UNIVERSIDAD CARLOS III DE MADRID
working papers

Working Paper 09-17
Economic Series (10)
February 2009

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 916249875

Dynamic binary outcome models with maximal heterogeneity*

Martin Browning
Department of Economics
University of Oxford
Martin.Browning@economics.ox.ac.uk

Jesus M. Carro
Departmento de Economia,
Universidad Carlos III de Madrid.
jcarro@eco.uc3m.es

First Draft: July 2007†
This Draft: February 2009

Abstract

Most econometric schemes to allow for heterogeneity in micro behaviour 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 modelling 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 the point identification of our heterogeneity structure and show how it depends on the length of the panel. A feasible ML estimation procedure is developed. Tests for a variety of subsidiary hypotheses such as the assumption that marginal dynamic effects are homogeneous are developed.

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.

JEL classification: C23, C24, J64

Keywords: discrete choice, Markov processes, nonparametric identification, unemployment dynamics.

* For comments and useful suggestions, we thank Whitney Newey, Sara Ayo, Ivan Fernandez-Val, 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, and by grant number SEJ2006-05710/ECON from the Spanish Minister of Education.

† The first draft was titled "Identification of the dynamic discrete choice model" and was presented at the CAM Summer Workshop (University of Copenhagen) in July 2007.

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: labour 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) and Ham and Shore-Sheppard (2005)). The usual time-homogeneous first order Markov model for unit i $(= 1, ..N)$ in period t $(t = 0, ..T)$ 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' η_i is widely used but it does have problems; Browning and Carro (2006a) discuss these but it is worth repeating the objections.

The first problem is that the imposition of common slope parameters (α and β) 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:

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

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:

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

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 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 the both the 'intercept' and 'slope' parameters in (1.1). A situation where this is the case can be found in Browning and Carro (2006b). Additional evidence will be provided in the empirical illustration

in this paper. Here the binary variable is 'having a spell of unemployment in a given year' (see Hyslop (1999)).

Model (1.1) with maximal heterogeneity has:

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

In addition to the homogeneity restrictions, model (1.1) is imposing two kind 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 all throughout the paper.[1] A nonparametric time-homogeneous first order Markov process with maximal heterogeneity will look directly at the transition probabilities allowing them to be different for each individual:

$$\Pr\left(y_{it}=1 \mid y_{i,t-1}, x_{it}\right)=p_{ixy_{-1}} \tag{1.5}$$

where we have one parameter to be estimated for each i and 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 (HFOM). 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 but not only, the average marginal effect (also known as average partial effect, APE) of a explanatory variable over the outcome y_{it}. This is important since different studies and questions require us to obtain different parameters of interest. Moreover, the average may not be a very informative measure because of the discrete nature of the problem. For instance, the APE could be found to be very small only because of a group in the population for which a change of a variable does not have enough effect as to change their y_{it} given their other observable and unobservable circumstances. In this case the APE will not be informative about other parts of the population for

[1] 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.

which the impact can be very large because they are close to the margin that make them change their y_{it}. In this situation measures like the median marginal effect are more informative. Also, even if we look only at mean effects, there is more than one that could be of interest: the mean effect for a randomly drawn individual (see Chamberlain, 1984) or ATE in the treatment effect literature, the average marginal effect of x when $x = x_1$ only for those with $x = x_1$ (see Altonji and Matzkin, 2005), the 'average treatment on treated', etc.. 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, Smith and Clements (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 (2006b) 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 (η, α) can be taken as the empirical distribution of these estimates (or some smoothed version of it).

In Browning and Carro (2006b), identification and estimation of (1.5) without imposing any restriction on the distribution of (η, α) 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.[2] This negative result is our starting point in this paper: identification from the cross section is our goal since we do 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 mixture distributions for the joint set of unknown heterogenous parameters. We refer to this as the *nonparametric discrete scheme* since no restriction is imposed other than there is a finite and discrete

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

number of points of support on this distribution. An advantage of using this discrete distribution is that it allows us to go from the full homogeneous case (one point of support) to the totally unrestricted case (as many points of support as N) within the same scheme. 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 our context 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.

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. Chernozhukov, Fernandez-Val, Hahn and Newey (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 extended to the more interesting case with covariates that is taken up in section 5. In sections 2 and 3 consider restrictions from the model and identification respectively. In section 4 we consider estimation and testing. Furthermore, the case without covariates will be a worst case reference in terms of identification; as we will show, having an exogenous x that is not constant across individuals facilitates identification. In Section 6 we apply the techniques we develop to a long 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. These conditions are simple and fast to check.

- Assuming the data has been generated by a HFOM process (both with and without covariates), we provide the limits (necessary and sufficient conditions)

of point identification for two types of distributions for the unobserved heterogeneity: parametric continuous and nonparametric discrete. In the latter case, it is shown that we can have a much richer distribution than the two point distribution usually found 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.

- 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 $\{y_{i0}, y_{i1}, ...y_{iT}\}_{i=1,2,...N}$ 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:

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

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

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

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

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 (α_i, η_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 kind of parametric restrictions: (i) the parametric form implied by the linear index and (ii) 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 heterogenous 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 effect

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

where $F_{(G,H)}\left(G_i, H_i\right)$ 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. As explained in the introduction, there is more than one parameter of interest and identifying the whole HFOM model will allow to obtain any of them, including the entire distribution of M_i in the population.

Given this, our research question is: given a large-H, 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:

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

where n_{01}^j is the number of $0 \to 1$ 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 $t = 0, 2...T$. Thus the first path is always 00..00, the second path is always 00..01 and the last path is always 11..11.

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

Table 2.1: Outcomes for three periods (T=2)

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 Γ outcomes. Denote the population values by π_j for $j = 1, 2...\Gamma$. Let (P,G,H) be distributed over $[0,1]^3$ with an unknown density $f(P,G,H)$. The population proportions are given by the integral equations:

$$\pi_j = \int_0^1 \int_0^1 \int_0^1 p_j(P,G,H) f(P,G,H) \, dP dG dH, \; j = 1, 2...\Gamma \tag{2.8}$$

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

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

The econometric issues are:

1. Given a set of observed π_j's for $j = 1,..2^T$, 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 π_j's, is it unique?
3. If we can find a unique inverse function, is the inverse mapping a continuous function of the values π_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 π_j's).

2.4. Restrictions.

Turning to the first question, we ask whether any observed π_j's that sum to unity could be generated by a HFOM process. The answer is clearly going to be 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*

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

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,

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

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 π_2 is actually equal to 0.25 then $P = 0$ and $G = 0.5$ which in turn imposes $\pi_1 = 0.25$. Although we have not been able to characterize the full set of necessary and sufficient conditions for a given π vector to be generated by a HFOM process, we show below how to test for them.

Using the Lemma above we can calculate the number of paths that are the same for any T, without considering the distribution $f(.)$. For small T this calculation can be done by generating all the possible paths and counting with a computer. However, the following proposition gives an 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* $\pi = (\pi_1, ..., \pi_j, ..., \pi_\Gamma)'$ *whose* π_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.

Table 2.2 presents the results for sample lengths of up to 16 and for 24 (the number used in our empirical example below). The values in the column headed r_T give the number of 'independent' values of the vector π and the column headed

# periods	T	$\Gamma = 2^{T+1}$	r_T	R_T
3	2	8	8	0
4	3	16	14	2
5	4	32	22	10
6	5	64	32	32
7	6	128	44	84
8	7	256	58	198
9	8	512	74	438
10	9	1024	92	932
11	10	2048	112	1936
12	11	4096	134	3962
13	12	8192	158	8034
14	13	16384	184	16200
15	14	32768	212	32556
16	15	65536	242	65294
24	23	$\sim 16.8 \times 10^6$	554	$\sim 16.8 \times 10^6$

Table 2.2: Numbers of posible paths, number of independent cases and number of restrictions

R_T gives the number of restrictions. For medium sized panels the reduction in the number of equations is quite dramatic. For example, for $T = 6$ we have 128 equations and 84 restrictions. This simply highlights that the first order and time-homogeneity assumptions impose strong restrictions if we have several periods of observations.

It is convenient to partition paths into groups based on their having the same probabilities. Define groups $k = 1, 2, ..r_T$ 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:

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

Thus for $T = 5$, for example, we have 32 equations if the HFOM implications are not rejected. Below we shall present a maximum likelihood estimator for our model. When we do this, we shall show how to test for the restrictions implicit in the assumption that our finite sample data are generated by a HFOM process. 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.[3] Since we are here interested in identification without imposing arbitrary homogeneity schemes, this option is not open to us. This leaves us with two broad alternatives.

3.1. Nonparametric identification of the parametric distribution.

The first broad alternative is take a known *parametric distribution function* $f(P, G, H; \beta)$ where β is an unknown L-vector. Thus:

$$\pi_k(\beta) = n_k \int_0^1 \int_0^1 \int_0^1 p_k(P, G, H) f(P, G, H; \beta) \, dP dG dH, \; k = 1, 2...r_T \tag{3.1}$$

The identification issue is to ask whether we can identify the vector of parameters β. The Jacobian is the matrix:

$$J = \left[\frac{\partial \pi_k(\beta)}{\partial \beta_l} \right]_{k=1,..r_T, l=1..L} \tag{3.2}$$

In general we require that this matrix has a rank L, so that a necessary condition for (local) identification is $L \leq r_T$. For example, if we take a 9 parameter distribution for $f(P, G, H; \beta)$ (three means, three variances and three covariances) then we could not point identify with $T = 2$ ($r_T = 8$) without imposing at least one restriction; for example that P is uncorrelated with (G, H). If we take a mixture of two such distributions we have 19 parameters (the two sets of distributional parameters and the mixing probability) which would require $T \geq 4$. If we have a long panel then many components are allowed; for example, with $T = 23$ we could theoretically identify the parameters of a parametric model with 55 component nine parameter distributions. Given the order condition $L \leq r_T$, the rank of (3.2) would need to be checked for the particular parametric form chosen.

[3] Subject to support restrictions that allow us to drive any probability to the limits of 0 or 1.

