Dynamic binary outcome models with maximal heterogeneity.*

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

Jesus Carro
Departmento de Economia,
Universidad Carlos III de Madrid,
Madrid 28903,
Getafe (Madrid), Spain
jcarro@eco.uc3m.es

June 2008
JEL classification: C23, C24, J64
Keywords: discrete choice, Markov processes, nonparametric identification, unemployment dynamics,

This draft is incomplete. Please do not quote.

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 present an estimator for the time homogeneous first order Markov (HFOM) model that allows for maximal heterogeneity. That is, the modelling of the heterogeneity does

*For comments and suggestions, we thank participants at seminars at Nuffield (Oxford); IFS (London); CEMFI; Manchester; Columbia and a conference in the Tinbergen Institute.

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 two transition probabilities is unrestricted. We establish exact conditions for the point identification of our heterogeneity structure and show how it depends on the length of the panel. We derive tests for a variety of subsidiary hypotheses such as the conventional assumption that marginal dynamic effects are homogeneous. We apply the techniques we develop to a long panel of Danish workers who are very homogeneous in terms of observables. We show that individual unemployment dynamics are very heterogeneous, even for such a homogeneous group. We also show that the impact of cyclical variables on individual unemployment probabilities differs widely across workers. Some workers have unemployment dynamics that are independent of the cycle whereas others are highly sensitive to macro shocks.

1. Introduction.

Models with a binary outcome that depends in part on previous realisations 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 (Ham and Shore-Sheppard (2005)). Ignoring covariates, the usual time-homogeneous first order Markov model for unit i $(= 1, ..H)$ in period t $(t = 0, ..T)$ is:

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

where $F(.)$ is a probability distribution function and y_{it} is an indicator for person i having some unemployment in period t. This 'one latent heterogeneous variable' model which only allows for a heterogeneous 'intercept' is widely used but it does have problems; Browning and Carro (2006) discuss these but it is worth repeating the objections.

The first problem is that the imposition of a common slope parameter (α) restricts the class of structural models that are consistent with the reduced form (1.1). Consider, for example, two people a and b for whom a has a lower proba-

bility of being unemployed if they were employed in the previous year:

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

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

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

This rules out, for example, that a's caution leads her to spend more time looking for a 'safe' job, so that her probability of remaining unemployed is *higher* than b's. Thus the choice of a statistical 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 ($T > 20$, say), we do find substantial heterogeneity in the two parameters in (1.4). This will be illustrated in our empirical illustration using Danish data. Here the binary variable is 'having a spell of unemployment in a given year' (see Hyslop (1999)).

The time-homogeneous first order form with maximal heterogeneity has:

$$pr(y_{it} = 1 \mid y_{i,t-1}) = F(\eta_i + \alpha_i y_{it-1}) \tag{1.4}$$

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) and it will fit any data that is generated by heterogenous workers with time-homogeneous first order Markov process (HFOM) processes.

Given the difficulties in estimating (1.1) with small and fixed T (see Arellano and Honoré (2001)), tackling (1.4) is a formidable task. In Browning and Carro (2008) we suggested two estimation methods that rely on reducing the bias or rms 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 (2008), identification and estimation of (1.4) without imposing any restriction on the distribution of (η, α), nor on the initial condition, rely on the T dimension. In this paper we propose an alternative approach that relies on large$-H$. The starting point here is that this model is not nonpara-

metrically identified only from the cross-section observations, which is what we would like to do since we do not have a large number of periods. Nevertheless, this negative result on identification does not imply that we can not learnt anything from a cross-section of paths with a given T. In general, some restrictions will have to be imposed on the distribution of the heterogeneity to get point identification. The interesting question is how many restrictions we have to impose, or how much information about our model with maximal heterogeneity we can identified from a cross-section of paths of length T.]

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.

- [If the conditions are satisfied, we provide the limits 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 like the initial condition or the correlation between η and α.]

- We provide maximum likelihood procedures to estimate the distributions if the point identification conditions are imposed.

- We provide likelihood ratio tests for important special cases: the conventional 'homogeneous slope' assumption and that the initial distribution is the long run distribution.

- We provide an LM test against the obvious general alternative, a time-homogeneous second order process.

- We show how to include covariates with maximal heterogeneity.

