Journal of Econometrics 178 (2014) 805–823

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Dynamic binary outcome models with maximal heterogeneity[☆]

Martin Browning[a], Jesus M. Carro[b,*]

[a] Department of Economics, University of Oxford, United Kingdom

[b] Departamento de Economia, Universidad Carlos III de Madrid, Spain

ARTICLE INFO

*Article history:*
Received 8 January 2011
Received in revised form
15 October 2013
Accepted 12 November 2013
Available online 21 November 2013

*JEL classification:*
C23
C24
J64

*Keywords:*
Discrete choice
Markov processes
Nonparametric identification
Unemployment dynamics

ABSTRACT

Most econometric schemes to allow for heterogeneity in micro behavior have two drawbacks: they do not fit the data and they rule out interesting economic models. In this paper we consider the time homogeneous first order Markov (HFOM) model that allows for maximal heterogeneity. That is, the modeling of the heterogeneity does not impose anything on the data (except the HFOM assumption for each agent) and it allows for any theory model (that gives a HFOM process for an individual observable variable). 'Maximal' means that the joint distribution of initial values and the transition probabilities is unrestricted.

We establish necessary and sufficient conditions for generic local point identification of our heterogeneity structure that are very easy to check, and we show how it depends on the length of the panel.

We apply our techniques to a long panel of Danish workers who are very homogeneous in terms of observables. We show that individual unemployment dynamics are very heterogeneous, even for such a homogeneous group. We also show that the impact of cyclical variables on individual unemployment probabilities differs widely across workers. Some workers have unemployment dynamics that are independent of the cycle whereas others are highly sensitive to macro shocks.

## 1. Introduction

Models with a binary outcome that depends in part on previous realizations of the outcome – dynamic binary outcome models – are common in applied microeconometrics. Some examples include: labor force participation (Heckman, 1981; Hyslop, 1999); smoking (Becker et al., 1994); firms exporting (Bernard and Jensen, 2004); stock market participation (Alessie et al., 2004) and taking up a welfare program (Gottschalk and Moffitt, 1994; Ham and Shore-Sheppard, 2005). The usual time-homogeneous first order Markov model for unit $i$ $(=1, \ldots, N)$ in period $t$ $(t = 0, \ldots, T)$ is:

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right) = F\left(\eta_i + \alpha y_{it-1} + \beta x_{it}\right) \tag{1.1}$$

where $F(.)$ is a probability distribution function and $y_{it}$ is a binary variable indicating, for example, that person $i$ had some unemployment in period $t$. This 'linear index model' which only allows for a heterogeneous 'intercept' $\eta_i$ is widely used but it does have problems; Browning and Carro (2007) discuss these but it is worth repeating the objections.

The first problem is that the imposition of common slope parameters ($\alpha$ and $\beta$) restricts the class of structural models that are consistent with the reduced form (1.1). For example, consider two people, $a$ and $b$, with the same value of the $x$ variables (so we can ignore them), and for whom $a$ has a lower probability of being unemployed if they were employed in the previous year:

$$F(\eta_a) < F(\eta_b). \tag{1.2}$$

For example, $a$ might choose a 'safer' job than $b$. Now suppose we impose the 'same slope' homogeneity assumption $\alpha_a = \alpha_b = \alpha$. This implies:

$$F(\eta_a + \alpha) < F(\eta_b + \alpha). \tag{1.3}$$

This rules out, for example, that $a$'s caution leads her to spend more time looking for a 'safe' job, so that her probability of remaining unemployed is *higher* than $b$'s. Thus the choice of a statistical

[☆] For comments and useful suggestions, we thank three referees, Enrique Sentana, Whitney Newey, Ivan Fernandez-Val, Sara Ayo, and participants at seminars at Boston University; MIT/Harvard; Yale University; Nuffield (Oxford); IFS (London); CEMFI; Manchester; Columbia, CAM (Copenhagen) and a conference at the Tinbergen Institute. The second author gratefully acknowledges that this research was supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme, by grants ECO2012-31358, ECO2009-11165 and SEJ2006-05710 from the Spanish Minister of Education, MCINN (Consolider-Ingenio2010), Consejería de Educación de la Comunidad de Madrid (Excelecon project); he also thanks the Department of Economics at the MIT at which he conducted part of this research as visiting scholar.

[*] Corresponding author. Tel.: +34 91 624 9586.
*E-mail addresses:* Martin.Browning@economics.ox.ac.uk (M. Browning), jcarro@eco.uc3m.es (J.M. Carro).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.11.005

scheme for dealing with heterogeneity has substantive restrictions on the set of admissible structural models.

The second problem with the conventional approach is that whenever we have long enough panels to estimate the model for each unit individually with minimal bias, we do find substantial heterogeneity in both the 'intercept' and 'slope' parameters in (1.1). A situation where this is the case can be found in Browning and Carro (2010). Additional evidence will be provided in the empirical illustration in this paper.

Model (1.1) with maximal heterogeneity has[1]:

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right) = F\left(\eta_i + \alpha_i y_{it-1} + \beta_i x_{it}\right). \tag{1.4}$$

In addition to the homogeneity restrictions, model (1.1) is imposing two kinds of parametric restrictions: the parametric form implied by the linear index and the probability distribution function $F(.)$. In this paper, we consider not only a semiparametric form but also the nonparametric case as well as having maximal heterogeneity throughout the paper.[2] The nonparametric time-homogeneous first order Markov process (HFOM) with maximal heterogeneity allowing that the transition probabilities can be different for each individual can be written:

$$\begin{aligned} &\Pr\left(y_{it} \mid y_{it-1} = y_{-1}, y_{it-2}, \ldots, y_{i0}, x_{it} = x\right) \\ &\quad = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = y_{-1}, x_{it} = x\right) = p_{i,x,y_{-1}} \end{aligned} \tag{1.5}$$

where the first equality for all $t$ is what characterizes a HFOM, and we have one parameter to be estimated for each $i$ and the value of $x$ and the lag of $y$. This does not impose any restrictions on the structural model (except, of course, for the assumption of time invariance and no effects higher than the first order that define the model considered in this paper) and it will fit any data that is generated by a time-homogeneous first order Markov process. For the simpler case without $x$ variables there is a one to one correspondence between (1.4) and (1.5) and, therefore, any $F(.)$ will give the same transition probabilities. For the general case with $x$ variables, a semiparametric form assuming a function $F(.)$ in (1.4) will impose some parametric restrictions that are not imposed in (1.5).

Identifying and estimating the whole set of transition probabilities in (1.5) – the whole set of parameters if we consider (1.4) – or their distribution over the population, allows us to obtain any parameter of interest in this problem, including the average marginal effects (also known as average partial effects, APE) and the median marginal effect of a explanatory variable over the outcome $y_{it}$. Furthermore, identifying and estimating the whole HFOM model will allow to obtain the entire distribution in the population of the effect of a variable over the outcome. In a program evaluation context, Heckman et al. (1997) present situations in which the entire distribution, and not only the mean effect, is the policy parameter of interest. In the IO literature it is also of interest to identify the entire distribution of the individual price elasticities when estimating demand functions; see for example Nevo (2001).

Given the difficulties in estimating (1.1) with small and fixed $T$ (see Arellano and Honoré, 2001), tackling (1.5) or (1.4) is a formidable task. In Browning and Carro (2010) we suggested two estimation methods for the simple case without $x$ variables, that rely on reducing the bias or RMSE for estimates based on each unit. This gives estimates for each unit and then the distribution for $(\eta, \alpha)$ can be taken as the empirical distribution of these estimates (or some smoothed version of it).

In Browning and Carro (2010), identification and estimation of (1.5) without imposing any restriction on the distribution of $(\eta, \alpha)$ nor on the initial condition, relies on the $T$ dimension; that is, it is only consistent when $T \to \infty$. In this paper we propose an alternative approach that relies on large $N$. In general the model is not nonparametrically identified from a cross section of observations of fixed length $T$.[3] This negative result is our starting point in this paper: identification from the cross section is our goal since we typically do not have panels with a very large number of periods. Nevertheless, this negative result on identification does not imply that we cannot learn anything from a cross section of paths with a fixed $T$. In general, some restrictions will have to be imposed on the distribution of the heterogeneity to achieve point identification. The interesting question is the nature of the restrictions we have to impose, or how much information about our model with maximal heterogeneity we can identify from a cross section of length $T$. To answer this question we use finite discrete mixture distributions for the joint set of unknown heterogeneous parameters. We refer to this as the *flexible discrete scheme* since no restriction is imposed other than there is a finite and discrete number of points of support on this distribution.

An advantage of this discrete scheme is that it allows us to go from the homogeneous case (one point of support) to the totally unrestricted case (as many points of support as $N$) within the same scheme. Also, given the discrete nature of problem and the finite number of possible observations, it is clear that we cannot nonparametrically identify a continuous distribution. So, the *flexible discrete scheme* is our route to study nonparametric point identification.[4]

The identification issue in this scheme will be: how many points of support can we take for a given $T$? A major gain from looking at models identified from a cross section with fixed $T$ is that there is no incidental parameters problem nor finite sample bias problem from not having a large number of periods.

Kasahara and Shimotsu (2009) take a different approach to a more general problem that includes the model we consider here, as well as other models. One of the examples included in their paper to illustrate their results is model (1.4) without $x$ variables. However, for this case they do not give identification conditions for an arbitrary number of periods. For example, their most important result for this model (Proposition 7 in Kasahara and Shimotsu (2009)) requires $T \geq 8$. Also they give stronger sufficient conditions than the conditions derived in this paper, whereas here we derive sufficient and necessary conditions for identification. Moreover, their conditions are nontrivial to check in actual data, whereas our conditions are simple to check.

A different and interesting analysis is to look at set identification for the cases that are not point identified. In particular to derive bounds in the non-identified situation when no restriction or distribution is assumed for the heterogeneous parameters.

[1] Model (1.4) can be seen as part of the larger literature on random coefficients model. In that literature there are some cases whose identification and estimation has been studied. An example is Gautier and Kitamura (2013) that considers the estimation of random coefficient static models with continuous covariates. Also, in contrast with us, they assume that the distribution of the unobserved heterogeneous $\beta$ coefficients is independent of the covariates.

[2] Notice also that in (1.1) an extra homogeneity assumption is imposed by assuming all $i$ have the same $F(.)$. In our nonparametric approach this homogeneity assumption is not imposed either.

[3] In general, not even the restrictive model (1.1) with only one fixed effect is identified; see Honorè and Tamer (2006).

[4] We note that our use of a discrete distribution to capture heterogeneity is different to that suggested by Heckman and Singer (1984). They show that the distribution of a continuous latent variable is nonparametrically identified for a particular parametric duration model. They then suggest that the continuous distribution can be reasonably approximated by a discrete distribution with a small number of support points. In contrast, in our scheme the continuous distribution is *not* nonparametrically identified, and any continuous distribution can be perfectly approximated by discrete finite mixtures (see Lemma A.1 in Ghosal and van der Vaart (2001)).

Chernozhukov et al. (2009) do this for the average marginal effect in models such as the ones considered here; they derive results showing that bounds can shrink and converge as $T$ grows.

In Sections 2–4 we study in detail the simpler dynamic HFOM model without $x$ covariates. Studying the model without $x$ covariates helps understanding the problem, and all the results derived for this case will be the base to the more interesting case with covariates that is taken up in Section 5. Sections 2 and 3 consider restrictions from the model and identification respectively. In Section 4 we consider estimation. In Section 6 we apply the techniques we develop to a panel of Danish workers who are very homogeneous in terms of observables. Section 7 concludes.

The principal contributions of paper are:

- We provide necessary nonparametric conditions for any panel data set with binary outcomes to be consistent with a time-homogeneous first order Markov (HFOM) process.
- Assuming the data has been generated by a HFOM process (both with and without covariates), we study identification for flexible discrete distributions of the unobserved heterogeneity. It is shown that we can have a much richer distribution than the two point distribution often used in applied work and still keep unrestricted important features of the distribution of the heterogeneity such as the initial condition or the correlation between the transition probabilities. Our main result provide necessary and sufficient conditions for generic local point identification.
- We give exact results on how identification depends on the length of the panel and on the covariates.
- We provide a framework that allows that macro variables have different effects for different agents.

## 2. HFOM model restrictions

### 2.1. The research question

We consider first a dynamic discrete choice model with no covariates in order to more easily study and understand the problem. The results derived for this case will be very useful for the case with covariates. The data consist of paths $\{y_{i0}, y_{i1}, \ldots, y_{iT}\}_{i=1,2,\ldots,N}$ where $y_{it}$ is the value of a binary variable for unit $i$. We assume a time-homogeneous first order Markov (HFOM) process for each unit and define transition probabilities (1.5) in this case:

$$G_i = \text{pr}\left(y_{it} = 1 \mid y_{i,t-1} = 0\right) \tag{2.1}$$

$$H_i = \text{pr}\left(y_{it} = 1 \mid y_{i,t-1} = 1\right) \tag{2.2}$$

and the unconditional probability of a unit value for the initial observation:

$$P_i = \text{pr}\left(y_{i0} = 1\right). \tag{2.3}$$

This direct formulation is much more convenient to work with than the usual econometric specification given in (1.4) for two reasons. The first reason is that we do not have to specify any probability distribution function $F(.)$, so we are nonparametric in modeling this HFOM. This reason does not have much consequences in this simpler model because allowing for maximal heterogeneity is enough to fit any data that is generated by a HFOM process when there is no $x$ covariates. There is a one to one correspondence between $(\alpha_i, \eta_i)$ and $(G_i, H_i)$ and, therefore, any $F$ will give the same $(G_i, H_i)$ transition probabilities. However in case with covariates the semiparametric form (1.4) will be imposing two kinds of parametric restrictions: the parametric form implied by the linear index and the probability distribution function $F(.)$.

The second reason for this direct formulation is that parameters of (1.4) do not have any meaning on their own, apart from being different from zero or their sign. In contrast, $(P_i, G_i, H_i)$ are probabilities and have a clear interpretation. Nevertheless the values of the parameters $(P_i, G_i, H_i)$ are not usually of primary interest; rather they can be used to generate any other 'outcomes or parameters of interest'. There are several candidates but the most widely considered for this model without covariates are the *marginal dynamic effects*:

$$\begin{aligned} M_i &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1\right) - \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right) \\ &= H_i - G_i \end{aligned} \tag{2.4}$$

and *the long run proportion of unit values*:

$$\begin{aligned} L_i &= \frac{\Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right)}{\Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0\right) + \Pr\left(y_{it} = 0 \mid y_{i,t-1} = 1\right)} \\ &= \frac{G_i}{1 + G_i - H_i}. \end{aligned} \tag{2.5}$$

Given that these values are heterogeneous in $i$, their distribution over the population or some moments of them are the parameters of interest. An example, though not necessarily the most informative measure, is the *average marginal dynamic effect*:

$$E\left[M_i\right] = \iint \left(H_i - G_i\right) dF_{(G,H)}\left(G_i, H_i\right) \tag{2.6}$$

where $F_{(G,H)}(G_i, H_i)$ is the joint distribution of $G$ and $H$ we want to identify. Another common object of interest is the probability that $y_{it} = 1$ in any given period $t$; this is given by the Chapman–Kolmogorov equations applied to the initial probability and the transition probabilities.[5] There is more than one parameter of interest. Identifying the whole HFOM model will allow to obtain any of them, including the entire distribution of $M_i$ in the population, as explained in Section 1.

Given this, our research question is: given a large-$N$, fixed-$T$ panel, what can we (point) identify about the distribution of $(P, G, H)$ over the population?

### 2.2. Enumerating paths

For the moment we can drop the $i$ subscript. There are $\Gamma = 2^{T+1}$ possible paths. The probability of a path $j$ is given by:

$$\begin{aligned} p_j(P, G, H) = {} & P^{y_0^j}(1-P)^{\left(1-y_0^j\right)} G^{n_{01}^j}(1-G)^{n_{00}^j} \\ & \times H^{n_{11}^j}(1-H)^{n_{10}^j} \end{aligned} \tag{2.7}$$

where $n_{01}^j$ is the number of $0 \rightarrow 1$ transitions for path $j$ and similarly for the other three transitions. We shall often use the $T = 2$ case to illustrate general points; Table 2.1 gives the probabilities for the eight possible paths. In all that follows we shall always order paths using a binary representation for ordering the elements for $t = 0, 2, \ldots, T$. Thus the first path is always 00..00, the second path is always 00..01 and the last path is always 11..11.