3.2. Identification for the nonparametric discrete scheme.

The second broad alternative assumption is that we have a *discrete finite mixture distribution* for (P, G, H). For this, we consider nonparametric identification. We take S distinct points of support $\{(P_1, G_1, H_1), ... (P_S, G_S, H_S)\}$ with probabilities given by the $(S \times 1)$ vector θ with non-negative individual values, θ_s, that sum to unity. The discrete analogue to (2.8) is:

$$\pi_j = \sum_{s=1}^{S} p_j (P_s, G_s, H_s) \theta_s \quad j = 1, 2, ...\Gamma \tag{3.3}$$

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

$$A_{js} = p_j (P_s, G_s, H_s), \ j = 1, 2..., 2^{T+1}, \ s = 1, 2...S \tag{3.4}$$

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

$$\pi = \mathbf{A}\theta \tag{3.5}$$

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 θ. We refer to this as the *nonparametric discrete scheme*. The identification issue is: how many points of support can we take for a given T?

Certainly not any discrete distribution with finite points of support will be identified from π. For example, the following two distributions of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S = 3$

$$(P_s, G_s, H_s) = \begin{cases} (0.1, 0.4, 0.4) & \text{with Pr}.\theta_1 = 0.25 \\ (0.1, 0.5, 0.5) & \text{with Pr}.\theta_2 = 0.50 \\ (0.1, 0.6, 0.6) & \text{with Pr}.\theta_3 = 0.25 \end{cases} \quad \text{and}$$

$$(P_s, G_s, H_s) = \begin{cases} (0.1, 0.3, 0.3) & \text{with Pr}.\theta_1 = 0.0625 \\ (0.1, 0.5, 0.5) & \text{with Pr}.\theta_2 = 0.875 \\ (0.1, 0.7, 0.7) & \text{with Pr}.\theta_3 = 0.0625 \end{cases}$$

give the same proportions with $T = 2$:

$$\pi = \{0.2295, 0.2205, 0.2205, 0.2295, 0.0255, 0.0245, 0.0245, 0.0255\}$$

Therefore we cannot identify the distribution of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ with $S = 3$, from the π we observe when $T = 2$.

From (3.5), for given S, we have a mapping from unobservables to observables

given by:

$$\pi\left(\mathbf{P},\mathbf{G},\mathbf{H},\theta_1,..\theta_S\right)=\mathbf{A}\left(\mathbf{P},\mathbf{G},\mathbf{H}\right)\theta$$

where the S-vector θ is normalized to sum to unity. The Jacobian of this is a $\Gamma\times(4S-1)$ matrix which we denote $J\left(T,S\right)$. For local point identification we require that the rank of $J\left(T,S\right)$ is greater than or equal to the number of parameters. In Appendix A.3 we show that, generically:

$$\min\left(r_T-1,4S-1\right)\leq rank(J)\leq\min\left(r_T,4S-1\right) \tag{3.6}$$

Although we are unable to prove it, we conjecture[4] that this bound could be tightened to:

$$rank\left(J\right)=\min\left(r_T,4S-1\right) \tag{3.7}$$

If we have S points of support then we have $4S-1$ free parameters (one θ_s is determined by the others). The parameters of these support points and their probabilities can only be point identified if the number of parameters is not greater than the rank of J; using (3.7), this requires:

$$S\leq\frac{r_T+1}{4}=\Upsilon_T \tag{3.8}$$

The final row of Table (3.1) gives the values for the maximum number of points of support for a given T, denoted Υ_T. Since we have non-integer values for Υ_T we can take S equal to the integer above Υ_T and impose a small number of 'common value' restrictions on the (G_s,H_s) values and/or on the probabilities. For example, for $T=2$ we have $\Upsilon_T=2.25$ so that we could take:

$$\left(P_1,G_1,H_1\right),\left(P_2,G_2,H_1\right),\left(P_1,G_1,H_2\right) \tag{3.9}$$

and 2 unrestricted values for the mixing probabilities; this gives a total of 8 unknown parameters. As can be seen from Table (3.1), if we have a reasonably long panel ($T=7$, for example) then we can have a relatively rich distribution with 14 independent points of support. Even with a short panel ($T=4$, for example) we can do better than the two point distribution that is commonly used in applied work.

All the previous results can be summarized in the following propositions:

Proposition 3.1. *For local identification of* $\{\mathbf{P},\mathbf{G},\mathbf{H}\}$ *and* θ *in the system (3.5) for a given* T*:*

[4]Based on a great number of simulations.

T	2	3	4	5	6	7	8	9	....	23
r_T	8	14	22	32	44	58	74	92	....	554
Υ_T	2.25	3.75	5.75	8.25	11.25	14.75	18.75	23.25	....	138.75

Table 3.1: Rank of the Jacobian and maximum number of points of support

(i) *A necessary condition is that the number of unknowns be smaller than* $r_T = T(T+1)+2$.

(ii) *A sufficient condition is that the number of unknowns be smaller than* $r_T - 1 = T(T+1)+1$, *except for the particular cases with* $P_s = 0.5$ *for all* $s = 1, ..., S$.

Proposition 3.2. *If we consider only the identification of the all the parameters of the nonparametric discrete distribution of* $\{\mathbf{P}, \mathbf{G}, \mathbf{H}\}$ *implied by a number* S *of points of support in this distribution, then a necessary and sufficient condition for local identification in the system (3.5) for a given* T *is*

$$S \leq \text{integer}\left[\frac{r_T + 1}{4}\right]$$

except for the particular cases with $P_s = 0.5$ *for all* $s = 1, ..., S$..

The proofs are given in Appendix.

Finally we note that our use of a discrete distribution to capture heterogeneity is fundamentally 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 the recourse to a discrete distribution is one route to nonparametric point identification.

3.2.1. Quadrature discrete approach.

For completeness, we also discuss an alternative to the finite mixture model which is to take a *quadrature* approach in which we pre-specify S grid points

$$\{(P_1, G_1, H_1), ... (P_S, G_S, H_S)\}$$

and then estimate the weights θ. For the identification in this quadrature scheme, we require $S \leq r_T + 1$ to identify the vector θ. Given S points of support we can construct the $\Gamma \times S$ A matrix as in (3.4) and (3.5). Note that here the relevant rank

is $rank(A)$ instead of $rank(J)$. The support points should be chosen so that the A matrix has rank equal to r_T. Estimation then proceeds by using standard methods to find a value θ that satisfies:

$$\begin{bmatrix} \pi \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{e}' \end{bmatrix} \theta \tag{3.10}$$

$$\theta \geq \mathbf{0} \tag{3.11}$$

where $\mathbf{e}$ is a vector of ones.

Bajari, Fox, Kim, and Ryan (2008) study in detail this approach and its properties in static random coefficients discrete choice models for demand estimation in IO. As they explain, the main advantage of this approach is that it is simpler to estimate because our nonlinear system of equations is now linear. Furthermore, we know that the values of the parameters for which we have to take grid points always lie in the unit cube. The main problem is that this crucially depends on either being able to take very many grid points or on the values you pre-specify. Given that we have a panel with a not very large number of periods, the number of grid points is not going to be very large for a fully characterization of the parameter space.[5] With respect to the pre-specify values, we do not have any other information that helps in choosing them. And using any additional method to help choosing the best grid points will break the simplicity that motivates this approach.

We tried this scheme in our empirical application, taking the maximum number of grid points for our sample with $T = 23$. Taking 555 more or less equally spaced points in the unit cube we found the fit to be very poor compared with using a finite discrete mixture distribution. Apart from that, there are some efficiency problems when comparing the linear estimator for the quadrature scheme with the MLE for the mixture distributions we use in next sections.

4. Estimation and Testing against alternative models.

4.1. The time-homogeneous first order Markov model.

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.

[5] For example, with $T = 8$ we can not take more than 75 points to try to characterize the joint distribution of (P, G, H) in the three dimensional parameter space with possible complex correlations between the three parameters.

In practice, it is much 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 subsection 2.4).

Take the full heterogeneity model with $S = \Upsilon_T$ so that we have a just identified model. Using the first form in (3.3), the structural model is:

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

Define a indicator $\delta_{ij} = 1$ if unit i has path j and zero otherwise. For given parameters, the likelihood of a sample $\{y_{i0}, y_{i1}, ...y_{iT}\}_{i=1,2,...N}$ 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. Denote the sample proportions for path j $c_j = n_j/N$. The log-likelihood function for the mixture model is:

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

$$= N\sum_{j=1}^{\Gamma} c_j \log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right) \tag{4.4}$$

Note that N is irrelevant for the maximization. With an iid random sample $c_j \to \pi_j$ as $N \to \infty$. from the structural model. The advantage of using the likelihood framework for estimation is that we know how to use all the information on the sample, how to make inference and how to test different models.

4.2. The unrestricted model.

A natural benchmark against which to test the HFOM model is the saturated model with:

$$\begin{aligned} &S = \Gamma, \mathbf{A} = I, \theta = \pi \text{ or} \\ &S = 1, \mathbf{A} = \pi \end{aligned} \tag{4.5}$$

These both give the likelihood value:

$$\ell_{sat} = \sum_{i=1}^{N} \sum_{j=1}^{\Gamma} \delta_{ij} \log(c_j) = N \sum_{j=1}^{\Gamma} c_j \log(c_j). \tag{4.6}$$

This can be used to derive a likelihood ratio statistic for the test of the Markov model against the unrestricted alternative. In particular, if we do not reject the restriction from (4.6) to (4.4) then we cannot reject that we have a time-homogeneous first order model. In practice, the large number of zeros for most paths if T is moderately sized leads to a distribution for the LR statistic that is very far from a χ^2distribution with degrees of freedom equal to the number of restrictions (R_T in table 2.2). In this case, we should simulate the distribution of the LR statistic to calculate the true the correct probability of the observed LR statistic.