- 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 a dynamic discrete choice model with no covariates. The data consist of paths $\{y_{i0}, y_{i1}, ...y_{iT}\}_{i=1,2,...H}$ 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:

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

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

$$P_i = pr\left(y_{i0} = 1\right) \tag{2.2}$$

This direct formulation is much more convenient to work with than the usual econometric specification given in (1.4). The values of the parameters (P_i, G_i, H_i) are not usually of primary interest; rather they can be used to generate other 'outcomes of interest'. There are several candidates but the most widely considered are the *marginal dynamic effects*:

$$\begin{aligned} M_i &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x\right) - \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x\right) \\ &= H_i - G_i \end{aligned} \tag{2.3}$$

and *the long run proportion of unit values*:

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

The other 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. 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:

$$p_j(P,G,H) = P^{y_0^j}(1-P)^{\left(1-y_0^j\right)} G^{n_{01}^j}(1-G)^{n_{00}^j} H^{n_{11}^j}(1-H)^{n_{10}^j} \tag{2.5}$$

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 ordering for the elements for $t = 1, 2...T$. 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:

$$\pi_j = \int_0^1 \int_0^1 \int_0^1 p_j(P,G,H) f(P,G,H) dP dG dH, \; j = 1, 2...\Gamma \tag{2.6}$$

In the terminology of integral equations, the probabilities $p_j(P,G,H)$ are known as kernels. Since the $p_j's$ and the π_j's sum to unity, $f(.)$ will be a well defined

density:

$$1 = \sum_{j=1}^{\Gamma} \pi_j = \int_0^1 \int_0^1 \int_0^1 \sum_{j=1}^{\Gamma} p_j (P, G, H) f (p, G, H) \, dP dG dH$$
$$= \int_0^1 \int_0^1 \int_0^1 f (P, G, H) \, dP dG dH \tag{2.7}$$

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 (2.6) 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 (2.5) and (2.6) we have the following equality restrictions:

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

$$y_0^j = y_0^{j'}, n_{00}^j = n_{00}^{j'}, n_{01}^j = n_{01}^{j'}, n_{10}^j = n_{10}^{j'}, n_{11}^j = n_{11}^{j'} \tag{2.8}$$

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 number of transitions and hence the same probability,

$$\pi_{0010} = \pi_{0100} =$$
$$\int_0^1 \int_0^1 \int_0^1 \left((1-P)(1-G) HG f(P,G,H) \right) dP dG dH, \; j = 1, 2...\Gamma \tag{2.9}$$

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

$$p_2(P,G,H) = (1-P)(1-G)G \leq 0.25 \tag{2.10}$$

Thus we have:

$$\pi_2 = \int_0^1 \int_0^1 \int_0^1 p_2(P,G,H) f(P,G,H) dP dG dH \leq 0.25 \tag{2.11}$$

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 characterise 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 $f(.)$. 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 R_T gives the number of restrictions. For the values of T given, we have $r_T = T(T+1) + 2$. 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

# 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

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 (2.6) as:

$$\pi_k = n_k \int_0^1 \int_0^1 \int_0^1 p_k(P, G, H) f(P, G, H) dP dG dH, \; k = 1, 2...r_T \tag{2.12}$$

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 will be 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 (2.12). 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.[1] 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:

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

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

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

In general we require that this matrix has a rank of at least L, so that a necessary condition for 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). Alternatively, we could set identify the nine parameters. 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.

3.2. Identification for the nonparametric discrete scheme.

The second broad alternative assumption is that we have a *discrete 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

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

by the $(S \times 1)$ vector θ with non-negative individual values, θ_s, that sum to unity. The discrete analogue to (2.6) is:

$$\begin{aligned} \pi_j &= \textstyle\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s \quad j = 1, 2, ...\Gamma \\ &= \textstyle\sum_{s=1}^{S} n_k p_k\left(P_s, G_s, H_s\right)\theta_s \quad k = 1, 2, ...r_T \end{aligned} \tag{3.3}$$

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

$$A_{ks} = n_k p_k\left(P_s, G_s, H_s\right), \; k = 1, 2...r_T, \; s = 1, 2...S \tag{3.4}$$

so that (2.12) 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.[2] 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?

From (3.5) and (A.2), 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 normalised 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 the Appendix A.1 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, a great number of simulations with random values for $(G_s, H_s, P_s, \theta_s)$ suggest that this bound could be tightened to:

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

If we have S points of support then we have $4S - 1$ free parameters (one θ_s

[2] An alternative that does not sem to work very well is to take grid points for $\{(P_1, G_1, H_1), ...\left(P_S, G_S, H_S\right)\}$ and then to estimate the weights θ_s.

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

is determined by the others). The parameters of these support points and their probabilities can only be point identified if we have fewer parameters 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:

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

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.

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 suggests 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 the one route to nonparametric point identification.

