Identification and estimation of non-stationary Hidden Markov Models

Martín García Vázquez

May 28, 2021

Abstract

This paper provides a novel constructive identification proof for non-stationary Hidden Markov models. The identification result establishes that only two periods of time are required if one wants to identify transition probabilities between those two periods. This is achieved by using three conditionally independent noisy measures of the hidden state. The paper also provides a novel estimator for non-stationary hidden Markov models based on the identification proof. Montecarlo experiments show that this estimator is faster to compute than maximum likelihood, and almost as precise for large enough samples. Moreover, I show how my identification proof and my estimator can be used in two different relevant applications: Identification and estimation of Conditional Choice Probabilities, initial conditions and laws of motion in dynamic discrete choice models when there is an unobservable state; and identification and estimation of the production function of cognitive skills in a child development context when skills and investment are unobserved.

1 Introduction

Economists often have to deal with unobservable heterogeneity that evolves over time. Near the end of the life-cycle, heterogeneity in health is an important driver of economic decisions of older workers and retirees. One's health evolves with age and is measured with error. In the middle stages of the life-cycle, skills of workers change with time due to the combined forces of investment and depreciation, and they are also measured with error. Finally, at the beginning of the life-cycle, unobserved children's skills are shaped by imperfectly measured parental investments, which in turn affect their economic outcomes later in life.

One way to model such processes that are unobservable and stochastically evolve with time is by using hidden Markov models. Hidden Markov models are used in

I am specially grateful to my advisors Mariacristina De Nardi, Jeremy Lise and Joseph Mullins for their continuous encouragement and support, as well as for very thoughtful discussions. I am also grateful to Elena Pastorino and Jacob Adembaum for very useful discussions, and to Stephane Bonhomme for very useful comments. All errors are my own.

many sub-fields of economics. In financial economics, they have been used to model regime switches in asset characteristics (Ang and Bekaert (2002b), Ang and Bekaert (2002a), Guidolin and Timmermann (2007)). In labor economics, they have been used to find the extent to which people misreport unemployment (Biemer and Bushery (2000)), or to flexibly capture labor market dynamics (Shibata (2019), Hall and Kudlyak (2019)). In health economics, hidden Markov models have been used to classify individuals in latent health groups according to a battery of health observables (Amengual, Bueren and Crego (2021)).

While some identification results are available for stationary hidden Markov models, not much is known about the identifiability of non-stationary hidden Markov models. For some of these applications it may be reasonable to assume that the underlying Markov process is stationary, but for some other applications that assumption is too strong. For example, assuming that labor market flows (employment to unemployment, unemployment to out of the labor force, etc) do not change with the business cycle seems like an unreasonable assumption, and it contradicts basic economic models such as the textbook search and matching model (see for example Shimer (2005)). As another example, because health tends to deteriorate in adulthood, assuming a stationary hidden Markov model for health is also implausible.

This paper fills this gap by establishing identifiability of non-stationary hidden Markov models under mild data requirements. This identification result additionally yields a new set of sufficient conditions for identifiability of *stationary* hidden Markov models as a by-product. In particular, using my identification argument one can show that with access to three conditionally independent measures of the state, a stationary hidden Markov model is identified with two periods. This is the minimum number of periods required to identify parameters if the markovian state was indeed observed.

The idea of the identification result consist in dividing the problem of identification in two steps. In the first step, three noisy measures are used to achieve identification of what I call cross-sectional parameters. These are the parameters that determine the distribution of the cross-section of the data, and include the distribution of the observed state at each point in time and the conditional distributions of the measures given the state. In the second step, the one-period transition probabilities of the hidden state are identified exploiting knowledge of the cross-sectional parameters and the observation of one of the noisy measures in two consecutive periods. Since only one noisy measure is required to infer the dynamics of the underlying state *given knowledge of cross-sectional parameters*, the identification proof does not require assumptions on the dynamics of the other two measures. In particular, those two measures can have arbitrary correlation structures with their own leads and lags, and with the leads and lags of the unobserved state.

In addition, I show that the identification result generalizes without change to a more general set of models. In particular, I show that it is not necessary to assume that the hidden state is first-order Markov in order to identify its one-period transition probabilities. Together with the fact that the dynamics of all the noisy measures but one can be left unspecified, this implies that the identification result is robust to two potentially important features of the data: Serial dependence for the unobserved state

more general than first-order Markov, and arbitrarily complicated dynamics for all the noisy measures but one.

Inspired by the identification result, I propose a novel estimator for non-stationary hidden Markov models that is root-n consistent and asymptotically normal. This estimator can be used when the researcher has access to longitudinal data and at least three measures that are assumed to be noisy signals of the true underlying markovian state. Note that for many relevant applications, having three measures of the state is not a strong requirement. For instance, having access to three measures in a health economics application is often feasible because commonly used datasets, such as HRS, ELSA and SHARE, contain a large battery of measures of health. Similarly, CPS contains questions that can be informative of labor force status besides the usual self-reported measure, so the identification and estimation results in this paper can also be used in a labor market application similar to the ones mentioned above[1].

Using a set of Montecarlo experiments, I show that this new estimator behaves well in finite samples. Additionally, I compare this new estimator to maximum likelihood. I find that my estimator is computationally less demanding (as measured by computing time) but less precise. However, for large samples the loss in precision is negligible, while the decrease in computational cost is substantial.

Moreover, I show that my identification and estimation results can be used in two different settings of importance to economists. More concretely, my identification result can be used to show constructively non-parametric identification of conditional choice probabilities, initial conditions and laws of motion of the observed and unobserved state in the context of a dynamic discrete choice model. In this application, the evolution of the hidden state is allowed to be influenced by observable choices and states, in the spirit of Hu and Shum (2012) and Hwang (2021). Moreover, I also show that my identification and estimation results can be used to establish identification and to estimate the production technology of cognitive skills in a child development context. In this application, investment and skills are assumed to be discrete and unobserved.

Taken together, these two applications illustrate how the results in this paper for non-stationary hidden Markov models can also handle cases in which the evolution of the unobservable state is endogenously affected by observed and unobserved variables.

The rest of the paper is organized as follows: Section 2 discusses the related literature. Section 3 presents the identification result. Section 4 presents the two-step estimator for non-stationary hidden Markov models and discusses the Montecarlo experiments. Section 5 reviews the two applications. Section 6 concludes.

2 Related literature

This paper contributes to the economics literature that applies hidden Markov models to different substantive problems, as well as the econometric and statistical literature that studies identification and estimation of hidden Markov models. Moreover, the two

[1]For an excellent review of datasets that contain many noisy measures of unobserved variables that are relevant for economics see Hwang (2021)

applications in the paper contribute to two additional strands of the literature. The first application provides identification and estimation results that are useful for the dynamic discrete choice literature, while the second application does the same for the child development literature.

Hidden Markov models have been succesfully used in the macro-labor literature. For instance, Shibata (2019) argues that a hidden Markov model represents better labor market dynamics than the commonly used First Order Markov (FOM) model (in which transition probabilities between labor force status categories depend only on current labor force status). One of the facts that motivates Shibata (2019) to adopt a hidden Markov model is that the FOM model cannot match the observed state dependence in job-finding rates. This is because job-finding rates decrease with duration in the data, while in the FOM model they stay constant by assumption. However, the stationary hidden Markov model that he considers does not allow the shape of this state dependence to change with the business cycle, while Kroft, Lange and Notowidigdo (2013) find, using cross-sectional evidence, that the shape of the duration-callback rate profile changes with aggregate labor market conditions. If this difference in callback rates is translated to a difference in job-finding rates, a stationary hidden Markov model would fail to capture this difference, whereas a non-stationary one would not. Hall and Kudlyak (2019) use a stationary hidden Markov model to model the observed heterogeneity in labor market transitions. They assume that workers can belong to finite number of time invariant types. In addition, at each point in time, workers can be in four different time varying states. The identification result in this paper implies that one can identify a version of Hall and Kudlyak (2019) model where transition probabilities between unobserved states change with the business cycle. Note that the presence of types is not a problem, since a hidden Markov model with L types and r states can be viewed as a model with $L \times r$ states and a block-diagonal transition matrix, hence my identification result covers the case with types provided that one has access to appropriate noisy measures for them. Hidden Markov models have also been used in labor to handle measurement error in self-reported labor market status. An example of this is Biemer and Bushery (2000). However, Biemer and Bushery (2000) assume stationary transition probabilities for the state. That is, they assume that flows between different labor force status are stationary. Therefore, their results could be biased if the transition probabilities between labor force status categories changes with the business cycle. A closely related paper to Biemer and Bushery (2000) is Feng and Hu (2013). They also correct for measurement error in self-reported labor market status, and find that the stock of unemployed workers in the US is underestimated when calculated using self-reported labor-market status in the CPS. One advantage in their framework is that they do not rely on the stationarity of the transition probabilities between unobserved labor market status categories, nor on labor-market status being first order Markov. However, they do not establish identification of the transition probabilities between those labor market categories. This paper is able to establish identification of those one period transition probabilities, without assuming that labor-market status is stationary. Moreover, even though I am framing all the results as they apply to hidden Markov models, my identification proof can be used to show identification of the one period

transition probabilities, even if the true underlying state is not first order Markov.

Another strand of the economics literature in which hidden Markov models have been applied is health economics. Amengual et al. (2021) use a non-stationary hidden Markov model to classify individuals in unobserved health groups using information on health measures from HRS. In order to estimate their model they use Markov Chain Montecarlo techniques. My paper ensures identification of their model, and provides a simpler estimation procedure.

The first application discussed in the paper contributes to the dynamic discrete choice literature. From Hotz and Miller (1993), we know that the estimation of many dynamic discrete choice models can be simplified by first estimating reduced-form conditional choice probabilities (CCP), the law of motion of the states and initial conditions.Arcidiacono and Miller (2011) show how to extend the estimation ideas from Hotz and Miller (1993) to models in which there is an unobservable state that is allowed to evolve according to a Markov chain or permanent unobserved heterogeneity. In Arcidiacono and Miller (2011), the time-varying unobservable state is exogenous, in the sense that its evolution is not allowed to be influenced by other variables in the model. Moreover, it is stationary, in the sense that its law of motion does not change with time. Hu and Shum (2012) show that that CCP's, laws of motion of the states and initial conditions can be identified in a model with an endogenous and non-stationary hidden state. Their identification result requires five periods of data for a non-stationary model, and four for a stationary one. It is important to note that in Hu and Shum (2012) the observable measure of the state is allowed to be an endogenous variable that partially determines the evolution of the unobserved state, which expands the set of noisy measures that can be used for identification. Hwang (2021) shows that with three conditionally independent measures of the state, CCP's initial conditions and laws of motion can be identified with two periods of data. Hence, transition probabilities for $T-1$ periods can be identified with T periods of data. Moreover, she proposes a two-step estimator for structural parameters, where the first step of the estimator amounts to maximizing an empirical log-likelihood using the EM algorithm. My identification result has similar requirements to the ones in Hwang (2021) (three conditionally independent measures, 2 periods of data for a stationary model). However, my identification result has some advantages with respect to the one in Hwang (2021). First, the identification result in this paper is fully constructive, whereas Hwang's result is not. This is not only conceptually nice, but it also has the advantage of leading naturally to an estimator of the reduced-form CCP's, initial conditions and laws of motion. Second, the estimator proposed here is computationally less-expensive than maximum lilkelihood, which implies that using it as a first-stage towards estimating structural parameters in a dynamic discrete choice model can reduce computational cost. Third, the rank condition on the conditional distribution of the measures given the state required in this paper has testable implications, whereas the rank condition required by Hwang (2021) seems difficult to test [2]. Finally, when no noisy measure of the state is observed twice, my identification result generalizes without restricting

[2]See footnote 7 in Hwang (2021)

the law of motion of the unobserved state to be stationary and without restrictions on transition probabilities of the observed state given the unobserved state, which strengthens the results in Hwang (2021) (see Appendix B in this paper for more on this).

In addition, my identification argument yields identification of all the reduced-form probabilities of interest without assuming limited feedback, which is assumed in Hwang (2021) and Hu and Shum (2012). Furthermore, I establish identification of observable choices and states tomorrow given observable choices and states today and the unobserved state today and tomorrow, which opens the possibility to testing the limited feedback assumption.

The second application in this paper contributes to the literature in child-development. The results presented in this paper can be used to establish non-parametric identification of the production technology of children's skills. The identification result presented here assumes that all the relevant unobserved variables (skills and investments) are discrete. Note that existent non-parametric identification result for the technology of children's skill formation assume that investments and skills are continuous, but in doing so they require their noisy measures to be continuous too (see for example Cunha, Heckman and Schennach (2010)). The result presented in this paper acknowledges that the discreteness of the measures imposes a restriction on the level of granularity one can allow for when identifying the production function of cognitive development. Moreover, the estimator presented in this paper can be used to estimate the technology of cognitive and non-cognitive achievement in a tractable way.

This paper also relates to the literature on identification of hidden Markov models. Possibly the first identification result for Hidden Markov models is due to Petrie (1969). However, one limitation of his result is that it requires the whole distribution of the observed measures for identification, that is, the distribution of the whole infinite history of observed signals emitted by the markovian state. As noted by Allman, Matias and Rhodes (2009), one can lower this requirement by invoking the result from Paz (1971) that the distribution of a stationary hidden Markov model with r unobserved states is fully characterized by $2r-1$ observations. Allman et al. (2009) provide a new set of sufficient conditions for identification in stationary hidden Markov models using algebraic results due to Kruskal (1976) and Kruskal (1977). The minimum number of consecutive observations required to get identification using their result decreases with the cardinality of the support of the observed measure. Recent results from Gassiat, Cleynen and Robin (2013) and Bonhomme, Jochmans and Robin (2016) lower the number of required observed periods even further to three. My paper establishes that for a stationary hidden Markov model we can lower the number of periods required for identification to two, provided that one has access to three conditionally independent measures of the state. In the context of a non-stationary hidden Markov model, Bonhomme et al. (2017) establish identification of the joint distribution of the hidden Markov state in periods two and three and the noisy measure of the state in periods two and three, under the condition that four periods of data are observed and there is one noisy measure of the hidden Markovian state. This paper shows that a stationary hidden Markov model can be identified with two periods of data if three noisy measures of the state are avaialable at each period. Hence, a emission matrices

and $T-1$ transition probabilities can be identified for a non-stationary hidden Markov model with T periods of data. Another contribution by this paper to the statistical literature is noting that results in the dynamic discrete choice literature, such as the ones in Hu and Shum (2012), Hwang (2021) and this paper, can be useful to users of hidden Markov models outside of economics to establish identification of their models.

Finally, this paper contributes to the literature on estimation of hidden Markov models. First, because inference on parameter estimates for non-stationary hidden Markov models only makes sense if the true parameters are identified. Second, because this paper proposes a novel estimator for non-stationary hidden Markov models. As I show below, this estimator is faster than the Baum-Welch algorithm when the data is a balanced panel. Moreover, as I will argue later , when the data contains refreshment samples the Baum-Welch algorithm becomes computationally very involved, while my estimator is as tractable as in a balanced panel.