4.3. The unrestricted HFOM model or restricted saturated model.

We can also write a closed form expression for the model with the HFOM equality restrictions form subsection 2.4 imposed, using equation (2.15). Let $k(j)$ denote the group (running from $k = 1, ..r_T$) that path j belongs to. Then define predicted probabilities for path $j = 1, ..\Gamma$ by:

$$\hat{c}_j = \frac{1}{n_{k(j)}} \sum\nolimits_{j \in k(j)} c_j \tag{4.7}$$

That is, we replace the unrestricted proportions for each path by the mean for the group.[6] The likelihood function is then given by:

$$\ell_{res\_sat} = \sum_{i=1}^{N} \sum_{j=1}^{\Gamma} \delta_{ij} \log(\hat{c}_j) = N \sum_{j=1}^{\Gamma} c_j \log(\hat{c}_j) \tag{4.8}$$

This likelihood function also plays an important role in the estimation and choice of the mixing model. If we take a mixture with the maximal number of components, Υ_T in Table 3.1 then it has a log likelihood value that is bounded above by ℓ_{res_sat}. The mixture model will only attain this likelihood value if the observed $\hat{\mathbf{c}}$ vector satisfies the inequality constraints discussed in subsection 2.4. Given the difficulties of finding global maxima when we have many components, having a benchmark value is a considerable advantage. Denote the likelihood value of this mixture model by ℓ^{Υ}_{mix}. Now consider a model with fewer than the maximum number of points of

[6] 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, $\hat{c}_3 = \hat{c}_5 = \frac{c_3 + c_5}{2}$; $\hat{c}_{12} = \hat{c}_{14} = \frac{c_{12} + c_{14}}{2}$; $\hat{c}_j = c_j$, all other j.

support: $S < \Upsilon_T$. We have the following ordering for the likelihood function values:

$$\ell_{sat} \geq \ell_{res\_sat} \geq \ell_{mix}^{\Upsilon} \geq \ell_{mix}^{S} \tag{4.9}$$

As already discussed, the likelihood ratio statistic does not have a known general distribution (see chapter 6.4 of McLachlan and Peel (2004)) but a test of the model with a smaller number of points of support than Υ_T can be constructed based on the simulated distribution for the LR statistic, taking the restricted model as the null.[7]

If we reject the first order time-homogeneous model, we have a number of alternatives. We could try a time-homogeneous second order model; this would give rise to similar calculations to those made above. Alternatively, we could continue to maintain that the model is a first order Markov chain but with time-inhomogeneous transition probabilities. One variant would be to assume a structural break. A second variant has that the transition probabilities depend on observable time-varying covariates. We consider that in the next section.

4.4. Testing for a second order Markov process

Although the test of the HFOM model against the saturated model allows for any alternative, it may lack power since the alternative is not specified. The obvious alternative is a time-homogeneous second order process. Given the estimates of the first order process, we can derive a standard LM test for this. The log-likelihood of a time-homogeneous second order Markov process has the following form for the predicted probabilities:

$$\begin{gathered} p_j\left(P_{00s}, P_{01s}, P_{10s}, G_{00s}, G_{10s}, H_{01s}, H_{11s}\right) = \\ P_{00s}^{1\left(y_0^j=0, y_1^j=0\right)} P_{01s}^{1\left(y_0^j=0, y_1^j=1\right)} P_{10s}^{1\left(y_0^j=1, y_1^j=0\right)} * \\ \left(1-P_{00s}-P_{01s}-P_{10s}\right)^{1\left(y_0^j=1, y_1^j=1\right)} G_{00}^{n_{001}^j}\left(1-G_{00}\right)^{n_{000}^j} G_{10}^{n_{101}^j} * \\ \left(1-G_{10}\right)^{n_{100}^j} H_{01}^{n_{011}^j}\left(1-H_{01}\right)^{n_{010}^j} H_{11}^{n_{111}^j}\left(1-H_{11}\right)^{n_{110}^j} \end{gathered} \tag{4.10}$$

[7] Given that we have a fully parametric model, simulating the distribution of the LR statistic under the null seems preferable to subsampling methods.

where $1(.)$ is the indicator function and:

$$P_{01} = \Pr(y_{i0} = 0, y_{i1} = 1), \tag{4.11}$$

$$G_{10} = \Pr(y_{it} = 1 \mid y_{it-2} = 1, y_{it-1} = 0), \tag{4.12}$$

$$H_{01} = \Pr(y_{it} = 1 \mid y_{it-2} = 0, y_{it-1} = 1), \tag{4.13}$$

$$...$$

This has seven parameters per type s, instead of three. Three of them are to account for the initial conditions, since now we have to condition on two previous observations. The other four are the transition probabilities given by the second order Markov process, what imposes less restrictions on the data than the first order process. Therefore, the log-likelihood now depends on $8S-1$ parameters.[8] To perform the LM test we have to:

1. Derive the log-likelihood with respect to $\{P_{00s}, P_{01s}, P_{10s}, G_{00s}, G_{10s}, H_{01s}, H_{11s}\}_{s=1}^{S}$ and $\{\theta_s\}_{s=1}^{S-1}$. This gives the score vector denoted by $g(.)$, and allows us to calculate the outer-product of the score, denoted by $h(.)$.

2. Evaluate $g(.)$ and $h(.)$ at the estimated values of the parameters of the first order Markov model $\left(\left\{\widehat{P}_s, \widehat{G}_s, \widehat{H}_s\right\}_{s=}^{S}, \left\{\widehat{\theta}_s\right\}_{s=1}^{S-1}\right)$. This means that we evaluate $g(.)$ and $h(.)$ at $P_{00s} = \left(1-\widehat{P}_s\right)\left(1-\widehat{G}_s\right)$, $P_{01s} = \left(1-\widehat{P}_s\right)\widehat{G}_s$, $P_{10s} = \widehat{P}_s\left(1-\widehat{H}_s\right)$, $G_{00s} = G_{10s} = \widehat{G}_s$, $H_{01s} = H_{11s} = \widehat{H}_s$ for $s = 1, ..., S$, and $\left\{\widehat{\theta}_s\right\}_{s=1}^{S-1}$. Denote the values we get from this by $\widehat{g}$ and $\widehat{h}$.

3. Then, the test statistic is

$$LM = \widehat{g}'\widehat{h}^{-1}\widehat{g} \tag{4.14}$$

Under the standard regularity conditions this test statistic is asymptotically distributed as $\varkappa_b^2$. The degrees of freedom are

$$b = (7S + S - 1) - (3S + S - 1) = 4S \tag{4.15}$$

4.5. Homogenous marginal dynamic effect.

We shall not consider the homogeneous case with (G, H, P) the same for everyone, since it is hardly considered a possibility. A less restricted model than the

[8] This means that we are keeping S constant. Related with this, it is important to notice that to point identify a first order Markov model with S points of support does not imply that a second order Markov model with S points of support can also be point identified.

homogeneous case is the usual 'fixed effect' case which only allows for one source of unobservable heterogeneity. The latter is usually in the intercept of the index in (1.1). A close analogue here is that we have a homogeneous dynamic marginal effect:

$$H_i = M + G_i \text{ for some constant } M \in [-1, 1] \tag{4.16}$$

This test can be done using a standard LR test statistic of the $(S-1)$ restrictions imposed.

4.6. Testing for time homogeneity.

As well as testing against a specific time homogeneous model, we can also derive a test for time homogeneity. To do this, we split the sample into an estimation sample $\{y_{i0}, y_{i1}, ...y_{iE}\}$ and a hold-out sample $\{y_{iE+1}, y_{iE+2}, ...y_{iT}\}$. We estimate the mixture model on the estimation subsample and test whether the predictions for the hold-out subsample fit. To do this we take the same transition probabilities for the hold-out subsample. To generate the distribution for period $E+1$ (the initial period for the hold-out sample) we use the estimated probabilities and the Chapman-Kolmogorov equations to generate the relevant distribution. An alternative procedure is split the sample into two equal subsamples in terms of term (E close to $(T+1)/2$), estimate on each subsample separately and then test whether the two sets of estimates are statistically different. A particularly simple variant of a stability test of this sort this will be given in the empirical section.

5. Allowing for covariates.

5.1. Model and Parameters of Interest

In the presence of covariates in the model, our estimation is conditional on the covariates, which are assumed to be strictly exogenous. As before, we look directly at 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_x are defined for each value x of x_{it}, and at the unconditional probability of a unit value for the initial observation:

$$P_{xi} = \Pr\left(y_{i0} = 1 | x_{i0} = x\right) \tag{5.1}$$

In addition to the marginal dynamic effect and the long run proportion of unit values mentioned for the model without covariates, (P_{xi}, G_{xi}, H_{xi}) can be used to generate any other outcomes or parameters of interest. There are several candidates but the most widely considered are those informing about the *marginal effects* of the change in a explanatory variable x over the probability of y_{it} being equal to 1, (where for notational convenience we consider only one covariate):

$$\begin{aligned} M_{x'i} &= \Pr\left(y_{it}=1 \mid y_{i,t-1}, x_{it}=x''=x'+1\right) - \Pr\left(y_{it}=1 \mid y_{i,t-1}, x_{it}=x'\right) \\ &= \begin{cases} H_{x''i} - H_{x'i} & \text{if } y_{it-1}=1 \\ G_{x''i} - G_{x'i} & \text{if } y_{it-1}=0 \end{cases} \end{aligned} \tag{5.2}$$

Given that the marginal effects are heterogenous across individuals in the population, the interest is usually in knowing their distribution or some moments. There are many possible measures that could be considered. For example, there is more than one mean effect that could be of interest. Here we mention just two of them.

The first is expected effect on the probability of $y=1$ of a change in variable x given the distribution of the unobservables conditional on $x=x'$:

$$\begin{aligned} E_{(G,H)|x}\left[M_i|x'\right] = \int\int [(H_{x''i} - H_{x'i}) \Pr(y_{it-1}=1|x_{it}=x'; G_{xi}, H_{xi}) \\ + (G_{x''i} - G_{x'i}) \Pr(y_{it-1}=0|x_{it}=x'; G_{xi}, H_{xi})] dF_{(G,H)|x}\left(G_{xi}, H_{xi}|x'\right) \end{aligned} \tag{5.3}$$

This is equivalent to the parameter of interest estimated in Altonji and Matzkin (2005). If x were a treatment indicator variable with $x'=0$ and $x''=1$, then (5.3) would give the average Treatment on the Untreated effect.