4. Testing against alternative models.

4.1. Homogenous marginal dynamic effect.

We shall not consider the homogeneous case with (G, H, P) the same for everyone, since it is never considered a possibility. A less restricted model than the 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 (**??**). A close analogue here is that we have a homogeneous dynamic marginal effect:

$$H_i = M + G_i \text{ for some constant } M \in [-1, 1] \tag{4.1}$$

We term this the homogeneous marginal effect model (HME). This has $3S$ parameters (the P_s's, the H_s's, M and $S-1$ probabilities, θ_s) which give stronger restrictions than the nonparametric discrete , model for $S > 1$. The rank of the matrix J for the HME model is $\min(3S-1, r_T)$, where r_T is in Table 3.1. This is as in (3.7) of the unrestricted case, but accounting for the fact that here we only need $3S$ parameters for S points of support, rather than $4S-1$.

4.2. Long run initial value.

If the process has been running for some time before we first observe it, we might wish to impose that the probability of a unit initial value is given by the long run probability:

$$P_s = \frac{G_s}{1 + G_s - H_s} \tag{4.2}$$

This gives revised probabilities for path j of:

$$p_j(P, G, H) = \left(\frac{G}{1+G-H}\right)^{y_0^j} \left(\frac{1-H}{1+G-H}\right)^{\left(1-y_0^j\right)} * G^{n_{01}^j} (1-G)^{n_{00}^j} H^{n_{11}^j} (1-H)^{n_{10}^j} \tag{4.3}$$

This is testable against the unrestricted form.

In the Appendix A.2 we show that for most of the possible values of $\{\mathbf{G}, \mathbf{H}, \theta\}$

$$\min(r_T - 1, 3S-1) \leq rank(J) \leq \min(r_T, 3S-1) \tag{4.4}$$

where r_T is in table 4.1. It is important to note that r_T is different from the previous case where the distribution of the initial observation was unrestricted. Table 4.1 presents the maximum number of support points that can be point

T	2	3	4	5	6	7	8	9	10
Γ	8	16	32	64	128	256	512	1024	2048
r_T	5	8	12	17	23	30	38	47	57
$\Upsilon(T)$	2	3	4.3	6	8	10.3	13	16	19.3

Table 4.1: Identification with a long-run initial value

identified, Υ_T, in this case.

4.3. 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:

$$\begin{aligned} &p_j\left(P_{00s}, P_{01s}, P_{10s}, G_{00s}, G_{10s}, H_{01s}, H_{11s}\right) = \\ &P_{00s}^{1\left(y_0^j=0, y_1^j=0\right)} P_{01s}^{1\left(y_0^j=0, y_1^j=1\right)} P_{10s}^{1\left(y_0^j=1, y_1^j=0\right)} * \\ &\left(1-P_{00s}-P_{01s}-P_{10s}\right)^{1\left(y_0^j=1, y_1^j=1\right)} G_{00}^{n_{001}^j}\left(1-G_{00}\right)^{n_{000}^j} G_{10}^{n_{101}^j} * \\ &\left(1-G_{10}\right)^{n_{100}^j} H_{01}^{n_{011}^j}\left(1-H_{01}\right)^{n_{010}^j} H_{11}^{n_{111}^j}\left(1-H_{11}\right)^{n_{110}^j} \end{aligned} \tag{4.5}$$

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

$$P_{01} = \Pr\left(y_{i0}=0, y_{i1}=1\right), \tag{4.6}$$

$$G_{10} = \Pr\left(y_{it}=1 \mid y_{it-2}=1, y_{it-1}=0\right), \tag{4.7}$$

$$H_{01} = \Pr\left(y_{it}=1 \mid y_{it-2}=0, y_{it-1}=1\right), \tag{4.8}$$

$$...$$

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 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.[3] To

[3]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.

perform the LM test we have to:

1. Derive the log-likelihood with respect to

$$\{P_{00s}, P_{01s}, P_{10s}, G_{00s}, G_{10s}, H_{01s}, H_{11s}\}_{s=1}^{S} \text{ and } \{\theta_s\}_{s=1}^{S-1} \tag{4.9}$$

   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 $\left(\{P_s, G_s, H_s\}_{s=}^{S}, \{\theta_s\}_{s=1}^{S-1}\right)$. This means that we evaluate $g(.)$ and $h(.)$ at

$$\begin{aligned} P_{00s} &= \left(1-\widehat{P}_s\right)\left(1-\widehat{G}_s\right) \\ P_{01s} &= \left(1-\widehat{P}_s\right)\widehat{G}_s \\ P_{10s} &= \widehat{P}_s\left(1-\widehat{H}_s\right) \\ G_{00s} &= G_{10s} = \widehat{G}_s \\ H_{01s} &= H_{11s} = \widehat{H}_s \end{aligned} \tag{4.10}$$

   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

$$LM = \widehat{g}'\widehat{h}^{-1}\widehat{g} \tag{4.11}$$

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

$$b = (7S + S - 1) - (3S + S - 1) = 4S \tag{4.12}$$

   which is the number of restrictions a first order model is imposing with respect to a second order model.