### 2.3. The general problem

To consider the restrictions from the model and identification we assume that we are given population values for the probabilities of each of the $\Gamma$ outcomes. Denote the population values by $\pi_j$ for $j = 1, 2, \ldots, \Gamma$. Let $(P, G, H)$ be distributed over $[0, 1]^3$ with an

[5] Making those calculations for our case, we obtain $\Pr(y_{it} = 1) = (H_i - G_i)^t P_i + \sum_{k=0}^{t-1} G_i (H_i - G_i)^k$.

**Table 2.1**
Outcomes for three periods ($T = 2$).

| Case | Path | $n_{00}$ | $n_{01}$ | $n_{10}$ | $n_{11}$ | Probability of case $j$, $p_j$ |
|---|---|---|---|---|---|---|
| 1 | 000 | 2 | 0 | 0 | 0 | $(1-P)(1-G)(1-G)$ |
| 2 | 001 | 1 | 1 | 0 | 0 | $(1-P)(1-G)G$ |
| 3 | 010 | 0 | 1 | 1 | 0 | $(1-P)G(1-H)$ |
| 4 | 011 | 0 | 1 | 0 | 1 | $(1-P)GH$ |
| 5 | 100 | 1 | 0 | 1 | 0 | $P(1-H)(1-G)$ |
| 6 | 101 | 0 | 1 | 1 | 0 | $P(1-H)G$ |
| 7 | 110 | 0 | 0 | 1 | 1 | $PH(1-H)$ |
| 8 | 111 | 0 | 0 | 0 | 2 | $PHH$ |

unknown density $f(P, G, H)$. The population proportions are given by the integral equations:

$$\pi_j = \int_0^1 \int_0^1 \int_0^1 p_j(P, G, H) f(P, G, H)\, dPdGdH, \quad j = 1, 2, \ldots, \Gamma. \tag{2.8}$$

Since the $p_j'$s and the $\pi_j$'s sum to unity, $f(.)$ will be a well defined density:

$$1 = \sum_{j=1}^{\Gamma} \pi_j = \int_0^1 \int_0^1 \int_0^1 \sum_{j=1}^{\Gamma} p_j(P, G, H) f(p, G, H)\, dPdGdH = \int_0^1 \int_0^1 \int_0^1 f(P, G, H)\, dPdGdH. \tag{2.9}$$

The econometric issues are:

1. Given a set of observed $\pi_j$'s for $j = 1, \ldots, 2^{T+1}$, can we find a density function $f(P, G, H)$ such that (2.8) holds?
2. If we can find such a function for a given set of $\pi_j$'s, is it unique?
3. If we can find a unique inverse function, is the inverse mapping a continuous function of the values $\pi_j$?

These are the usual set of conditions for a well posed inverse problem. The first condition asks if the model choice (in this case the form of the $p_j(P, G, H)$ functions due to the HFOM assumption) imposes any restrictions on observables. The second is the classical identification condition: given that the data are consistent with the model, can we recover unique estimates of the unknowns, in this case, the density $f(P, G, H)$. The final condition requires that the estimate of the unknown is 'stable' in the sense that small changes in the distribution of observables lead to small changes in the inferred unknowns. The continuity of the inverse mapping is also useful for estimation since we can recover consistent estimates of the structural form (in this case, $f(.)$) from consistent estimates of the reduced forms (the $\pi_j$'s).

### 2.4. Restrictions

Turning to the first question, we ask whether any observed $\pi_j$'s that sum to unity could be generated by a HFOM process. The answer is clearly negative, since the data might have been generated by, for example, a time-homogeneous second order Markov scheme or a time inhomogeneous first order process (or even more general models). Thus the time-homogeneity first order assumption will usually impose restrictions. The restrictions are a combination of equality restrictions and inequality restrictions. Considering (2.7) and (2.8) we have the following equality restrictions:

**Lemma 2.1.** *Given two paths $j$ and $j'$, if*

$$y_0^j = y_0^{j'}, \quad n_{00}^j = n_{00}^{j'}, \quad n_{01}^j = n_{01}^{j'}, \quad n_{10}^j = n_{10}^{j'}, \quad n_{11}^j = n_{11}^{j'} \tag{2.10}$$

*then $\pi_j = \pi_{j'}$.*

Thus two population proportions will be equal if they have the initial value and the same number of transitions. For example, for $T = 3$ (that is, four periods of observation) the two paths 0010 and 0100 have the same initial value and the same number of transitions and hence the same probability,

$$\pi_{0010} = \pi_{0100} = \int_0^1 \int_0^1 \int_0^1 ((1-P)(1-G) \times HGf(P, G, H))\, dPdGdH, \quad j = 1, 2, \ldots \Gamma. \tag{2.11}$$

These are necessary conditions. There are further inequality restrictions. Consider, for example, the case of $T = 2$; see Table 2.1. There are no equality restrictions of the kind described in the lemma. However, the restriction that $G \in [0, 1]$ imposes that

$$p_2(P, G, H) = (1-P)(1-G)G \leq 0.25. \tag{2.12}$$

Thus we have:

$$\pi_2 = \int_0^1 \int_0^1 \int_0^1 p_2(P, G, H) f(P, G, H)\, dPdGdH \leq 0.25. \tag{2.13}$$

Moreover, if $\pi_2$ is actually equal to 0.25 then $P = 0$ and $G = 0.5$ which in turn imposes $\pi_1 = 0.25$. Although we are not able to characterize the full set of necessary and sufficient conditions for a given $\pi$ vector to be generated by a HFOM process, we show below how to write the likelihood of models that separate the equality and inequality restrictions. These likelihoods may be used for testing those restrictions.

Using the lemma above we can calculate the number of paths that are the same for any $T$, without considering the distribution $f(.)$. For example, for $T = 6$ we have 128 equations and 84 restrictions (i.e. number of paths that are restricted to be the same). This simply highlights that the first order and time-homogeneity assumptions impose strong restrictions if we have several periods of observations. For small $T$ this calculation can be done by generating all the possible paths and counting with a computer. However, the following proposition gives a simple analytic formula for the number of different paths for any $T$, denoted by $r_T$.

**Proposition 2.2.** *The number of different paths in values of the vector $\pi = (\pi_1, \ldots, \pi_\Gamma)'$ whose $\pi_j$ elements are defined in (2.8) is*

$$r_T = T(T+1) + 2. \tag{2.14}$$

The proof is given in Appendix A.1.

It is convenient to partition paths into groups based on them having the same probabilities. Define groups $k = 1, 2, \ldots, r_T$ with $\pi_j = \pi_{j'}$ implying that $j$ and $j'$ are in the same group. Let $n_k$ denote the number of members of group $k$ and re-write (2.8) as:

$$\pi_k = n_k \int_0^1 \int_0^1 \int_0^1 p_k(P, G, H) f(P, G, H)\, dPdGdH, \quad k = 1, 2, \ldots, r_T. \tag{2.15}$$

We turn now to identification.

## 3. Identification

Suppose the restrictions for the HFOM model developed in the previous section are not rejected. It is clear that with a finite set of path probabilities we cannot nonparametrically identify a continuous density $f(P, G, H)$ from the finite set of equations (2.15). If we had a continuous covariate and allowed that it had a homogeneous marginal effect on the parameters we could potentially identify the continuous distribution.[6] Since we

[6] Subject to support restrictions that allow us to drive any probability to the limits of zero and unity.

**Table 3.1**
Rank of the Jacobian and minimum number of periods.

| $S$ | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 14 | 25 | 50 | 100 | 138 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_T$ | 4 | 8 | 14 | 22 | 22 | 32 | 32 | 58 | 112 | 212 | 422 | 544 |
| $\min T+1$ | 2 | 3 | 4[a] | 5[a] | 5[a] | 6[a] | 6 | 8 | 11[a] | 15[a] | 21[a] | 24 |
| $\min T+1$ in Browning and Carro (2013) | 2 | 4 | 6 | 8 | 10 | 12 | 16 | 28 | 30 | 50 | 200 | 276 |

[a] Over-identified.

are here interested in identification without imposing arbitrary homogeneity schemes, this option is not open to us. Another option could be identification when $T$ is infinity. The idea is that, given that we will be able to consider discrete distributions with larger number of support points the larger the $T$, as $T$ tends to infinity we should be able to consider models with continuously distributed unobserved heterogeneity. This option is not open to us either since we are considering a situation with a finite $T$.

### 3.1. Identification for the flexible discrete scheme

We consider a *discrete finite mixture distribution* for $(P, G, H)$; we refer to this as the *flexible discrete scheme*. We take $S$ distinct points of support $\{(P_1, G_1, H_1), \dots, (P_S, G_S, H_S)\}$ with probabilities given by the $(S \times 1)$ vector $\theta$ with non-negative individual values, $\theta_s$, that sum to unity. The discrete analogue to (2.8) is:

$$\pi_j = \sum_{s=1}^{S} p_j (P_s, G_s, H_s)\, \theta_s \quad j = 1, 2, \dots, \Gamma. \tag{3.1}$$

Define the $(\Gamma \times S)$ matrix $A$ by:

$$A_{js} = p_j (P_s, G_s, H_s), \quad j = 1, 2, \dots, 2^{T+1},\ s = 1, 2, \dots, S \tag{3.2}$$

so that (2.15) can be written in matrix form as:

$$\pi = \mathbf{A}\theta. \tag{3.3}$$

We take the support points and the probabilities to be unknown so that we have to solve for the values of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ (the vectors of parameters) and $\theta$. $\mathbf{P} = (P_1, \dots, P_S)'$, $\mathbf{G} = (G_1, \dots, G_S)'$, and $\mathbf{H} = (H_1, \dots, H_S)'$. The identification issue we pose is: how many periods we need to identify a distribution with $S$ points of support?

Certainly not any discrete distribution with finite points of support will be identified from $\pi$. For example, it is easy to see that there are many distributions of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S = 8$ that will give the same proportions with $T = 2$.[7] Therefore we cannot identify the distribution of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S = 8$, from the $\pi$ we observe when $T = 2$. We need more periods to identify it.

From (3.3), for given $S$, we have a mapping from $(4S - 1)$ unobservables to observables given by:

$$\pi\,(\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta_1, \dots, \theta_S) = \mathbf{A}\,(\mathbf{P}, \mathbf{G}, \mathbf{H})\,\theta \tag{3.4}$$

where the $S$-vector $\theta$ is normalized to sum to unity. In Appendix B.1 we show that identification in this system is equivalent to studying identification in the system conditional on the first observation. We denote the Jacobian matrix of the system conditional on the first observation by $\mathbf{J}_r\,(T, S)$. For local point identification we require that the rank of $\mathbf{J}_r\,(T, S)$ is greater than or equal to the number of parameters. In Appendix B we show that, in general (that is, except on a set of measure zero), the rank of $\mathbf{J}_r$ is:

$$\text{rank}\,(\mathbf{J}_r) = \min\,(r_T - 2, 4S - 2)\,. \tag{3.5}$$

The parameters of $S$ support points and their probabilities can only be point identified if the number of parameters is not greater than the rank of $\mathbf{J}_r$; from (3.5), this requires:

$$S \le \frac{r_T}{4} = \frac{T(T+1)+2}{4} = \Upsilon_T. \tag{3.6}$$

Note that the maximum $S$ increases quadratically with $T$. From this condition we can calculate the minimum $T$ needed to point identify a model with $S$ number of points on support of the distribution of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$:

$$\min T = \left\lceil -\frac{1}{2} + \sqrt{4S - \frac{7}{4}} \right\rceil \tag{3.7}$$

where $\lceil x \rceil$ gives the smallest integer greater than or equal to $x$. Table 3.1 presents that minimum number of periods, $\min T + 1$, for some cases.[8] As can be seen from Table 3.1, to identify a relatively rich distribution with 14 different points of support we only need a relatively short panel ($T = 7$). Even a short panel ($T = 4$, for example) is more than we need to identify a distribution with more than the two points commonly used in applied work.

This condition based on the rank of the Jacobian is not only sufficient for local identification but it is also necessary (almost everywhere) because they are regular points.[9] The only non-regular points are those in the set of measure zero at which $\text{rank}\,(\mathbf{J}_r) < \min\,(r_T - 2, 4S - 2)$, because, as shown in Appendix B.5, the rank is constant almost everywhere. All the previous results are summarized in the following proposition that gives necessary and sufficient conditions to locally identify our model almost everywhere:

**Proposition 3.1.** *The joint distribution of* $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ *with S points of support in system* (3.3) *is locally identified almost everywhere if and only if:*

$$T \ge -\frac{1}{2} + \sqrt{4S - \frac{7}{4}}. \tag{3.8}$$

The proof is given in Appendix B.

The condition in this proposition is weaker than the condition derived in Browning and Carro (2013), since we have shown that smaller $T$ is required in general.[10] Furthermore, the number of periods we need to identify the system in general here increases at a square root rate with the number of points of support of the distribution we seek to identify, as opposed to increasing our need of periods linearly (that is, at the same rate) as in Browning and Carro (2013). Numbers in the last row of Table 3.1 illustrate how much stronger are the requirements of the condition in Browning and Carro (2013). The reason for these differences is that Browning

[7] To see this, take two different sets of values of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S = 8$ such that $\mathbf{A}$ is invertible. Then there is a $\theta = \mathbf{A}^{-1}\pi$ for each set that defines two different sets of values of the parameters that imply the same $\pi$ with $T = 2$.

[8] The last row of the Table will be explained shortly.

[9] A point is defined as regular when for all points in a sufficiently small neighborhood of it the Jacobian has the same rank as in the point (see Definition 5.A.1 in Fisher, 1966).

[10] The result in Browning and Carro (2013) states that $T \ge 2S - 1$ is sufficient for identification of the joint distribution of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S$ points of support.

and Carro (2013) do not use moment conditions in which both $G$ and $H$ interact.[11]

Generic identification results such as these are useful in practice because they mean that, if we take random values of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta\}$ the probability of finding a value that is a non-regular point (so that condition (3.8) is not necessary and sufficient for identification) is zero. Nevertheless, we search for and study these non-regular points in Appendix C.

### 3.1.1. Global identification

Even if we are usually interested in global identification, the previous results are still useful because local identification is necessary for global identification. However, it is not sufficient in general and we still may want to obtain conditions that guarantee global identification. The problem is that, as explained in Rothenberg (1971), it is much more difficult to prove, and there are few global identification results.

We have not been able to show global identification, even for the simplest case with $T = 2, S = 2$. In that case, a computer program using symbolic calculus is able to invert (3.3) for specific values of $\pi$. For over $10^6$ simulations we found that all the locally identified points were also globally identified. Given this failure to find a numerical counter-example, whether condition (3.8) in Proposition 3.1 is also a condition for generic global identification remains an open question. Thus the only global identification sufficient conditions in this context are those given in Browning and Carro (2013) that, as explained, require many more periods than those given above.

## 4. Estimation

### 4.1. ML estimator

The identification analysis above suggests the following estimation procedure. First, estimate the proportions for each path and test for the model restrictions. If these are not rejected, then impose the conditions and solve for the unknown parameters using the identification conditions. In practice, it is better and more efficient to combine the two steps in a maximum likelihood analysis. This is particularly the case given we cannot derive analytically the inequality constraints that the HFOM imposes (see the discussion in Section 2.4).