The second example is the average marginal effect without conditioning on x:

$$\begin{aligned} E_{(G,H)}\left[M_i\right] = \int\int [(H_{x''i} - H_{x'i}) \Pr(y_{it-1}=1|G_{xi}, H_{xi}) \\ + (G_{x''i} - G_{x'i}) \Pr(y_{it-1}=0|G_{xi}, H_{xi})] dF_{(G,H)}\left(G_{xi}, H_{xi}\right) \end{aligned} \tag{5.4}$$

This is equivalent to the parameter of interest proposed by Chamberlain (1984) defined there as the mean effect for a randomly drawn individual. If x were a treatment indicator variable with $x'=0$ and $x''=1$, then (5.4) would give the Average Treatment Effect.

Equations (5.4) and (5.3) are the answer to different questions and with more explanatory variables, averages over different distributions could be considered. On the other hand, as explained in the introduction, average effects may not be very informative in nonlinear models such as this. In such a case other moments of the

individual marginal effects such as the median are more informative. Furthermore, there are cases where the entire distribution of the marginal effect over the population is the object of interest; see Heckman, Smith and Clements (1997). In the IO literature the object of interest is the entire distribution of the individual price elasticities; see, for example, Nevo (2001). Identifying and estimating the distribution of (P_i, G_{xi}, H_{xi}) allow us to obtain any possible parameter of interest since it fully characterize the HFOM model.[9]

Adding covariates not only changes the number of transition probabilities we have to identify, but also introduces the possibility of dependence between the probability of being of each unobserved type and the covariates. We start with the simplest case: a binary covariate that is constant over time. From there we move to covariates that vary over time but are common to all individuals. Then, the case with covariates that vary both over i and t is considered. A summarizing table with numbers for representative cases can be found at the end of this section.

5.2. Covariates constant over time

We begin with the case in which we only have an x variable that is constant over time and only varies across individuals; examples include year of birth and education. That is, $x_{it} = x_i$ for all t, and our data set is $\{x_i, y_{i0},..., y_{iT}\}_{i=1}^N$. Here, it is conceptually simple to extend our analysis. Continuing with similar notation, for a binary x_i, the time homogenous first order Markov model is fully characterized by:

$$\begin{aligned}
&P_{0i} = \Pr(y_{i0} = 1 \mid x_i = 0)\,;\; P_{1i} = \Pr(y_{i0} = 1 \mid x_i = 1) \\
&G_{0i} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_i = 0)\,;\; H_{0i} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_i = 0) \\
&G_{1i} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_i = 1)\,;\; H_{1i} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_i = 1)
\end{aligned} \tag{5.5}$$

As before, we consider a nonparametric discrete distribution for $(P_{0i}, G_{0i}, H_{0i},\ P_{1i}, G_{1i}, H_{1i})$ with S distinct points of support $\{P_{0s}, G_{0s}, H_{0s}, P_{1s}, G_{1s}, H_{1s}\}_{s=1}^S$ with probabilities given by the $(S \times 1)$ vector θ_x with non-negative values θ_{xs} that sum to unity.

$$\theta_x = \begin{cases} (\theta_{01}, ..., \theta_{0S})' & \text{if } x = 0 \\ (\theta_{11}, ..., \theta_{1S})' & \text{if } x = 1 \end{cases} \tag{5.6}$$

where $\theta_{xS} = 1 - \sum_{s=1}^S \theta_{xS}$. 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

[9] Since the x variables are assumed to be exogenous, there is no problem in obtaining their distribution from a random sample when needed.

calculated from this conditional distribution and the distribution of x, which can be obtained from the data.

The possible number of $\{x_i, y_{i0,}..., y_{iT}\}$ paths we can observe is $2 * 2^{T+1}$. This is equal to the 2^{T+1} paths of $\{x_i = 0, y_{i0,}..., y_{iT}\}$ plus the 2^{T+1} paths of $\{x_i = 1, y_{i0,}..., y_{iT}\}$. The probability of a path j given x and $(P_{0s}, G_{0s}, H_{0s}, P_{1s}, G_{1s}, H_{1s})$ is $p_j(x; P_{0s}, G_{0s}, H_{0s}, P_{1s}, G_{1s}, H_{1s}) =$

$$= \begin{cases} P_{0s}^{y_0^j} (1 - P_{0s})^{\left(1-y_0^j\right)} G_{0s}^{n_{01}^j} (1 - G_{0s})^{n_{00}^j} H_{0s}^{n_{11}^j} (1 - H_{0s})^{n_{10}^j} & \text{if } x = 0 \\ P_{1s}^{y_0^j} (1 - P_{1s})^{\left(1-y_0^j\right)} G_{1s}^{n_{01}^j} (1 - G_{1s})^{n_{00}^j} H_{1s}^{n_{11}^j} (1 - H_{1s})^{n_{10}^j} & \text{if } x = 1 \end{cases} \tag{5.7}$$

So, we stratify the sample, making one strata for each value of x and we have the same exact problem as in the case without covariates for each strata. As a consequence, the rank of the Jacobian is $\min(2(4S-1), 2r_T)$, where r_T was defined in (2.14). We have doubled the rank, but also the number of parameters. Then, we can identify distributions with the same number of points of support S as in the case without covariates, Υ_T defined in (3.8).

If instead of a binary covariate we have a general discrete x variable that takes N_x different values, we can easily repeat the same analysis and arrive at the same conclusion. The number of parameters is now $N_x(4S-1)$, and the rank of the Jacobian is $\min(N_x(4S-1), N_x r_T)$. This implies that the maximum number of points of support we can identify is the same as in the case without covariates.

$$\Upsilon_{Nx,T} = \frac{N_x(r_T+1)}{4N_x} = \frac{r_T+1}{4} = \Upsilon_T \tag{5.8}$$

If we have a continuous covariates, we can always discretise it on very many N_x grid points and use this result in (5.8). Therefore, with covariates constant over time we can nonparametrically identify as many points of support as in the case without covariates.

Independence between θ and x. If the probabilities of the S points of support of $(P_{0i}, G_{0i}, H_{0i}, P_{1i}, G_{1i}, H_{1i})$ do not depend on x, then $\theta_x = (\theta_1, ..., \theta_S)'$ for all values of x. This reduces the number of parameters, but not the number of equations. There are $(3SN_x + (S-1))$ parameters and $N_x r_T$ 'independent' equations. Therefore, the maximum number of points of support we can identify is

$$\frac{N_x r_T + 1}{3N_x + 1} > \frac{r_T + 1}{4} \tag{5.9}$$

The independence assumption allows us to identify distributions with higher number of points of support.

5.2.1. Semiparametric model

In the previous analysis we have not only allowed for maximal (nonparametric 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_i affects y_{it}. Nevertheless, if x_i 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:

$$\begin{aligned} P_{si} &= F_0(p_{s0} + p_{s1}x_i) \\ G_{si} &= F(g_{s0} + g_{s1}x_i) \\ H_{si} &= F(h_{s0} + h_{s1}x_i) \\ \theta_{si} &= F_\theta(d_{s0} + d_{s1}x_i) \end{aligned} \tag{5.10}$$

where F_0, F and F_θ are known cdf functions, such as the standard normal cdf or the standard logistic function. The heterogenous parameters that have S points of support and conditional probabilities θ_{si} are now $(p_{s0}, p_{s1}, g_{s0}, g_{s1}, h_{s0}, h_{s1})$. This is equivalent to the representation

$$\Pr\left(y_{it} = 1 \mid y_{i,t-1}, x_i\right) = F\left(\eta_i + \alpha_i y_{it-1} + \beta_i x_i + \delta_i x_i y_{it-1}\right) \tag{5.11}$$

where $(\eta_i, \alpha_i, \beta_i, \delta_i)$ follow a discrete distribution with S points of support.[10]

Therefore, the number of parameters here is $2 * (4S - 1)$ $(= 8S - 2)$, and it does not depend on the number of values of x_i, N_x. However, the number of equations and the rank of the Jacobian still depends on N_x. As previously shown this multiplies r_T by N_x. Therefore, the maximum number of points of support we can identify is

$$S \leq \frac{N_x r_T + 2}{8} \tag{5.12}$$

$\frac{N_x r_T + 2}{8}$ is equal to $\frac{r_T + 1}{4}$ if $N_x = 2$ (the case with a binary x), but it is greater than $\frac{r_T + 1}{4}$ for any $N_x > 2$.

According with (5.12), the more values x_i takes, or the more we discretise a continuous x_i, the richer the distribution we can point identify. Given that

[10] Note that all possible interactions between y_{it-1} and x_{it} are being considered. The number of parameters could be reduced even more if those interactions were not included.

$\lim_{N_x \to \infty} \frac{N_x r_T + 2}{8} = \infty$, we could *potentially* identify as many points of support as we wish when we have a continuous covariate.

5.3. Covariates that only vary over time.

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 are 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 heterogenous parameters over the i population, that is, $\Pr\left(s | \{x_t\}_{t=1}^T\right) = \Pr(s) = \theta_s$.

5.3.1. Binary variable

If x_t takes only two values 0 and 1, for each point of support s we have

$$\begin{aligned} P_s &= \Pr(y_{i0} = 1 \mid x, s) = \Pr(y_{i0} = 1 \mid s) \\ G_{0s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_t = 0, s)\,;\; H_{0s} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_t = 0, s) \\ G_{1s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_t = 1, s)\,;\; H_{1s} = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_t = 1, s) \end{aligned} \tag{5.13}$$

where the probability of s is given by the $(S \times 1)$ vector θ with non-negative individual values, θ_s, that sum to unity and it is independent of $\{x_t\}_{t=1}^T$. This makes $6S - 1$ parameters.