3 Identification result

3.1 Non-stationary hidden Markov model.

Consider a Hidden Markov model of the following form: The underlying state, call it S_t, can take r values, that is:

$$S_t \in \{S_1, S_2, ..., S_r\}$$

Moreover, S_t follows a Markov process. Let K_t be the following matrix:

$$(K_t)_{i,j} = \mathbb{P}(S_{t+1} = j | S_t = i)$$

Note that I allow the transition matrix to depend on time in an unrestricted fashion. At each point in time, the econometrician can observe three measures that are conditionally independent given the contemporaneous state. Call this measures Y_t^1, Y_t^2, Y_t^3. The conditional distribution of Y_t^m is given by the matrix P^m for $m = 1,2,3$:

$$P^m = \begin{pmatrix} p_1^m, & p_2^m, & .... & , p_r^m \end{pmatrix}$$

where

$$p_c^m = \mathbb{P}(Y_t^m | S_t = c)$$

is the distribution of Y_t^m given $S_t = c$ Moreover, denote by π_t the distribution of S_t , that is:

$$\pi_t = \begin{pmatrix} \mathbb{P}(S_t = 1), & \mathbb{P}(S_t = 2), & .... & \mathbb{P}(S_t = r) \end{pmatrix}$$

3.2 Identification of the non-stationary Hidden Markov model

The identification result proceeds in two steps. In the first step, which I call the cross-sectional step, I establish identification of P^m for $m = 1,2,3$ and π_t, the cross-sectional

distribution of the underlying state at time t.[3]. This can be done just by making sure that the sufficient conditions in Bonhomme et al. (2016) are satisfied.

Once identification of $P^m, m = 1,2,3$ and $\pi_t, t = 0,...T$ are secured, I establish the identification of the transition matrices $\{K_t\}_{t=0}^{T-1}$ for the hidden markovian state in what I call the longitudinal step. This is done by noticing that under appropriate conditional serial independence assumption for one of the measures, say Y^1, one can write the joint distribution of Y_t^1 and Y_{t+1}^1, which is observed, as a function of π_t, π_{t+1}, P^1 and K_t. Under the same conditions required for the cross-sectional identification step, one can invert this relation in closed-form for K_t as a function of the joint distribution of Y_t^1 and Y_{t+1}^1 and parameters that were identified in the cross-sectional step.

The two conditional serial independence assumption that are needed are the following:

Assumption 1. *(i)* $\mathbb{P}(S_{t+1}|S_t, Y_t^1) = \mathbb{P}(S_{t+1}|S_t)$ *for* $t = 0,...T-1$
(ii) $\mathbb{P}(Y_t^1|S_{t+1}, Y_{t+1}^1) = \mathbb{P}(Y_t^1|S_{t+1})$ *for* $t = 1,...T$

Assumption 1 (*i*) says that Y^1 today does not help predicting the state tomorrow once the state today is known. Assumption 1 (*ii*) says that Y^1 tomorrow does not help predicting Y^1 today once the state tomorrow is known. Note that both assumptions are implied by B1 and are therefore weaker:

Assumption B1. $\mathbb{P}(Y_t^1|Y_0^1,...Y_{t-1}^1, Y_{t+1}^1,...Y_T^1, S_0,...,S_T) = \mathbb{P}(Y_t^1|S_t)$ *for all* $t = 0,....T$

Assumption B1 is assumed in estimation of hidden Markov models. The identification results presented here build on Theorems 1-3 in Bonhomme et al. (2016) for the identification of multivariate latent-structure models applied to the particular case of finite-mixture models with discrete measurements. It is well-known that this models are only identified up to a joint permutation of the columns of the mixing proportions and the component distributions. Hence, for point-identification we need to make sure that a unique re-labeling of the states is available from the emission matrix of some of the measures. This is ensured by the following assumption:

Assumption 2. *There exist* i *and* m^* *known by the researcher such that for row* i *of matrix* P^{m^*} *we have* $P_{i,j}^{m^*} \neq P_{i,j'}^{m^*}$ *for all column* $j \neq j'$ *and either:*
i) $P_{i,j}^{m^*}$ *is increasing in* j *or*
ii) $P_{i,j}^{m^*}$ *is decreasing in* j.

Note that Assumption 2 is implied by some common assumptions in the literature, such as Assumption 2 i) in Hwang (2021). Assumption 2 will be true if the unobserved state and one of the measures have a natural ordering, and a monotonicity condition between them holds. For example, if the unobserved state is health, and the noisy measure is number of limitations with Activities of Daily living, Assumption 2 will hold if the probability of suffering from limitations with **all** Activities of Daily Living decreases as health gets better.

Now I present the identification result:

[3]Note that this cross-sectional distribution of the state at t is implied by π_0 and K_t and therefore is not a "primitive" of the model for $t > 0$

Theorem 1. *Suppose that* P^m *for* $m = 1,2,3$ *is full column rank and* $\pi_t(c) > 0$ *for* $c = 1,...r$. *Under Assumption 1 and 2 the model is identified.*

Proof. I will make the identification argument for K_0, π_0, π_1 , $\{P^m\}_{m=1,2,3}$. The argument for the remaining periods is a trivial extension of this argument.
From Bonhomme et al. (2016) Theorems 1-3, π_0, π_1 , $\{P^m\}_{m=1,2,3}$ are identified up to a joint permutation. Moreover, there is only one permutation consistent with 2, so π_0, π_1 , $\{P^m\}_{m=1,2,3}$ are point-identified. Hence, it only remains to show that K_0 is identified. To see that this is the case, note that the joint distribution of Y_0^1 and Y_1^1 is given by:

$$\mathbb{P}(Y_0^1, Y_1^1) = P^1 \Pi_0 K_0 \Pi^{1-1} (P^1 \Pi_1)'$$

where

$$\mathbb{P}(Y_0^1, Y_1^1)_{i,j} = \mathbb{P}(Y_0^1 = i, Y_1^1 = j)$$

and

$$\Pi_t = \text{diag}(\pi_t)$$

Let $\Omega_0 = P^1 \Pi_0$ and $\Omega_1 = \Pi_1^{-1} (P^1 \Pi_1)'$.
Since P^1 is full column-rank by assumption:

$$K_0 = (\Omega_0' \Omega_0)^{-1} \Omega_0' \mathbb{P}(Y_0^1, Y_1^1) \Omega_1' (\Omega_1' \Omega_1)^{-1}$$

Since $\mathbb{P}(Y_0^1, Y_1^1)$ is observable and Π_1, Π_0 and P^1 are identified this completes the proof. □

It is worth noting that one of the sufficient conditions in the previous proposition involves $\{\pi_t\}_{t=1}^T$, which are in themselves functions of π_0 and $\{K_t\}_{t=0}^{T-1}$. To see what the condition of π_t for $t > 0$ implies in terms of π_0 and $\{K_t\}_{t=0}^{T-1}$ note that I can write the i-th element of π_1 as:

$$\pi_1(i) = \sum_{j=1}^{r} K_0(j,i) \pi_0(j)$$

Since $\pi_0(j) > 0$ by assumption, a sufficient condition for $\pi_1(i)$ to be non-zero is $K_0(j,i) > 0$ for some j. Hence, it is sufficient for all the columns of K_0 to be non-zero. By induction, we have that if $\pi_0(j) > 0$ for all j and K_t has non-zero columns for all t, then π_t has non-zero elements for $t > 0$.

Since the cross-sectional step uses Theorems 1-3 in Bonhomme et al. (2016), which are proven constructively, and since the longitudinal step of the proof is also constructive, the proof of Theorem 1 in this paper is fully constructive.

Note that the cardinality of the noisy measures limits the cardinality of the unobserved state via the full-column rank assumption. More precisely, in order to apply 1, we need three noisy measures of the unobserved state with as many points of support as the unobserved state. Note that the previous identification result only requires one of the measures (Y^1) to comply with the usual conditional serial independence assumption for Hidden Markov Models. This has various implications. First, the weakness of this requirement expands the set of measures that can be used for identification and estimation of non-stationary Hidden Markov models. It also implies that estimation

strategies for non-stationary Hidden Markov models that rely on a full-information likelihood approach are unnecessarily restrictive.
Moreover, if more than one noisy measure Y^m, say Y^1 and Y^2, can plausibly satisfy the conditional serial independence assumption, this opens the door for an overidentification test. This is because one can identify K_t from Y^1 and Y^2 and compare the results.

Also, note that in order to identify K_t we only need data from t and $t+1$. This implies that we can identify a *stationary* hidden Markov model with only two observed periods. To the best of my knowledge, this is a new identification result. This shows that the identification approach in this paper trades-off data requirements in two dimensions: by requiring three noisy measures of the hidden Markov state at each point in time one can reduce the number of periods needed to identify the stationary hidden Markov model.

Another important observation is that the full-column rank assumption of the emission matrices is testable. As noted by Bonhomme et al. (2016), the rank assumptions on P^1 and P^2 imply that $\mathbb{P}(Y_t^1, Y_t^2)$ has rank r. The same applies to a pair of variables including Y^3. Hence, the full-column rank assumptions on P^1, P^2 and P^3 are testable.

Finally, the identification proof does not uses that S_t is first-order Markov [4]. Hence, my identification result can be used to ensure identification of the one period transition probabilities of the state even if S_t is not first order Markov. This may be relevant in some applications. For example, even if one is not willing to assume that the true unobserved labor force status is first order Markov (maybe because the whole history of employment for a particular worker affects physical or human capital of workers, which in turn affects the probability of moving out of unemployment) the identification result in this paper tells us that one can still identify labor market flows from noisy measures of labor market attachment.

The closed form expression for K_t presented in the proof of 1 can be used to prove constructively identification of non-stationary hidden Markov models under a different set of data requirements. More concretely, with at least three time periods ($T \geqslant 3$) we can use Y_{t-1}, Y_t, Y_{t+1} as noisy measures of the unobserved state S_t, given that in a hidden Markov model they are conditionally independent given S_t. Then, we can identify the conditional distribution of $Y_t|S_t$, call it P, using the results from Bonhomme et al. (2016). If this conditional distribution is time-invariant, then we can infer the cross-sectional distribution of the unobserved state at each t from the knowledge of P and the observed cross-sectional distribution of Y_t at each t, that is π_t in the notation of this paper. These pieces of information, P and π_t, are then enough to invert for K_t if we have longitudinal observations at t and $t+1$. This is formalized in the following Theorem:

Theorem 2. *Let the data generating process be given by the one in Section 3.1 with* $m = 1$*, and let the noisy measure of the state be denoted by* Y_t *and its corresponding emission matrix be denoted by* P*. Suppose that 2 holds for some row of* P*. Suppose that* $\{Y_t\}_{t=1}^T$ *is observed and that* $T \geqslant 3$*. Moreover, suppose that* P *is full column rank, and that* π_t *has no zero elements for any* t*. If* $Y_{t-1} \perp Y_t \perp Y_{t+1}|S_t$ *for some* t*, then* P*,* $\{\pi_t\}_{t=0}^T$ *and* $\{K_t\}_{t=1}^{T-1}$ *are identified.*

[4] More on this on Appendix A

Proof. Since Y_t, Y_{t-1}, Y_{t+1} are conditionally independent given S_t, P is full column rank and π_t has no zero elements, P is identified from Theorem 2 in Bonhomme et al. (2016) up to a re-ordering of its columns. Since 2 holds for P there is only one possible re-ordering and P is point-identified. Since P is identified and the cross-sectional distribution of Y_t, $\mathbb{P}(Y_t)$, is observed, π_t for each t is identified from:

$$\pi_t = (P'P)^{-1}P'\mathbb{P}(Y_t)$$

(see Bonhomme et al. (2016) Theorem 3). From knowledge of π_t, P and $\mathbb{P}(Y_t, Y_{t+1})$, we can write K_t as:

$$K_t = (\Omega_t'\Omega_t)^{-1}\Omega_t'\mathbb{P}(Y_t, Y_{t+1})\Omega_{t+1}'(\Omega_{t+1}'\Omega_{t+1})^{-1}$$

where $\Omega_t = P\Pi_t$ and $\Omega_{t+1} = \Pi_{t+1}^{-1}(P\Pi_{t+1})'$ □

Note that again this proof is entirely constructive, given that the identification of P comes from the constructive proofs of Theorems 1 and 2 in Bonhomme et al. (2016), and π_t and K_t are identified constructively in closed-form.

Again, it should be noted that the generality of this result extends beyond non-stationary hidden Markov models. First, the requirement of Y_{t-1}, Y_t, Y_{t+1} for some t is weaker that what is commonly assumed for hidden Markov models. Moreover, even if S_t is not markovian we will be able to identify K_t as the one period transition probabilities.

It should be noted that the result that we can identify emission matrices and cross-sectional distributions of the state with $T = 3$ under assumptions similar to the ones here is not new. In fact, that result is due to Feng and Hu (2013). What is new here with respect to their paper is the identification of the transition probabilities.

Next, I turn my attention to estimation of first-order non-stationary hidden Markov models.

4 Two-step estimator for Hidden Markov Models

The proofs of Theorem 1 shows that the problem of identifying a non-stationary hidden Markov model can be divided into two sub-problems. In the first sub-problem, we identify parameters that determine the cross-sectional distribution of measures. In the second sub-problem, we identify the period-t transition matrix from the joint distribution of measures in t and $t+1$ and parameters identified in the first sub-problem. I now show how to apply the same principle to parameter estimation.

Since the cross-section of measures at t is generated by a finite-mixture model with mixing proportions π_t and component distributions P^1, P^2, P^3, one can estimate those parameters using one of the existing consistent and asymptotically normal estimators for finite mixture models with discrete measurements, such as Maximum Likelihood or the joint diagonalization estimator of Bonhomme et al. (2016). This results in $T+1$ estimates of P^m and in estimates of $\{\pi_t\}_{t=0}^T$. Each of those is a consistent and asymptotically normal. Alternatively, a convex combination with arbitrary weights is also a consistent and asymtptotically normal estimator for P^m. Then, K_t can be estimated by calculating

the sample analogue of the closed-form expression for K_t found in Proposition 1. The consistency and asymptotic normality of the estimator for K_t then follows from the consistency and asymptotic normality of the estimates for the cross-sectional parameters and the Continuous Mapping Theorem.

4.1 Estimator

Given a panel of measures drawn from the model in Section 2 that contains N independent units observed for T periods of time, the two step estimator works as follows:

1. For each $t = 1, ...T$ estimate π_t (the distribution of the state at t) and $\{P_t^m\}_{m=1,2,3}$ using some $\sqrt{n}$ consistent and asymptotically normal estimator. Examples of $\sqrt{n}$ consistent and asymptotically normal estimators at this stage are the MLE (that can be found using direct maximization of the likelihood or the EM algorithm) or the joint approximate diagonalization estimator proposed in Bonhomme et al. (2016). Re-lable the estimates for P^m and π_t according to Assumption 2. Call the corresponding estimates $\hat{\pi}_t$ and $\hat{P}_t^m$. Then, the estimator for π_0 is given by: $\hat{\pi}_0$ and the estimator for P^m is given by:

$$\hat{P}^m = \sum_{t=1}^{T} \alpha_t^m \hat{P}_t^m$$

with

$$\sum_{t=1}^{T} \alpha_t^m = 1$$

2. For each $t = 1, ...T-1$ estimate K_t as:

$$\hat{K}_t = (\hat{\Omega}_{0t}'\hat{\Omega}_{0t})^{-1}\hat{\Omega}_{0t}'\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)\hat{\Omega}_{1t}'(\hat{\Omega}_{1t}'\hat{\Omega}_{1t})^{-1}$$

where

$$\hat{\Omega}_{0t} = \hat{P}^m\hat{\Pi}_t$$

$$\hat{\Omega}_{1t} = \hat{\Pi}_{t+1}^{-1}(\hat{P}_1\hat{\Pi}_{t+1})'$$

where

$$\hat{\Pi}_t = diag(\hat{\pi}_t) \text{ for } t = 0, 1, ...T$$

and where $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)$ is the joint frequency estimator of $\mathbb{P}(Y_t^1, Y_{t+1}^1)$

Note that α_t^m is chosen by the researcher. As long as α_t^m is chosen so that $\sum_{t=1}^{m} \alpha_t^m = 1$, asymptotic normality and consistency will be preserved. However, the finite sample properties and the asymptotic variance of the estimator may depend on the particular choice of $\{\alpha_t^m\}_{t=0}^{T}$.