Take the full heterogeneity model with $S = \Upsilon_T$ so that we have a just identified model. From (3.1), the structural model is:

$$\pi_j = \sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s \quad j = 1, 2, \ldots, \Gamma. \tag{4.1}$$

Define an indicator $\delta_{ij} = 1$ if unit $i$ has path $j$ and zero otherwise. For given parameters, the likelihood of a sample $\{y_{i0}, y_{i1}, \ldots, y_{iT}\}_{i=1,2,\ldots,N}$ is:

$$\prod_{i=1}^{N}\prod_{j=1}^{\Gamma}\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right)^{\delta_{ij}}$$
$$= \prod_{j=1}^{\Gamma}\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right)^{n_j} \tag{4.2}$$

where $n_j$ is the number of times a sequence $j$ appears in the sample (i.e., $n_j = \sum_{i=1}^{N}\delta_{ij}$). Denote the sample proportions for path $j$ $c_j = n_j/N$. The log-likelihood function for the mixture model is:

$$\ell_{\text{mix}} = \sum_{i=1}^{N}\sum_{j=1}^{\Gamma}\delta_{ij}\log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right) \tag{4.3}$$
$$= N\sum_{j=1}^{\Gamma} c_j\log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right). \tag{4.4}$$

Note that $N$ is irrelevant for the maximization. With an i.i.d. random sample $c_j \to \pi_j$ as $N \to \infty$. The advantage of using the likelihood framework for estimation is that we know how to use all the information on the sample, its asymptotic properties and how to make inference.

### 4.2. Asymptotic properties

*Consistency*. We show how this estimator satisfies the conditions of the unconditional ML version of Proposition 7.5 in Hayashi (2000), so it is a consistent estimator. Let $\mathbf{Y}_i = (y_{i0}, y_{i1}, \ldots, y_{iT})$ be a realization of a (discrete) random variable with probability density function given by

$$\pi\left(\mathbf{Y}_i; \beta_0\right) = \prod_{j=1}^{\Gamma}\left(\pi_j\right)^{\delta_{ij}}$$

where $\pi_j$ is defined in (4.1), $\beta_0 = \left(\mathbf{P}_0, \mathbf{G}_0, \mathbf{H}_0, \theta_{01}, \ldots, \theta_{0,S-1}\right)$ denotes the true value of the parameters and $\beta_0$ is in the interior of the parameter space $\mathbf{B} = [0, 1]^{4S-1}$ so that the model is correctly specified. Note that this parameter space is a compact subset of $\mathbb{R}^{4S-1}$. $\pi\left(\mathbf{Y}_i; \beta\right)$ is continuous in $\mathbf{B}$ for all $\mathbf{Y}_i$ and it is measurable in $\mathbf{Y}_i$ for all $\beta \in \mathbf{B}$. Let $\{y_{i0}, y_{i1}, \ldots, y_{iT}\}_{i=1,2,\ldots,N}$ be a random sample from that variable (that is, $N$ i.i.d. realizations of that random variable) and let $\widehat{\beta} = \left(\left\{\widehat{P}_s, \widehat{G}_s, \widehat{H}_s\right\}_{s=1}^{S}, \left\{\widehat{\theta}_s\right\}_{s=1}^{S-1}\right)$ be the ML estimator, which maximizes the average log-likelihood:

$$\widehat{\beta} = \arg\max_{\beta\in\mathbf{B}}\frac{1}{N}\sum_{i=1}^{N}\log\pi\left(\mathbf{Y}_i; \beta_0\right)$$
$$= \arg\max_{\beta\in\mathbf{B}}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{\Gamma}\delta_{ij}\log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right).$$

It is trivial to see that $E\left[\sup_{\beta\in\mathbf{B}}\left|\log\pi\left(\mathbf{Y}_i; \beta_0\right)\right|\right] < \infty$ so that the dominance condition is satisfied. Finally, the crucial assumption is the identification assumption: $\Pr\left[\pi\left(\mathbf{Y}_i; \beta\right) \neq \pi\left(\mathbf{Y}_i; \beta_0\right)\right] > 0$ for all $\beta \neq \beta_0$ in $\mathbf{B}$. This assumption is going to be satisfied only for those cases for which we have shown in Proposition 3.1 in Section 3.1 that they are identified. Then, given all these conditions, Proposition 7.5 in Hayashi (2000) implies $\widehat{\beta} \to_p \beta_0$ as $N \to \infty$.

*Asymptotic normality*. For asymptotic normality we employ the unconditional ML version of Proposition 7.9 in Hayashi (2000) to show asymptotic normality of $\widehat{\beta}$. In addition to the conditions for consistency, we require the following assumptions:

(i) $\pi\left(\mathbf{Y}_i; \beta\right)$ is twice continuously differentiable in $\beta$ for all $\mathbf{Y}_i$.

(ii) $\mathrm{E}\left[\mathbf{s}\left(\mathbf{Y}_i; \beta_0\right)\right] = \mathbf{0}$ and $\mathrm{E}\left[\mathbf{s}\left(\mathbf{Y}_i; \beta_0\right)\mathbf{s}\left(\mathbf{Y}_i; \beta_0\right)'\right] = -\mathrm{E}\left[\mathbf{H}\left(\mathbf{Y}_i; \beta\right)\right]$ where

$$\mathbf{s}\left(\mathbf{Y}_i; \beta\right) = \frac{\partial\left(\log\pi\left(\mathbf{Y}_i; \beta\right)\right)}{\partial\beta}$$
$$\mathbf{H}\left(\mathbf{Y}_i; \beta\right) = \frac{\partial^2\left(\log\pi\left(\mathbf{Y}_i; \beta\right)\right)}{\partial\beta\partial\beta'}$$

(iii) for some neighborhood $\mathcal{N}$ of $\beta_0$, $\mathrm{E}\left[\sup_{\beta\in\mathcal{N}}\left\|\mathbf{H}\left(\mathbf{Y}_i; \beta\right)\right\|\right] < \infty$.

(iv) $\mathrm{E}\left[\mathbf{H}\left(\mathbf{Y}_i; \beta_0\right)\right]$ is non-singular.

[11] On the other hand, the condition in Browning and Carro (2013) yields global identification.

These conditions hold generally in the case of ML estimators and our case is not an exception to this. Again the crucial condition to satisfy these assumptions is the identification condition we have studied in previous sections. Then, given all these conditions, Proposition 7.9 in Hayashi (2000) implies $\widehat{\beta}$ is asymptotically normal with Avar $(\widehat{\beta}) = \{-\mathrm{E}[\mathbf{H}(\mathbf{Y}_i; \beta_0)]\}^{-1}$.

This asymptotic normal distribution can be used for testing, along with the standard equivalent tests such as the LR test. However, these asymptotic properties have been derived under the assumption of a correctly specified model, including a correct number of points $S$ in the discrete mixture. Any test statistic of a hypothesis that involves testing for the number of points $S$ will *not* have a standard asymptotic distribution because it will imply testing a parameter on the boundary of the parameter space (for example, testing that $\theta_s = 0$ for a given $s$) and under the null some parameters will not be identified (the $(P_s, G_s, H_s)$ associated with the extra points of support under the alternative).

### 4.3. Computation of the MLE

We compute the ML estimates using standard constrained optimization routines. To restrict probabilities going to the boundary, we require that all probabilities be between 0.01 and 0.99. Aside from the problem of hitting the boundaries, the principal computational issue is to find a global maximum. Alternatives to conventional optimizers, such as EM algorithms (or extensions such as Arcidiacono and Jones, 2003) share the same problem. In this computationally intensive search for a global maxima, having a benchmark value for the log-likelihood is of some value. The benchmark value we take is the likelihood of the unrestricted HFOM model, which is easy to compute. First take the saturated model with as many proportions to be estimated as different paths we can observe; this likelihood is:

$$\ell_{\text{sat}} = \sum_{i=1}^{N} \sum_{j=1}^{\Gamma} \delta_{ij} \log(c_j) = N \sum_{j=1}^{\Gamma} c_j \log(c_j). \tag{4.5}$$

We then impose the HFOM equality restrictions from Section 2.4, using Eq. (2.15). Let $k(j)$ denote the group (running from $k = 1, \ldots, r_T$) to which path $j$ belongs. Then define predicted probabilities for path $j = 1, \ldots, \Gamma$ by:

$$\hat{c}_j = \frac{1}{n_{k(j)}} \sum_{j \in k(j)} c_j. \tag{4.6}$$

That is, we replace the unrestricted proportions for each path by the mean of the group.[12] The likelihood function is then given by:

$$\ell_{\text{res\_sat}} = \sum_{i=1}^{N} \sum_{j=1}^{\Gamma} \delta_{ij} \log(\hat{c}_j) = N \sum_{j=1}^{\Gamma} c_j \log(\hat{c}_j). \tag{4.7}$$

If we take a mixture with the maximal number of components, $\Upsilon_T$ from Eq. (3.6), then it has a log likelihood value that is bounded above by $\ell_{\text{res\_sat}}$. The mixture model will only attain this likelihood value if the observed $\hat{\mathbf{c}}$ vector satisfies the inequality constraints discussed in Section 2.4. Denote the likelihood value of this mixture model by $\ell_{\text{mix}}^{\Upsilon}$. Now consider a model with fewer than the maximum number of points of support: $S < \Upsilon_T$. We have the following ordering for the likelihood function values:

$$\ell_{\text{sat}} \geq \ell_{\text{res\_sat}} \geq \ell_{\text{mix}}^{\Upsilon} \geq \ell_{\text{mix}}^{S}. \tag{4.8}$$

## 5. Allowing for covariates

In the presence of covariates in the model, our estimation is conditional on the covariates, $x_{it}$. The covariates are assumed to be strictly exogenous; that is:

$$\begin{aligned} &\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{i0}, \ldots, x_{it}, \ldots, x_{iT}\right) \\ &\quad = \Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right). \end{aligned} \tag{5.1}$$

We consider directly the conditional probabilities:

$$H_{xi} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = x\right)$$
$$G_{xi} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = x\right)$$

where $H_{xi}$ and $G_{xi}$ are defined for each value $x$ of $x_{it}$, and at the unconditional probability of a unit value for the initial observation:

$$P_{xi} = \Pr\left(y_{i0} = 1 | x_{i0} = x\right). \tag{5.2}$$

The number of periods needed to identify a model with $S$ points of support depends on the variation the covariates add to the data (such as whether the covariates are constant over time or they vary exogenously in both $i$ and $t$), and on the assumptions about the relation between the probability of being of each unobserved type and the covariates. These assumptions can go from independence to arbitrary correlation with some middle ground cases that we will cover too. We begin with the simplest case, a binary covariate that is constant over time, as an introduction to the case with a general $x_{it}$. The special case of covariates that only vary with time (that is, in each $t$ they take a common value for all $i$) is explicitly discussed, including the case with time dummies. A summarizing table with numbers for representative cases can be found at the end of this section.

### 5.1. A time invariant binary covariate

If we only have an $x$ variable that is constant over time and only varies across individuals (for example, year of birth or education), it is straightforward to extend our identification result in Section 3.1. For a binary $x_i$, the time homogeneous first order Markov model is fully characterized by:

$$\begin{aligned} &P_{0i} = \Pr\left(y_{i0} = 1 \mid x_i = 0\right); \qquad P_{1i} = \Pr\left(y_{i0} = 1 \mid x_i = 1\right) \\ &G_{0i} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_i = 0\right); \\ &H_{0i} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_i = 0\right) \\ &G_{1i} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_i = 1\right); \\ &H_{1i} = \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_i = 1\right). \end{aligned} \tag{5.3}$$

As before, we consider a flexible discrete distribution for $(P_{0i}, G_{0i}, H_{0i}, P_{1i}, G_{1i}, H_{1i})$ with $S$ distinct points of support $\{P_{0s}, G_{0s}, H_{0s}, P_{1s}, G_{1s}, H_{1s}\}_{s=1}^{S}$.

*Arbitrary correlation between $\theta$ and $x_i$.* Allowing here for arbitrary correlation between types and covariates, the probabilities of each point of support $s$ are given by the $(S \times 1)$ vector $\theta_x$:

$$\theta_x = \begin{cases} (\theta_{01}, \ldots, \theta_{0S})' & \text{if } x = 0 \\ (\theta_{11}, \ldots, \theta_{1S})' & \text{if } x = 1 \end{cases} \tag{5.4}$$

where each vector sum one and all their elements take positive values.[13]

[12] To illustrate, consider the case $T = 3$. Paths 3 (0010) and 5 (0100) are restricted in the HFOM model to have the same probability and so are paths 12 and 14. Therefore, $\hat{c}_3 = \hat{c}_5 = \frac{c_3+c_5}{2}$; $\hat{c}_{12} = \hat{c}_{14} = \frac{c_{12}+c_{14}}{2}$; $\hat{c}_j = c_j$, for all other $j$.

[13] The analysis and estimation is made conditional on $X$, and therefore we are specifying and obtaining the distribution of the individual parameters conditional on $x$. Nevertheless, the unconditional distribution can be calculated from this conditional distribution and the distribution of $x$, which can be obtained from the data.

**Table 5.1**
Number of independent paths. Discrete covariate.

| $T$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $r_{xit}(T,2)$ | 60 | 184 | 472 | 1 056 | 2 132 | 3 976 | 6 964 |
| $r_{xit}(T,4)$ | 464 | 2 656 | 12 088 | 45 888 | 151 456 | 447 648 | 1 210 032 |
| $r_{xit}(T,6)$ | 1548 | 12 984 | 84 852 | 454 104 | 2 079 840 | | |

Here we simply divide the observations into two groups (one with $x_i = 0$ and the other $x_i = 1$) and do the identification analysis and estimation for each group. Each group contains the same number of parameters to identify and the same moment conditions as the problem without covariates. Therefore the number of periods necessary and sufficient to identify $(P_{0i}, G_{0i}, H_{0i}, P_{1i}, G_{1i}, H_{1i})$ with $S$ points of support almost everywhere is the same as to identify $(P_i, G_i, H_i)$ with $S$ points of support in the case without covariates in Section 3.1. If $x_i$ takes $N_x$ values we stratify the sample based on the value of $x_i$ and everything is the same as with a binary covariate.[14]

*Assuming independence between $\theta$ and $x_i$.* If the probability of each type is assumed to be independent of $x_i$, that is $\theta_x = (\theta_1, \dots, \theta_S)'$ for all values of $x$, the number of parameters is reduced, but not the number of equations. There are $(3SN_x + (S-1))$ parameters instead of $(N_x(4S-1))$. Therefore, to identify $S$ points of support in this case we require:

$$S \leq \frac{N_x\,(r_T - 1) + 1}{3N_x + 1}. \tag{5.5}$$

This is greater than $\frac{r_T}{4}$ which is the condition without covariates or arbitrary correlation here.

### 5.2. Covariates that vary in both i and t

We now consider the case of $x_{it}$ covariates that have positive probability of taking any value of their support at any $i$ and $t$. For each point of support $s$:

$$\begin{aligned} P_{x_0 s} &= \Pr\left(y_{i0} = 1 \mid x_{i0} = x_0, s\right) \\ G_{xs} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = x, s\right) \\ H_{xs} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = x, s\right). \end{aligned} \tag{5.6}$$

It is conceptually simple to extend our model if the additional covariates are discrete. We denote the number of values that a discrete $x_{it}$ can take by $N_x$. The probability of a path $j$ given $X_i = X$ and $(P_{xs}, G_{xs}, H_{xs})$ is, where $X_i \equiv (x_{i0}, \dots, x_{iT})$:

$$\begin{aligned} p_{js|X} = {} & P_{x_{i0}s}^{y_0^j}\left(1 - P_{x_{i0}s}\right)^{\left(1-y_0^j\right)} \\ & \times \prod_x G_{xs}^{n^j_{01|x}}\,(1 - G_{xs})^{n^j_{00|x}}\, H_{xs}^{n^j_{11|x}}\,(1 - H_{xs})^{n^j_{10|x}} \end{aligned} \tag{5.7}$$

where $X_i \equiv (x_{i0}, \dots, x_{iT})$, $X$ is a vector of realization of $X_i$, $\prod_x$ denotes the product over the $N_x$ values $x_{it}$ can take, $n^j_{01|x}$ is the number of $y_{it-1} = 0 \to y_{it} = 1$ transitions given $x_{it} = x$ for path $j$, and so on. If, for example, $N_x = 2$, the number of possible equations in our system is $2^{2(T+1)}$, because we have $2^{T+1}$ possible paths of $\{y_{it}\}_{t=0}^{T+1}$ given each one of the $2^{T+1}$ possible observations of $\{x_{it}\}_{t=0}^{T+1}$. As in other cases, some of those paths will give the same equation. We denote the number of different equations by $r_{xit}(T, N_x)$ whose specific expression is in Eq. (A.2), given and proved in Appendix A. Table 5.1 shows this number for some $T$ and $N_x$. Notice that $r_{xit}(T, N_x)$ grows very fast with $N_x$.

