_2'}^{-1} \tag{2.5}$$

follows by the same argument. Combining (2.4) and (2.5) yields

$$(\boldsymbol{M}_{x_1',x_2}\boldsymbol{M}_{x_1',x_2'}^{-1})(\boldsymbol{M}_{x_1,x_2'}\boldsymbol{M}_{x_1,x_2}^{-1}) = \boldsymbol{Q}_{x_2}(\boldsymbol{G}_{x_2,x_2'}^{x_1'}\boldsymbol{G}_{x_2',x_2}^{x_1})\boldsymbol{Q}_{x_2}^{-1}.$$

This is an eigendecomposition with $\boldsymbol{Q}_{x_2}$ being the matrix of eigenvectors.

The above discussion motivates the following assumption. We let

$$\boldsymbol{\mathcal{X}}_x \coloneqq \{(x_1, x_1', x') : x_1' \neq x_1, x' \neq x, \operatorname{rank}(\boldsymbol{K}_{x_1,x}) = q, \operatorname{rank}(\boldsymbol{K}_{x_1',x'}) = q\}$$

for each x.

Assumption 2. *For each x,*

(i) $d_x \coloneqq |\boldsymbol{\mathcal{X}}_x| \geq 1$; and

(ii) the elements $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{d_x}$ of the set of vectors

$$\boldsymbol{\mathcal{H}}_x \coloneqq \left\{\operatorname{diag}(\boldsymbol{G}_{x,x'}^{x_1'}\boldsymbol{G}_{x',x}^{x_1}) : (x_1, x_1', x') \in \boldsymbol{\mathcal{X}}_x\right\}$$

are such that all the rows of the $q \times d_x$ matrix $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{d_x})$ are distinct.

Part (i) of Assumption 2 ensures that at least one matrix exists that can be diagonalized by $\boldsymbol{Q}_x$ for each x. When multiple such matrices exist, $\boldsymbol{Q}_x$ is a joint diagonalizer. Part (ii) is a necessary and sufficient condition for the (joint) diagonalizer $\boldsymbol{Q}_x$ to be unique up to scaling and ordering of its columns; this follows from De Lathauwer, De Moor and Vandewalle (2004, Theorem 6.1). Thus, Assumption 2 enables us to identify

$$\tilde{\boldsymbol{Q}}_x \coloneqq \boldsymbol{Q}_x\boldsymbol{\Omega}_x\boldsymbol{\Delta}_x,$$

where $\boldsymbol{\Omega}_x$ is a diagonal scaling matrix and $\boldsymbol{\Delta}_x$ is a permutation matrix, for each x.

The matrix $\boldsymbol{\Omega}_x$ can be recovered, up to permutation of the entries on its main diagonal, from restrictions on the r-vector

$$(\boldsymbol{p}_x)_i \coloneqq \mathbb{P}\{X_0 = i, X_1 = x\}.$$

Indeed, because $\boldsymbol{p}_x = \boldsymbol{B}_x \boldsymbol{\iota}_q$ for $\boldsymbol{\iota}_q$ the q-vector of ones, we have that

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

and, hence, it holds that

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

where we have used the fact that each row of any permutation matrix sums to unity to make the last transition. Further, because $\boldsymbol{\Delta}_x^{-1} \boldsymbol{\Omega}_x \boldsymbol{\Delta}_x$ is a diagonal matrix, this equation yields

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

for all x.

To recover a common permutation matrix, we impose our third and final assumption.

Assumption 3. *There exists a value x_0 such that, for all x, $\boldsymbol{G}_{x,x_0}^{x'}$ is invertible for some x'.*

With x_0 as in Assumption 3, by (2.4),

$$\tilde{\boldsymbol{Q}}_x^{-1} \boldsymbol{M}_{x',x} \boldsymbol{M}_{x',x_0}^{-1} \tilde{\boldsymbol{Q}}_{x_0} = \boldsymbol{\Delta}_x^{-1} \boldsymbol{\Delta}_{x_0} (\boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\Omega}_x^{-1} \boldsymbol{G}_{x,x_0}^{x'} \boldsymbol{\Omega}_{x_0} \boldsymbol{\Delta}_{x_0}) \tag{2.6}$$