4.2 Asymptotic Theory.

The previous estimator starts with consistent and asymptotically normal estimates for P^m, $m = 1,2,3$ and π_t for $t = 1,2,...T$. Under the full-rank assumption for P^m and the assumption that π_t has no zero entries for any t, consistent and asymptotically normal estimates can be found using MLE or the joint-diagonalization estimator of Bonhomme et al. (2016). The estimate for K_t inherits these properties because it is given by a continuous function of consistent and asymptotically normal sample statistics. This is formalized in the next two propositions:

Proposition 1. *Suppose Assumptions 1 and 2 hold. Suppose that* $\{\pi\}_{t=0}^{T}$ *has all elements greater than zero. Suppose that* P^m *has full column rank for all* m*. Then the two-step procedure described above produces consistent estimates of* π_0, P^m *for* $m = 1,2,3$ *and* K_t *for* $t = 1,...T-1$*. That is:*

$$\hat{\pi}_0 \to_p \pi_0$$

$$\hat{P}^m \to_p P^m$$

$$\hat{K}_t \to_p K_t$$

Proof. $\hat{\pi}_0$ is consistent by assumption. $\hat{P}^m$ is a convex combination of consistent estimators of P, so by Lemma 2.3 a) in Hayashi (2000) (the Continuous Mapping Theorem for probability limits) is consistent. For the consistency of K note that

$$\hat{\Omega}_{0t} \to_p \Omega_{0t}$$

again by the continuous mapping theorem, where Ω_{0t} is defined as in the identification section.
The same argument applies to $\hat{\Omega}_{1t}$ and Ω_{1t}.
Finally, $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)$ is consistent for $\mathbb{P}(Y_t^1, Y_{t+1}^1)$, since it is a frequency estimator. Hence, by the Continuous Mapping Theorem and the fact that K_t is given by the expression provided in the identification section, $\hat{K}_t$ is consistent for K_t □

Proposition 2. *Suppose Assumptions 1 and 2 hold. .Suppose that* $\{\pi\}_{t=0}^{T}$ *has all elements greater than zero. Suppose that* P^m *has full column rank for all* m*. Then the two-step procedure described above produces asymptotically normal estimates of* π_0, P^m *for* $m = 1,2,3$ *and* K_t *for* $t = 1,...T-1$*.*

Proof. The proof follows the same step as the previous one replacing convergence in probability by convergence to a normal random variable and replacing the Continuous Mapping Theorem for probability limits by the Delta Method. □

The precise asymptotic variance for $\hat{K}_t$ can be calculated from knowledge of the asymptotic variance-covariance matrix of the cross-sectional parameters and $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)$ using the Delta Method. Note that the delta method can be applied here since the mapping from the vector of cross-sectional parameter estimates and the joint-frequency of Y^1 at t and $t+1$ is differentiable. However, this approach is impractical, since it depends on the covariances of the cross-sectional parameters and $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)$. Moreover,

it depends on the particular method chosen to conduct the cross-sectional estimation step. However, since the two-step estimator is fast to compute, a researcher interested in conducting inference can do so using bootstrap.

The desirable asymptotic properties of the two-step estimator follow from the fact that the longitudinal estimation step is given by the sample analogue of a known closed form expression for K_t in the population. Hence, the asymptotic properties of the estimator for K_t will go through as long as the population expression for K_t remains valid. This implies that the two-step estimator is robust to at least two relevant deviations from assumptions commonly used in the literature.

First, as discussed in the identification section, the population expression for the one period transition probabilities for the unobserved state remains valid even if the state is not first-order Markov.

Second, note that in order to find K_t in the population I did not have to make assumptions on the distribution of Y_t^2 and Y_t^3 given their own past and future and the past and future of the state. The only thing that is required from Y_t^2 and Y_t^3 is to be conditionally independent (jointly with Y_t^1) given the contemporaneous unobserved state S_t. This is because once the cross-sectional parameters are identified, the identification of K_t comes from the one-period-ahead dynamics of Y_t^1.

This robustness should be contrasted with Maximum Likelihood Estimation, which is commonly used in the estimation of hidden Markov models, and it requires specifying the whole distribution of the data.

4.3 Imposing restrictions on the estimates for K_t

One issue with the estimator presented above is that it does not impose that the resulting estimate for K_t is a valid transition matrix. Hence, due to sample noise it wont typically be. One easy fix for this is to divide each element of $\hat{K}_t$ by the sum of elements in its row. This normalization will yield a valid transition matrix if all of the elements of $\hat{K}_t$ are non-negative. This does not seem to be an issue in the Montecarlo experiments. Moreover, since this normalization is a continuous function that leaves K_t unaltered at the true population value, it retains consistency and asymptotic normality.

If the negativity of some element of $\hat{K}_t$ turns out to be an issue in a particular application, there are some alternative procedures that can be used to estimate the parameters of the model. For instance, one can impose parametric restrictions on K_t, and estimate the parameters of K_t by simulated minimum distance using $\hat{K}_t$ from the previous two-step estimator as targets in the minimum distance objective. Since $\hat{K}_t$ is consistent and asymptotically normal and K_t is identified, this minimum distance procedure should yield consistent and asymptotically normal estimates. Another way of imposing the restriction that the estimate for K_t is a stochastic matrix is by using the following two-step procedure:

1. This step is the same as the first step for the two-step estimator introduced above. For each $t = 1,...T$ estimate π_t (the distribution of the state at t) and $\{P_t^m\}_{m=1,2,3}$ using some consistent and asymptotically normal estimator. Call the

corresponding estimates $\hat{\pi}_t$ and $\hat{P}_t^m$. Then, the estimator for π_0 is given by: $\hat{\pi}_0$ and the estimator for P^m is given by:

$$\hat{P}^m = \sum_{t=1}^{T} \alpha_t^m \hat{P}_t^m$$

with

$$\sum_{t=1}^{T} \alpha_t^m = 1$$

2. Iterate between the following two steps until convergence:

   - **E step:** Given estimates for $\pi_t, \pi_{t+1}, P^1, P^2, P^3$, $\{Y^1_{i,\tau}\}_{\tau=t,t+1}$ and a guess for $K_t^{(h)}$ calculate the filtered probabilities:

     $$\hat{v}_{i,k,j} := \mathbb{P}(S_{i,t+1} = j, S_{i,t} = k | Y^1_{i,t}, Y^1_{i,t+1}, \{\hat{\pi}_\tau\}_{\tau=t,t+1}, \hat{P}^1 K_t^h)$$

     These filtered probabilities can be calculated as follows:

     $$\hat{v}_{i,k,j} = \frac{\hat{P}^1(y^1_{i,t}, k)\hat{\pi}_t(k)K^{(h)}(k,j)\hat{P}^1(y^1_{i,t+1}, j)}{\sum_{j=1}^{r}\sum_{k=1}^{r}\hat{P}^1(y^1_{i,t}, k)\hat{\pi}_t(k)K^{(h)}(k,j)\hat{P}^1(y^1_{i,t+1}, j)}$$

   - **M step:** Calculate the new guess $K_t^{(h+1)}$ as:

     $$K_t^{(h+1)} = \arg\max_{K} \sum_{i=1}^{N} \left\{ \sum_{k=1}^{r}\sum_{j=1}^{r} \hat{v}_{ikj} \log(K(k,j)) \right\}$$
     $$\text{s.t} \sum_{j=1}^{r} K_t(k,j) = 1 \text{ for all } k$$
     $$\sum_{j=1}^{r} K(j,c)\hat{\pi}_t(j) = \hat{\pi}_{t+1}(c)$$

The second step of the procedure is an EM algorithm that seeks to maximize the log-likelihood of the dataset given by $\{Y^1_\tau, Y^2_\tau, Y^3_\tau\}_{t=t,t+1}$ taking as known the first-stage estimation parameters and imposing that the distributions in t and $t+1$ are consistent with K_t. Hence, the estimator for K_t will be consistent and asymptotically normal since it can be viewed as a two-stage M estimator and since K_t is identified from $\{Y^1_\tau, Y^2_\tau, Y^3_\tau\}_{t=t,t+1}$ given the cross-sectional parameters (see Section 12.4 in Wooldridge (2010)). Moreover, note that the computation of the M step is numerically tractable, since it amounts to solving a program with a concave objective and convex constraint set.

4.4 Two-step estimator vs Maximum Likelihood: Computing time and precision.

We are going to compare the performance of the two-step estimator with the performance of Maximum Likelihood in terms of precision and computing time. The details of this computations can be found in Appendix B. We are going to measure precision using the Mean absolute error of each parameter. More precisely, since P^m, K_t and π_0 are not scalar parameters, for expositional purposes we are going to summarize the precision of the two estimators using the sup-norm of the mean absolute error for each of those parameters. That is, the measure of precision for $\hat{P}^1$ is going to be given by:

$$\text{Maximum MAE}(P_1) = \max_{i,j} \frac{1}{20} \sum_{m=1}^{20} |\hat{P}_1^m(i,j) - P_1(i,j)|$$

In order to find the Maximum-Likelihood estimator, we are going to use the Baum-Welch algorithm. The Baum-Welch algorithm is a popular algorithm to find maximum likelihood estimates in hidden Markov models. It can be viewed as a particular application of the EM algorithm. See Baum, Petrie, Soules and Weiss (1970) for a classic reference on the Baum-Welch algorithm. In Figure 1 we can see the sup-norm of the mean absolute error as a function of the number of cross-sectional units for estimates of the cross-sectional parameters calculated using the maximum likelihood estimator and the two step estimator respectively. In Figure 2 we can see an analogous comparison for the transition parameters in $K = [K_1, K_2]$. As we can see, the Maximum Likelihood estimator seems to be more precise than the two-step estimator (as one would expect). However, Figure 3 also suggests that it is more expensive in terms of computing time.

It is worth mentioning that in Figures 1, 2 and 3 the first step of the two-step estimator, and the full-likelihood maximization for the maximum likelihood estimator, are conducted using a few local search of the EM algorithm. Since the EM algorithm is not guaranteed to converge to a global maximum of the sample log-likelihood, the sup-norm of the mean absolute error is not only going to reflect sampling error, but also computational error coming from the failure of finding a global maximum for some Montecarlo artificial samples. In order to asses to which extent the difference in precision comes from this computational error, as well as to provide a more meaningful comparison between the estimators, we compare computing time and precision for different sample sizes for many numbers of local searches. [5] Figures 1, 2, 4, 5, 6, 7 provide the comparison in precision between the two estimators as a function of the number of cross-sectional units, and figures 3, 8 and 9 compare the computing time as a function of the number of cross-sectional units for the two estimators. As we can see, as the number of local maximizations increases, the precision of the two estimators becomes more similar. Moreover, as the sample size increases, the precision of the two-step estimator converges to the precision of the full-information maximum-likelihood estimator when the number of local maximizations is reasonably high. Moreover, the

[5] By local searches I mean the number of times we run the EM algorithm to find a local maximum. This is done for the two-step estimator to estimate cross-sectional parameters and the distribution of the hidden state at each t, and in the Baum-Welch algorithm to estimate all the parameters simultaneously.

computing time of both estimators increases with sample size, but the computational cost of the Baum-Welch algorithm grows faster with sample size than the computational cost of the two-step estimator. Putting this evidence together, this suggests that for small samples the Baum-Welch estimator may be preferable, whereas for larger samples the two-step estimator may be more practical since it achieves a computational efficiency higher than the one of the Baum-Welch algorithm at a small finite-sample precision cost.

It is worth emphasizing that the two-step estimator requires fewer assumptions than full-information maximum likelihood estimation. In other words, the two-step estimator is more robust. In particular, the two step estimator requires only Assumptions 1 , whereas full-information maximum likelihood requires Assumption B1 for not only Y^1, but also Y^2 and Y^3. Therefore, the two-step estimator is applicable in situations where maximum-likelihood is not.

Moreover, in the previous Montecarlo exercises we only used information on Y^1 in the longitudinal estimation step of the two-step estimator. That is, we calculated K_t using only $\hat{\mathbb{P}}(Y^1_t, Y^1_{t+1})$ Under the same assumptions as in full-maximum likelihood one could combine estimates of K using information on Y^1, Y^2, Y^3. This could improve the finite-sample properties of the two-step estimator and make them more similar to the ones of full-information maximum likelihood. [6]

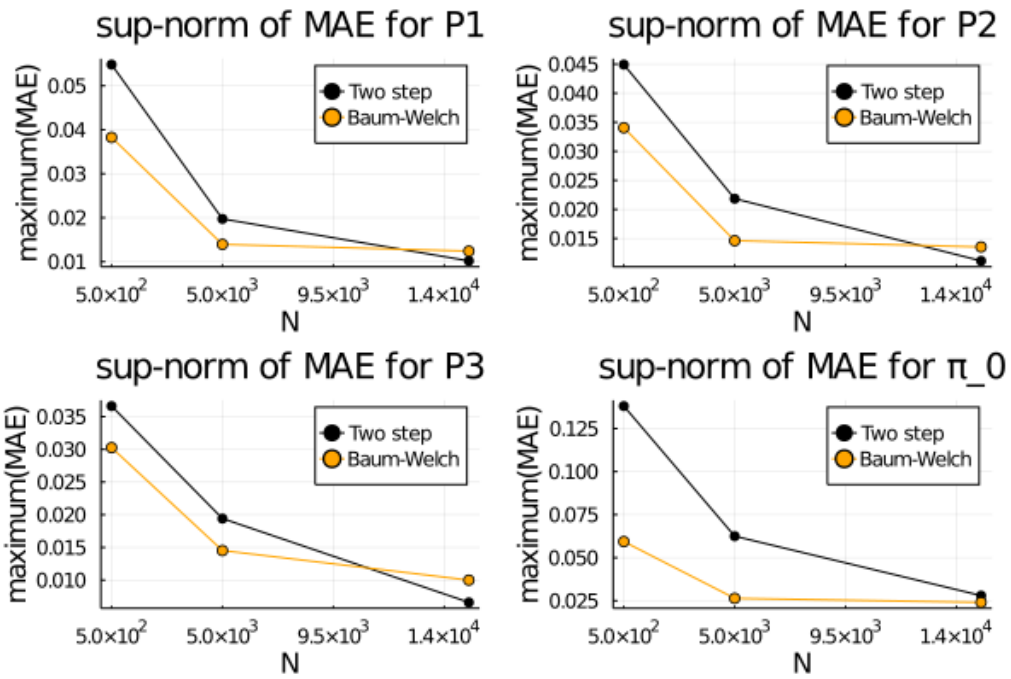

Figure 1: Mean absolute error for different sample sizes. Number of local searches $= 20$

[6] If one wants to apply the two-step estimator and does not impose the conditional serial independence assumption on Y^2 and Y^3, one should allow P^2 and P^3 to change with time. This doesn't add any computational complication.