The probability of a path j given $\{x_t\}_{t=1}^T$ and $(P_s, G_{0s}, H_{0s}, G_{1s}, H_{1s})$ is:

$$\begin{aligned} p_{js} = {} & P_s^{y_0^j} (1 - P_s)^{\left(1 - y_0^j\right)} G_{0s}^{n_{01|0}^j} (1 - G_{0s})^{n_{00|0}^j} H_{0s}^{n_{11|0}^j} \\ & (1 - H_{0s})^{n_{10|0}^j} G_{1s}^{n_{01|1}^j} (1 - G_{1s})^{n_{00|1}^j} H_{1s}^{n_{11|1}^j} (1 - H_{1s})^{n_{10|1}^j} \end{aligned} \tag{5.14}$$

where $n_{01|0}^j$ is the number of $y_{t-1} = 0 \to y_t = 1$ transitions for path j given $x_t = 0$, $n_{01|1}^j$ is the number of $y_{t-1} = 0 \to y_t = 1$ transitions for path j given $x_t = 1$, and similarly for the other three transitions. Here the probability of observing a path j given $\{x_t\}_{t=1}^T$ is $\pi_{xj} = \sum_{s=1}^S p_{js}\theta_{xs}$ for $j = 1, 2, ...2^{T+1}$. If two paths have the same

$\left(y_0^j, n_{00}^j, n_{01}^j, n_{10}^j, n_{11}^j\right)$ they will also have the same $(y_0^j, n_{00|0}^j, n_{01|0}^j, n_{10|0}^j, n_{11|0}^j, n_{00|1}^j,$ $n_{01|1}^j,\ n_{10|1}^j, n_{11|1}^j)$, simply because we are dividing each n_{yz}^j in two $n_{yz|x}^j$ using the same $\{x_t\}_{t=1}^T$ to divide both paths. For the same reason if they are different they will be different here too. So the number of different equations is $r_T = T(T+1)+2$ as without covariates.

Thus we can identify distributions with a smaller number of points of support S than in the case without covariates. Specifically:

$$S \leq \frac{r_T+1}{6} < \frac{r_T+1}{4} = \Upsilon_T \tag{5.15}$$

5.3.2. Discrete covariates with N_x values.

As an example of a discrete common covariate, consider the national unemployment rate (notionally a continuous variable between zero and 100) with rates rounded to the nearest 0.1% ($N_x = 1001$). If we have a general discrete x variable that can take N_x different values, we can easily repeat the same analysis. The only change is in the number of parameters, which is equal to $(2+2N_x)S-1$, provided $N_x \leq T$. If $N_x > T$, the number of parameters is $(2+2T)S-1$, because even though x_t could take more values, we only can observe in our population $(T+1)$ different values at most; and x_0 does not affect our model for the reasons already explained. Then,

$$\begin{aligned} S &\leq \frac{r_T+1}{2(1+N_x)} \text{ if } N_x \leq T \\ S &\leq \frac{r_T+1}{2(1+T)} \text{ if } N_x > T \end{aligned} \tag{5.16}$$

5.3.3. K covariates: time dummies

Adding more x_t covariates will only increase the number of parameters and we can extend the previous analysis. However, this is not a very interesting case to consider, because it is impossible to nonparametrically identify the marginal effects of an x_{1t} variable given another continuous x_{2t} variable, since we never observe any individual with variation in x_{1t} while keeping constant the value of x_{2t}, nor will we have other combinations of values of (x_{1t}, x_{2t}) than that in our fixed T population. This problem is different than being nonparametric in the distribution of the unobserved heterogeneity. A solution to it is the use of a semiparametric model of the effects of the covariates. We consider this in next subsection.

On the other hand a situation often found in practice that is relevant to consider here, 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 other 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. So, we have $(2 + 2T)S - 1$ parameters: one G and H for each time dummy. Then,

$$S \leq \frac{r_T + 1}{2 + 2T} \tag{5.17}$$

This implies a very small number S for each T unless the number of periods is large. For example, we need $T \geq 8$ for the identification conditions of a model with $S = 4$ to be satisfied. With $T = 23$ we cannot identify more than $S = 11$.

5.3.4. A semiparametric model

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

$$\begin{aligned} G_s &= F(g_{s0} + \sum\nolimits_{k=1}^{K} g_{sk} x_{kt}) \\ H_s &= F(h_{s0} + \sum\nolimits_{k=1}^{K} h_{sk} x_{kt}) \end{aligned} \tag{5.18}$$

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

$$S \leq \frac{r_T + 1}{2 + 2(K + 1)} \tag{5.19}$$

For example, if $K = 2$ and $T = 8$, then $S \leq 9.375$; or if $K = 2$ and with $T = 23$, then $S \leq 69.375$. This and values for other cases can be found in table 5.2.

5.4. Covariates that vary in both i and t

Finally we 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:

$$\begin{aligned} P_{xs} &= \Pr\left(y_{i0} = 1 \mid x_{i0}, s\right) \\ G_{xs} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it}, s\right) \\ H_{xs} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it}, s\right) \\ \theta_{Xs} &= \Pr\left(s | x_{i0}, ..., x_{iT}\right) \end{aligned} \tag{5.20}$$

With respect to θ_{si} we can have:

Case 1. Independence between θ_{si} and x_{it}: $\theta_{si} = \Pr\left(s | x_{i0}, ..., x_{iT}\right) = \Pr\left(s\right) = \theta_s$. Here there are $S - 1$ parameters θ to estimate, as in the case without covariates.

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.

Case 2. θ_{si} depends only on the first observation x_{i0}: $\theta_{si} = \Pr(s|x_{i0}, ..., x_{iT}) = \Pr(s|x_{i0})$. This case corresponds with the assumptions made about permanent unobserved heterogeneity in papers such as Keane and Wolpin (1997) and Carro and Mira (2006). If we do not place any restrictions on this probability, with a discrete x_{i0} variable that can take N_x values, there are $(S-1)*N_x$ parameters θ_{si}.

Case 3. θ_{si} depends on all the $T+1$ observations of x_{it}: $\theta_{si} = \Pr(s|x_{i0}, ..., x_{iT}) = F_\theta(d_{s0} + \sum_{t=0}^{T} d_{s1t}x_{it})$ where F_θ is a known cdf. Here there are $(S-1)*(T+2)$ parameters with one x_{it} variable. Hyslop (1999) is an example where this is the assumption made about unobserved heterogeneity.

Notice that in Case 3 we are using a semiparametric form. If we did not place any restrictions of this kind, 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 equation.

It is conceptually simple to extend our model if the additional covariates are discrete. For a single binomial covariate we have:

$$\begin{aligned}
P_{0s} &= \Pr(y_{i0} = 1 \mid x_{i0} = 0, s) \\
P_{1s} &= \Pr(y_{i0} = 1 \mid x_{i0} = 1, s) \\
G_{0s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = 0, s) \\
H_{0s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = 0, s) \\
G_{1s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = 1, s) \\
H_{1s} &= \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = 1, s)
\end{aligned} \tag{5.21}$$

These are 6 parameters of each of the S points of support. Additionally, we have a number of θ parameters which varies depending on which of the three possible cases mentioned we have. The probability of a path j given $\{x_{it}\}_{t=1}^{T}$ and

T	2	3	4	5	6	7	8
$r_{xit}(T,2)$	60	184	472	1056	2132	3976	6964
$r_{xit}(T,4)$	464	2656	12088	45888	151456	447648	1210032
$r_{xit}(T,6)$	1548	12984	84852	454104	2079840		

Table 5.1: Number of independent paths. Discrete covariate

$(P_s, G_{0s}, H_{0s}, G_{1s}, H_{1s})$ is:

$$p_{js} = P_{0s}^{y_0^j\left(1-x_0^j\right)}\left(1-P_{0s}\right)^{\left(1-y_0^j\right)\left(1-x_0^j\right)} P_{1s}^{y_0^j x_0^j}\left(1-P_{1s}\right)^{\left(1-y_0^j\right)x_0^j}$$
$$G_{0s}^{n_{01|0}^j}\left(1-G_{0s}\right)^{n_{00|0}^j} H_{0s}^{n_{11|0}^j}\left(1-H_{0s}\right)^{n_{10|0}^j} G_{1s}^{n_{01|1}^j}\left(1-G_{1s}\right)^{n_{00|1}^j} H_{1s}^{n_{11|1}^j}\left(1-H_{1s}\right)^{n_{10|1}^j} \tag{5.22}$$

where $n_{01|0}^j$ is the number of $y_{it-1}=0 \rightarrow y_{it}=1$ transitions given $x_{it}=0$ for path j, $n_{01|1}^j$ is the number of $y_{it-1}=0 \rightarrow y_{it}=1$ transitions for path j given $x_{it}=1$ for path j, and similarly for the other transitions. The number of possible paths 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. The number of different equations is

$$r_{xit}(T,2) = 4\left[(T+1) + \sum_{m=1}^{T}\sum_{q=0}^{m-1}(T-m+1)\left(\left\lceil\frac{m-q}{2}\right\rceil+1\right)\left(\left\lfloor\frac{m-q}{2}\right\rfloor+1\right)(q+1)\right] \tag{5.23}$$

where $\lceil x\rceil$ gives gives the smallest integer greater than or equal to x and $\lfloor x\rfloor$ gives the largest integer less than or equal to x. (5.23) is a particular case of (5.24) with $N_x=2$. In the appendix we proof the more general formula (5.24).

Table 5.1 shows this number for some T. Notice that $r_{xit}(T,2) \leq 2^{T+1} * r_T$.

Generalizing this to the case with a discrete covariate that takes N_x values, we have that the number of possible paths is $2^{N_x(T+1)}$. The number of different equations is

$$r_{xit}(T,N_x) = 2N_x\frac{(T+N_x-1)!}{T!\,(N_x-1)!} + 2N_x\sum_{m=1}^{T}\sum_{q=0}^{m-1}\frac{(T-m+N_x-1)!}{(T-m)!\,(N_x-1)!}\frac{\left(\left\lceil\frac{m-q}{2}\right\rceil+N_x-1\right)!}{\left(\left\lceil\frac{m-q}{2}\right\rceil\right)!\,(N_x-1)!}$$
$$\frac{\left(\left\lfloor\frac{m-q}{2}\right\rfloor+N_x-1\right)!}{\left(\left\lfloor\frac{m-q}{2}\right\rfloor\right)!\,(N_x-1)!}\frac{(q+N_x-1)!}{q!\,(N_x-1)!} \tag{5.24}$$

Table 5.1 shows this number for $N_x = 2$, 4, and 6 for several T. Notice that $r_{xit}(T,N_x)$ grows very fast with N_x.