for all x and some x'. First, notice that $\boldsymbol{\Delta}_x^{-1} \boldsymbol{\Delta}_{x_0}$ is a permutation matrix and that the matrix in parentheses on the right-hand side is diagonal. Therefore, the latter is equal to the matrix on the left-hand side up to ordering of its rows. It then follows that the column sums of $\tilde{\boldsymbol{Q}}_x^{-1} \boldsymbol{M}_{x',x} \boldsymbol{M}_{x',x_0}^{-1} \tilde{\boldsymbol{Q}}_{x_0}$ identify $\boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\Omega}_x^{-1} \boldsymbol{G}_{x,x_0}^{x'} \boldsymbol{\Omega}_{x_0} \boldsymbol{\Delta}_{x_0}$. Next, because Assumption 3 implies that these matrices are invertible, we can solve for the permutation matrix in (2.6) to get

$$\boldsymbol{S}_{x,x_0} \coloneqq \boldsymbol{\Delta}_x^{-1} \boldsymbol{\Delta}_{x_0} = (\tilde{\boldsymbol{Q}}_x^{-1} \boldsymbol{M}_{x',x} \boldsymbol{M}_{x',x_0}^{-1} \tilde{\boldsymbol{Q}}_{x_0})(\boldsymbol{\Delta}_{x_0}^{-1} \boldsymbol{\Omega}_x^{-1} \boldsymbol{G}_{x,x_0}^{x'} \boldsymbol{\Omega}_{x_0} \boldsymbol{\Delta}_{x_0})^{-1}$$

for each x. We may then compute

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

and recover

$$\boldsymbol{Q}^*_x \coloneqq \tilde{\boldsymbol{Q}}_x \boldsymbol{S}_{x,x_0} (\boldsymbol{\Omega}^*_x)^{-1} = \boldsymbol{Q}_x \boldsymbol{\Delta}_{x_0}$$

for all x. Thus, we have identified the matrices of eigenvectors $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ up to a common column permutation.

We may now recover the transition probabilities of the Markov chain from

$$\boldsymbol{K}^*_{x_1,x_2} \coloneqq (\boldsymbol{Q^*}_{x_2})^{-1} \boldsymbol{M}_{x_1,x_2} \boldsymbol{Q^*}_{x_1} = \boldsymbol{\Delta}^{-1}_{x_0} \boldsymbol{K}_{x_1,x_2} \boldsymbol{\Delta}_{x_0}.$$

It then only remains to identify the distribution of the initial condition, $Y_0 = (X_0, Z_0)$. This distribution is the $q \times r$ table $(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_r)$, with columns

$$(\boldsymbol{\pi}_x)_z \coloneqq \mathbb{P}\{X_0 = x, Z_0 = z\}.$$

To see how this may be done, note that

$$\boldsymbol{p}'_x = \boldsymbol{A}_x \boldsymbol{\pi}_x, \qquad (\boldsymbol{p}'_x)_i \coloneqq \mathbb{P}\{X_1 = i, X_0 = x\}.$$

Observe that the ith row of the $r \times q$ matrix $\boldsymbol{A}_x$ is equal to the column sum of $\boldsymbol{K}_{i,x}$, i.e., $(\boldsymbol{A}_x)_{i,z} = (\boldsymbol{\iota}_q^\top \boldsymbol{K}_{i,x})_z$. Performing the same operation on $\boldsymbol{K}^*_{i,x}$ yields $\boldsymbol{A}^*_x \coloneqq \boldsymbol{A}_x \boldsymbol{\Delta}_{x_0}$ for all x. Assumption 1 implies that this latter matrix has full column rank for each x. Therefore,

$$\boldsymbol{\pi}^*_x \coloneqq ((\boldsymbol{A}^*_x)^\top \boldsymbol{A}^*_x)^{-1} (\boldsymbol{A}^*_x)^\top \boldsymbol{p}'_x = \boldsymbol{\Delta}^{-1}_{x_0} \boldsymbol{\pi}_x$$

is identified. With all unknowns recovered up to a common permutation our argument is complete.

Theorem 1. *Under Assumptions 1–3 the distribution of $Y_0 = (X_0, Z_0)$ and the transition kernel of $Y_t = (X_t, Z_t)$ are identified from the distribution of X_0, X_1, X_2, X_3.*

3 Comments

3.1 Discussion on the required conditions

The assumptions underlying Theorem 1 are in line with those used in related work on the identification of multivariate mixture models.