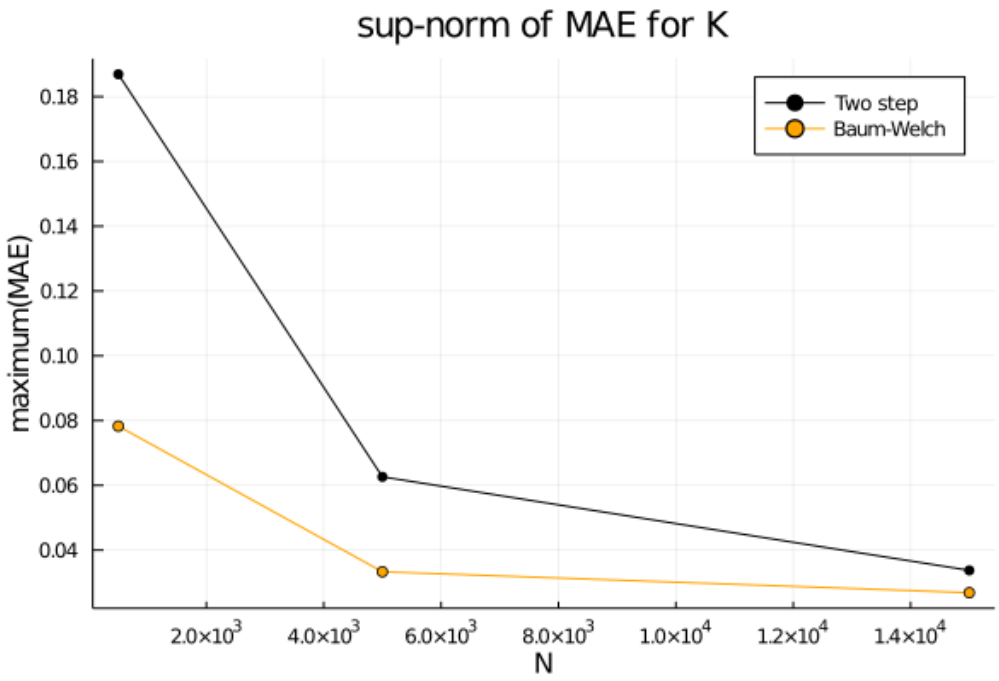

Figure 2: Mean absolute error for different sample sizes. Number of local searches = 20

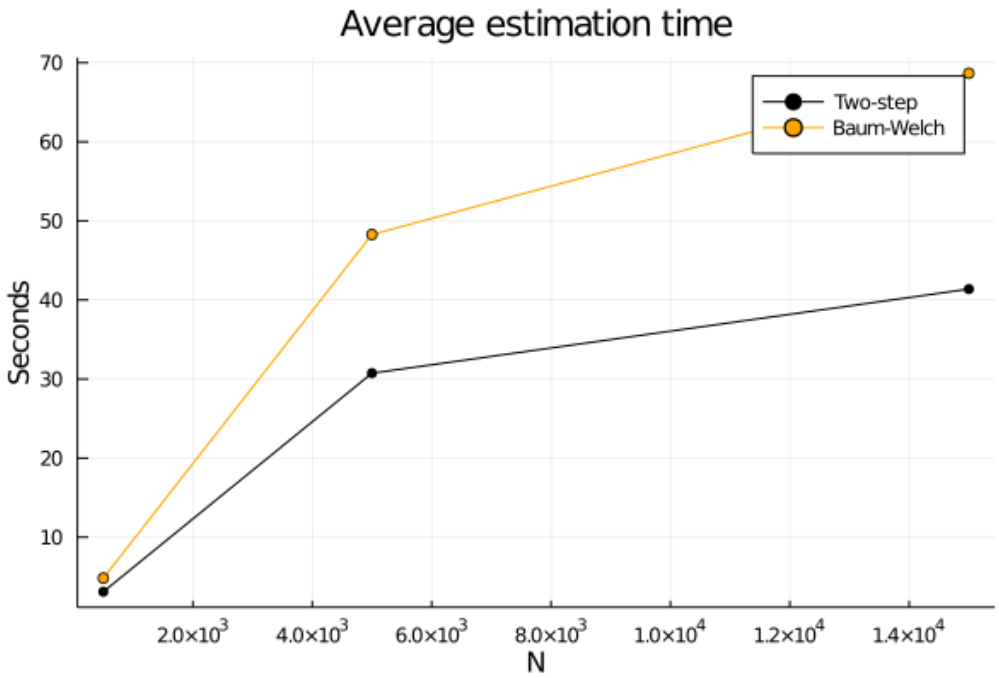

Figure 3: Computing time for different number of units. Number of local searches = 20

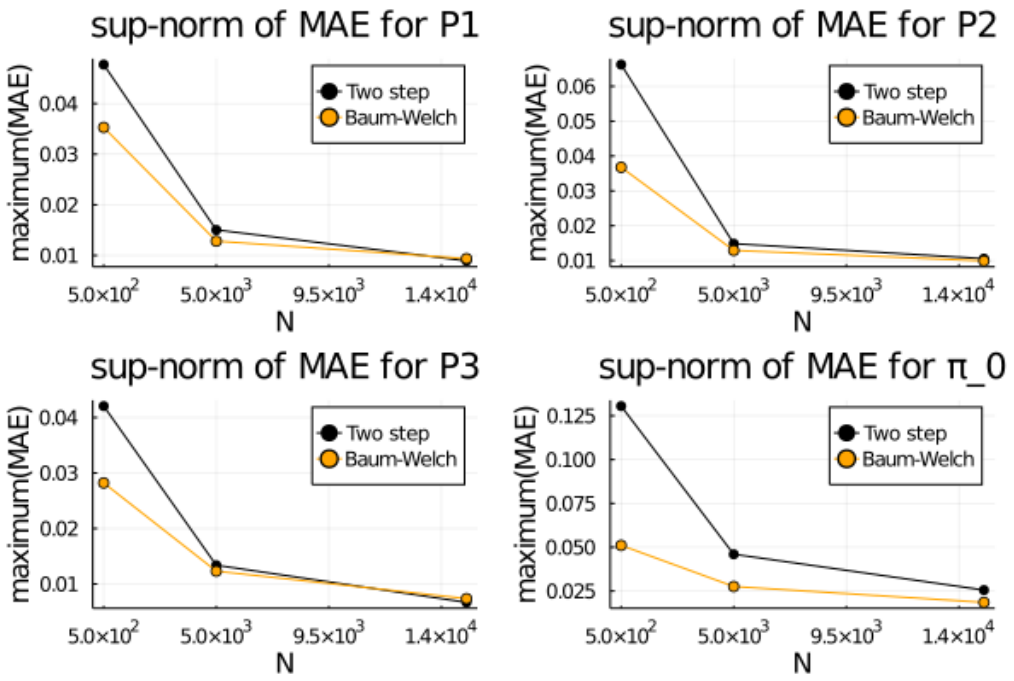

Figure 4: Mean absolute error for different sample sizes. Number of local searches = 50

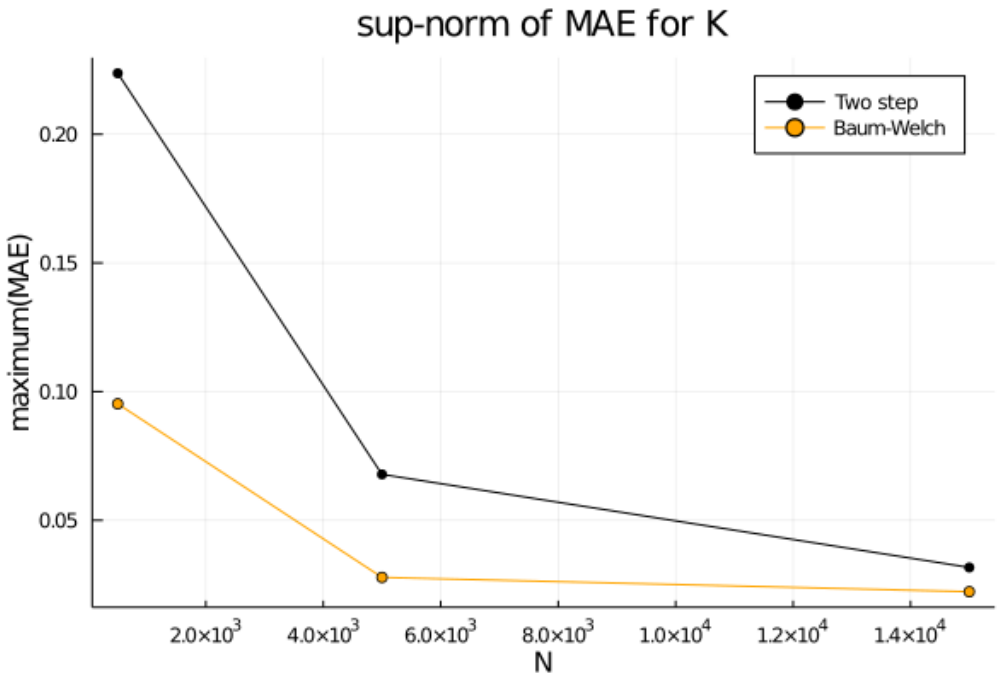

Figure 5: Mean absolute error for different sample sizes. Number of local searches = 50

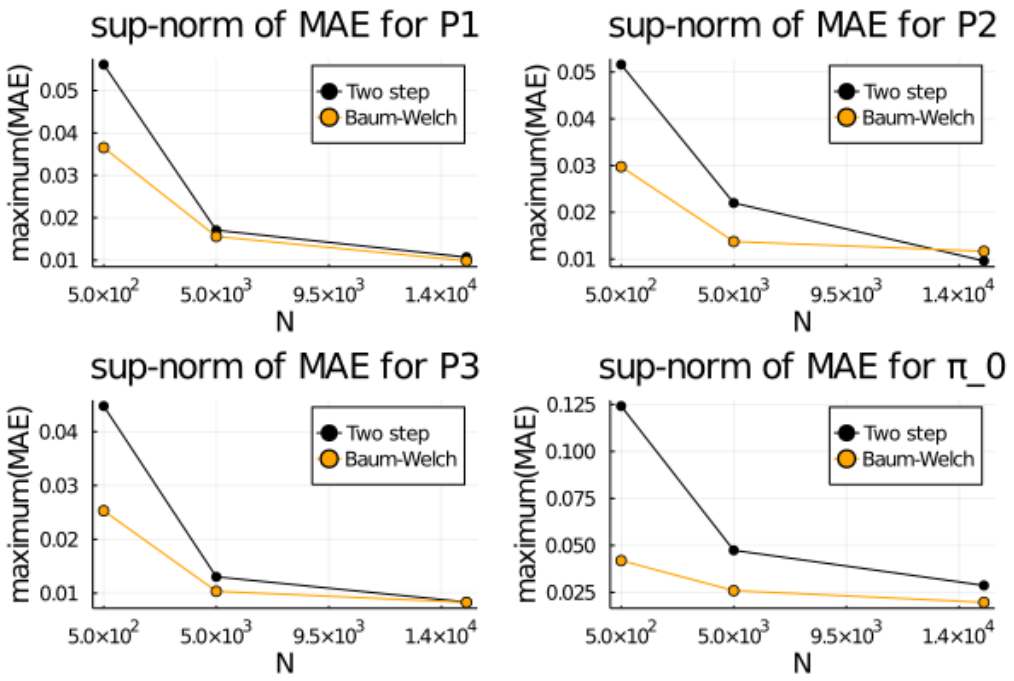

Figure 6: Mean absolute error for different sample sizes. Number of local searches = 80

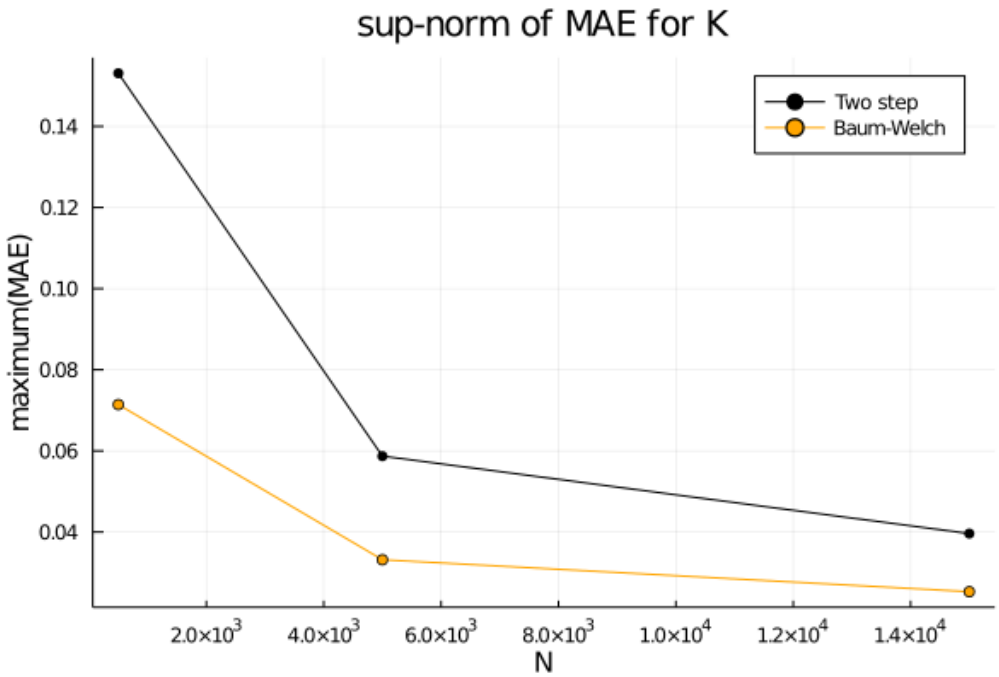

Figure 7: Mean absolute error for different sample sizes. Number of local searches = 80

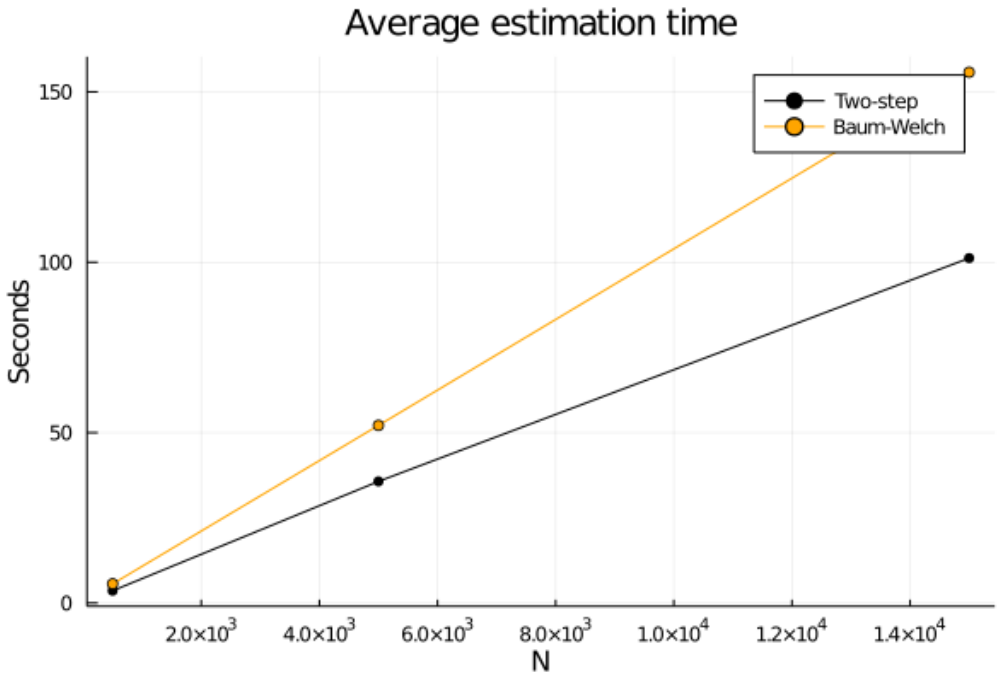

Figure 8: Computing time for different number of units. Number of local searches = 50