4.4. 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. Estimation.

5.1. 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. That is, solve the set of nonlinear equations:

$$\pi = \mathbf{A}\theta \tag{5.1}$$

In practice, it is much better 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:

$$\pi_j = \sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s \quad j = 1, 2, ...\Gamma \tag{5.2}$$

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:

$$\prod_{i=1}^{N}\prod_{j=1}^{\Gamma}\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right)^{\delta_{ij}} = \prod_{j=1}^{\Gamma}\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right)^{n_j} \tag{5.3}$$

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:

$$\ell_{mix} = \sum_{i=1}^{N}\sum_{j=1}^{\Gamma} \delta_{ij} \log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right) \tag{5.4}$$

$$= N\sum_{j=1}^{\Gamma} c_j \log\left(\sum_{s=1}^{S} p_j\left(P_s, G_s, H_s\right)\theta_s\right) \tag{5.5}$$

Note that N is irrelevant for the maximization. With an iid random sample in N, as $N \to \infty$ this $c_j \to \pi_j$ 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, how to test different models.

5.2. The unrestricted model.

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

$$\begin{aligned} &S = \Gamma, \mathbf{A} = I, \theta = \pi \text{ or} \\ &S = 1, \mathbf{A} = \pi \end{aligned} \tag{5.6}$$

These both give the likelihood value:

$$\begin{aligned} \ell_{sat} &= \sum_{i=1}^{N}\sum_{j=1}^{\Gamma} \delta_{ij} \log\left(c_j\right) \\ &= N\sum_{j=1}^{\Gamma} c_j \log\left(c_j\right). \end{aligned} \tag{5.7}$$

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 5.7 to 5.5 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 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.

5.3. The unrestricted HFOM model.

We can also write a closed form expression for the model with the HFOM equality restrictions form subsection (2.4) imposed, using equation (2.12). 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:

$$\hat{c}_j = \frac{1}{n_{k(j)}} \sum\nolimits_{j \in k(j)} c_j \tag{5.8}$$

That is, we replace the unrestricted proportions for each path by the mean for the group. 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:

$$\hat{c}_3 = \hat{c}_5 = \frac{c_3 + c_5}{2} \tag{5.9}$$

$$\hat{c}_{12} = \hat{c}_{14} = \frac{c_{12} + c_{14}}{2} \tag{5.10}$$

$$\hat{c}_j = c_j, \text{ all other } j \tag{5.11}$$

The likelihood function is then given by:

$$\begin{aligned} \ell_{hfom} &= \sum_{i=1}^{N} \sum_{j=1}^{\Gamma} \delta_{ij} \log(\hat{c}_j) \\ &= N \sum_{j=1}^{\Gamma} c_j \log(\hat{c}_j) \end{aligned} \tag{5.12}$$

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_{hfom}. 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 the this mixture model by ℓ^{Υ}_{mix}. Now consider a model with fewer than the maximum number of points of support: $S < \Upsilon_T$. We have the following ordering for the likelihood function values:

$$\ell_{sat} \geq \ell_{hfom} \geq \ell^{\Upsilon}_{mix} \geq \ell^{S}_{mix} \tag{5.13}$$

Once again, 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.[4]

If we do 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. For example, particularly bad news about smoking might induce permanent changes in the both the propensity to start smoking and the propensity to stop. If this news were common across agents then we could test whether the first order conditions are satisfied for $T = 3$, $T = 4$ and so. A second variant has that the transition probabilities depend on observable time-varying covariates. We consider that in the next section.

5.4. Computational issues.

The foregoing models are all relatively simple to estimate. First define an $N \times \Gamma$ matrix Δ that has $\Delta_{ij} = 1$ if person i has path j and zero otherwise. For the unrestricted (saturated) model we have:

$$l_{sat} = sum_{i=1}^{N} \left(\Delta \ln \left(\mathbf{c} \right) \right) \tag{5.14}$$

where $\ln(\mathbf{c})$ is the Γ-vector of observed proportions for each possible path. The saturated HFOM model we have:

$$l_{sfom} = sum_{i=1}^{N} \left(\Delta \ln \left(\hat{\mathbf{c}} \right) \right) \tag{5.15}$$

where the vector of c_j's defined before equation (5.12). Finally the mixture model log likelihood is written:

$$\ell_{mix} \left(\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta \right) = sum_{i=1}^{N} \left(\Delta \ln \left(\sum_{s=1}^{S} p_j \left(P_s, G_s, H_s \right) \theta_s \right) \right) \tag{5.16}$$

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

These are all very simple to program and use in standard maximum likelihood routines.