Assumption 1 is equivalent to demanding that the matrices $\boldsymbol{A}_x$ and $\boldsymbol{B}_x$ have maximal column rank for each x. Recall that

$$(\boldsymbol{A}_x)_{i,z} = \mathbb{P}\{X_t = i | X_{t-1} = x, Z_{t-1} = z\}.$$

That is, the zth column of $\boldsymbol{A}_x$ contains the distribution of X_t given $X_{t-1} = x$ and $Z_t = z$. The matrix $\boldsymbol{B}_x$, in turn, factors as $\boldsymbol{C}_x \boldsymbol{\Lambda}_x$, where

$$(\boldsymbol{C}_x)_{i,z} := \mathbb{P}\{X_1 = i | X_2 = x, Z_2 = z\},$$

i.e., the zth column of $\boldsymbol{C}_x$ contains the conditional distribution of X_1 given $X_2 = x$ and $Z_2 = z$, and $\boldsymbol{\Lambda}_x$ is a diagonal matrix with

$$(\boldsymbol{\Lambda}_x)_{z,z} := \mathbb{P}\{X_2 = x, Z_2 = z\}.$$

Thus, the rank conditions in Assumption 1 amount to linear-independence requirements on certain conditional distributions of the Markov chain, together with a full support condition on $Y_2 = (X_2, Z_2)$. Such linear-independence conditions imply irreducibility of the mixture model and are standard in the analysis of multivariate latent-variable models (Hu 2008, Allman, Matias and Rhodes 2009, Bonhomme, Jochmans and Robin 2016a,b, Vandermeulen and Scott 2020).

Consider Part (i) of Assumption 2 next. It demands invertibility of $\boldsymbol{K}_{x_1,x_2} = \boldsymbol{D}_{x_1,x_2} \boldsymbol{T}_{x_1}$ for several pairs of values (x_1, x_2). Obviously, this requires that both the diagonal matrix $\boldsymbol{D}_{x_1,x_2}$ and the square matrix $\boldsymbol{T}_{x_1}$ are invertible. The former requirement is equivalent to

$$\mathbb{P}\{X_t = x_2 | X_{t-1} = x_1, Z_t = z\} > 0$$

holding for all z, i.e., state x_2 being reachable from state x_1 for all values z. The latter requirement will hold if the distributions $\mathbb{P}\{Z_t = z_2 | Z_{t-1} = z_1, X_{t-1} = x_1\}$, when seen as a function of z_1, are linearly dependent. This is a rank condition that is again familiar from the literature on hidden Markov models (as in Gassiat, Cleynen and Robin 2016 and Bonhomme, Jochmans and Robin 2016a). Our result only requires invertibility of the matrix $\boldsymbol{K}_{x_1,x_2}$ for some pairs (x_1, x_2), and not necessarily for all pairs. Moreover, Theorem

1 covers cases where some states in the observable chain may not be reachable for certain states of the latent variable, and can accommodate situations in which the transition kernel of the hidden chain is singular for certain values of the observables. Part (i) of Assumption 2 ensures that the matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ can be characterized as eigenvectors.

Part (ii) of Assumption 2 is a condition on the associated eigenvalues. If $d_x = 1$ then $\boldsymbol{Q}_x$ diagonalizes a single matrix and the assumption demands the eigenvalues to be unique. More generally, $\boldsymbol{Q}_x$ is the joint diagonalizer of a collection of d_x matrices, and Part (ii) of the assumption is a necessary and sufficient condition for uniqueness of the matrix of joint eigenvectors (up to scale and permutation)

Assumption 3, finally, is the requirement that there is a value x_0 such that for each x there exists a value x' for which

$$\mathbb{P}\{X_t = x_0 | X_{t-1} = x', Z_t = z\} > 0, \qquad \mathbb{P}\{X_t = x \,| X_{t-1} = x', Z_t = z\} > 0,$$

for all z. Moreover, for each state x we can find a state x' from which both x and x_0 are reachable for all z. The need for this requirement arises from the dynamics on the observable chain. It is fundamental in recovering the transition probabilities from $Y_{t-1} = (X_{t-1}, Z_{t-1})$ to $Y_t = (X_t, Z_t)$ up to an (arbitrary) labelling of the support of the sequence of the latent sequence $\{Z_t\}$.