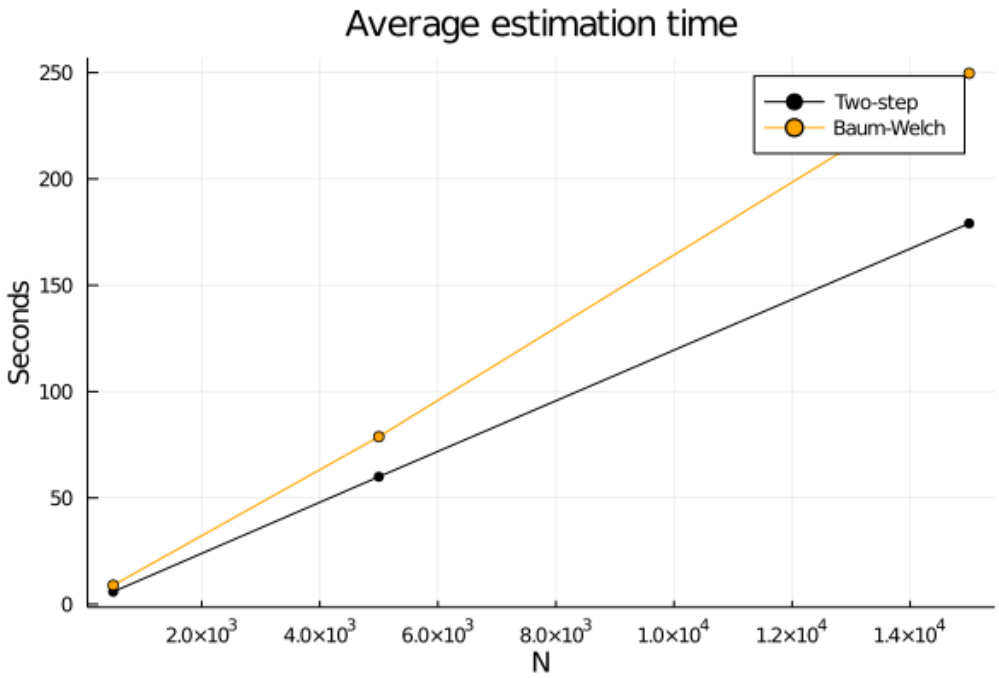

Figure 9: Computing time for different number of units. Number of local searches = 80

4.5 Two-step estimator with refreshment samples.

We describe now the two-step estimator when applied to a dataset that contains refreshment samples. Refreshment samples are of interest in economics because they are common in panel survey datasets, such us for example HRS,ELSA or SHARE. Moreover, there are times in which unbalanced panels with refreshment samples arise from balanced panels. Think of a situation in which a sample of N individuals is sampled for T consecutive periods. Moreover, assume that each individual is initially sampled at a different age. If a researcher wants to estimate a non-stationary hidden Markov model in which the relevant time dimension is age, as opposed to chronological date, she will have to deal with the resulting unbalanced panel with refreshment samples. Estimating this model using the popular Baum-Welch algorithm is going to be computationally expensive, given that with non-stationarity and refreshment samples the maximization step is no longer available in closed form. This happens because the transition parameters in $\{K_\tau\}_\tau^{t-1}$ now affect the probability that a given individual is sampled at state j at time t. In contrast, the two step algorithm is just as simple as in the balanced panel case. The researcher estimates the cross-sectional parameters using all available observations at time t and then it calculates K_t in the same way as before, using as input the empirical joint-frequency of Y_t and Y_{t+1} calculated in the subsample of individuals observed in t and $t+1$. This is formalized below:

1. For each $t = 1, ...T$ estimate π_t (the distribution of the state at t) and $\{P_t^m\}_{m=1,...q}$ using some consistent and asymptotically normal estimator. **Use all available observations at time** t. Re-lable the estimates for P^m and π_t according to Assumption 2 Call the corresponding estimates $\hat{\pi}_t$ and $\hat{P}_t^m$. Then, the estimator for π_0 is given by: $\hat{\pi}_0$ and the estimator for P^m is given by:

$$\hat{P}^m = \sum_{t=1}^{T} \alpha_t^m \hat{P}_t^m$$

with

$$\sum_{t=1}^{T} \alpha_t^m = 1$$

2. For each $t = 1, ...T-1$ estimate K_t as:

$$\hat{K}_t = (\hat{\Omega}_{0t}'\hat{\Omega}_{0t})^{-1}\hat{\Omega}_{0t}'\hat{\mathbb{P}}(Yt^1, Y_{t+1}^1)\hat{\Omega}_{1t}'(\hat{\Omega}_{1t}'\hat{\Omega}_{1t})^{-1}$$

where

$$\hat{\Omega}_{0t} = \hat{P}^m\hat{\Pi}_t$$

$$\hat{\Omega}_{1t} = \hat{\Pi}_{t+1}^{-1}(\hat{P}_1\hat{\Pi}_{t+1})'$$

where

$$\hat{\Pi}_t = diag(\hat{\pi}_t) \text{ for } t = 0, 1, ...T$$

and where $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1)$ is the joint frequency estimator of $\mathbb{P}(Y_t^1, Y_{t+1}^1)$ **calculated on the subsample of individuals observed in** t **and** $t+1$

As it can be seen from the previous description of the estimator, the computation of the two-step estimator with refreshment samples is almost identical to the computation of the two-step estimator in a balanced panel. Hence, if in the balanced panel case the two-step estimator was already faster than Baum-Welch, *a fortiori* the difference will be even larger with refreshment samples, since now the maximization step of the Baum-Welch algorithm has to be carried numerically.

5 Two applications

In this section I show how to exploit my identification and estimation results in two different contexts. First, I will illustrate how my identification argument can be applied to a dynamic discrete choice model with an endogenous unobserved state to establish constructively the identification of the law of motion of that state, the initial distribution of the unobserved state given the observed state and the conditional choice probabilities (CCP) of interest. Moreover, I am going to argue that my two-step estimator can be used to estimate CCP's, the law of motion of the unobserved state and the initial distribution of the unobserved state conditional on the observable state. These objects can be used as a first step towards estimation of structural parameters in a dynamic discrete choice model, using for example the nested pseudo-likelihood estimator proposed by Aguirregabiria and Mira (2002). My identification result for the CCP's, intial distribution of the latent state, and laws of motion for the states has similar data requirements to Hwang (2021). Like Hwang (2021), I am also able to establish identification of those objects in a stationary model with $T = 2$ and three independent noisy measures of the state. However, my full column rank assumption for the emission matrices is testable. On the other hand, Hwang (2021) requires a condition on the *Kruskal rank* [7] of the emission matrices, which as she recognizes in footnote 7, it may be difficult to test in practice. Most importantly, my identification proof of CCP's, laws of motion of the states and initial distribution of the unobservable state is constructive, and leads naturally to an estimator of those objects that is conceptually straightforward and easy to compute.

Second, I am going to apply my results to a child development context to show that my identification proof provides non-parametric identification of the production function of cognitive skills when the underlying investments and skills, and their corresponding noisy measurements, are discrete. The case with discrete investments and skills is relevant for many reasons. First, it is used in practice in structural models of child development (see for example Gayle, Golan and Soytas (2015)). Second, existing non-parametric identification results for continuous investments and skills in the child development literature (Cunha et al. (2010)) assume that measures are continuous too. However, commonly used noisy measures of investments and skills are discrete random variables with very few points of support[8]. Finally, the discrete case can be

[7] See Hwang (2021) or Allman, Matias and Rhodes (2009) for a formal definition of the Kruskal rank of a matrix.

[8] For example, commonly used measures of cognitive skills from C-NLSY 79 include the "Memory for locations" score and the "Knowledge of body parts" score, each of them with 11 points of support.

regarded as an approximation to the continuous case.

5.1 Application to a dynamic discrete choice model

The discrete choice model presented here follows closely the one presented in Hwang (2021). Let S_{it} and X_{it} be the unobservable and observable state respectively of agent i at time t. The unobservable state S_{it} is assumed to be discrete, and its support is assumed to have cardinality r. That is:

$$S_{it} \in \{1, 2, \ldots, r\}$$

Denote as

$$\Omega_{it} := [S_{it}, X_{it}]'$$

the vector of all states of agent i at time t. At each period t each agent i makes an observable discrete choice c_{it} to maximize her lifetime expected discounted utility V_t, that is:

$$V_t(\Omega_{it}) = \max_c \{v_c(\Omega_{it}) + \epsilon_{cit}\}$$

$$v_c(\Omega_{it}) = u(c_{it}, \Omega_{it}) + \beta \mathbb{E} V_{t+1}(\Omega_{it+1} | \Omega_{it}, c_{it})$$

where ϵ_{cit} are choice-specific taste shocks.

I assume that we have access to at least three noisy measures of the unobservable state $\{Y_{it}^m\}_{m=1}^3$. Each of the Y^m is assumed to be independent of its own past and future, the present past and future of other noisy measures; the observable state and choices; and the past and future of the unobservable state conditional on the unobservable state today. This is formalized in the following assumption:

Assumption 3. *For each* $t = 1, \ldots T$ *and for each* $m = 1, 2, 3$ *the following is true:*

$$Y_t^m \perp \{\{Y_\tau^{m'}\}_{\tau=1}^T\}_{m' \neq m}, \{Y_\tau^m, S_\tau\}_{\tau \neq t}, \{X_t, c_t\}_{t=1}^T \mid S_t$$

The conditional distribution of Y^m given the unobserved state S_t are recorded in the emission matrices P^m, which again are assumed to have full-column rank.

5.1.1 Identification

On top of the full-column rank assumption on the emission matrices P^m, we are going to need that for each t the distribution of S_t conditional on X_t, c_t has full support. Moreover, for each $t = 1, \ldots, T-1$ the distribution of S_t conditional on X_{t+1}, X_t, c_t has full support. Provided that this is true, the identification of the CCP's of interest, the initial distribution of S_{it} given X_{it} and the laws of motion for the states follows naturally from the identification argument in Section 3.

First, note that we can apply the cross-sectional identification step conditional on X_t, c_t to identify $\mathbb{P}(S_t | c_t, X_t)$ for each t. Conditioning only on X_t we can identify

The number of points of support is usually even smaller for measures of investment. Examples include the number of books the child has, with 4 points of support, or whether or not the family owns a CD or record player, which is binary.

$\mathbb{P}(S_t|X_t)$, again applying the cross-sectional step of the identification proof. Hence, the conditional choice probability $\mathbb{P}(c_t|S_t, X_t)$ is identified since:

$$\mathbb{P}(c_t|S_t, X_t) = \frac{\mathbb{P}(S_t|X_t, c_t)\mathbb{P}(c_t|X_t)}{\mathbb{P}(S_t|X_t)}$$

The initial type distribution $\mathbb{P}(S_{i0}|X_{i0})$ is identified again from applying the cross-sectional step conditional on X_{i0} for the cross-section at $t = 0$.

Furthermore, one can show how to achieve identification of the law of motion $\mathbb{P}(S_{i,t+1}|S_{it}, X_t, c_t)$. Let $K_t^{X_t,c_t}$ be given by:

$$(K_t^{X_t,c_t})_{k,j} = \mathbb{P}(S_{t+1} = j|S_t = k, X_t, c_t)$$

Moreover, let

$$\pi_t^{X_t,c_t}(j) = \mathbb{P}(S_t = j|X_t, c_t)$$

and let $\Pi_t^{X_t,c_t} = diag(\pi_t^{X_t,c_t})$. Moreover, let

$$\pi_{t+1}^{X_t,c_t}(j) = \mathbb{P}(S_{t+1} = j|X_t, c_t)$$

and let $\Pi_{t+1}^{X_t,c_t} = diag(\pi_{t+1}^{X_t,c_t})$, and note that this object is also identified (we can apply the cross-sectional step at $t+1$ conditional on the observables X_t, c_t). Then, applying the longitudinal step of the identification proof in section 3 we can write $K_t^{X_t,c_t}$ as:

$$K_t^{X_t,c_t} = ((\Omega_t^{X_t,c_t})'\Omega_t^{X_t,c_t})^{-1}(\Omega_t^{X_t,c_t})'\mathbb{P}(Y_t^1, Y_{t+1}^1|X_t, c_t)(\Omega_{t+1}^{X_t,c_t})'((\Omega_{t+1}^{X_t,c_t})'\Omega_{t+1}^{X_t,c_t})^{-1}$$

where

$$\Omega_t^{X_t,c_t} = P_1\Pi_t^{X_t,c_t}$$

and

$$\Omega_t^{X_t,c_t} = (\Pi_{t+1}^{X_t,c_t})^{-1}P_1\Pi_t^{X_t,c_t}$$

We can also identify the law of motion of the observed state $\mathbb{P}(X_{t+1}|c_t, X_t, S_{t+1})$. Note that applying the cross-sectional step at $t+1$ conditional on X_{t+1}, X_t, c_t we get $\mathbb{P}(S_{t+1}|X_{t+1}, X_t, c_t)$. As argued before, $\mathbb{P}(S_{t+1}|X_t, c_t)$ is identified. Hence, we can identify:

$$\mathbb{P}(X_{t+1}|S_{t+1}, X_t, c_t) = \frac{\mathbb{P}(S_{t+1}|X_{t+1}, X_t, c_t)\mathbb{P}(X_{t+1}|X_t, c_t)}{\mathbb{P}(S_{t+1}|X_t, c_t)}$$

Finally, if the limited feedback assumption in Hwang (2021) and Hu and Shum (2012) does not hold, we may also be interested in identifying $\mathbb{P}(c_{t+1}, X_{t+1}|S_t, S_{t+1}, c_t, X_t)$. Under Assumption 3 this is straightforward to do. To see why, note that we can condition on S_t, S_{t+1}, c_t, X_t since they are observable. From Theorem 1 the we can identify $\mathbb{P}(S_{t+1}|S_t, c_t, X_t, c_{t+1}, X_{t+1})$ (again, this is simply K_t conditional on the corresponding observables). Conditioning on $X_{t+1}, c_{t+1}, c_t, X_t$ and applying a cross-sectional step at t we get $\mathbb{P}(S_t|X_{t+1}, c_{t+1}, X_t, c_t)$. Hence

$$\mathbb{P}(S_{t+1}, S_t, X_{t+1}, c_{t+1}, X_t, c_t) = \mathbb{P}(S_{t+1}|S_t, c_t, X_t, c_{t+1}, X_{t+1})\mathbb{P}(S_t|X_{t+1}, c_{t+1}, X_t, c_t)$$

is identified. Conditioning on X_t, c_t only and applying similar arguments one can show that $\mathbb{P}(S_{t+1}, S_t, X_t, c_t)$ is identified. Hence, $\mathbb{P}(X_{t+1}, c_{t+1}|S_{t+1}, S_t, X_t, c_t)$ is identified as:

$$\mathbb{P}(X_{t+1}, c_{t+1}|S_{t+1}, S_t, X_t, c_t) = \frac{\mathbb{P}(S_{t+1}, S_t, X_{t+1}, c_{t+1}, X_t, c_t)}{\mathbb{P}(S_{t+1}, S_t, X_t, c_t)}$$

Note that establishing that limited feeback is indeed an overidentifying restriction when there are "pure" noisy measures of the state (and the appropiate full-column rank assumptions and full support conditions hold) opens the possibility to testing this assumption.