*Assuming independence between $\theta$ and $x_{it}$.* Assume independence between the probability of each type and $x_{it}$ : $\theta_{si} = \Pr(s|x_{i0}, \dots, x_{iT}) = \Pr(s) = \theta_s$. Crawford and Shum (2005) is an example of an analysis in which permanent unobserved heterogeneity is assumed to be independent of the covariates. This case corresponds also with the assumption made in many papers using random coefficients discrete choice models. For each point of support $s$ we have to estimate a $\theta_s$ parameter in addition to the $(P_{xs}, G_{xs}, H_{xs})$ parameters. Thus, the number of parameters to identify is $(3N_x + 1)\,S - 1$. By the same arguments used in the case without covariates, the number of periods needed to identify $S$ points of support almost everywhere is given by the following condition

$$S \leq \frac{r_{xit}(T, N_x) - N_x^{T+1} + 1}{3N_x + 1} \tag{5.8}$$

obtained from comparing the number of parameters with the rank of the Jacobian of the system. Table 5.2 gives in each column the highest value of $S$ for which $T$ in that column is min $T$ for identification of that $S$ (that is, the maximum integer value of $S$ such that (5.8) is satisfied for each $T$).

*Assuming $\theta$ depends on the first observation $x_{i0}$.* If we assume that $\theta_{si}$ depends on the first observation $x_{i0}$ but it is independent of the rest observations of $x_{it}$, then $\theta_{si} = \Pr(s|x_{i0}, \dots, x_{iT}) = \Pr(s|x_{i0})$. This case corresponds with the assumptions made about permanent unobserved heterogeneity in papers such as Keane and Wolpin (1997) and Carro and Mira (2006). With a discrete $x_{i0}$ variable that can take $N_x$ values, if we do not place any parametric restriction on this probability there are $N_x(S-1)$ parameters $\theta_{si}$, plus the $3N_xS$ parameters $(P_{xs}, G_{xs}, H_{xs})$. Therefore, the number of periods needed to identify $S$ points of support almost everywhere is given by the following condition:

$$S \leq \frac{r_{xit}(T, N_x) - N_x^{T+1} + N_x}{4N_x}. \tag{5.9}$$

*Parametric dependence between $\theta$ and $x_{it}$.* Assume that $\theta_{si}$ depends on all the $T+1$ observations of $x_{it}$. Then, $\theta_{si} = \Pr(s|x_{i0}, \dots, x_{iT}) = F_\theta(d_{s0} + \sum_{t=0}^{T} d_{s1t}x_{it})$ where $F_\theta$ is a known cdf. Hyslop (1999) is an example where this is the assumption made about the relation between unobserved heterogeneity and covariates. If we did not place any restriction in the relation between $\theta$ and $x_{it}$, we would be allowing any new $x_{iT+1}$ observation to unrestrictedly affect the probability of $i$ being type $s$ even though the type $s$ is a constant characteristic of $i$. Furthermore, we would be treating differently the same value of $x_{it}$ if it were observed in different periods. This extreme flexibility would break solving the identification problem by having $T \to \infty$, because more periods would imply more (incidental) parameters to be estimated, with the number of parameters growing faster with $T$ than the identifying equations. The parametric restriction we have place through $F\,()$ avoids that problem. With one covariate, the number of $\theta_{si}$ parameters is $(T+2)(S-1)$, which lead to a total of $(3N_x + T + 2)S - (T+2)$ parameters to be identified. Therefore, the number of periods needed to identify $S$ points of support almost everywhere is given by the following condition:

$$S \leq \frac{r_{xit}(T, N_x) - N_x^{T+1} + T + 2}{(3N_x + T + 2)}. \tag{5.10}$$

[14] With a continuous nonparametric distribution, it is known that permanent unobserved heterogeneity cannot be separated from covariates when covariates do not vary over time. However, here the discrete scheme is imposing some restrictions, so, in some cases, it is still possible to achieve point identification if the necessary and sufficient conditions indicated above are satisfied.

**Table 5.2**
Maximum number of points of support for some representative cases.

| $T$ | 2 | 3 | 4 | 5 | 6 | 7 | 23 |
|---|---|---|---|---|---|---|---|
| $\Upsilon_T$: No covariates | 2 | 3 | 5 | 8 | 11 | 14 | **138** |
| Covariate constant over time ($x_{it} = x_i$ for all $t$) | | | | | | | |
| Any $N_x$, free relation with $\theta$ | 2 | 3 | 5 | 8 | 11 | 14 | 138 |
| $N_x = 10$, independence of $\theta$ | 2 | 4 | 6 | 10 | 13 | 18 | 178 |
| $N_x = 10$, semiparametric | 9 | 16 | 26 | 39 | 54 | 71 | 691 |
| Covariates $x_{it} = x_t$ for all $i$ | | | | | | | |
| Time dummies | 1 | 1 | 2 | 2 | 3 | 3 | 11 |
| 2 continuous $x_t$, semiparam. | 1 | 1 | 2 | 4 | 5 | 7 | **69** |
| Covariate that varies in both $i$ and $t$ | | | | | | | |
| $N_x = 2$, independent of $\theta$ | 7 | 24 | 63 | 141 | 286 | 531 | |
| $N_x = 4$, independent of $\theta$ | 30 | 184 | 851 | 3 214 | 10 390 | 29 393 | |
| $N_x = 4$, semiparametric | 40 | 218 | 992 | 3 215 | 9 648 | 25 474 | |
| $N_x = 6$, semiparametric | 133 | 1063 | 6423 | 31 342 | 128 565 | | |

$N_x$ is the number of possible values $x$ can take. Where semiparametrically is not specifically mentioned, a flexible HFOM model with the indicated covariates is being considered.

*Covariates with large support.* If $x_{it}$ is a covariate with large support, like a continuous variable, we can have an arbitrary large number $N_x$ of points and use the previous results. Since (5.8)–(5.10) are increasing with $N_x$, this means we can potentially nonparametrically identify as many points of support as we wish simply by discretizing the continuous covariate in as many points as needed; see Remark 2(iv) in Kasahara and Shimotsu (2009). Two caveats should be made with respect to this result. The first is that there is a numerical limit in the way we can discretize a continuous variable. Each discrete group we create should contain enough observations for estimation. If we discretize too much, we may have groups without any or only one observation. The second caveat is that there is curse of dimensionality problem here. We are trying to describe a higher dimensional distribution; and the same number of points of support are less informative about a higher dimensional distribution. The same caveats may arise with the number of values a discrete covariate takes and with the result on the inclusion of covariates independent of $\theta$.

### 5.3. Semiparametric model

In the previous analysis we have not only allowed for maximal (flexible discrete) heterogeneity across $i$, but also we are not restricting our HFOM model to have a particular functional form. In particular we have not imposed any restriction on the way different values of $x_{it}$ affects $y_{it}$. Nevertheless, if $x_{it}$ is continuous, or a cardinal discrete variable that takes many values, such as year of birth, then the effect of different values of $x$ is usually restricted by a parametric form. The obvious example is a linear index model:

$$P_{si} = F_0(p_{s0} + p_{s1}x_{i0})$$

$$G_{sit} = F(g_{s0} + g_{s1}x_{it})$$

$$H_{sit} = F(h_{s0} + h_{s1}x_{it})$$

$$\theta_{si} = F_\theta\left(d_{s0} + \sum_{l=0}^{L} d_{s1l}x_{il}\right) \tag{5.11}$$

$F_0$, $F$ and $F_\theta$ are known cdf functions, such as the standard normal cdf or the standard logistic function. This is equivalent to the representation

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_{it}\right) = F\left(\eta_i + \alpha_i y_{it-1} + \beta_i x_{it} + \delta_i x_{it} y_{it-1}\right)$$

where $(\eta_i, \alpha_i, \beta_i, \delta_i)$ follow a discrete distribution with $S$ points of support.

Eq. (5.11) allows for dependence between $\theta$ and $x$, and includes the independent case ($L = -1$ and $d_{s1l} = 0$ for $l = 0, \ldots, T$), a case with correlation only with the observation of the initial period ($L = 0$ and $d_{s1l} = 0$ for $l = 1, \ldots, T$) and the case of correlation with all the observations of $x_{it}$ ($L = T$ and $d_{s1l} \neq 0$ for $l = 0, \ldots, T$).

The number of parameters is now $(8 + L)S - (L + 2)$. It does not depend on the number of values $x_{it}$ can take. This reduces the number of parameters to identify with respect to the nonparametric case without altering the number of different moment conditions, $r_{xit}(T, N_x)$. The latter value still depends on $N_x$ and it is given by Eq. (A.2). This implies that the maximum number of points of support for which $T$ periods are required for identification is

$$\frac{r_{xit}(T, N_x) - N_x^{T+1} + L + 2}{8 + L} \tag{5.12}$$

which is greater than (5.10). Assuming independence between $\theta$ and $x$ ($L = -1$), this number is $\frac{r_{xit}(T,N_x)-N_x^{T+1}+1}{7}$, which is also greater than (5.8). These reflect the important gains due to the semiparametric assumption.

### 5.4. Time dummies and common variables

Finally, we consider the situation in which we add a covariate that it is common to all individuals and only varies across periods: $x_{it} = x_t$ for all $i$. For instance, this is the case with aggregate variables being used in a micro study, or with time dummy variables. Since we are studying identification over the $i$ population for a fixed $T$, we are only going to observe a given and fixed realization of $\{x_t\}_{t=1}^T$. This implies we only have the $2^{T+1}$ possible paths given $\{x_t\}_{t=1}^T$ that arises from the possible combinations of $\{y_{it}\}_{t=1}^T$ we can observe over the population of $i$. Then, the number of equations in our system here is the same as in the case without covariates and the rank of the Jacobian also depends on $r_T$. For the same reason, $x_t$ is not going to be an informative variable for the probability of $y_{i0}$, nor for the distribution of the heterogeneous parameters over the $i$ population, that is, $\theta$ is independent of $x_t$ ($\Pr(s|\{x_t\}_{t=1}^T) = \Pr(s) = \theta_s$). However, this covariate increases the number of parameters to be identified. Therefore, in this case, in contrast with the results in previous subsections, more periods are required for identification than in the case without covariates.

A situation often found in practice is the use of time dummies. These variables take deterministic values, and, while treated as separate variables, the only meaningful situation is where one of them takes value one and all the others take value zero. If we add time dummies to the model, we have $K = T$ variables $x_t$ that can take $N_x = 2$ values each, but in a deterministic way. Thus we have $(2 + 2T)S - 1$ parameters: one $G$ and $H$ for each time dummy.

**Table 6.1**
Incidence of unemployment.

| | Number | Proportion |
|---|---|---|
| No unemployment | 936 | 36.4 |
| At most 1 year with unemployment | 1141 | 44.4 |
| At most 2 years with unemployment | 1291 | 50.2 |
| At most 3 years with unemployment | 1435 | 55.8 |
| At most 5 years with unemployment | 1710 | 66.5 |
| At most 10 years with unemployment | 2188 | 85.1 |
| At most 20 years with unemployment | 2519 | 98.0 |
| Unemployment in all years | 16 | 0.6 |
| Total sample size | 2571 | – |

**Table 6.2**
Fit for different numbers of support points.

| $S$ | df | LR stat | $\#\theta$'s $= 0.01$ |
|---|---|---|---|
| 2 | 547 | 1063 | 0 |
| 3 | 543 | 701 | 0 |
| 4 | 539 | 605 | 0 |
| 5 | 535 | 536 | 0 |
| 6 | 531 | 512 | 0 |
| 7 | 527 | 500 | 0 |
| 8 | 523 | 494 | 0 |
| 9 | 519 | 491 | 1 |
| 10 | 515 | 491 | 2 |

**Table 6.3**
Parameter estimates with five support points.

| Group | Probabilities ($\times 100$) | | | | |
|---|---|---|---|---|---|
| | **P**<br>$p\,(y_0 = U)$ | **G**<br>$p\,(U \mid E)$ | **H**<br>$p\,(U \mid U)$ | **M**<br>$H - G$ | $\theta$<br>Proportion |
| 1 | 26.9<br>(4.5) | 0.3<br>(0.1) | 86.8<br>(1.2) | 86.5<br>(1.3) | 33.9<br>(5.3) |
| 2 | 63.9<br>(5.7) | 9.8<br>(1.1) | 68.8<br>(1.6) | 59.0<br>(1.7) | 28.4<br>(3.1) |
| 3 | 0.8<br>(5.5) | 2.7<br>(0.9) | 48.3<br>(6.8) | 45.6<br>(6.8) | 23.4<br>(5.2) |
| 4 | 73.4<br>(4.2) | 35.8<br>(3.1) | 81.6<br>(1.2) | 45.8<br>(2.7) | 7.8<br>(1.3) |
| 5 | 25.1<br>(5.6) | 18.6<br>(1.8) | 34.5<br>(4.1) | 15.9<br>(4.6) | 6.5<br>(1.5) |

Standard errors given in brackets.

Then,

$$S \le \frac{r_T}{2 + 2T}. \tag{5.13}$$

This implies a much larger $T$ to identify a given $S$. For example, we need $T \ge 8$ for the identification of a model with $S = 4$. Similarly, the minimum $T$ to identify with $S = 11$ is $T = 23$.

If, on the other hand, $\mathbf{X}_t$ contains $K$ discrete variables taking many values or continuous variables, then we can use a semiparametric model to capture the effect of $X$. For each point of support $s$:

$$G_{st} = F\left(g_{s0} + \sum_{k=1}^{K} g_{sk} x_{kt}\right)$$
$$H_{st} = F\left(h_{s0} + \sum_{k=1}^{K} h_{sk} x_{kt}\right) \tag{5.14}$$

where $F$ is a known cdf, such as the logistic. In this case the number of parameters is $(2 + 2(K + 1))\,S - 1$, and

$$S \le \frac{r_T}{2 + 2(K + 1)}. \tag{5.15}$$

For example, if $K = 2$ and $S = 9$, then $\min T = 8$; or if $K = 2$ and $S = 69$, then $\min T = 23$. These and values for other cases can be found in Table 5.2.

## 6. An empirical illustration

### 6.1. Sample selection

We consider the incidence of unemployment in a year for workers in Denmark from 1980 to 2003 (so that $T = 23$). We draw a sample of male workers with high school education who were aged 25 at the beginning of 1980 and who are continuously married to the same wife for all 24 years that we follow them. This is thus a *very* homogeneous sample in terms of observables; we do this so that our finding of considerable heterogeneity cannot be attributed to insufficient allowance for observable heterogeneity. In all, we have 2571 such workers.[15] We create a dummy variable $y_{it}$ which is set to unity if worker $i$ has any unemployment in year $t$ (and zero otherwise). Table 6.1 gives some statistics for the sample.

### 6.2. The model without covariates

The indicator variable $y_{it}$ is unity if worker $i$ had a spell of unemployment in year $t$. We begin with the model without covariates. The likelihood function value for the saturated model, $\ell_{\text{sat}}$ (4.5), is $-12{,}252$. The value for the saturated HFOM model, $\ell_{\text{res\_sat}}$, (4.7), is $-17{,}449$. The likelihood ratio statistic, $2\,(\ell_{\text{sat}} - \ell_{\text{res\_sat}})$, is thus 10,395.[16] When estimating the mixture model we restrict the mixing probabilities $\theta_s \ge 0.01$ and we restrict $G_s$, $H_s$ and $P_s$ to be between 0.01 and 0.99 to ensure that we do not assign zero probability to any path. The maximum number of support points we could have for the HFOM model is 138 (see Table 3.1). In practice, we cannot find more than a much smaller number than this; see Table 6.2. For ease of reading, we present all likelihood function values for mixture models in LR terms relative to the value for $\ell_{\text{res\_sat}}$; that is, the LR statistic shown is $2\left(\ell_{\text{res\_sat}} - \ell^{S}_{\text{mix}}\right)$. We also show how many mixing parameters are at the imposed minimum of 0.01. As can be seen, it does not seem to be possible to estimate with more than eight components; that is, $\ell^{10}_{\text{mix}} \simeq \ell^{\Upsilon}_{\text{mix}}$.