The total number of P, G, H, and θ parameters to be identified in the three cases considered are:

Case 1. $3N_xS + S - 1 = (3N_x + 1)S - 1$

Case 2. $3N_xS + N_x(S-1) = 4N_xS - N_x$

Case 3. $3N_xS + (T+2)(S-1) = (3N_x + T + 2)S - (T+2)$

Therefore, the maximum number of points of support for the three different relations between θ and x are:

$$\frac{r_{xit}(T, N_x) + 1}{(3N_x + 1)} \tag{5.25}$$

$$\frac{r_{xit}(T, N_x) + N_x}{4N_x} \tag{5.26}$$

$$\frac{r_{xit}(T, N_x) + T + 2}{(3N_x + T + 2)} \tag{5.27}$$

Looking at (5.25), (5.26), and (5.27), the more values x_{it} takes or the more we discretise a continuous x_{it}, the richer the distribution we can point identify. Given that the limit of these expression goes to infinity as N_x grows, we could potentially identify as many points of support as we wish when we have a continuous covariate by discretising it in as many intervals as needed.

5.4.1. Semiparametric model

If x_{it} is a continuous covariate, or discrete taking many values, it is usually restricted with a parametric form the way different values of x_{it} affect the probabilities of $y_{it} = 1$. For example, for each point of support s:

$$\begin{aligned} P_{si} &= F_0(p_{s0} + p_{s1}x_{i0}) \\ G_{sit} &= F(g_{s0} + g_{s1}x_{it}) \\ H_{sit} &= F(h_{s0} + h_{s1}x_{it}) \end{aligned}$$

and

$$\begin{aligned} \theta_{si} &= F_\theta(d_{s0} + d_{s1}x_{i0}) \text{ or} \\ \theta_{si} &= F_\theta(d_{s0} + \sum_{t=0}^{T} d_{s1t}x_{it}) \end{aligned}$$

depending in whether we are in case 2 or 3 in the relation between θ and x. F_0, F and F_θ are known cdf functions, like 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.

The number of parameters does not depend on the number of values x_{it} can take:

Case 1. $6S + S - 1 = 7S - 1$

Case 2. $6S + 2(S-1) = 8S - 2$

Case 3. $6S + (T+2)(S-1) = (8+T)S - (T+2)$

The number of equations $r_{xit}(T, N_x)$ still depends on N_x and it is given by equation (5.24). The maximum number of points of support we can identify is

$$\frac{r_{xit}(T, N_x) + 1}{7} \tag{5.28}$$

$$\frac{r_{xit}(T, N_x) + 2}{8} \tag{5.29}$$

$$\frac{r_{xit}(T, N_x) + (T+2)}{(8+T)} \tag{5.30}$$

Therefore, with a continuous variable in this semiparametric model, we could potentially identify as many points of support as we wish, and for a given N_x there are important gains from the semiparametric assumption.

If we have K covariates, then

$$\begin{aligned} P_{si} &= F_0(p_{s0} + \sum\nolimits_{k=1}^{K} p_{sk} x_{ji0}) \\ G_{sit} &= F(g_{s0} + \sum\nolimits_{k=1}^{K} g_{sk} x_{kit}) \\ H_{sit} &= F(h_{s0} + \sum\nolimits_{k=1}^{K} h_{sk} x_{kit}) \end{aligned}$$

and similarly for θ_{si}. In terms of identification, a covariate $x_{it} = x_t$ for all i is the additional covariate that will help the least. This extra covariate x_t in a model with a continuous x_{it}, will imply two extra parameters in this setting. However it will not change the number of equations, which can be as large as we want. This means the previous result does not change when having more covariates.

T	2	3	4	5	6	7	8	...	23
Υ_T : No covariates	2.25	3.75	5.75	8.25	11.25	14.75	18.75		**138.75**
Covariate constant over time ($x_{it} = x_i$ for all t)									
Any N_x, free relation with θ	2.25	3.75	5.75	8.25	11.25	14.75	18.75		138.75
$N_x = 100$, independence of θ	2.66	4.65	7.31	10.63	14.62	19.27	24.59		184.06
$N_x = 100$, semiparametric	100.25	175.25	275.25	400.25	550.25	725.25	925.25		6925.3
Covariates $x_{it} = x_t$ for all i									
Time dummies	1.5	1.875	2.3	2.75	3.21	3.69	4.17		11.56
2 continuous x_t, semiparametric	1.125	1.875	2.875	4.125	5.625	7.375	9.375		**69.375**
Covariate that varies in both i and t									
$N_x = 2$, independent of θ	8.71	26.43	67.57	151	403.71	568.14	995		
$N_x = 2$, θ depends on x_{i0}	7.75	23.25	59.25	132.25	266.75	497.25	870.75		
$N_x = 2$, Case 3	6.4	17.18	39.83	81.77	152.86	265.67	435.88		
$N_x = 4$, independent of θ	35.77	204.38	929.9	3529.9	11650.5	34434.5	93079.5		
$N_x = 4$ and θ depends on x_{i0}	29.25	166.25	755.75	2868.25	9466.25	27987.25	75627.25		
$N_x = 4$, Case 3	29.25	156.53	671.89	2415.53	7573.2	21327	55001.9		
$N_x = 4$, semiparametric, Case 3	46.8	241.91	1007.83	3530.38	10818.86	29843.8	75627.6		
$N_x = 6$, semiparametric, Case 3	155.2	1180.8	7071.5	34931.6	148560.6				

Table 5.2: Maximum number of points of support for some representative cases

As in previous subsections, N_x is the number of possible values x can take. Case 3 has been described at the beginning of subsection 5.4. It refers to a situation where θ depends on $\{x_{it}\}_{t=1}^{T}$. Where semiparametrically is not specifically mentioned, a nonparametric first HFOM model with the indicated covariates is being considered.

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.[11] We create a dummy variable y_{it} which is set to unity if worker i has any unemployment in year t (and zero otherwise). The following Table gives some statistics for the sample.

	Number	Proportion
Total sample size	2571	–
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

Table 6.1: Incidence of unemployment

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, ℓ_{sat} (4.6), is $-12,252$. The value for the saturated HFOM model, ℓ_{res_sat}, (4.8), is $-17,449$. The likelihood ratio statistic, $2\left(\ell_{sat} - \ell_{res\_sat}\right)$, is thus $10,395$.[12] When estimating the mixture model we restrict the mixing probabilities $\theta_s \geq 0.01$ and we restrict G_s, H_s and P_s to be between 0.01and 0.99 to ensure that

[11] 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.

[12] 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.

S	df	LR stat	# $\theta'_s = 0.01$
2	547	1,063	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 6.2: Fit for different numbers of support points

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 ℓ_{res_sat}; that is, the LR statistic shown is $2\left(\ell_{res\_sat} - \ell^S_{mix}\right)$. 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 nine components; that is, $\ell^{10}_{mix} \simeq \ell^{\Upsilon}_{mix}$.

Since we are concerned to illustrate the mechanics of our method, we shall sidestep the issue of the distribution of the LR statistics and simply take a convenient value, $S = 5$. Table 6.3 presents the estimates for the model with 5 points of support. These display a number of features. First, all groups display positive state dependence ($H_s > G_s$). Second, the marginal dynamic effects ($H_s - G_s$) vary quite considerably across groups. The LR statistic for the hypothesis of a homogeneous marginal dynamic effect,

$$H_s = G_s + (H_1 - G_1) \text{ for } s = 2, ..5 \tag{6.1}$$

is 421; this is distributed as a $\chi^2(4)$ and represents a decisive rejection of this homogeneity assumption. Moreover the (weighted) correlation between G and H is -0.35; the conventional 'one fixed effect' assumption imposes that the correlation is positive so that even the qualitative implication is wrong for the homogeneous model.

To see the substantive implications of the estimates it is best to graph the implied paths for the probability of being unemployed at some time during the year. This is shown in the left panel of Figure 6.1 which graphs the probabilities implied by the Chapman-Kolomogrov equations for the five groups against age (or year, since all

	Probabilities				
Group	**P**	**G**	**H**	**M**	θ
	$p(y_0 = U)$	$p(U \mid E)$	$p(U \mid U)$	$H - G$	Proportion
1	0.27	0.01	0.87	0.86	0.34
2	0.64	0.10	0.69	0.59	0.28
3	0.01	0.03	0.48	0.46	0.24
4	0.73	0.36	0.82	0.46	0.08
5	0.25	0.18	0.34	0.16	0.06

Table 6.3: Parameter estimates with five support points

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 figure suggests a fascinating mix of workers who rarely experience unemployment (group 3), those who are very prone to unemployment (group 4) and those who start off badly, but quickly 'find their feet' (groups 2 and 1). However, there is evidence that the HFOM model does not fit the data well. This is shown in the right panel of the figure which shows the average proportions of unemployed for each year and the predicted mean from the model. 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 for the later period using. 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 Figure 6.1. To capture this time-inhomogeneity we turn to estimation adding the covariates to the model.

6.3. Model with covariates.

The right panel of Figure 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 (5.18).[13, 14] We continue to keep

[13]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.

[14]Other factors that we could take into account are other macro variables such as changes in the UI system; individual time varying factors such as health or marital status and individual time invariant factors such as parental background. Note that in this empirical illustration we have taken account of the time invariant factor, cohort.

S fixed at 5. 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 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 see, age and the aggregate unemployment rate are individually and jointly highly significant. Moreover, 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.

Test against SFOM		
Model	df	χ^2
Age and cycle	20	808
Age only	10	766
Cycle only	10	163

Table 6.4: Tests for 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 figure **??**. 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 group who have very low probabilities are hardly affected at all. However, the next prone group (with a starting value of 0.22) display considerable cyclical variation. However, the group who 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 heterogeneity would mask this effect.

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 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 nonparametric discrete mixture distribu-

tions. We establish necessary and sufficient conditions for point identification of our heterogeneity structure 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. The identification is strengthened if we have continuous covariates in the model. 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 size with fixed panel length are well known: it is consistent and efficient. We apply the techniques we study to a long 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.

References

[1] Alessie, R.; Hochguertel, S. and Soest, A. (2004): "Ownership of Stocks and Mutual Funds: A Panel Data Analysis." *Review of Economics and Statistics*, 86(3), pp. 783-96.