3.2 Longer time series

With more than four time-series observations we can obtain Theorem 1 under a weaker version of Assumption 1. Say we have access to the joint distribution of $X_0, X_1, \ldots, X_T$. Let $\lfloor \cdot \rfloor$ be the floor function. We can redefine $\boldsymbol{P}_x$ to be the table of $X_0, X_1, \ldots, X_{T-1}$ at $X_{\lfloor T/2 \rfloor} = x$ and $\boldsymbol{P}_{x_1,x_2}$ the table of $X_0, X_1, \ldots, X_T$ at $X_{\lfloor T/2 \rfloor} = x_1$ and $X_{\lfloor T/2 \rfloor + 1} = x_2$, both unfolded into matrices. These matrices admit the same type of factorization as performed above. The proof of identification then generalizes in a straightforward manner. The matrices $\boldsymbol{P}_x$ are now 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. Clearly, a larger value of T can only be beneficial in making requirement that $\operatorname{rank}(\boldsymbol{P}_x) = q$ easier to satisfy. As such, having access to longer panel data has the same type of identifying power as working with observables that live on a richer state space.

3.3 Comparison to existing work

Hu and Shum (2012) established an identification result for the transition probabilities in our model for the case $r = q$. The assumptions they rely on are similar to ours (although our Assumption 2 is weaker than their corresponding Assumption 3) except for the last one. We use Assumption 3 to show how the matrices $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_r$ are recoverable up to a common permutation of their columns. This is crucial to be able to identify the matrices $\boldsymbol{K}_{x_1,x_2}$ up to a re-arrangement of their entries that is independent of the pair (x_1, x_2). When $r = q$ Assumption 1 implies that $\boldsymbol{A}_x$ and $\boldsymbol{B}_x$ are square and invertible. Hence, transforming the $r \times r$ matrices $\boldsymbol{P}_{x_1,x_2}$ to the $q \times q$ matrices $\boldsymbol{M}_{x_1,x_2}$ does not achieve any dimension reduction and we can proceed without it. Furthermore, in the arguments of Hu and Shum (2012), $\boldsymbol{A}_x$ then plays a similar role to our $\boldsymbol{Q}_x$. In contrast to the columns of $\boldsymbol{Q}_x$, the columns of $\boldsymbol{A}_x$ are probability mass functions. As these are known to sum to one, their scale is readily recovered. However, the model restrictions used by Hu and Shum (2012) only yield the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ up to different (arbitrary) permutations of their columns. This is insufficient to identify the transition kernel of the Markov process. To arrive at the same conclusion as us, instead of exploiting dynamic restrictions implied by the Markov chain, they impose that for each x there exists a known functional of the distribution of X_t given $X_{t-1} = x$ and $Z_{t-1} = z$ that is strictly monotonic in z. Such a condition implies that the columns of the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ can be re-arranged such that they satisfy this monotonicity requirement. Once this has been done,

$$\boldsymbol{P}_{x_1,x_2}\boldsymbol{P}_{x_1}^{-1} = \boldsymbol{A}_{x_2}\boldsymbol{K}_{x_1,x_2}\boldsymbol{A}_{x_1}^{-1}$$

can be used to recover the matrix of transition probabilities of the Markov chain; recall that, when $r = q$, $\boldsymbol{P}_x$ is invertible for all x by Assumption 1. Our results states that invoking monotonicity conditions to claim identification is not needed. This is useful as such conditions may be difficult to justify and can prove difficult to deal with when turning to estimation.

4 Conclusion

This paper has considered a generalization of the hidden Markov model where the aim is to recover the initial condition and transition kernel of a bivariate Markov chains in which one of the variables is latent. Primitive conditions on the process were given under which this can be achieved from short longitudinal data with as little as four waves.

References

Ailliot, P. and F. Pène (2015). Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models. *ESAIM: Probability and Statistics 19*, 268–292.

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

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

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

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

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.

Gassiat, E., A. Cleynen, and S. Robin (2016). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing 26*, 61–71.

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

Higgins, A. and K. Jochmans (2021). Identification of mixtures of dynamic discrete choices. TSE Working Paper n. 21-1272.

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.

Krogh, A., M. Brown, I. S. Mian, K. Sjölander, and D. Haussler (1994). Hidden Markov models in computational biology: Applications to protein modelling. *Journal of Molecular Biology 235*, 1501–1531.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy 71*, 1565–1578.

Pouzo, D., Z. Psaradakis, and M. Sola (2022). Maximum likelihood estimation in Markov regime-switching models with covariate-dependent transition probabilities. Forthcoming in *Econometrica*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 257–285.

Vandermeulen, R. A. and C. D. Scott (2020). An operator theoretic approach to nonparametric mixture models. *Annals of Statistics 47*, 2704–2733.