It is worth noting that Assumption 3 is stronger than needed. In principle, in order to apply the cross-sectional identification steps, one only needs Y_t^1, Y_t^2, Y_t^3 to be independent given X_{t+1}, c_t, X_t, S_t, given X_t, c_t, S_t, and given X_{t-1}, c_{t-1}, S_t. For the initial cross-sectional step (to identify the distribution of S_0 given X_0) one only needs Y_0^1, Y_0^2, Y_0^3 to be idenpendent given S_0, X_0. And to apply the longitudinal step of the identification results in section 3 one only needs Assumption 1 to hold conditional on X_t, c_t for every X_t, c_t and for every $t = 1, ...T-1$. Presenting the identification results under a more restrictive assumption is done for the sake of clarity.

Finally, it is worth noting that even if no noisy measure of the unobserved state is observed twice (see Appendix B), we can still identify all the reduced-form probabilities of interest, without assuming stationarity of the unobserved state or without placing further restrictions on $\mathbb{P}(X_{t+1}|X_t, c_t, S_t)$. This strengthens the results in Hwang (2021).

5.1.2 Estimation

The logic behind this constructive identification proof can be applied to estimation of CCP's, to the initial distribution of the unobserved state given the observed state and to the laws of motion of the unobserved state and the observed state. More concretely, the two-step estimator presented before can be used to estimate those objects at a low computational cost.

For estimation, I am going to assume that the variables contained in the observable state vector X_{it} are discrete. As before, $\hat{\mathbb{P}}()$ is going to denote a frequency estimator.
Estimating the conditional distributions of the noisy measures $\{P^m\}_{m=1}^3$ Estimates of the P^m's, call them $\hat{P}^m$ can be obtained by performing the first step of the cross-sectional step of the two-step estimator presented in section 4.

Estimating the initial distribution of the unobservable state $\mathbb{P}(S_0|X_0)$

Restrict the sample at $t = 0$ to observations with observable state X_0. Perform a cross-sectional estimation step, using Maximum Likelihood and the EM algorithm or the approximate joint-diagonalization estimator of Bonhomme et al. (2016). The resulting estimated probability distribution for the unobservable state is $\sqrt{n}$ consistent and asymptotically normal estimator for $\mathbb{P}(S_0|X_0)$

Estimating CCP's $\mathbb{P}(c_t|S_t, X_t)$

Restrict the cross-sectional sample at t to observations with observable choices and states c_t, X_t. Perform a cross-sectional estimation step as discussed above. The associated estimator of the probability distribution of the unobservable state gives

an estimator for $\mathbb{P}(S_t|c_t, X_t)$, call it $\tilde{\mathbb{P}}(S_t|c_t, X_t)$. Do the same, now restricting attention to observations with observable state X_t (that is, do not use the observable choice to restrict the sample). The associated estimate for the distribution of the unobservable state yields an estimator for $\mathbb{P}(S_t|X_t)$, call it $\tilde{\mathbb{P}}(S_t|X_t)$. Finally, calculate the frequency estimator $\hat{\mathbb{P}}(c_t|X_t)$. The estimator for the CCP's is given by:

$$\tilde{\mathbb{P}}(c_t|X_t, S_t) = \frac{\tilde{\mathbb{P}}(S_t|X_t, c_t)\hat{\mathbb{P}}(c_t|X_t)}{\tilde{\mathbb{P}}(S_t|X_t)}$$

Estimating $\mathbb{P}(S_{t+1}|S_t, X_t, c_t)$

Restrict the cross-sectional samples at t and $t+1$ to observations with observable state vector X_t and observable choice c_t. Calculate $\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1|X_t, c_t)$, the joint sample frequency of Y_t^1, Y_{t+1}^1 in this sub-sample. Perform a cross-sectional step in these two sub-samples and get the estimate cross-sectional distributions of the unobserved state $\hat{\pi}_t^{X_t,c_t}$ and $\hat{\pi}_{t+1}^{X_t,c_t}$. Calculate:

$$\hat{K}_t^{X_t,c_t} = ((\hat{\Omega}_t^{X_t,c_t})'\hat{\Omega}_t^{X_t,c_t})^{-1}(\hat{\Omega}_t^{X_t,c_t})'\hat{\mathbb{P}}(Y_t^1, Y_{t+1}^1|X_t, c_t)(\hat{\Omega}_{t+1}^{X_t,c_t})'((\hat{\Omega}_{t+1}^{X_t,c_t})'\hat{\Omega}_{t+1}^{X_t,c_t})^{-1}$$

where

$$\hat{\Omega}_t^{X_t,c_t} = \hat{P}_1\hat{\Pi}_t^{X_t,c_t}$$

and

$$\hat{\Omega}_t^{X_t,c_t} = (\hat{\Pi}_{t+1}^{X_t,c_t})^{-1}\hat{P}_1\hat{\Pi}_t^{X_t,c_t}$$

and where

$$\hat{\Pi}_t^{X_t,c_t} = diag(\hat{\pi}_t^{X_t,c_t})$$

The (i,j)-th element of that matrix contains a $\sqrt{n}$ consistent and asymptotically normal estimator of $\mathbb{P}(S_{i,t+1} = j|S_{i,t} = i, X_t, c_t)$.

Note that $\hat{P}^1$ can be obtained in many ways. For example, it can be obtained by performing a cross-sectional step in the whole sample or by performing a cross-sectional step conditional on observable states or choices.

Estimating the law of motion of the observable state $\mathbb{P}(X_{t+1}|c_t, X_t, S_{t+1})$

Restrict the sample to observations with observable state at $t+1$, observable state at t and observable choice at t given by (X_{t+1}, X_t, c_t). Perform a cross-sectional step at $t+1$ on this subsample to get an estimate of $\mathbb{P}(S_{t+1}|X_{t+1}, X_t, c_t)$, call it $\tilde{\mathbb{P}}(S_{t+1}|X_{t+1}, X_t, c_t)$. Estimate $\mathbb{P}(X_{t+1}|c_t, X_t, S_{t+1})$ as:

$$\tilde{\mathbb{P}}(X_{t+1}|S_{t+1}, X_t, c_t) = \frac{\tilde{\mathbb{P}}(S_{t+1}|X_{t+1}, X_t, c_t)\hat{\mathbb{P}}(X_{t+1}|X_t, c_t)}{\tilde{\mathbb{P}}(S_{t+1}|X_t, c_t)}$$

Note that $\hat{\mathbb{P}}(X_{t+1}|X_t, c_t)$ is just a frequency estimator, and the estimation of $\tilde{\mathbb{P}}(S_{t+1}|X_t, c_t)$ was discussed before.

Estimating $\mathbb{P}(X_{t+1}, c_{t+1}|S_{t+1}, S_t, X_t, c_t)$ **if limited feedback is not assumed to hold.** Restrict the sample to inidviduals with a particular value of $X_{t+1}, c_{t+1}, X_t, c_t$. Applying the cross-sectional estimation step at $t+1$ yields $\tilde{\mathbb{P}}(S_{t+1}|X_{t+1}, c_{t+1}, X_t, c_t)$, an estimator of $\mathbb{P}(S_{t+1}|X_{t+1}, c_{t+1}, X_t, c_t)$. Applying the longitudinal step yields $\tilde{\mathbb{P}}(S_{t+1}|S_t, X_{t+1}, c_{t+1}, X_t, c_t)$,

an estimator of $\mathbb{P}(S_{t+1}|S_t, X_{t+1}, c_{t+1}, X_t, c_t)$. Restricting the sample to individuals with values X_t, c_t and doing the same yields $\tilde{\mathbb{P}}(S_{t+1}|S_t, c_t, X_t)$ and $\tilde{\mathbb{P}}(S_{t+1}|c_t, X_t)$ respectively. The estimator for $\mathbb{P}(X_{t+1}, c_{t+1}|S_{t+1}, X_t, c_t, S_t)$ is given by:

$$\tilde{\mathbb{P}}(X_{t+1}, c_{t+1}|S_{t+1}, X_t, c_t, S_t) = \frac{\tilde{\mathbb{P}}(S_{t+1}|X_{t+1}, c_{t+1}, S_t, c_t, X_t)\tilde{\mathbb{P}}(S_{t+1}|X_{t+1}, c_{t+1}, c_t, X_t)}{\tilde{\mathbb{P}}(S_{t+1}|, S_t, c_t, X_t)\tilde{\mathbb{P}}(S_{t+1}|, c_t, X_t)}$$

5.2 Identification and estimation of a production function for cognitive skills

5.2.1 Setting

Children are observed for $T \geqslant 2$ developmental periods. For each period t, three discrete noisy measures of cognitive ability $\tilde{\theta}_t^1, \tilde{\theta}_t^2, \tilde{\theta}_t^3$ and three discrete noisy measures of investment $\tilde{I}_t^1, \tilde{I}_t^2, \tilde{I}_t^3$ are observed.

These measures of cognition and investment are assumed to be "pure" noisy measures of true investment and cognition respectively. This is formalized in the following two assumptions:

Assumption 4.

$$\tilde{I}_t^m \perp \{\{\tilde{I}_\tau^{m'}\}_{\tau=1}^T\}_{m' \neq m}, \{\{\tilde{\theta}^m\}_{\tau=1}^T\}_{m=1,2,3}, \{\theta\}_{\tau=1}^T, \{\tilde{I}_\tau^m\}_{\tau \neq t} \ |I$$

Assumption 5.

$$\tilde{\theta}_t^m \perp \{\{\tilde{\theta}_\tau^{m'}\}_{\tau=1}^T\}_{m' \neq m}, \{\{\tilde{\theta}^m\}_{\tau=1}^T\}_{m=1,2,3}, \{I\}_{\tau=1}^T, \{\tilde{\theta}_\tau^m\}_{\tau \neq t} \ |\theta$$

That is, conditional on true investment, the noisy measures for investment are independent of everything else, including their own past and future; other present, past and future noisy measures of investment and skills; and the present past and future of true unobserved skills. Noisy measures of skills are also assumed to be pure noisy measures of skills, and hence a similar conditional independence condition apply to them. The support of noisy measure $\tilde{\theta}^m$ has cardinality κ_θ^m. Similarly, the support of noisy measure $\tilde{I}^m$ has cardinality κ_I^m.

Let $(P_I^m)_{i,j} = \mathbb{P}(\tilde{I}^m = i|I = j)$ and $(P_\theta^m)_{i,j} = \mathbb{P}(\tilde{\theta}^m = i|\theta = j)$. Both P_I^m and P_θ^m are assumed to have full column rank.

Both true unobserved cognitive skill θ and true unobserved investment I are assumed to be discrete too, and have support with cardinality r^θ and r^I respectively.

Finally, the relationship between $\tilde{I}^1$ and $\tilde{\theta}^1$ and I and θ respectively is assumed to be monotonic, in a sense clarified by the following two assumptions:

Assumption 6. $\mathbb{P}(\tilde{\theta}^1 = \kappa_\theta^1|\theta = j) > \mathbb{P}(\tilde{\theta}^1 = \kappa_\theta^1|\theta = j') \ \forall \ j' < j$

Assumption 7. $\mathbb{P}(\tilde{I}^1 = \kappa_I^1|I = j) > \mathbb{P}(\tilde{I}^1 = \kappa_I^1|I = j') \ \forall \ j' < j$

Assumption 6 says that the probability of observing the highest possible value of $\tilde{\theta}^1$ is increasing in true cognitive skill θ. Assumption 7 has a similar interpretation.

Note that Assumptions 6 and 7 are unnecesarily strong, given that an assumption like Assumption 2 will still be enough for identification in this context. I explain the results using 6 and 7 for the sake of clarity.

Finally, it is worth noting that not all the noisy measures need to be available at each developmental period for identification. This is shown in Appendix B.

5.2.2 Identification

We are interested in identifying $\mathbb{P}(\theta_{t+1}|\theta_t, I_t)$ for $t = 1, ...T-1$ and $\mathbb{P}(\theta_t, I_t)$ for $t = 1, ...T$ from the probability distribution of $\{\{\tilde{\theta}^m, \tilde{I}^m\}_{m=1,2,3}\}_{t=1}^{T}$. It turns out that one can use the constructive identification proof presented in Section 3 to establish the identification of the objects of interest.

First, note that we can map the pair of unobserved states (θ, I) to a uni-dimensional unobservable state as follows:

$$\{(1,1),(1,2),\dots(r^\theta, r^I)\} \longrightarrow^{G^S} \{1,2,\dots r^\theta \times r^I\}$$

where the mapping G^S is given by:

$$G^S(\theta, I) = (\theta - 1) \times r^I + I$$

Note that G^S is one-to-one and its inverse is given by:

$$(G^S)^{-1}(n) = \left(1 + \max\left\{0, \left\lfloor \frac{n}{r^I} \right\rfloor\right\},\ n - \max\left\{0, \left\lfloor \frac{n}{r^I} \right\rfloor\right\} \times r^I\right)$$

where $\lfloor x \rfloor$ stands for the integer part of x.

Using the mapping G^S we can write the pair of unobservable states (θ, I) as a uni-dimensional state:

$$S_t = G^S(\theta, I)$$

Since we want to have three noisy measures of the unobservable state, we can form pairs of noisy measures as follows:

$$(\tilde{\theta}^1, \tilde{I}^1), (\tilde{\theta}^2, \tilde{I}^2), (\tilde{\theta}^3, \tilde{I}^3)$$

Applying to each of those pairs a mapping similar to the one applied to the true unobservable state we get three noisy measures Y^1, Y^2, Y^3.