[2] Altonji J. G. and R. L. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors", *Econometrica*, 73(4), 1053-1112.

[3] Arellano, M. and B. H. Honoré (2001): "Panel Data Models: Some Recent Developments." *Handbook of Econometrics*, chapter 5, pp. 3229-96.

[4] Bajari B., J. T. Fox, K. I. Kim, and S. Ryan (2008): "A Simple Nonparametric Estimator for the Distribution of Random Coefficients in Discrete Choice Models", *unpublished manuscript*.

[5] Becker, G. S.; Grossman, M. and Murphy, K. M. (1994) "An Empirical Analysis of Cigarette Addiction." *American Economic Review*, 84(3), pp. 396-418.

[6] Bernard, A. B. and Jensen, J. B. (2004) "Why Some Firms Export." *Review of Economics and Statistics*, 86(2), pp. 561-69.

[7] Browning, M. and J. M. Carro (2006a). "Heterogeneity and Microeconometrics Modelling." Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, vol. 3.

[8] Browning, M. and J. M. Carro. (2006b): "Heterogeneity in Dynamic Discrete Choice Models.," Discussion Papers Series, number 207, Department of Economics, University of Oxford.

[9] Carro J. M. and P. Mira (2006). "A dynamic model of contraceptive choice of Spanish couples", *Journal of Applied Econometrics*, 21, 955-980.

[10] Chamberlain, G. (1984): "Panel Data", in Griliches, Z. and M.D. Intriligator (eds.) *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

[11] Chernozhukov V., I. Fernandez-Val, J. Hahn and W. K. Newey (2009), "Identification and Estimation of Marginal Effects in Nonlinear Panel Data", *unpublished manuscript*.

[12] Crawford G. S. and M. Shum (2005): "Uncertainty and Learning in Pharmaceutical Demand.", *Econometrica*, 73(4), 1137-1173.

[13] Feng, Z. D. and C. E. McCulloch (1996). "Using Bootstrap Likelihood Methods in Finite Mixture Models", *Journal of the Royal Statistical Society, Series B*, 58(3), 609-617.

[14] Fisher, F. M. (1966). *The Identification Problem in Econometrics*. New York. McGraw-Hill.

[15] Gottschalk, P. and R. A. Moffitt (1994): "Welfare Dependence - Concepts, Measures, and Trends." *American Economic Review*, 84(2), pp. 38-42.

[16] Ham, J. C. and L. Shore-Sheppard (2005): "The Effect of Medicaid Expansions for Low-Income Children on Medicaid Participation and Private Insurance Coverage: Evidence from the Sipp." *Journal of Public Economics*, 89(1), pp. 57-83.

[17] Heckman, J. J. (1981) "Heterogeneity and State Dependence." *Studies in Labor Markets*, ch. 31, pp. 91-140.

[18] Heckman, J. J. and Singer, B. (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric-Models for Duration Data." *Econometrica*, 52(2), pp. 271-320.

[19] Heckman J. , J. Smith and N. Clements (1997):"Making the most out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts", *Review of Economic Studies*, 64, 487-535.

[20] Honorè, B. E. and E. Tamer (2006), "Bounds on Parameters in Panel Dynamic Discrete Choice Models", *Econometrica*, 74(3), pp. 611-629.

[21] Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women", *Econometrica*, 67, 1255-1294.

[22] Kasahara, H. and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices" *Econometrica*, 77(1), 135-175

[23] Keane M. P. and K. I. Wolpin (1997): "The Career Decisions of Young Men", *Journal of Political Economy*, 105, 473-521.

[24] McLachlan, G. and D. Peel (2004), *Finite Mixture Models*, Wiley-Interscience.

[25] Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry.", *Econometrica*, 69(2), 307-342.

A. Proofs.

A.1. Number of 'independent' equations

Here we proof equation (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 $j = 1, ..., 2^{T+1}$ possible 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 this 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 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 $m = 0$, i.e. $\{(0, T, 0, 0, 0)\}$. Note that for this proof it is not necessary to write the 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 proof equation (5.24), 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

$$r_{xit}(T, N_x) = 2N_x \frac{(T+N_x-1)!}{T!\,(N_x-1)!} + 2N_x \sum_{m=1}^{T} \sum_{q=0}^{m-1} \frac{(T-m+N_x-1)!}{(T-m)!\,(N_x-1)!} \frac{\left(\left\lceil \frac{m-q}{2} \right\rceil + N_x - 1\right)!}{\left(\left\lceil \frac{m-q}{2} \right\rceil\right)!\,(N_x-1)!}$$
$$\frac{\left(\left\lfloor \frac{m-q}{2} \right\rfloor + N_x - 1\right)!}{\left(\left\lfloor \frac{m-q}{2} \right\rfloor\right)!\,(N_x-1)!} \frac{(q+N_x-1)!}{q!\,(N_x-1)!}$$

It can be seen in (5.22) that now we have to count the number of different sets $\left\{y_0^j, x_0^j, n_{00|1}^j, ..., n_{00|N_x}^j, n_{01|1}^j, ..., n_{01|N_x}^j, n_{10|1}^j, ..., n_{10|N_x}^j, n_{11|1}^j, ..., n_{11|N_x}^j\right\}$ that the $j = 1, ..., 2^{N_x(T+1)}$ 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 $n_{00|1}^j, ...,$ 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:

$$\frac{(n + N_x - 1)!}{n!\,(N_x - 1)!} \tag{A.2}$$

(A.2) 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.2) 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 equation (A.1), the first half of the r_T sets of $\{y_0, n_{00}, n_{01}, n_{10}, n_{11}\}$ are

$$\left\{ (0, T, 0, 0, 0)\,, \left\{ \left\{ \left(0, T - m, \left\lceil \frac{m-q}{2} \right\rceil, \left\lfloor \frac{m-q}{2} \right\rfloor, q \right) \right\}_{q=0}^{m-1} \right\}_{m=1}^{T} \right\} \tag{A.3}$$

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.2) and (A.3) we have that the number $r_{xit}(T, N_x)$ of possible sets of $\left\{y_0, x_0, n_{00|1}, ..., n_{00|N_x}, n_{01|1}, ..., n_{01|N_x}, n_{10|1}, ..., n_{10|N_x}, n_{11|1}, ..., n_{11|N_x}\right\}$ is given by equation (5.24) 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.

A.3. Rank of J matrix.

A.3.1. Decomposition of matrix A

From equations (2.7) and (3.4), any element of a row j of matrix $\mathbf{A}$ is given by $G^{n_{01}^j}(1-G)^{n_{00}^j} H^{n_{11}^j}(1-H)^{n_{10}^j}$ multiplied by $(1-P)$ for $j = 1, ..., \frac{\Gamma}{2}$ and multiplied by P for $j = \frac{\Gamma}{2} + 1, ..., \Gamma$. From the binomial theorem we have that

$$G^{n_{01}^j}(1-G)^{n_{00}^j} H^{n_{11}^j}(1-H)^{n_{10}^j} = \sum_{z=0}^{n_{10}^j} \sum_{x=0}^{n_{00}^j} (-1)^x (-1)^z \binom{n_{00}^j}{x} \binom{n_{10}^j}{z} G^{(x+n_{01}^j)} H^{(z+n_{11}^j)} \tag{A.4}$$

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

$$\mathbf{A} = \mathbf{CE} \tag{A.5}$$

where $\mathbf{C}$ will contain the coefficients $\left((-1)^x(-1)^z\binom{n_{00}^j}{x}\binom{n_{10}^j}{z}\right)$ of (A.4) and $\mathbf{E}$ will contain the corresponding G, H and P terms. The matrix C does not depend on the value of the parameters and, therefore, it will be unique for a given T.

$\mathbf{E}$ is the following $2e_T \times S$ matrix:

$$\mathbf{E} = \begin{bmatrix} (1-P_1)\mathbf{E}_1 & (1-P_2)\mathbf{E}_2 & ... & (1-P_S)\mathbf{E}_S \\ P_1\mathbf{E}_1 & P_2\mathbf{E}_2 & ... & P_S\mathbf{E}_S \end{bmatrix} \tag{A.6}$$

where

$$\mathbf{E}'_s = \begin{bmatrix} 1 & G_s & .. & G_s^T & H_s & G_sH_s & .. & G_s^{T-1}H_s & H_s^2 & .. & G_s^{T-2}H_s^2 & . & . & . & H_s^{T-1} & G_sH_s^{T-1} & H_s^T \end{bmatrix} \tag{A.7}$$

is a vector of dimension

$$e_T = \frac{(T+1)(T+2)}{2} \tag{A.8}$$

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 have the binomial coefficients from the path (i.e. the binary number with $T+1$ digits) that correspond with the decimal number $(j-1) : j = 1, ..., \frac{\Gamma}{2}$. 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 (A.4) plus the zeros needed to filling the rest of the row. 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

$$(Z + n_{11}^j)(T+2) - \frac{(z+n_{11}^j)(z+n_{11}^j+1)}{2} + x + 1 + n_{01}^j \tag{A.9}$$

of matrix $\mathbf{C}_0$.

Define $\mathbf{C}_1$ the same way as $\mathbf{C}_0$, but $j = \frac{\Gamma}{2}+1, ..., T$. Each coefficient of (A.4) is in column given by (A.9) and row $j - \frac{\Gamma}{2}$. Then,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{bmatrix} \tag{A.10}$$

The dimension of $\mathbf{C}$ is $\Gamma \times 2e_T$. From (A.4) and (A.9) matrix $\mathbf{C}$ can be easily

computed for any given T. For example, with $T = 2$

$$\mathbf{C} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.11}$$

with dimension 8×12.

A.3.2. The rank of A.

It is important to note that $\mathbf{C}$ does not depend on S, G, H or any other unknown value. It only depends on T, so we can calculate $rank(\mathbf{C})$ for any given T, using (A.4) and (A.9). Table 2.2 reports the $rank(\mathbf{C})$, for $T = 2, ..., 23$. For all those values of T, the rank of $\mathbf{C}$ is the number of equations that are different in the system, r_T:

$$r_T = T(T+1) + 2 \tag{A.12}$$

We now can use the following two results about the rank of a product of two matrices:

$$rank(\mathbf{A}) \leq \min(rank(\mathbf{C}), rank(\mathbf{E})) \leq \min(rank(\mathbf{C}), 2e_T, S) = \min(r_T, S) \tag{A.13}$$

$$rank(\mathbf{A}) \geq rank(\mathbf{C}) + rank(\mathbf{E}) - 2e_T \tag{A.14}$$

where (A.14) comes from the Frobenius rank inequality. Note that $r_T = T(T+1)+2$ is smaller than $2e_T = (T+1)(T+2)$.