To illustrate the mechanics of our method, we take a value of $S = 5$.[17] Table 6.3 presents the estimates. These display a number of features. First, most (but not all) of the probabilities are precisely estimated. Second, the starting values fall into three categories: high (groups 2 and 4); medium (groups 1 and 5) and low (group 3). Third, all groups display positive state dependence ($H_s > G_s$). Finally, the marginal dynamic effects ($M_s = H_s - G_s$) are all significant and vary across groups. Finally we note that the conventional 'one fixed effect' assumption imposes that the correlation between $G$ and $H$ is positive. The (weighted) correlation calculated from our estimates is $-0.35$ (with a standard error of 0.19); so that even the qualitative implication is wrong for the standard model.

The substantive implications of the estimates are best seen graphically. The left panel of Fig. 6.1 graphs the probabilities

[15] Denmark has an administrative panel that follows *all* of the population of about five million from 1980 onwards. Consequently we can select very homogeneous strata without compromising sample size. Indeed, the sample drawn here is, in fact, the population of men who fulfilled the selection criteria.

[16] In an earlier version of this paper we developed a parametric bootstrap test for assessing whether the HFOM hypothesis is rejected and for choosing $S$ if it is not. Since this is controversial (see Feng and McCulloch, 1996) and takes us too far from the main theme of this paper, we do not present results here. In the next section we develop a valid test against an HFOM with covariates.

[17] This choice is partly motivated by statistical criteria for the models without and with covariates and partly for presentational clarity in the figures below.

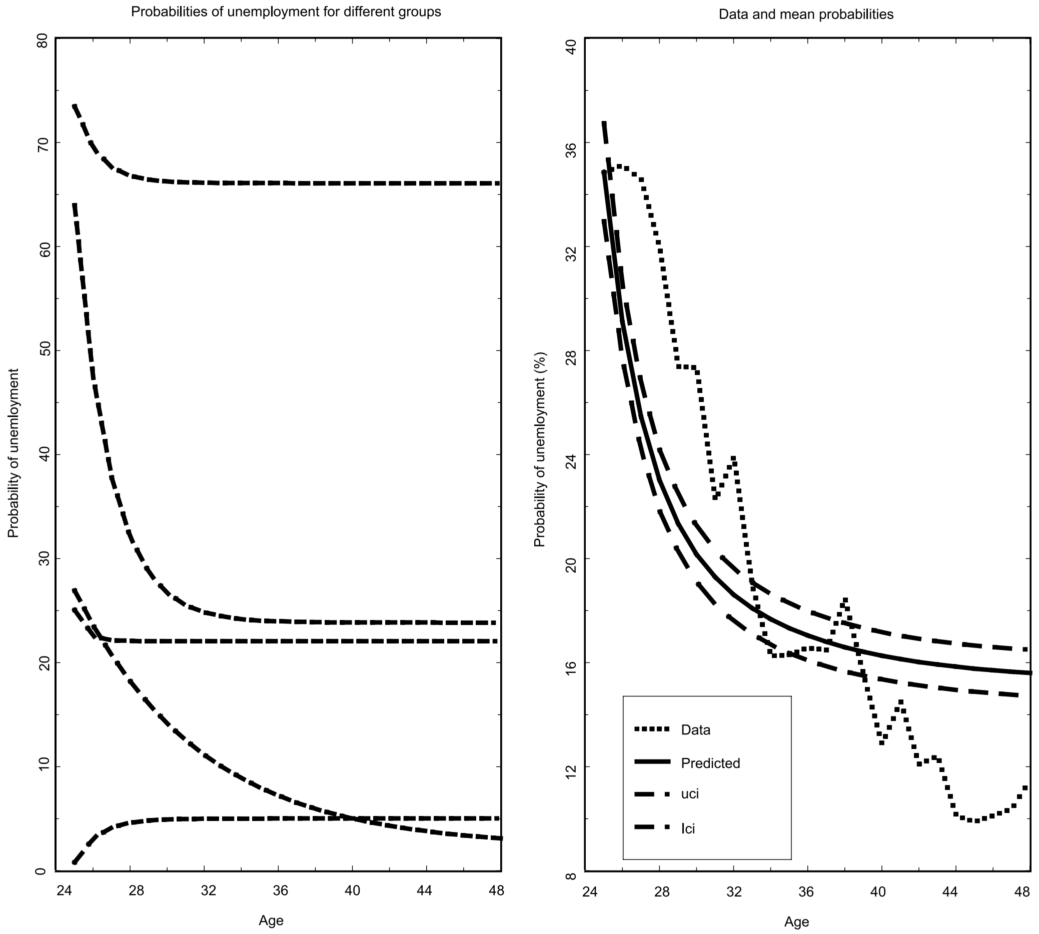

**Fig. 6.1.** Probabilities with 5 points of support.

implied by the Chapman–Kolmogorov equations for the five groups against age (or year, since all the workers in the sample are in the same birth cohort). The groups can be identified from their initial values given in Table 6.3. The largest group start with a relatively high probability of unemployment (26.9%) but have a very low transition probability from employment to unemployment; consequently members of this group have a very low probability of unemployment after age 40. The second largest group display a sharply declining probability but, in comparison to group 1, they start from a much higher probability (63.9%) and only fall to a level of 26% in later life. The third largest group rarely experience unemployment (group 3) and when they do, they have a high probability of being employed the next year. The fourth group are very prone to unemployment. Finally, the smallest group have an almost constant probability of about 0.25.

However, there is evidence that the HFOM model does not fit the data well. This is apparent in the right panel of the figure which shows the data mean proportions of unemployed for each year and the predicted mean from the model (along with confidence intervals). The estimation imposes that the two coincide at age 25 but they are conspicuously different thereafter. A formal test for parameter stability can be constructed by splitting the sample and estimating with dummy shifters for $H_s$ and $G_s$. If we do this with a dummy variable that is unity for the last 11 periods we have an LR statistic of 384; given that we have an extra parameter for each $H_s$ and $G_s$, this has a $\chi^2(10)$ distribution. This formally confirms the time inhomogeneity that we see in the right panel of Fig. 6.1. To capture this time inhomogeneity we turn to estimation adding covariates to the model.

**Table 6.4**
Tests for age and cyclical effects.

| Test against SFOM | | |
|---|---|---|
| Model | df | $\chi^2$ |
| Age and cycle | 20 | 808 |
| Age only | 10 | 766 |
| Cycle only | 10 | 163 |

### 6.3. Model with covariates

The right panel of Fig. 6.1 suggests that we need to allow for time inhomogeneity that is associated with age. There also seem to be cyclical deviations from a smooth age profile. To capture these we include age and the aggregate unemployment rate as covariates and the semiparametric specification in (5.14).[18,19] We first present likelihood ratio statistics for including the extra sets of variables. Since we have 5 points of support and we include regressors in the $G_s$ and $H_s$ transition probabilities, we have 10

[18] Note that aggregate unemployment rate is endogenous by definition, because the endogenous variable in our model is part of this explanatory variable. A solution to this is to construct an aggregate unemployment rate excluding from the population the group we are using. Since our group of workers represents less than 0.0001% of the working population, this will hardly have an impact on the estimates.

[19] Other factors that we could take into account are other macro variables such as changes in the unemployment insurance system; individual time invariant factors such as parental background and individual time varying factors such as health or the presence of children. Note that in this empirical illustration we have taken account of the time invariant cohort factor by taking only one birth cohort.

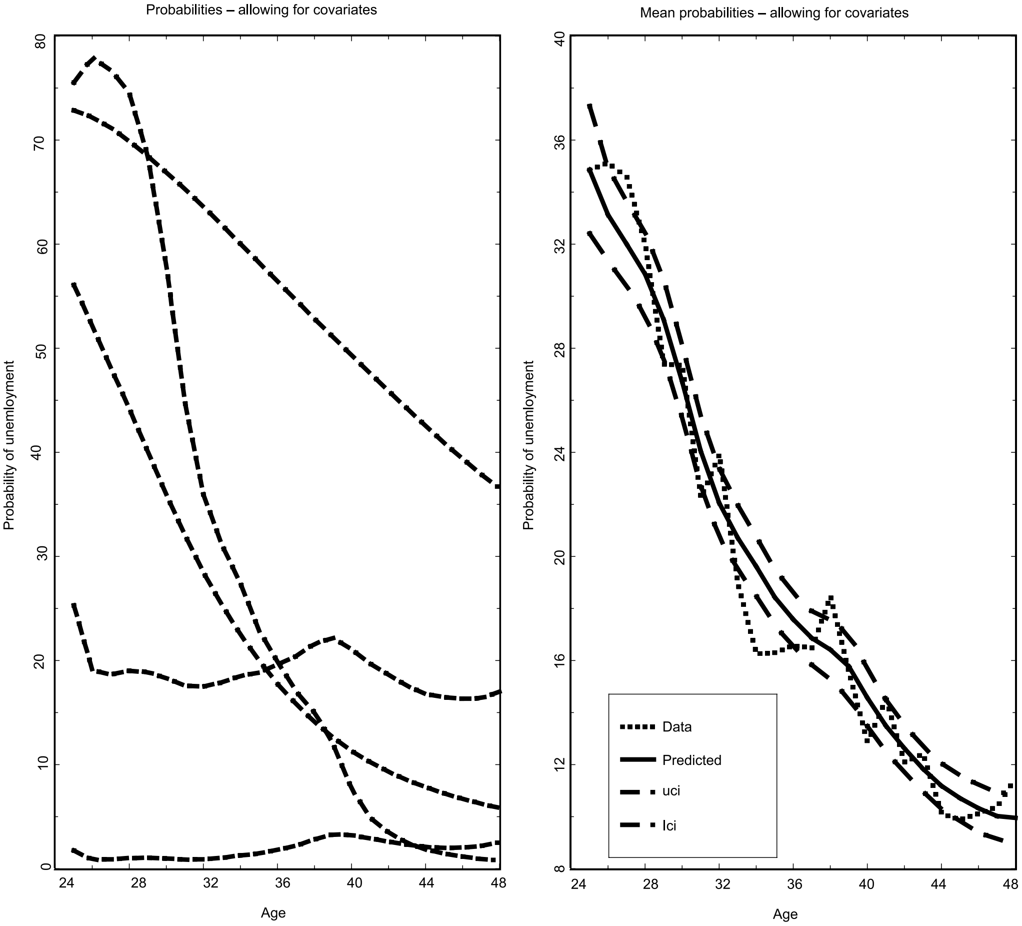

**Fig. 6.2.** Probabilities with age and cyclical effects.

extra parameters for each covariate. Table 6.4 presents the LR statistics against the model with 5 points of support and no covariates. As can be seen, age and the aggregate unemployment rate are individually and jointly highly significant. The $\chi^2(10)$ statistic for the stability test used in the previous subsection is 36; although formally this is a rejection, it is a considerable improvement on the model without age and cyclical effects.

As before, the implications of the estimates are most easily seen in figures of the unemployment sequences. These are given in Fig. 6.2. The right hand panel indicates that adding the age effects remedies most of the misfit seen in the earlier figure. The left hand panel shows that the impact of the business cycle is very heterogeneous. For example, the groups which have very low probabilities are hardly affected at all. However, the next prone group (with a starting value of 0.22) displays considerable cyclical variation. The group which have the highest propensity to be unemployed (the highest curve after age 32) also seem to be unaffected by the cycle. Thus the link between the propensity to be unemployed and the impact of the business cycle is not monotone. Estimates that did not allow for heterogeneous effects of covariates would mask this effect.

### 6.4. *Sample with small T*

Since one of the main advantages of our estimator is that it is a fixed-$T$ consistent estimator, it makes sense to consider a situation where $T$ is relatively small. We use the same $N = 2571$ workers as before but taking only the ages 35–40, giving six waves and $T = 5$. The maximum number of support points we could have for the HFOM model is 8 (see Table 3.1). We include age and the aggregate unemployment rate as covariates and employ the semiparametric specification in (5.14). Table 6.5 presents the likelihood function values for mixture models in LR terms relative to the value for $\ell_{\text{res\_sat}}$.

**Table 6.5**
Fit for different numbers of support points.

| $S$ | df | LR stat | $\#\theta$'s $= 0.01$ |
|---|---|---|---|
| 1 | 56 | 419.7 | 0 |
| 2 | 48 | 79.5 | 0 |
| 3 | 40 | 56.4 | 0 |
| 4 | 32 | 36.4 | 0 |
| 5 | 24 | 24.4 | 0 |
| 6 | 16 | 19.9 | 0 |
| 7 | 8 | 13.5 | 1 |
| 8 | 0 | 12.8 | 2 |

Given these results we take the model with $S = 4$. This is an example for which our identification result is very useful relative to the result in Browning and Carro (2013), since we have more than $(T + 1)/2$ types describing our data. The $\chi^2(16)$ likelihood ratio statistic for including the covariates is 67.2. Once again, it is most convenient to show the implications using a figure; see Fig. 6.3. The initial probabilities for the four groups are 0.01, 0.02, 0.57 and 0.99 with group proportions of 0.03, 0.77, 0.11 and 0.08, respectively. The first group (which starts very low and then goes up dramatically) catches the eye; this small group seems very sensitive to the business cycle. The second group is the largest group and always has a low probability with no apparent responsiveness to the cycle.

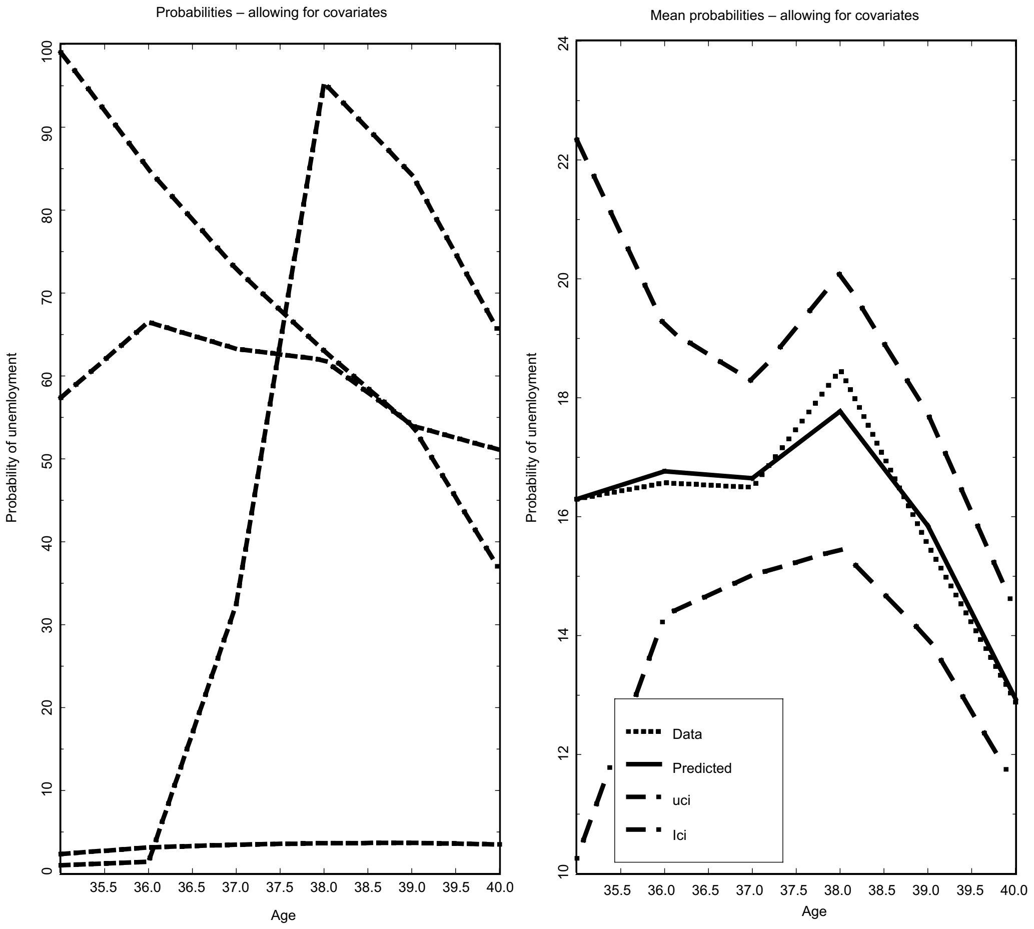

**Fig. 6.3.** Probabilities with six waves.

An additional remark[20] from this small-$T$ illustration is that the Fig. 6.3 does not seem to be a subplot of Fig. 6.2. This is an indication of parameter instability in the data with large $T$ that cannot be captured only by the time trends. Slicing the large $T$ sample in several small-$T$ samples could be a very flexible solution. This provides an additional motivation for having small-$T$ identification results. Turning to the mean fit (the right hand side panel) we see that the mean prediction tracks the data well, albeit with wide confidence intervals, reflecting the loss of precision from using only six waves.