More concretely, let G^m be given by:

$$G^m(\tilde{\theta}^m, \tilde{I}^m) = (\tilde{\theta}^m - 1) \times \kappa_I^m + \tilde{I}^m$$

Then Y^m is given by:

$$Y^m = G^m(\tilde{\theta}^m, \tilde{I}^m)$$

Again, note that G^m is one-to-one and its inverse is given by:

$$(G^m)^{-1}(n) = \left(1 + \max\left\{0, \left\lfloor \frac{n}{\kappa^I} \right\rfloor\right\},\ n - \max\left\{0, \left\lfloor \frac{n}{\kappa^I} \right\rfloor\right\} \times \kappa^I\right)$$

Assumptions 4 and 5 ensure that Y^1, Y^2, Y^3 are conditionally independent given the unobservable state S_t. Moreover, they also ensure that assumptions 1 and **??** in Section 3 are satisfied. Let P^m be defined as in Section 3. Since P_I^m and P_θ^m are full-column rank, P^m is full-column rank, given that:

$$P^m = P_\theta^m \otimes P_I^m$$

where $\otimes$ denotes the Kronecker product. Hence, the only thing that remains to check, in order to apply Proposition 1 is that for each t $\pi_t(c) > 0$ for every c, where again $\pi_t(c) > 0 := \mathbb{P}(S_t = c)$. Note that this assumption is not as strong as it may seem, given that the location of the unobserved skills can be thought of as being re-normalized at each age. This escapes the critique in Agostinelli and Wiswall (2016) because the technology of child development (in this case a transition function) is left completely unrestricted. Moreover, the requirement that given a level of skill, each level of investment has positive probability will hold if parents are subject to taste shocks to investment and these have support in all the positive real line.

If this condition holds, Proposition 1 ensures that π_t for $t = 1, ...T$ and $K_t[i,j] := \mathbb{P}(S_{t+1} = j | S_t = i)$ [9] for $t = 1, ..., T-1$ are identified. Note that this identifies the objects of interest since:

$$\mathbb{P}(\theta_t, I_t) = \pi_t(G^S(\theta_t, I_t))$$

and

$$\mathbb{P}(\theta_{t+1} = i'_\theta, I_{t+1} = i'_I | \theta_t = i_\theta, I_t = i_I) = K_t[G^S(i'_\theta, i'_I), G^S(i_\theta, i_I)]$$

$$\mathbb{P}(\theta_{t+1} | \theta_t, I_t) = \sum_{i'_I=1}^{r_I} \mathbb{P}(\theta_{t+1}, I_{t+1} = i'_I | \theta_t, I_t)$$

Finally, note that Assumptions 6 and 7 allow us to re-lable parameters according to P^1 in a way that is consistent with the ordering of θ and I.

As a concluding remark for this section, it is worth noting that the previous identification proof can acomodate more unobservable characteristics, such as non-cognitive ability and parental skills. The idea is that if those are discrete, and we have at least 3 conditionally independent noisy measures for them, all the unobservables can again be combined into one uni-dimensional unobservable state and the previous identification result follows through.

5.2.3 Estimation

Estimation of the technology of cognitive development $\mathbb{P}(\theta_{t+1} | \theta_t, I_t)$ and of $\mathbb{P}(\theta_t, I_t)$ follows essentially the same steps as identification.

Suppose that a researcher has access to a sample of noisy measures of cognitive skills and investment $\{\tilde{\theta}^m, \tilde{I}^m\}_{m=1,2,3}$. Note that if more than three noisy measures are available, then they can be combined into three. Estimation of the parameters of interest proceed as follows:

[9] For the sake of readibility, I am using now $A[i,j]$ to denote the i,j element of matrix A

1. Generate new noisy measures

$$Y^m = G^m(\tilde{\theta}^m, I^m)$$

where G^m has been defined before.

2. Apply the two-step estimator to the sample $\{\{Y_t^m\}_{m=1,2,3}\}_{t=1,...,T}$

3. Calculate estimates for the objects of interest as:

$$\hat{\mathbb{P}}(\theta_t, I_t) = \hat{\pi}_t(G^S(\theta_t, I_t))$$

and

$$\hat{\mathbb{P}}(\theta_{t+1} = i'_\theta, I_{t+1} = i'_I | \theta_t = i_\theta, I_t = i_I) = \hat{K}_t[G^S(i'_\theta, i'_I), G^S(i_\theta, i_I)]$$

$$\hat{\mathbb{P}}(\theta_{t+1} | \theta_t, I_t) = \sum_{i'_I=1}^{r_I} \hat{\mathbb{P}}(\theta_{t+1}, I_{t+1} = i'_I | \theta_t, I_t)$$

When applying the two-step estimator, it is important to re-lable the entries of $\hat{\pi}_t$ and the columns of $\hat{P}^m$ according to Assumptions 6 and 7.

6 Conclusion

This paper provides a novel identification proof for non-stationary hidden Markov models that is applicable when three or more noisy measures of the hidden markovian state are observed. The identification argument presented in the paper breaks the problem of identification into two sub-problems: First, some parameters are identified using existing identification results for finite mixture models with discrete measurements. Then, the transition parameters are identified using information from the first step and longitudinal observations. In particular, my result yields identification of a stationary hidden Markov model with two periods, the same number of periods necessary for identification if the hidden state was observable.

Additionally, this paper proposes a $\sqrt{n}$ consistent and asymptotically normal estimator for non-stationary hidden Markov models. This estimator is essentially a finite-sample analogue of the constructive identification proof. The estimator is easy to implement, faster than Baum-Welch (a popular estimation algorithm to find the MLE for hidden Markov models) , conceptually straightforward and more robust than MLE. Moreover, for large datasets the loss in precision of my estimator with respect to maximum likelihood seems to be negligible, and the gain in computing time considerable. Computational gains are likely to be even more important in a dataset that contains refreshment samples.

Finally, the paper shows how the identification and estimation results can be applied in two different contexts: dynamic discrete choice models and child-development.

When applied to the dynamic discrete choice models, the results in this paper provide a novel proof of identification of CCP's, initial conditions and laws of motion.

Moreover, the estimators presented in this paper allow to estimate these objects in a tractable way.

In the child development context, the identification result establishes non-parametric identification of a production function of children's skills using discrete measurements, when skills and investments are themselves discrete too. The estimators presented in this paper also provide a tractable way of estimating that technology.

A Writing $\mathbb{P}(Y_0^1, Y_1^1)$ in terms of parameters

We want to show that:

$$\mathbb{P}(Y_0^1, Y_1^1) = P^1 \Pi_0 K_0 \Pi_1^{-1} (P^1 \Pi_1)'$$

Remember that $\Pi_t = diag(\pi_t)$. Hence, note that:

$$P^1 \Pi_0 = \left(\mathbb{P}(Y_0^1 | S_0 = 1) \dots \mathbb{P}(Y_0^1 | S_0 = r) \right) \begin{pmatrix} \pi_0(1) & & \\ & \ddots & \\ & & \pi_0(r) \end{pmatrix}$$

$$= \left(\pi_0(1) \mathbb{P}(Y_0^1 | S_0 = 1) \dots \mathbb{P}(Y_0^1 | S_0 = r) \right) = \left(\mathbb{P}(Y_0^1, S_0 = 1) \dots \pi_0(r) \mathbb{P}(Y_0^1, S_0 = r) \right)$$

where $\mathbb{P}(Y_0^1 | S_0 = j)$ is a column vector that represents the distribution of Y_0^1 conditional on $S_0 = j$ and $\mathbb{P}(Y_0^1, S_0 = j)$ is a column vector whose i-th element gives the probability of $Y_0^1 = i$ and $S_0 = j$. From here we get:

$$\left(P^1 \Pi_0 K_0 \right)_{i,j} = \sum_{c=1}^{r} \mathbb{P}(Y_0^1 = i, S_0 = c) \mathbb{P}(S_1 = j | S_0 = c)$$

Hence, under **A1** we get:

$$\left(P^1 \Pi_0 K_0 \right)_{i,j} = \mathbb{P}(Y_0^1 = i, S_1 = j)$$

Multiplying by Π_1^{-1} we get:

$$\left(\frac{\mathbb{P}(Y_0^1, S_1 = 1)}{\pi_1(1)} \dots \frac{\mathbb{P}(Y_0^1, S_1 = r)}{\pi_1(r)} \right) = \left(\mathbb{P}(Y_0^1 | S_1 = 1) \dots \mathbb{P}(Y_0^1 | S_1 = r) \right)$$

Finally we have to multiply by $(P^1 \Pi_1)'$. Note that $P^1 \Pi_1$ is a matrix whose j-th column is the vector with i-th element equal to $\mathbb{P}(Y_1^1 = i, S_1 = j)$ as seen before for the analogous case of period 0. Therefore:

$$\left(P^1 \Pi_0 K_0 \Pi_1^{-1} (P^1 \Pi_1)' \right)_{i,j} = \sum_{c=1}^{r} \mathbb{P}(Y_0^1 = i | S_1 = c) \mathbb{P}(Y_1^1 = j, S_1 = c)$$

Under **A2** this is equivalent to:

$$\left(P^1 \Pi_0 K_0 \Pi_1^{-1} (P^1 \Pi_1)' \right)_{i,j} = \mathbb{P}(Y_0^1 = i, Y_1^1 = j)$$

as we wanted to show.

Note that in order to derive this expression we didn't use that the hidden stater is first order Markov. Rather, we have only use that K_t contains the first order transition probabilities. Hence, my identification result is useful to identify first order transition probabilities even if the hidden state is not first order Markov.

B Identification with no measure repeated twice

The identification result in Theorem 1 was derived under the assumption that the same three measures Y^1, Y^2, Y^3 are available at every period $t = 0, \dots T-1$. It turns out that this requirement is not essential. In fact, identification can be obtained even if no noisy measure is observed twice. Before proving that identification can be in this case, we need again assumptions on the dynamics of the measures given the true underlying state.

Assumption C1. *For each* $t = 0, \dots T-1$ *there are measures* $Y_t^{1,t}$ *and* $Y_{t+1}^{1,t+1}$ *such that:*
i) $\mathbb{P}(S_{t+1}|S_t, Y_t^{1,t}) = \mathbb{P}(S_{t+1}|S_t)$
ii) $\mathbb{P}(Y_t^{1,t}|S_{t+1}, Y_{t+1}^{1,t+1}) = \mathbb{P}(Y_t^{1,t}|S_{t+1})$

Note that measures $Y_t^{1,t}$ and $Y_{t+1}^{1,t+1}$ can be different measures (In particular, they can have different conditional distributions given their contemporaneous unobserved state). The first part of C1 says that once we know the unobserved state at t, we cannot better predict the unobserved state at $t+1$ by knowing Y^{1t} at t. The second part of the assumption says that knowing $Y^{1,t+1}$ at $t+1$ doesn't help us predicting $Y^{1,t}$ at t once we know S_{t+1}. Note that these assumptions generalize 1 to the case in which no measure is observed twice. 3 establishes identification when there are no repeated measures:

Theorem 3. *Suppose that at each* $t = 0 \dots T-1$ *three conditionally independent measures of the state* $Y_t^{1,t}, Y_t^{2,t}, Y_t^{3,t}$ *are observed. Moreover, suppose that for each* t*, the cross-sectional distribution of the state* π_t *has no zero elements and* $P^{m,t}$ *(the conditional distribution of* $Y_t^{m,t}$ *given* S_t*) has full-column rank. Then under C1 the model is identified.*

Proof. Again, from Theorems 1-3 in Bonhomme et al. (2016) π_t and $\{P^{m,t}\}_{m=1,2,3}$ are identified for every t.
From a reasoning similar to the one in Appendix A we get that:

$$\left(P^{1,0}\Pi_0 K_0\right)_{i,j} = \sum_{c=1}^{r} \mathbb{P}(Y_0^{10}, S_0 = c)\mathbb{P}(S_1 = j|S_0 = c)$$

Under Assumption C1 i) this is equivalent to:

$$\left(P^{1,0}\Pi_0 K_0\right)_{i,j} = \mathbb{P}(Y_0^{1,0}, S_1)$$

Post-multiplying by Π_1^{-1} we get:

$$\left(P^{1,0}\Pi_0 K_0 \Pi_1^{-1}\right)_{i,j} = \mathbb{P}(Y_0^{1,0} = i, S_1 = j)$$

for $i = 1, \dots \kappa_{1,0}$ and $j = 1 \dots r$, where $\kappa_{1,0}$ is the cardinality of $Y^{1,0}$. Therefore:

$$\left(P^{1,0}\Pi_0 K_0 \Pi_1^{-1}(P^{1,1}\Pi_1)'\right)_{i,j} = \sum_{c=1}^{r} \mathbb{P}(Y_0^{1,0} = i|S_1 = c)\mathbb{P}(Y_1^{1,1} = j, S_1 = c)$$

Under C1 ii) this is equivalent to:

$$\left(P^{1,0}\Pi_0 K_0 \Pi_1^{-1}(P^{1,1}\Pi_1)'\right)_{i,j} = \mathbb{P}(Y_0^{1,0} = i, Y_1^{1,1} = j)$$

Hence we have that:

$$\mathbb{P}(Y_0^{1,0}, Y_1^{1,1}) = P^{1,0}\Pi_0 K_0 \Pi_1^{-1}(P^{1,1}\Pi_1)'$$

Since π_0, π_1 have non-zero entries and $P^{1,0}$ and $P^{1,1}$ are full-column rank this implies:

$$K_0 = (\Omega_0'\Omega_0)^{-1}\Omega_0'\mathbb{P}(Y_0^{1,0}, Y_1^{1,1})\Omega_1'(\Omega_1'\Omega_1)^{-1}$$

where this time Ω_0 and Ω_1 are defined in the following way:

$$\Omega_0 = P^{1,0}\Pi_0$$

$$\Omega_1 = \Pi_1^{-1}(P^{1,1}\Pi_1)'$$

The proof for K_t with $t = 1, \ldots, T-1$ follows similar steps. □

C Two step vs Baum-Welch in a balanced panel

The data is generated according to the hidden Markov model described in the identification section. I choose $r = 2$ and $T = 3$. Moreover, I let K_1 and K_2 be different so that the data comes from a non-stationary hidden Markov model. More concretely, I let K_1 and K_2 be given by:

$$K_1 = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$$

$$K_2 = \begin{pmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{pmatrix}$$

Moreover, the emission matrices are given by:

$$P_1 = \begin{pmatrix} 0.1 & 0.5 \\ 0.7 & 0.3 \\ 0.2 & 0.2 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 0.2 & 0.4 \\ 0.6 & 0.2 \\ 0.2 & 0.4 \end{pmatrix}$$

$$P_3 = \begin{pmatrix} 0.25 & 0.35 \\ 0.55 & 0.25 \\ 0.2 & 0.4 \end{pmatrix}$$

The initial distribution of the state is given by:

$$\pi_0 = (0.3, 0.7)$$

Since emission matrices are full-column rank and K_t has non-zero columns for $t = 1, 2$, so the conditions for identification are met. All the conditions for identification are met.

I consider three different sample sizes: $N = 500, 5000, 15000$. For each sample size, parameters are estimated using the Baum-Welch algorithm and the two-step estimator respectively. The first step of the two-step estimator is conducted using Maximum Likelihood and the EM algorithm for $t = 1, 2, 3$. Initial guesses for the EM algorithm used in the cross-sectional step and for the Baum-Welch respectively are selected at random. I compare mean absolute errors and computing times for numbers of random initialization equal to: 1,20,50,80.

Mean absolute errors are calculated across $M = 25$ Montecarlo experiments for each N For example, the mean absolute error for $\hat{P}^1$ is calculated as:

$$MAE(P_1) = \max_{i,j} \frac{1}{20} \sum_{m=1}^{25} |\hat{P}_1^m(i,j) - P_1(i,j)|$$

D Montecarlo for two-step estimator with refreshment samples.