The problem is that $rank(\mathbf{E})$ depends on the values of the unknowns $\mathbf{P}, \mathbf{G}, \mathbf{H}$. For instance, for the special case with $P_1 = ... = P_S$ (S being large), we have the rank of $\mathbf{E}$ is reduced so that $rank(\mathbf{E}) = e_T$; and thus $r_T - e_T \leq rank(\mathbf{A}) \leq r_T$. However, for many of the possible values of $\{P_s, G_s, H_s\}_{s=1}^{S}$ the rank of $\mathbf{A}$ will be equal to $\min(r_T, S)$. Simulating many times the matrix $\mathbf{A}$ with large values of S ($S \geq \Gamma$) and random draws for the the P_s's, G_s's and H_s's we found that the rank of $\mathbf{A}$ is given by: $r_T = T(T+1) + 2$.

A.3.3. The rank of J

From (3.5) and (A.5), for given S we have a mapping from unobservables to observables given by:

$$\begin{aligned}\pi\left(\mathbf{P},\mathbf{G},\mathbf{H},\theta_1,..\theta_S\right) &= \mathbf{A}\left(\mathbf{P},\mathbf{G},\mathbf{H}\right)*\theta \\ &= \mathbf{C}*\mathbf{E}\left(\mathbf{P},\mathbf{G},\mathbf{H}\right)*\theta \end{aligned} \tag{A.15}$$

where the θ S-vector is normalized to sum to unity by setting the last value equal to the sum of the first $S-1$ values. The Jacobian of this is a $\Gamma \times (4S-1)$ matrix which we denote $J\left(T,S\right)$. For local point identification we require that the rank of $J\left(T,S\right)$ is greater than or equal to the number of parameters.

$(\mathbf{E}*\theta)$ is a column vector of dimension $2e_T$. The Jacobian J can be written as

$$J = \mathbf{C}*D(\mathbf{E}*\theta) \tag{A.16}$$

where $D(\mathbf{E}*\theta)$ is the Jacobian of $(\mathbf{E}*\theta)$. The dimension of $D(\mathbf{E}*\theta)$ is $2e_T \times 4S-1$. Then, from results about the rank of the product of two matrices we have:

$$rank(J) \leq \min\left(rank(\mathbf{C}), rank\left(D(\mathbf{E}*\theta)\right)\right) \leq \min\left(rank(\mathbf{C}), 2e_T, 4S-1\right) \tag{A.17}$$

$$rank(J) \geq rank(\mathbf{C}) + rank\left(D(\mathbf{E}*\theta)\right) - 2e_T \tag{A.18}$$

The general form of $D(\mathbf{E}*\theta)$ for a given T is

$$\left[\begin{array}{ccccccccc} ... & -\mathbf{E}_s\theta_s & ... & (1-P_s)\frac{\partial \mathbf{E}_s}{\partial G_s}\theta_s & ... & (1-P_s)\frac{\partial \mathbf{E}_s}{\partial H_s}\theta_s & ... & (1-P_l)\mathbf{E}_l - (1-P_S)\mathbf{E}_S & ... \\ ... & \mathbf{E}_s\theta_s & ... & P_s\frac{\partial \mathbf{E}_s}{\partial G_s}\theta_s & ... & P_s\frac{\partial \mathbf{E}_s}{\partial H_s}\theta_s & ... & P_l\mathbf{E}_l - P_S\mathbf{E}_S & ... \end{array}\right] \tag{A.19}$$

where $\mathbf{E}_s$ is in equation (A.7), $s = 1, ..., S$ and $l = 1, ..., S-1$.

The rank of $\mathbf{C}$ has already been calculated on previous subsection. For most of the possible values of $\{\mathbf{P},\mathbf{G},\mathbf{H},\theta\}$ the rank of $D(\mathbf{E}*\ \theta)$ is equal to $\min(2e_T - 1, 4S-1)$.[15] One exception is the case with $P_s = 1 - P_s = 0.5$ for all $s = 1, ..., S$, where $D(\mathbf{E}*\ \theta)$ has a smaller rank Compared with $rank(\mathbf{E})$, the condition that $P_1 = ... = P_S$ is not enough to give a reduced rank of $D(\mathbf{E}*\ \theta)$. Given this, from equations (A.17) and (A.18) and previous calculations of $rank(\mathbf{C})$ $(= r_T)$ we have that for most of the possible values of $\{\mathbf{P},\mathbf{G},\mathbf{H},\theta\}$

$$\min\left(r_T - 1, 4S-1\right) \leq rank(J) \leq \min\left(r_T, 4S-1\right) \tag{A.20}$$

[15] Notice that in $D(E*\ \theta)$ row $e(T)+1$ is minus the first row.

because $r_T = T(T+1)+2$ is strictly smaller than $2e_T - 1 = (T+1)(T+2) - 1$ for any $T \geq 1$. As a matter of fact simulations suggest a general form: $rank\,(J) = \min\,(4S-1, r\,(T))$ for any (S,T).

A.4. Identification for each T.

A.4.1. Proof of proposition 3.1

The sufficient condition (part (ii)) in proposition 3.1 is a direct application of the general inverse theorem. For local point identification (i.e. unique solution to system (3.5)) it requires that the rank of J be equal to the number of unknown parameters. According with the bounds we have found, the rank of J is greater than or equal to $\min\,(r_T - 1,$ number of unknown parameters). Therefore, the requirement for this case is that the number of unknowns be smaller than or equal to $r_T - 1$.

To obtain the necessary condition (part (i)) in proposition 3.1 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). From our calculation it can be seen that rank of J is the same for points considered in this theorem, and that this rank is smaller than or equal to r_T. Therefore, for local identification it is necessary that the number of unknowns be smaller than or equal to r_T.

A.4.2. Proof of proposition 3.2

Firstly note that with S points of support there are $4S-1$ unknown parameters to be identified from a system with a maximum rank of J between $r_T - 1$ and r_T. Secondly, note that $r_T + 1$ is always an odd number. This implies that

$$\text{integer}\left[\frac{r_T+1}{4}\right] = \text{integer}\left[\frac{r_T}{4}\right]$$

Then, Proposition 3.2 follows from Proposition 3.1.

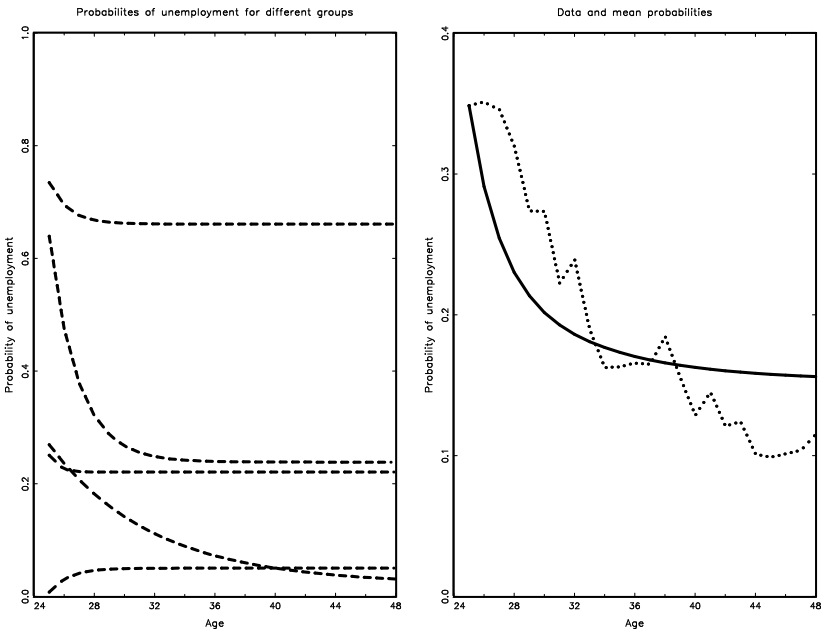

Figure 6.1: Probabilities with 5 points of support.

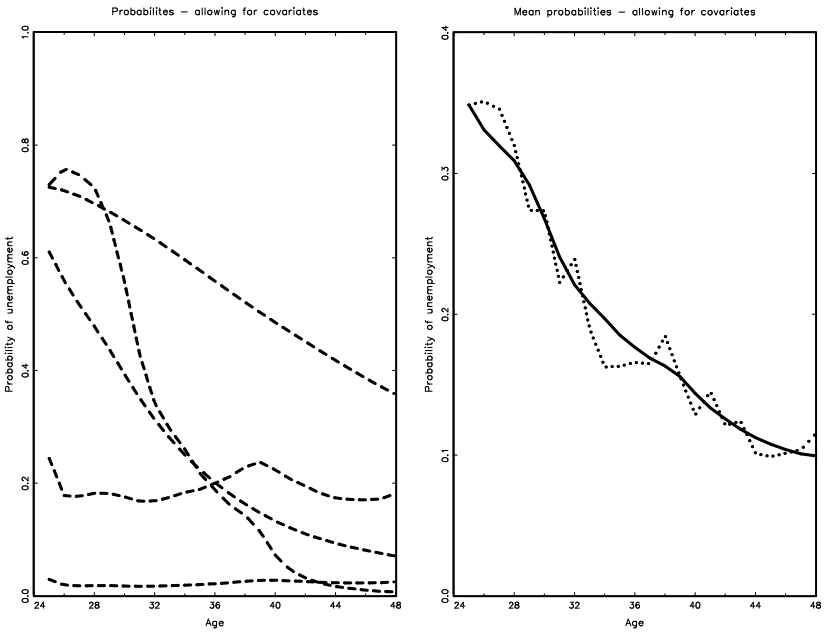

Figure 6.2: Probabilities with age and cyclical effects.