## 7. Conclusions

This paper studies identification from a panel with given $T$ of a non-parametric and a semiparametric dynamic binary choice model with maximal heterogeneity. The more traditional linear-index specification where only the constant term is individual specific is extended since the latter imposes undesired restrictions on the economic model and it does not generally fit the data. In contrast, our model allows variation in all of the parameters (and even the distribution function) across individuals. These models are not generally identified from a cross section of fixed-$T$ periods.

In our specification the joint distribution of the initial observation and the transition probabilities is unrestricted, using flexible discrete mixture distributions. We establish necessary and sufficient conditions for point identification of our heterogeneity structure that are very easy to check and show how it depends on the length of the panel.

A conclusion from this study is that a model with a very flexible distribution of the heterogeneity can be identified from a cross section of $T$ periods, even for $T$ as small as 3. So a model that allows for maximal heterogeneity with a very rich and flexible distribution can be point identified. With such flexibility, important features of the distribution of the heterogeneity such as dependencies of transition probabilities on initial condition are unrestricted.

We show how to estimate using Maximum Likelihood. The asymptotic properties of the estimator in sample sizes with fixed panel length are well known: it is consistent and efficient. We apply the techniques we study to a panel of Danish workers who are very homogeneous in terms of observables. One of our principal findings is that the impact of cyclical variations on unemployment for individual workers are heterogeneous with non-obvious relations. Findings in this application seems to us very illustrative of the potential usefulness of our approach for applied work.

## Appendix A. Proof of the number of different equations

### A.1. Number of 'independent' equations

Here we prove Eq. (2.14), that is, that the number of 'independent' equations in system (2.8) is

$$r_T = T(T+1) + 2.$$

By Lemma 2.1, all we have to do is to count the number of different sets $\left\{y_0^j, n_{00}^j, n_{01}^j, n_{10}^j, n_{11}^j\right\}$ that the $j = 1, \ldots, 2^{T+1}$ possible

[20] For which we thank a referee.

paths can generate. Before counting, note that half of the $r_T$ possible different paths have $y_0 = 0$ and the other half have $y_0 = 1$ and these two halves are symmetric, so we can count only paths with $y_0 = 0$ and multiply its number by two. Notice also that, for $y_0 = 0$ cases, $n_{00} + n_{01} > 0, n_{10} + n_{11} > 0$ only if $n_{01} > 0$, and that $n_{10} \in \{n_{01} - 1, n_{01}\}$. We set $n_{00}$ to count, starting with the maximum value it can take:

- If $n_{00} = T$, then there is only one possibility: $\{(y_0, n_{00}, n_{01}, n_{10}, n_{11})\} = \{(0, T, 0, 0, 0)\}$.
- If $n_{00} = T - 1$, then there is only 1 possibility: $\{(0, T - 1, 1, 0, 0)\}$.
- If $n_{00} = T - 2$, then there are 2 possibilities: $\{(0, T - 2, 1, 1, 0)\,, \,(0, T - 2, 1, 0, 1)\}$.
- If $n_{00} = T - 3$, then there are 3 possibilities: $\{(0, T - 3, 2, 1, 0)\,, (0, T - 3, 1, 1, 1)\,, (0, T - 3, 1, 0, 2)\}$.
- If $n_{00} = T - m$, then there are $m$ possibilities, which are:

$$\left\{\left(0, T - m, \left\lceil \frac{m-q}{2} \right\rceil, \left\lfloor \frac{m-q}{2} \right\rfloor, q\right)\right\}_{q=0}^{m-1} \tag{A.1}$$

where $\lceil x \rceil$ gives the smallest integer greater than or equal to $x$ and $\lfloor x \rfloor$ gives the largest integer less than or equal to $x$.

This goes until $m = T$. Therefore,

$$r_T = 2\left(1 + \sum_{m=1}^{T} m\right) = 2\left(1 + \frac{T\,(T+1)}{2}\right) = T(T+1) + 2$$

where the 1 in $\left(1 + \sum_{m=1}^{T} m\right)$ is accounting for the one case with $m = 0$, that is, $\{(0, T, 0, 0, 0)\}$. Note that for this proof it is not necessary to write all the possible different $\left\{y_0^j, n_{00}^j, n_{01}^j, n_{10}^j, n_{11}^j\right\}$ sets. We only wanted to count them. However, knowing (A.1) is going to be useful for the next proof.

### *A.2. Number of 'independent' equations with covariates: $r_{xit}(T, N_x)$*

Here we prove Eq. (A.2), that is, that the number of different equations in the case with $x_{it}$ covariate that takes $N_x$ values and varies both in $i$ and $t$ is

$$\begin{aligned} r_{xit}(T, N_x) = {} & 2N_x \frac{(T + N_x - 1)!}{T!\,(N_x - 1)!} \\ & + 2N_x \sum_{m=1}^{T} \sum_{q=0}^{m-1} \frac{(T - m + N_x - 1)!}{(T - m)!\,(N_x - 1)!} \\ & \times \frac{\left(\left\lceil \frac{m-q}{2} \right\rceil + N_x - 1\right)!}{\left(\left\lceil \frac{m-q}{2} \right\rceil\right)!\,(N_x - 1)!} \\ & \times \frac{\left(\left\lfloor \frac{m-q}{2} \right\rfloor + N_x - 1\right)!}{\left(\left\lfloor \frac{m-q}{2} \right\rfloor\right)!\,(N_x - 1)!}\,\frac{(q + N_x - 1)!}{q!\,(N_x - 1)!}. \end{aligned} \tag{A.2}$$

It can be seen in (5.7) that now we have to count the number of different sets $\{y_0^j, x_0^j, n_{00|1}^j, \ldots, n_{00|N_x}^j, n_{01|1}^j, \ldots, n_{01|N_x}^j, n_{10|1}^j, \ldots, n_{10|N_x}^j, n_{11|1}^j, \ldots, n_{11|N_x}^j\}$ that the $j = 1, \ldots, 2^{N_x(T+1)}$ possible paths can generate. $n_{01|l}^j$ is the number of $y_{t-1} = 0 \rightarrow y_t = 1$ transitions for path $j$ given $x_{it}$ takes the $l$-th value. Note that $\sum_{l=1}^{N_x} n_{00|l} = n_{00}$, so the number of 00 transitions we have for the $y_t$ are being divided between $n_{00|1}^j, \ldots,$ and $n_{00|N_x}^j$ depending on the value of $x_{it}$ for each particular path. Therefore, we first count the number of ways $n_{00}$ can be arranged into those $N_x$ possible transitions without any other restriction than that (this includes that $n_{00}$ transitions can be arranged in a way that some of the $N_x$ new transition counters are zero). For any given value of $n_{00} = n$ this number is:

$$\frac{(n + N_x - 1)!}{n!\,(N_x - 1)!}. \tag{A.3}$$

(A.3) gives the number for a given set with $n_{00} = n$. We now have to add this for all the possible values of $n_{00}$. The problem and formula (A.3) are the same for $n_{01}$, $n_{10}$, and $n_{11}$. The number of possible sets of $\{y_0, n_{00}, n_{01}, n_{10}, n_{11}\}$ and the sets have being derived in previous appendix. There are $r_T$ possible sets and, from Eq. (A.1), the first half of the $r_T$ sets of $\{y_0, n_{00}, n_{01}, n_{10}, n_{11}\}$ are

$$\left\{ (0, T, 0, 0, 0)\,, \left\{ \left\{ \left(0, T - m, \left\lceil \frac{m-q}{2} \right\rceil, \left\lfloor \frac{m-q}{2} \right\rfloor, q\right) \right\}_{q=0}^{m-1} \right\}_{m=1}^{T} \right\}. \tag{A.4}$$

The other half with $y_0 = 1$ can be obtained similarly, and the total number will be the number for $y_0 = 0$ multiplied by two.

Therefore, combining (A.3) and (A.4) we have that the number $r_{xit}(T, N_x)$ of possible sets of $\{y_0, x_0, n_{00|1}, \ldots, n_{00|N_x}, n_{01|1}, \ldots, n_{01|N_x}, n_{10|1}, \ldots, n_{10|N_x}, n_{11|1}, \ldots, n_{11|N_x}\}$ is given by Eq. (A.2) that has been written again in this appendix. The $N_x$ comes from the number of possible values of $x_0$ that will give other different combinations with everything else being equal.

## Appendix B. Proof of Proposition 3.1: conditions for local identification

Proving the local identification result in Proposition 3.1 is a direct implication of the rank of Jacobian of the system. The first section here shows that studying identification of the distribution of $(P, G, H)$ in (3.3) is equivalent to study identification of $(G, H)$ conditional on the first observation. Then we present several steps that simplify the system and matrices we need to analyze. This is done in order to obtain a tractable form of the Jacobian of the system. Then, we present the result and its proof about the rank of the system conditional on the first observation. Finally, using that result, we prove Proposition 3.1.

### *B.1. Breaking the problem in two: focusing on the process conditional on the first observation*

The system of equations that defines our problem (3.3) can be expressed in terms of that system conditional on the initial observation times the distribution of the initial observation. That is $\pi = \pi_{y_0} * \Pr(y_0)$, where $\pi_{y_0}$ contains the probability of each of the $\Gamma = 2^{T+1}$ paths conditional on the initial observation being $y_0$. The first $\frac{\Gamma}{2}$ rows of $\pi_{y_0}$ are the probabilities of the paths that start at $y_0 = 0$, given that $y_0 = 0$, and the last $\frac{\Gamma}{2}$ rows are the probabilities of the paths that start at $y_0 = 1$ conditional on $y_0 = 1$.[21] The system is, then:

$$\pi_{y_0} = \begin{bmatrix} \pi_{y_0=0} \\ \pi_{y_0=1} \end{bmatrix} = \mathbf{A}_{y_0} \theta_{y_0} = \begin{bmatrix} \mathbf{A}_{y_0=0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{y_0=1} \end{bmatrix} \begin{bmatrix} \theta_{y_0=0} \\ \theta_{y_0=1} \end{bmatrix} \tag{B.1}$$

where $\pi_{y_0=0}$ and $\pi_{y_0=1}$ are vectors of dimension $\frac{\Gamma}{2} \times 1$, $\mathbf{A}_{y_0}$ is a $\Gamma \times 2S$ matrix, $\theta_{y_0}$ is a vector of dimension $2S$, $\mathbf{A}_{y_0=0}$ and $\mathbf{A}_{y_0=0}$ are

[21] Notice that the probability of the first $\frac{\Gamma}{2}$ paths given $y_0 = 1$ is zero because these are the paths that start at $y_0 = 0$. For the same reason the probability of the last $\frac{\Gamma}{2}$ paths given $y_0 = 0$ is zero.

$\frac{\Gamma}{2} \times S$ matrices, $\mathbf{0}$ are $\frac{\Gamma}{2} \times S$ matrices whose elements are all zero, and $\theta_{y_0=1}$ and $\theta_{y_0=0}$ are vectors of dimension $S \times 1$. System (B.1) is simply a compact expression for two separate processes: one for those observation that start with 0 and those that start with 1. The $j$th elements of $\pi_{y_0=0}$ and $\pi_{y_0=1}$ are respectively:

$$\pi_{j|y_0=0} = \frac{\pi_j}{\sum_{k=1}^{2^T} \pi_k}, \quad j = 1, \ldots, 2^T$$

$$\pi_{j-2^T|y_0=1} = \frac{\pi_j}{\sum_{k=2^T+1}^{2^{T+1}} \pi_k}, \quad j = 2^T + 1, \ldots, 2^{T+1}$$

where $\pi_j$ and $\pi_k$ are the elements of $\pi$ in (3.3), that is, the unconditional proportions of each path. The elements of $\theta_{y_0=0}$ and $\theta_{y_0=1}$ give the probability of each type conditional on $y_0 = 0$ and $y_0 = 1$ respectively:

$$\theta_{y_0=0} = \left[\theta_{1|y_0=0}, \ldots, \theta_{S-1|y_0=0}, 1 - \sum_{s=1}^{S-1} \theta_{s|y_0=0}\right]'$$

$$\theta_{y_0=1} = \left[\theta_{1|y_0=1}, \ldots, \theta_{S-1|y_0=1}, 1 - \sum_{s=1}^{S-1} \theta_{s|y_0=1}\right]'$$

where

$$\begin{aligned}\theta_{s|y_0=0} &= \frac{\Pr(s, y_0 = 0)}{\Pr(y_0 = 0)} = \frac{\Pr(y_0 = 0|s) * \Pr(s)}{\sum_{k=1}^{S} \Pr(y_0 = 0|s) * \Pr(k)} \\ &= \frac{(1 - P_s) * \theta_s}{\sum_{k=1}^{S} (1 - P_k) * \theta_k} \quad \text{for } s = 1, \ldots, S - 1 \end{aligned} \tag{B.2}$$

$$\theta_{s|y_0=1} = \frac{P_s * \theta_s}{\sum_{k=1}^{S} P_k * \theta_k}, \quad \text{for } s = 1, \ldots, S - 1. \tag{B.3}$$

The system (B.1) contains all the information we can use to identify the distribution of $(G_s, H_s)$ conditional on the initial observation. Once we have recovered $\theta_{y_0=0}$ and $\theta_{y_0=1}$ from that system, the distribution of $P_s$ and the unconditional probability of each type $(\theta_s)$ can be uniquely recovered from (B.2), (B.3), and the unconditional probability of the initial observation (which is a proportion of ones that we observe):

$$\Pr(y_0 = 1) = \sum_{k=1}^{S} P_k * \theta_k. \tag{B.4}$$

Once we have $\theta_{y_0=0}$ and $\theta_{y_0=1}$ and $\Pr(y_0 = 1)$, (B.2)–(B.4) from a system of $2S - 1$ equations that uniquely identify the $2S - 1$ unknowns $(P_1, P_2, \ldots, P_S, \theta_1, \ldots, \theta_{S-1})$ for possible values of the parameters that are in the open interval (0, 1). Furthermore, these solutions have close forms. Substituting (B.4) in (B.2) and (B.3) implies

$$(1 - \Pr(y_0 = 1))\, \theta_{s|y_0=0} = \theta_s - P_s * \theta_s \tag{B.5}$$

$$P_s * \theta_s = \Pr(y_0 = 1)\, \theta_{s|y_0=1}. \tag{B.6}$$

Then, substituting (B.6) in (B.5), and doing some manipulations we obtain the solution for $\theta_s$

$$\theta_s = \theta_{s|y_0=0} * (1 - \Pr(y_0 = 1)) + \theta_{s|y_0=1} * \Pr(y_0 = 1), \quad \text{for } s = 1, \ldots, S - 1. \tag{B.7}$$

Replacing $\theta_s$ with its solution in (B.6) we obtain the solution

$$P_s = \frac{\theta_{s|y_0=1} * \Pr(y_0 = 1)}{\theta_{s|y_0=0} * (1 - \Pr(y_0 = 1)) + \theta_{s|y_0=1} * \Pr(y_0 = 1)}, \quad \text{for } s = 1, \ldots, S - 1. \tag{B.8}$$

Finally, (B.4) can be written as $\Pr(y_0 = 1) = \sum_{k=1}^{S-1} P_k * \theta_k + P_S * \left(1 - \sum_{k=1}^{S-1} \theta_k\right)$. Substituting (B.7) and (B.8) here we can recover the solution for $P_S$:

$$P_S = \frac{\Pr(y_0 = 1)\left(1 - \sum_{k=1}^{S-1} \theta_{k|y_0=1}\right)}{1 - (1 - \Pr(y_0 = 1)) \sum_{k=1}^{S-1} \theta_{k|y_0=0} - \Pr(y_0 = 1) \sum_{k=1}^{S-1} \theta_{k|y_0=1}}. \tag{B.9}$$

This uniqueness or global invertibility in (B.2), (B.3), and (B.4) means that any non-identification problem is going to be only in (B.1). That is, if we are able to identify the distribution of $(G, H)$ conditional on the first observation, we are also able to identify the unconditional distribution of $(P, G, H)$.