The data is generated according to the model described in the identification section with $r = 2$ and $T = 3$. In particular, I let K_1 and K_2 be different, so the data is generated according to a non-stationary process. More precisely, I let K_1 and K_2 be given by:

$$K_1 = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$$

$$K_2 = \begin{pmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{pmatrix}$$

Moreover, the emission matrices are given by:

$$P_1 = \begin{pmatrix} 0.1 & 0.5 \\ 0.7 & 0.3 \\ 0.2 & 0.2 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 0.2 & 0.4 \\ 0.6 & 0.2 \\ 0.2 & 0.4 \end{pmatrix}$$

$$P_3 = \begin{pmatrix} 0.25 & 0.35 \\ 0.55 & 0.25 \\ 0.2 & 0.4 \end{pmatrix}$$

The initial distribution of the state is given by:

$$\pi_0 = (0.3, 0.7)$$

Since emission matrices are full-column rank and K_t has non-zero columns for $t = 1, 2$, so the conditions for identification are met. Since we are interested in the case with refreshment sample, $8N$ individuals are sampled in the first period, $3N$ individuals in the second and N individuals in the third. In order to make the refreshment samples respresentative of the cross-section of individuals at $t = 2$ and $t = 3$ respectively, the individuals of the refreshment sample are generated from $t = 1$, but information prior to the period they are sample in is truncated and recorded as missing.

For the first stage of the two step estimator I use Maximum Likelihood and the EM algorithm to estimate the cross-sectional parameters at $t = 1, 2, 3$. I use equal weights when taking weighted averages of the period-specific estimates of the emission matrices. 10 shows the sup-norm of the Mean Absolute Error for the emission matrices and the initial distribution of the hidden state for $N = 100, 1000$ and 8000. In order to calculate the sup-norm of the Mean Absolute Error for say, P_1, I find the absolute difference between each element of $\hat{P}_1$ and the corresponding element of the true emission matrix P_1. Then, we find the average of this matrices of absolute differences across 20 montecarlo experiments, and I take the maximum of the resulting matrix. In other words, I calculate:

$$MAE(P_1) = \max_{i,j} \frac{1}{20} \sum_{m=1}^{20} |\hat{P}_1^m(i,j) - P_1(i,j)|$$

11 shows the sup-norm of the Mean absolute error for the estimates of $K = [K_1, K_2]$ (that is, the four-dimensional array that contains K_1 and K_2). As we can see in both graphs, the maximum mean absolute error decreases for all parameters as the sample-size gets bigger, which is in line with the consistency result.

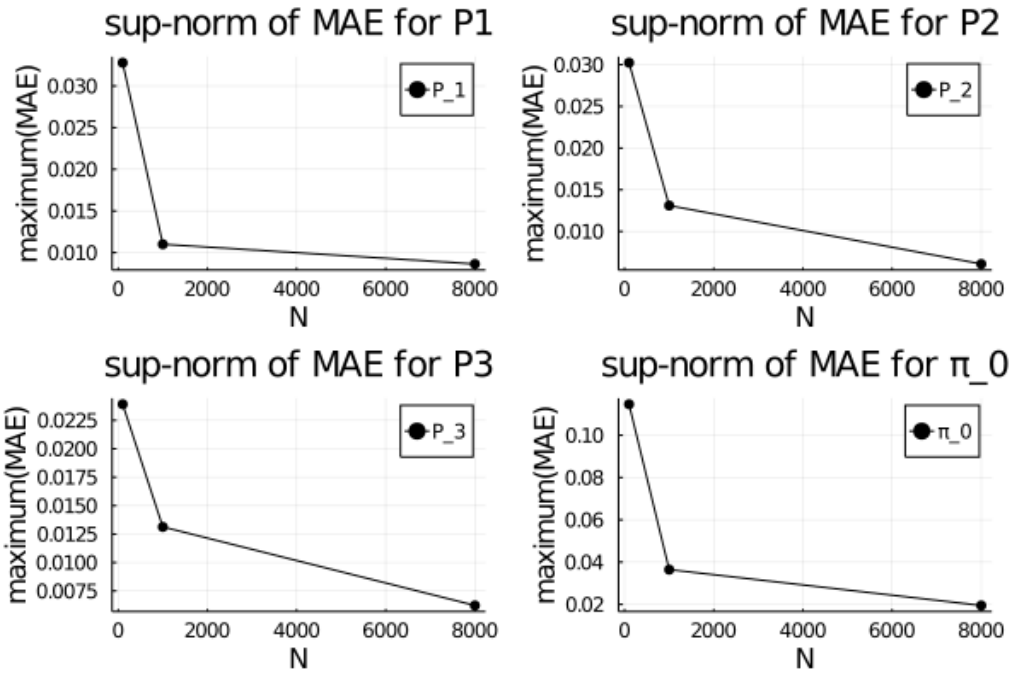

Figure 10: Mean absolute error for different sample sizes

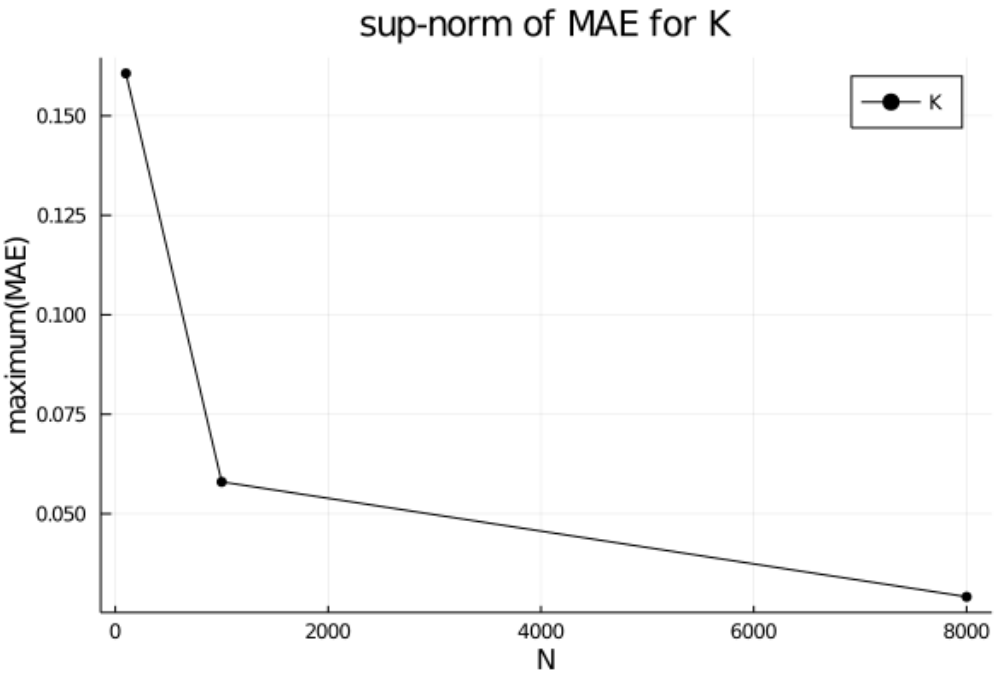

Figure 11: Mean absolute error for different sample sizes

E Baum-Welch algorithm

The Baum-Welch algorithm is mentioned throughout the text, and used as a benchmark for precision and computational cost for the two-step estimator. In order to make the paper more self-contained, I am going to describe the Baum-Welch here.

The Baum-Welch algorithm is a particular application of the popular EM algorithm to hidden Markov models. Hence, it has an expectation and a maximization step. In the expectation step, the filtered probability of being in each hidden state are calculated given a guess for the parameters. This is done efficiently by using an iterative procedure. In the maximization step, new guesses for the parameters are found by maximizing the expected log-likelihood, which can be done in closed form (although, as I explain in the text, this tractability breaks down with refreshment samples and non-stationarities). These two steps are repeated until convergence.

I explain this formally below:

Let

$$y_{it} = (y_{it}^1, y_{it}^2, y_{it}^3)$$

and let $\mathbf{y^T}_i$ be the whole history of y_{it} from 0 to T. Let $\mathbf{y^T} = \{\mathbf{y^T}_i\}_{i=1}^N$

Denote:

$$P(y_{it}, c) := P^1(y_{it}^1, c)P^2(y_{it}^2, c)P^3(y_{it}^3, c)$$

Start the algorithm with a guess of parameters $\theta^{(0)} = (\pi_0^{(0)}, \{(P^m)^0\}_{m=1,2,3}, \{K_t^0\}_{t=1}^{T-1})$

- **Expectation step** At iteration h the guess for parameters is given by $\theta^{(h)}$. Let $\alpha_{ir}(t) = \mathbb{P}(y_{i0}, ...y_{i,t}, S_{it} = r|\theta)$ be the "backward probabilities". These can be calculated as:

$$\alpha_{is}(0) = \pi_0(r)P(y_{0i}, s)$$

$$\alpha_{is}(t+1) = \sum_{c=1}^{m} \alpha_{ic}(t)K_t(c, s)P(y_{i,t+1}, s)$$

Moreover, define $\beta_{is}(t) = \mathbb{P}(y_{i,t+1}, ...y_{i,T}|S_t = s)$ This can be calculated as:

$$\beta_{is}(T) = 1$$

$$\beta_{is}(t) = \sum_{c=1}^{m} K_t(s, c)\beta_{ic}(t+1)P(c, y_{i,t+1})$$

Define

$$\hat{u}_{itj} = \mathbb{P}(S_{it} = j|\mathbf{y^T}_i)$$

$$\hat{v}_{itjk} = \mathbb{P}(S_{it-1} = j, S_{it} = k|\mathbf{y^T}_i)$$

Once we calculated β and α we can use them to calculate the posterior probabilities (the "hat" variables) as follows:

$$\hat{u}_{itj} = \frac{\mathbb{P}(S_{i,t}, \mathbf{y^T})}{\mathbb{P}(\mathbf{y^T})} = \frac{\alpha_{i,j}(t)\beta_{i,j}(t)}{\sum_{j=1}^{r} \alpha_{i,j}(t)\beta_{i,j}(t)}$$

$$\hat{v}_{itjk} = \frac{\mathbb{P}(S_{i,t-1}, S_{it}, \mathbf{y^T})}{\mathbb{P}(\mathbf{y^T})} = \frac{\alpha_{i,j}(t-1)K_{t-1}(j,k)P(y_{i,t}, k)\beta_{i,k}(t)}{\sum_{k=1}^{m}\sum_{j=1}^{r} \alpha_{i,j}(t-1)K(j,k)P(y_{i,t}, k)\beta_{i,k}(t)}$$

- **Maximization step** For a guess of parameters $\theta^{(0)}$, and the complete history of noisy measures $\mathbf{y}^T$, the expected complete log-likelihood is given by:

$$\mathbb{E}_{\theta^{(0)}}[\log L_c(\theta)|\mathbf{y}^T] = \sum_{i=1}^{N}\Big\{\sum_{j=1}^{r}\hat{u}_{i0j}\log\pi_0(j) + \sum_{t=1}^{T}\sum_{j=1}^{r}\sum_{k=1}^{r}\hat{v}_{itjk}\log K(j,k) + \sum_{t=0}^{T}\sum_{j=1}^{r}\sum_{m=1}^{3}\sum_{y=1}^{\kappa_m}\hat{u}_{itj}\mathbb{1}(y_{it}^m = y)\log P^m(y_{i,t},j)\Big\}$$

The new guess can be calculated in closed form as:

$$\pi_0^{(h+1)}(r) = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{i0r}$$

$$K^{(h+1)}(j,k) = \frac{\sum_{i=1}^{N}\sum_{t=0}^{T-1}\hat{v}_{itjk}}{\sum_{j=1}^{r}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\hat{v}_{itjk}}$$

$$(P^m)^{(h+1)}(j,k) = \frac{\sum_{i=1}^{N}\sum_{t=0}^{T}\mathbb{1}(y_{it}^m = j)\hat{u}_{itk}}{\sum_{i=1}^{N}\sum_{t=0}^{T}\hat{u}_{itk}}$$

This new guesses are used to find the filtered probabilities in the next step and the process is repeated until convergence.

References

Agostinelli, F. and Wiswall, M. (2016). Identification of dynamic latent factor models: The implications of re-normalization in a model of child development, *Technical report*, National Bureau of Economic Research.

Aguirregabiria, V. and Mira, P. (2002). Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models, *Econometrica* **70**(4): 1519–1543.

Allman, E. S., Matias, C. and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables, *The Annals of Statistics* **37**(6A): 3099–3132.

Amengual, D., Bueren, J. and Crego, J. A. (2021). Endogenous health groups and heterogeneous dynamics of the elderly, *Technical report*.

Ang, A. and Bekaert, G. (2002a). International asset allocation with regime shifts, *The review of financial studies* **15**(4): 1137–1187.

Ang, A. and Bekaert, G. (2002b). Regime switches in interest rates, *Journal of Business & Economic Statistics* **20**(2): 163–182.

Arcidiacono, P. and Miller, R. A. (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity, *Econometrica* **79**(6): 1823–1867.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains, *The annals of mathematical statistics* **41**(1): 164–171.

Biemer, P. P. and Bushery, J. M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data, *Survey Methodology* **26**(2): 139–152.

Bonhomme, S., Jochmans, K. and Robin, J.-M. (2016). Estimating multivariate latent-structure models, *Annals of Statistics* **44**(2): 540–563.

Bonhomme, S., Jochmans, K. and Robin, J.-M. (2017). Nonparametric estimation of non-exchangeable latent-variable models, *Journal of Econometrics* **201**(2): 237–248.

Cunha, F., Heckman, J. J. and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation, *Econometrica* **78**(3): 883–931.

Feng, S. and Hu, Y. (2013). Misclassification errors and the underestimation of the us unemployment rate, *American Economic Review* **103**(2): 1054–70.

Gassiat, E., Cleynen, A. and Robin, S. (2013). Finite state space non parametric hidden markov models are in general identifiable, *arXiv preprint arXiv:1306.4657* .

Gayle, G.-L., Golan, L. and Soytas, M. (2015). What accounts for the racial gap in time allocation and intergenerational transmission of human capital?

Guidolin, M. and Timmermann, A. (2007). Asset allocation under multivariate regime switching, *Journal of Economic Dynamics and Control* **31**(11): 3503–3544.

Hall, R. E. and Kudlyak, M. (2019). Job-finding and job-losing: A comprehensive model of heterogeneous individual labor-market dynamics, *Technical report*, National Bureau of Economic Research.

Hayashi, F. (2000). *Econometrics*, Princeton University Press, Princeton, New Jersey.

Hotz, V. J. and Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models, *The Review of Economic Studies* **60**(3): 497–529.

Hu, Y. and Shum, M. (2012). Nonparametric identification of dynamic models with unobserved state variables, *Journal of Econometrics* **171**(1): 32–44.

Hwang, Y. (2021). Identification and estimation of a dynamic discrete choice model with an endogenous time-varying unobservable state using proxies, *Working paper* .

Kroft, K., Lange, F. and Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment, *The Quarterly Journal of Economics* **128**(3): 1123–1167.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling, *Psychometrika* **41**(3): 281–293.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear algebra and its applications* **18**(2): 95–138.

Paz, A. (1971). *Introduction to probabilistic automata*, Academic Press, New York.

Petrie, T. (1969). Probabilistic functions of finite state markov chains, *The Annals of Mathematical Statistics* **40**(1): 97–115.

Shibata, I. (2019). Labor market dynamics: A hidden markov approach.

Shimer, R. (2005). The cyclical behavior of equilibrium unemployment and vacancies, *American economic review* **95**(1): 25–49.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, MIT press.