That one-to-one map from $\left(\{\theta_{s|y_0=0}, \theta_{s|y_0=1}\}_{s=1}^{S-1}, \Pr(y_0 = 1)\right)$ to $\left(\{\theta_s\}_{s=1}^{S-1}, \{P_s\}_{s=1}^{S}\right)$, also shows that we can identify different values of $P_s$, that is, an underlying distribution of the heterogeneity in the probability of the initial observation, due to its relation with the distribution of the heterogeneity in $(G, H)$. If they were independent, then $\theta_{s|y_0=0} = \theta_{s|y_0=1}$, and we could not identify different values of $P_s$ but the proportions of ones we observe in the first period. That is, $\theta_{s|y_0=0} = \theta_{s|y_0=1}$ imply in (B.8) and (B.9) that $P_s = \Pr(y_0 = 1)$ for all $s = 1, \ldots, S$.[22] On the other hand we most often only care about the distribution of the initial condition as long as it is correlated with the distribution of the rest of the periods and ignoring it leads to misleading conclusions. In such a situation we are only interested in the distribution of the heterogeneity in $(G, H)$.

### B.2. Decomposition of matrix $\mathbf{A}_{y_0}$

From Eqs. (2.7) and (3.2), without the probability of the initial observation since we have conditioned on it, any element of a row $j$ of matrices $\mathbf{A}_{y_0=0}$ and $\mathbf{A}_{y_0=1}$ is given by $G^{n^j_{01}}(1 - G)^{n^j_{00}} H^{n^j_{11}} (1 - H)^{n^j_{10}}$. From the binomial theorem we have that

$$G^{n^j_{01}}(1 - G)^{n^j_{00}} H^{n^j_{11}} (1 - H)^{n^j_{10}} = \sum_{z=0}^{n^j_{10}} \sum_{x=0}^{n^j_{00}} (-1)^x (-1)^z \binom{n^j_{00}}{x} \binom{n^j_{10}}{z} G^{(x+n^j_{01})} H^{(z+n^j_{11})}. \tag{B.10}$$

Based on this we can decompose matrix $\mathbf{A}_{y_0}$ as the product of two matrices:

$$\mathbf{A}_{y_0} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{bmatrix} \begin{bmatrix} \mathbf{E}_1 & \cdots & \mathbf{E}_S & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{E}_1 & \cdots & \mathbf{E}_S \end{bmatrix} \tag{B.11}$$

where $\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{bmatrix}$ will contain the coefficients $\left((-1)^x(-1)^z \binom{n^j_{00}}{x}\binom{n^j_{10}}{z}\right)$ of (B.10) and $\mathbf{E}$ will contain the corresponding $G, H$ and $P$ terms. The matrix $\mathbf{C}$ does not depend on the value of the parameters and, therefore, it will be unique for a given $T$.

$\mathbf{E}_s$ is the following vector

$$\begin{aligned}\mathbf{E}'_s = [1 \quad G_s \quad .. \quad G_s^T \quad H_s \quad G_s H_s \quad .. \quad G_s^{T-1} H_s \quad H_s^2 \\ .. \quad G_s^{T-2} H_s^2 \quad . \quad . \quad . \quad H_s^{T-1} \quad G_s H_s^{T-1} \quad H_s^T] \end{aligned} \tag{B.12}$$

[22] Another way of seeing this is to notice the following. If the initial condition is independent of the transition probabilities, the observations subsequent to the initial observation contain no information whatsoever about $P$. Therefore, we would have only one observation (the initial observation) to identify the distribution of $P$. With only one observation, only one proportion ($P_s = \Pr(y_0 = 1)$ for all $s = 1, \ldots, S$) can be identified.

of dimension

$$e_T = \frac{(T+1)(T+2)}{2}. \tag{B.13}$$

Notice that $e_T$ is the triangular number $(T+1)$. For instance, with $T = 2$

$$\mathbf{E}_s = \begin{bmatrix} 1 & G_s & G_s^2 & H_s & G_sH_s & H_s^2 \end{bmatrix}'.$$

Define $\mathbf{C}_0$ as $\frac{\Gamma}{2} \times e_T$ matrix whose row $j$ has the binomial coefficients from the path (the binary number with $T+1$ digits) that correspond with the decimal number $(j-1) : j = 1, \dots, \frac{\Gamma}{2}$. For instance, the third row with $T = 2$ corresponds with the path 010, which is the three-digit binary number that represents the decimal number 2. This way of using the corresponding decimal numbers to order the paths and rows of $\mathbf{C}_0$, also implies the order of the elements of vector $\mathbf{E}_s$. Each row $j$ in $\mathbf{C}_0$ contains the coefficients of the different terms of (B.10) plus the zeros needed to fill the rest of the cells for those elements in $\mathbf{E}_s$ that do not appear in the probability of path $j$. A coefficient $\left((-1)^x(-1)^z \binom{n_{00}^j}{x}\binom{n_{10}^j}{z}\right)$ is completely defined by $j$, $x$ and $z$, and it is in row $j$ and column

$$(Z + n_{11}^j)(T+2) - \frac{(z+n_{11}^j)(z+n_{11}^j+1)}{2} + x + 1 + n_{01}^j \tag{B.14}$$

of matrix $\mathbf{C}_0$.

Define $\mathbf{C}_1$ in the same way as $\mathbf{C}_0$, but $j = \frac{\Gamma}{2} + 1, \dots, T$. Each coefficient of (B.10) is in the column given by (B.14) and row $j - \frac{\Gamma}{2}$. Then,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{bmatrix}. \tag{B.15}$$

The dimension of $\mathbf{C}$ is $\Gamma \times 2e_T$ and the dimension of each sub-matrix $\mathbf{C}_0$ and $\mathbf{C}_1$ is $\frac{\Gamma}{2} \times e_T$. From (B.10) and (B.14) matrix $\mathbf{C}$ can be easily computed for any given $T$. For example, with $T = 2$

$$\mathbf{C} = \begin{bmatrix}
1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} \tag{B.16}$$

with dimension 8 × 12.

### B.3. Eliminating redundancies in $\mathbf{A}_{y_0}$

As stated in Proposition 2.2 only $r_T$ $(=T(T+1)+2)$ of the $\Gamma$ $(=2^{T+1})$ possible paths are distinct paths. Therefore, $\mathbf{A}_{y_0}$ cannot have a rank bigger than $r_T$ since $\Gamma - r_T$ rows in $\mathbf{A}_{y_0}$ are repetitions of rows whose paths are the same. Here eliminate those redundancies in $\mathbf{A}_{y_0}$ since it is the rank of it which will define the rank of the system and the rank of its Jacobian will be the rank of $J$. Let us denote the matrix without redundancies with the subscript $r$.

First we take from $\mathbf{C}$ those rows $j$ that correspond to a path $j$ that is not different from a previous path. This means that the number of rows of $\mathbf{C}_r$ is $r_T$. Secondly we reduce the number of columns in $\mathbf{C}$ that are zero or can be expressed as linear combinations of other columns. This means that we are eliminating $2T$ columns ($T$ in each sub-matrix $\mathbf{C}_0$ and $\mathbf{C}_1$) so that the number of columns of $\mathbf{C}_r$ equals $r_T$ too. This column reduction requires the corresponding adjustment in $E_s$.

In $\mathbf{C}_0$ this only requires to eliminate the $T$ columns that are zero. These columns correspond to the $T$ elements in $E_s$ that are only a function of $H_s$. That is, we eliminate $H_s, H_s^2, \dots,$ and $H_s^T$ from $E_s$ when using it in the part of the system that gives the probability for those paths starting with 0. These elements are part of (B.10) for the paths that start at $y_0 = 1$ but $H_s$ cannot be alone an element of (B.10) when $y_0 = 0$. Then,

$$\mathbf{E}'_{s,0r} = \begin{bmatrix} 1 & G_s & .. & G_s^T & G_sH_s & .. & G_s^{T-1}H_s & G_sH_s^2 \\ & .. & G_s^{T-2}H_s^2 & . & . & . & G_sH_s^{T-1} \end{bmatrix}. \tag{B.17}$$

In $\mathbf{C}_1$, in addition to a column that is zero and corresponds to element $G_s^T$ in vector $E_s$, there are $T-1$ columns that are linear combinations of other columns. These $T-1$ columns to be eliminated from $\mathbf{C}_1$ correspond to $\left\{G_s^{T-i}H^i\right\}_{i=1}^{T-1}$ in $E_s$. We eliminate $\left\{G_s^{T-i}H_s^i\right\}_{i=1}^{T-1}$ from $E_s$ and replace $\left\{\left\{G_s^{T-i}H_s^j\right\}_{j=0}^{i-1}\right\}_{i=1}^{T-1}$ by $\left\{\left\{G_s^{T-i}H_s^j\left(1-H_s^{i-j}\right)\right\}_{j=0}^{i-1}\right\}_{i=1}^{T-1}$. This reflects the fact that, for paths starting at $y_0 = 1$, $G_s^T$ cannot be part of (B.10), and $G_s$ with any exponent will only appear in (B.10) if there is at least a $(1-H)$, given that $G$ is $\Pr(y_t = 1|y_{t-1} = 0)$. Thus, the vector $\mathbf{E}_{s1r}$ is of dimension $\frac{r_T}{2}$, its typical element is $G_s^{T-i}H_s^j\left(1-H_s^{i-j}\right)$ for $i = 1, \dots, T-1$ and $j = 0, \dots, i-1$, which is in position $T + i + 1 + \mathbf{1}\{j > 0\}\sum_{k=0}^{j-1}(T-k)$ in the vector. That is,

$$\begin{aligned}\mathbf{E}'_{s,1r} = \big[1 \quad & G_s\left(1-H_s^{T-1}\right) \quad .. \quad G_s^{T-1}(1-H_s) \quad H_s \\ & G_sH_s\left(1-H_s^{T-2}\right) \quad .. \quad G_s^{T-2}H_s(1-H_s) \quad H_s^2 \\ & G_sH_s^2\left(1-H_s^{T-3}\right) \quad .. \quad G_s^{T-3}H_s^2(1-H_s) \quad . \quad . \quad . \\ & . \quad H_s^{T-2} \quad G_sH_s^{T-2}(1-H_s) \quad H_s^{T-1} \quad H_s^T\big].\end{aligned} \tag{B.18}$$

For example, with $T = 2$ and $S = 2$:

$$\begin{bmatrix} \pi_{1|y_0=0} \\ \pi_{2|y_0=0} \\ \pi_{3|y_0=0} \\ \pi_{4|y_0=0} \\ \pi_{5|y_0=1} \\ \pi_{6|y_0=1} \\ \pi_{7|y_0=1} \\ \pi_{8|y_0=1} \end{bmatrix}$$

$$= \begin{bmatrix}
1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

$$* \begin{bmatrix}
1 & 1 & 0 & 0 \\
G_1 & G_2 & 0 & 0 \\
G_1^2 & G_2^2 & 0 & 0 \\
H_1 & H_2 & 0 & 0 \\
G_1H_1 & G_2H_2 & 0 & 0 \\
H_1^2 & H_2^2 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & G_1 & G_2 \\
0 & 0 & G_1^2 & G_2^2 \\
0 & 0 & H_1 & H_2 \\
0 & 0 & G_1H_1 & G_2H_2 \\
0 & 0 & H_1^2 & H_2^2
\end{bmatrix}
\begin{bmatrix} \theta_{1|y_0=0} \\ 1-\theta_{1|y_0=0} \\ \theta_{1|y_0=1} \\ 1-\theta_{1|y_0=1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & 1 & 0 & 0 \\ G_1 & G_2 & 0 & 0 \\ G_1^2 & G_2^2 & 0 & 0 \\ G_1H_1 & G_2H_2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & G_1\left(1-H_1\right) & G_2\left(1-H_2\right) \\ 0 & 0 & H_1 & H_2 \\ 0 & 0 & H_1^2 & H_2^2 \end{bmatrix}$$

$$\times \begin{bmatrix} \theta_{1|y_0=0} \\ 1-\theta_{1|y_0=0} \\ \theta_{1|y_0=1} \\ 1-\theta_{1|y_0=1} \end{bmatrix} \tag{B.19}$$

where the three matrices in the last line are respectively denoted by $\mathbf{C}_r \left(= \begin{bmatrix} \mathbf{C}_{0r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{1r} \end{bmatrix}\right)$, $\mathbf{E}_{y_0r} \left(= \begin{bmatrix} \mathbf{E}_{1,0r} & \mathbf{E}_{2,0r} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}_{1,1r} & \mathbf{E}_{2,1r} \end{bmatrix}\right)$, and $\theta_{y_0}$.

### *B.4. Isolating the unknown parameters*

It is important to note that **C** and $\mathbf{C}_r$ do not depend on unknown parameters. We can construct **C** and $\mathbf{C}_r$ and calculate the rank of **C** for any given $T$, using (B.10), (B.14) and indications in Appendix B.3. Obviously, the rank of **C** is equal to the rank of $\mathbf{C}_r$. Table 3.1 reports rank(**C**), for $T = 2, \ldots, 23$. For all those values of $T$, the rank of **C** is the number of equations that are different in the system, $r_T$:

$$r_T = T(T+1) + 2 = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{C}_r). \tag{B.20}$$

The value $r_T$ is also the dimension of the square matrix $\mathbf{C}_r$. That is, $\mathbf{C}_r$ is a matrix of full rank and we can invert it. Then, our system conditional in the first observation:

$$\pi_{y_0r} = \mathbf{C}_r\mathbf{E}_{y_0r}\theta_{y_0}$$

is equivalent to

$$\mathbf{C}_r^{-1}\pi_{y_0r} = \mathbf{E}_{y_0r}\theta_{y_0} \tag{B.21}$$

where $\mathbf{C}_r^{-1} = \begin{bmatrix} \mathbf{C}_{0r}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{1r}^{-1} \end{bmatrix}$ and $\pi_{y_0r}$ are the proportions, conditional on $y_0$, for the $r_T$ paths that are different. The advantage of (B.21) is that the right hand side contains only unknown values. The two subsystems in (B.21), one conditional on $y_0 = 0$ and the other conditional on $y_0 = 1$, are

$$\mathbf{C}_{0r}^{-1}\pi_{y_0=0r} = \begin{bmatrix} 1 & \cdots & 1 \\ G_1 & \cdots & G_S \\ . & \cdots & . \\ G_1^{T-i}H_1^j & \cdots & G_S^{T-i}H_S^j \\ . & \cdots & . \\ G_1H_1^{T-1} & \cdots & G_SH_S^{T-1} \end{bmatrix} \times \begin{bmatrix} \theta_{1|y_0=0} \\ . \\ \theta_{S-1|y_0=0} \\ 1-\sum_{s=1}^{S-1}\theta_{s|y_0=0} \end{bmatrix} \tag{B.22}$$

$$\mathbf{C}_{1r}^{-1}\pi_{y_0=1r} = \begin{bmatrix} 1 & \cdots & 1 \\ G_1\left(1-H_1^{T-1}\right) & \cdots & G_S\left(1-H_S^{T-1}\right) \\ . & \cdots & . \\ G_1^{T-i}H_1^j\left(1-H_1^{i-j}\right) & \cdots & G_S^{T-i}H_S^j\left(1-H_S^{i-j}\right) \\ . & \cdots & . \\ H_1^T & \cdots & H_S^T \end{bmatrix} \times \begin{bmatrix} \theta_{1|y_0=1} \\ . \\ \theta_{S-1|y_0=1} \\ 1-\sum_{s=1}^{S-1}\theta_{s|y_0=1} \end{bmatrix}. \tag{B.23}$$

It is clear from a direct inspection of (B.22) and (B.23) that the first equation in each of these two systems is trivially satisfied for any value of the parameters since it is the sum of the probability of point of support $s$ which by definition is always equal to one. Therefore, although (B.21) contains $r_T$ different equations, only $r_T - 2$ equations restrict the value of the unknowns. This correspond to the fact that elements in $\pi_{y_0=0}$ and $\pi_{y_0=1}$ sum one, so one of them can be expressed as a linear combination of all the other elements inside each subsystem. Given this, in what follows when we refer to the these systems (B.21)–(B.23), and to matrix $\mathbf{E}_{y_0r}$ we are referring to their formulation without the first elements that trivially sum one. Then the dimension of $\mathbf{E}_{y_0r}$ is $(r_T - 2) \times S$.

### *B.5. The rank of* $\mathbf{J}_r$ *matrix*

The identification result that we try to prove is based on **J** (the Jacobian matrix of system (3.3)) having rank greater than or equal to the number of unknowns. This is equivalent to the Jacobian of (B.21) having rank greater than or equal to the number of unknowns in this system. We denote the latter by $\mathbf{J}_r$. $\mathbf{J}_r$ is a matrix of dimension $(r_T - 2) \times (4S - 2)$ composed of the following parts:

- First $S$ columns that contain the derivatives with respect to $G_1, \ldots, G_S$, whose general form is

$$\begin{bmatrix} \theta_{s|y_0=0} \\ . \\ (T-i)\,G_s^{T-i-1}H_s^j\theta_{s|y_0=0} \\ . \\ H_s^{T-1}\theta_{s|y_0=0} \\ \left(1-H_s^{T-1}\right)\theta_{s|y_0=1} \\ . \\ (T-i)\,G_s^{T-i-1}H_s^j\left(1-H_1^{i-j}\right)\theta_{s|y_0=1} \\ . \\ 0 \end{bmatrix}. \tag{B.24}$$

- Next $S$ columns that contain the derivatives with respect to $H_1, \ldots, H_S$, whose general form is

$$\begin{bmatrix} 0 \\ . \\ jG_s^{T-i}H_s^{j-1}\theta_{s|y_0=0} \\ . \\ (T-1)\,G_sH_s^{T-2}\theta_{s|y_0=0} \\ -\,(T-1)\,G_sH_s^{T-2}\theta_{s|y_0=1} \\ . \\ G_s^{T-i}\left(jH_s^{j-1}\left(1-H_s^{i-j}\right)-(i-j)\,H_s^{i-1}\right)\theta_{s|y_0=1} \\ . \\ TH_s^{T-1}\theta_{s|y_0=1} \end{bmatrix}. \tag{B.25}$$

$$\mathbf{J}_r = \begin{bmatrix} \theta_{1|y_0=0} & \left(1-\theta_{1|y_0=0}\right) & 0 & 0 & G_1-G_2 & 0 \\ 2G_1\theta_{1|y_0=0} & 2G_2\left(1-\theta_{1|y_0=0}\right) & 0 & 0 & G_1^2-G_2^2 & 0 \\ H_1\theta_{1|y_0=0} & H_2\left(1-\theta_{1|y_0=0}\right) & G_1\theta_{1|y_0=0} & G_2\left(1-\theta_{1|y_0=0}\right) & G_1H_1-G_2H_2 & 0 \\ \left(1-H_1\right)\theta_{1|y_0=1} & \left(1-H_2\right)\left(1-\theta_{1|y_0=1}\right) & -G_1\theta_{1|y_0=1} & -G_2\left(1-\theta_{1|y_0=1}\right) & 0 & * \\ 0 & 0 & \theta_{1|y_0=1} & \left(1-\theta_{1|y_0=1}\right) & 0 & H_1-H_2 \\ 0 & 0 & 2H_1\theta_{1|y_0=1} & 2H_2\left(1-\theta_{1|y_0=1}\right) & 0 & H_1^2-H_2^2 \end{bmatrix} \tag{B.27}$$

$$* = G_1\left(1-H_1\right) - G_2\left(1-H_2\right). \tag{B.28}$$

**Box I.**

- Last 2 $(S-1)$ columns that contain the derivatives with respect to $\theta_{1|y_0=0}, \ldots, \theta_{S-1|y_0=0}, \theta_{1|y_0=1}, \ldots, \theta_{S-1|y_0=1}$, whose general form is:

$$\begin{bmatrix} G_s - G_S \\ \cdot \\ G_s^{T-i}H_s^j - G_S^{T-i}H_S^j \\ \cdot \\ G_sH_s^{T-1} - G_SH_S^{T-1} \\ G_1\left(1-H_1^{T-1}\right) - G_S\left(1-H_S^{T-1}\right) \\ \cdot \\ G_s^{T-i}H_s^j\left(1-H_s^{i-j}\right) - G_S^{T-i}H_S^j\left(1-H_S^{i-j}\right) \\ \cdot \\ H_s^T - H_S^T \end{bmatrix}. \tag{B.26}$$

For example, with $T = 2$ and $S = 2$, we have $\mathbf{J}_r$ given in Box I

Since we are trying to derive the minimum number of periods needed for identifying a distribution with $S$ points of support, we first look at the case where $S$ is not limiting the rank of the matrix. Therefore, we consider here a case where $\mathbf{J}_r$ is a squared matrix: $4S-2 = r_T-2$. If squared $\mathbf{J}_r$ matrix has full rank, this will give the identification condition.

$\mathbf{J}_r$ depends on the value of the unknown parameters, and so does the determinant of it $\det(\mathbf{J}_r)$. Therefore, by simply looking at its general form we cannot conclude whether $\det(\mathbf{J}_r)$ is different from zero for all the possible values of the parameters. However, it is not difficult to see that if we evaluate $\det(\mathbf{J}_r)$ at values of the parameters where there is no special relations between the different parameters and points of support all the rows and columns in $\mathbf{J}_r$ are linearly independent and, therefore $\det(\mathbf{J}_r) \neq 0$ when evaluated at those values. Furthermore, simulating many times the matrix $\mathbf{J}_r$ with random draws for the $P_s$'s, $G_s$'s and $H_s$'s we found for all those values that squared $\mathbf{J}_r$ has $\det(\mathbf{J}_r) \neq 0$ that is, full rank and, therefore the rank of $\mathbf{J}_r$ is given by: $r_T-2$. Of course this only shows that $\mathbf{J}_r$ has full rank for those particular numbers tried on the simulations. However finding even only one point for which $\det(\mathbf{J}_r) \neq 0$ is going to be crucial to prove a result about the rank of $\mathbf{J}_r$ in general. The argument is as follows.

Firstly, it is important to note that the equations in system (B.21) are polynomial functions and, therefore $\det(\mathbf{J}_r)$ is also a polynomial function $\mathbf{R}^{4S-2} \longrightarrow \mathbf{R}$. A polynomial function is either identically zero; that is, it is zero for all values at which that function is evaluated, or the set of values at which is zero (its roots), is of measure zero in $\mathbf{R}^{4S-2}$. This result is proved in Lemma 1.1 of Eisenfeld (1986). According to this result, if the polynomial function is not identically zero, then it is different from zero almost everywhere. Therefore, using Lemma 1.1 of Eisenfeld (1986), it is enough to have found a value of the parameters such that $\det(\mathbf{J}_r) \neq 0$ at that particular point to conclude that the $\det(\mathbf{J}_r) \neq 0$ almost everywhere. Putting it in different words, given that there are points at which $\mathbf{J}_r$ has full rank, the set of values of the unknown parameters for which squared $\mathbf{J}_r$ does not have full rank is a set of measure zero.

Thus, we can conclude that for any given $S$,

$$\operatorname{rank}(\mathbf{J}_r) = \min\left(r_T-2, 4S-2\right)$$

almost everywhere.

### B.6. Proof of Proposition 3.1

That (3.8) in Proposition 3.1 is a sufficient condition for identification is a direct application of the general inverse function theorem and the result about the rank of the Jacobian we have shown above. For local point identification of $(G, H)\,|y_0$ the inverse function theorem requires that the rank of $\mathbf{J}_r$ be equal to the number of unknown parameters. As shown in B.5, the rank of $\mathbf{J}_r$ is equal to min ($r_T-2$, number of unknown parameters of the distribution conditional on the first observation). Therefore, the requirement for this case is that the number of unknowns be smaller than or equal to $r_T-2$, that is $4S-2 \leq T(T+1)$. From here we obtain the sufficient condition to identify a distribution of $(G, H)\,|y_0$ in Eq. (3.8). To prove that this condition is also sufficient for identification of the distribution of $(P, G, H)$ with $S$ points of support, it is enough to recall that $(G, H)\,|y_0$ and the observed probability of the initial observation define a unique value for the parameters of the unconditional distribution of $(P, G, H)$, as shown in B.1.

To prove the necessity of condition (3.8) we use Theorem 5.A.1 in Appendix to Chapter 5 in Fisher (1966). That theorem states that having the rank being equal to the number of unknowns is a necessary condition for a local identification of a solution if that solution is a regular point. A point is defined as regular when for all points in a sufficiently small neighborhood of it the Jacobian has the same rank as in the point (see Definition 5.A.1 in Appendix to Chapter 5 in Fisher (1966)). As shown in B.5 the rank of the Jacobian is constant for all points for which (3.8) is a sufficient condition (that is, all points except a set of points of zero measure). Therefore, for those points it is also a necessary condition for identification.

## Appendix C. Non-regular points

When a point is not regular, the condition on the rank of the Jacobian is sufficient but it is not necessary. Thus points that are in the zero measure set of non-regular points may also be locally identified. As it turns out, all of the non-regular points we have found are not locally identified. Moreover, it is potentially fruitful to investigate the properties of non-regular points.

Given there is no explicit solution to system (3.3) and it contains many non-linear equation and unknowns, we have worked with the simpler case $T = 2, S = 2$ to locate non-regular points and study identification. Recall that non-identification means that if we have a $\pi$ $(\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta_1, \ldots, \theta_S)$ in (3.4) that is generated from one of these points, then we will not be able to recover a unique value of $(\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta_1, \ldots, \theta_S)$ from those $\pi'$s. The most interesting cases in terms of its economic interpretation are straightforward to generalize to higher $T$ and $S$. Some of the other points we have located are only for that simpler case. These other cases usually do not have an economic interpretation in our model but impose very particular restrictions between the different points of support of the unobserved heterogeneity.[23] In practical terms we are not generally interested in such cases and they have measure zero. Aside for the obvious cases where any of the parameters is at the boundaries of the parameters' space (that is, it is zero or unity), the following is the list of the interesting non-regular points we have found[24]:

1. $P_s = G_s = H_s$ for all $s = 1, \ldots, S$. In this case the model is not a Markov chain but a static model where each period, including the initial observation, are independent realizations of mixtures of identical Bernoulli distributions.
2. $P_s$ is from the steady state. That is: $P_s = \frac{G_s}{1+G_s-H_s}$, for all $s = 1, \ldots, S$. If we knew that our initial observation is from the steady state and incorporated this to the model we try to identify, then (3.8) will again be a sufficient condition for identification.
3. $G_1 = G_2 = \cdots = G_S$, or $H_1 = H_2 = \cdots = H_S$, or $P_1 = P_2 = \cdots = P_S$. Here, the $S$ points of support are not distinct points in all the three dimensions. This is a violation of the assumption that the model we tried to identify has $S$ points of support. In practice, this is the non-identified case with probably easier solution because we only need to adjust $S$ to the actual (smaller) number of distinct points of support. The problems coming from not having distinct points in $P_s$ are particularly easy to avoid since we can focus on identifying $G$ and $H$ conditioning on the first observation, as done in Appendix B.

## References

Alessie, R., Hochguertel, S., Soest, A., 2004. Ownership of stocks and mutual funds: a panel data analysis. Rev. Econ. Stat. 86 (3), 783–796.

Arcidiacono, P., Jones, J.B., 2003. Finite mixture distributions, sequential likelihood and the EM algorithm. Econometrica 71 (3), 933–946.

Arellano, M., Honoré, B.H., 2001. Panel data models: some recent developments. In: Handbook of Econometrics. pp. 3229–3296 (chapter 5).

Becker, G.S., Grossman, M., Murphy, K.M., 1994. An empirical analysis of cigarette addiction. Amer. Econ. Rev. 84 (3), 396–418.

Bernard, A.B., Jensen, J.B., 2004. Why some firms export. Rev. Econ. Stat. 86 (2), 561–569.

Browning, M., Carro, J.M., 2007. Heterogeneity and microeconometrics modelling. In: Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. 3. Cambridge University Press.

Browning, M., Carro, J.M., 2010. Heterogeneity in dynamic discrete choice models. Econom. J. 13 (1), 1–39.

Browning, M., Carro, J.M., 2013. The identification of a mixture of first order binary Markov chains. Oxf. Bull. Econ. Stat. 75 (3), 455–459.

Carro, J.M., Mira, P., 2006. A dynamic model of contraceptive choice of Spanish couples. J. Appl. Econometrics 21, 955–980.

Chernozhukov, V., Fernandez-Val, I., Hahn, J., Newey, W.K., 2009. Identification and estimation of marginal effects in nonlinear panel data. CeMMAP Working Papers CWP05/09.

Crawford, G.S., Shum, M., 2005. Uncertainty and learning in pharmaceutical demand. Econometrica 73 (4), 1137–1173.

Eisenfeld, J., 1986. A simple solution to the compartmental structural-identifiability problem. Math. Biosci. 79, 209–220.

Feng, Z.D., McCulloch, C.E., 1996. Using bootstrap likelihood methods in finite mixture models. J. R. Stat. Soc. Ser. B 58 (3), 609–617.

Fisher, F.M., 1966. The Identification Problem in Econometrics. McGraw-Hill, New York.

Gautier, E., Kitamura, Y., 2013. Nonparametric estimation in random coefficients binary choice models. Econometrica 81 (2), 581–607.

Ghosal, S., van der Vaart, A.W., 2001. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Statist. 5, 1233–1263.

Gottschalk, P., Moffitt, R.A., 1994. Welfare dependence—concepts, measures, and trends. Amer. Econ. Rev. 84 (2), 38–42.

Ham, J.C., Shore-Sheppard, L., 2005. The effect of medicaid expansions for low-income children on medicaid participation and private insurance coverage: evidence from the SIPP. J. Public Econ. 89 (1), 57–83.

Hayashi, F., 2000. Econometrics. Princeton University Press.

Heckman, J.J., 1981. Heterogeneity and state dependence. In: Studies in Labor Markets. pp. 91–140 (chapter 31).

Heckman, J.J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric-models for duration data. Econometrica 52 (2), 271–320.

Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev. Econom. Stud. 64, 487–535.

Honorè, B.E., Tamer, E., 2006. Bounds on parameters in panel dynamic discrete choice models. Econometrica 74 (3), 611–629.

Hyslop, D.R., 1999. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. Econometrica 67, 1255–1294.

Kasahara, H., Shimotsu, K., 2009. Nonparametric identification of finite mixture models of dynamic discrete choices. Econometrica 77 (1), 135–175.

Keane, M.P., Wolpin, K.I., 1997. The career decisions of young men. J. Polit. Econ. 105, 473–521.

Nevo, A., 2001. Measuring market power in the ready-to-eat cereal industry. Econometrica 69 (2), 307–342.

Rothenberg, T.J., 1971. Identification in parametric models. Econometrica 39 (3), 577–591.

[23] An example of this are non-regular points that require $G_1 = (1 - H_2), G_2 = (1 - H_1)$, plus other restrictions on the $P$'s and $\theta$.

[24] Regardless of the effort made to cover all possibilities (at least in the $T = 2, S = 2$ case), we cannot show that this list is exhaustive.