6. Allowing for covariates.

6.1. Discrete covariates.

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

$$\begin{aligned} P_0 &= \Pr\left(y_{i0} = 1 \mid x_{i0} = 0\right) \\ P_1 &= \Pr\left(y_{i0} = 1 \mid x_{i0} = 1\right) \\ G_{0ti} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = 0\right) \\ H_{0ti} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = 0\right) \\ G_{1ti} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 0, x_{it} = 1\right) \\ H_{1ti} &= \Pr\left(y_{it} = 1 \mid y_{i,t-1} = 1, x_{it} = 1\right) \end{aligned} \tag{6.1}$$

For more discrete covariates or for covariates that take on more than two values, we need many more periods to identify the discrete distribution. An alternative would be to impose some homogeneity across units in the effect of the covariates. This leads onto the next subsection.

6.2. Semiparametric estimation with covariates.

If we have K covariates for unit i in period t, $(x_{1ti}, x_{2ti}, ..x_{Kti})$, we use the following semiparametric form with time varying transition probabilities for each point of support:

$$\begin{aligned} G_{sti} &= \frac{\exp\left(g_{s0} + \sum_{k=1}^{K} g_{sk} x_{1ti}\right)}{1 + \exp\left(g_{s0} + \sum_{k=1}^{K} g_{sk} x_{1ti}\right)} \\ H_{sti} &= \frac{\exp\left(h_{s0} + \sum_{k=1}^{K} h_{sk} x_{1ti}\right)}{1 + \exp\left(h_{s0} + \sum_{k=1}^{K} h_{sk} x_{1ti}\right)} \end{aligned}$$

Since we have time varying transitions, the number of transitions n_{00}^{i} are no longer sufficient statistics. To write the likelihood function, first define the time

specific transitions by:

$$Q_{sti} = (G_{sti})^{1(y_{it}=0,y_{it}=1)} (1 - G_{sti})^{1(y_{it}=0,y_{it}=0)} (H_{sti})^{1(y_{it}=1,y_{it}=1)} (1 - H_{sti})^{1(y_{it}=1,y_{it}=0)} \tag{6.2}$$

Then the log likelihood for S points of support is given by:

$$\ell_{cov}^{S} = \sum\nolimits_{i=1}^{N} \ln \left[\sum\nolimits_{s=1}^{S} \left((P_s)^{yi0} (1 - P_s)^{1-yi0} \prod\nolimits_{t=1}^{T} Q_{sti} \right) \right] \tag{6.3}$$

We present a specific example in the empirical illustration that follows.

7. An empirical illustration.

7.1. Sample selection.

We consider the incidence of unemployment in a year for workers in Denmark from 1980 to 2003. 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.[5] 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 7.1: Incidence of unemployment

[5] 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.

S	df	LR stat	# $\theta_s = 0.01$
2	547	1063	0
3	543	701	0
4	539	604	0
5	535	534	0
6	531	509	0
7	527	499	0
8	523	492	0
9	519	489	1
10	515	489	2

Table 7.2: Fit for different numbers of support points

7.2. 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 maximum number of support points we could take is 138 (see Table 3.1). For computational convenience, we restrict the mixing probabilities $\theta_s \geq 0.01$ to ensure that we do not assign zero probability to any path so that the practical limit is 100 points of support.[6] In practice, we cannot find more than a much smaller number than this; see Table 7.2. In this table we provide the LR statistic for mixture models relative to the upper bound, ℓ_{hfom}, as given by equation (5.12). 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. As already discussed, the differences between successive values are *not* distributed as a $\chi^2(4)$ so that choosing a value for S is controversial. Since we are concerned to illustrate the mechanics of our method, we shall sidestep these complications and simply take a convenient value, $S = 5$.

Table 7.3 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,

$$H_s = G_s + (H_1 - G_1) \text{ for } s = 2, ..5 \tag{7.1}$$

is 421; this is distributed as a $\chi^2(4)$ and represents a decisive rejection of the con-

[6] We also restrict G_s, H_s and P_s to between 0.01and 0.99.

ventional homogeneity assumption. Moreover the (weighted) correlation between G and H is -0.35; the conventional 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 7.1 which graphs the probabilities implied by the Chapman-Kolomogrov equations for the five groups against age (or age, since all the workers in the sample are in the same birth cohort). 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 (6.3). 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 7.1. To capture this time-inhomogeneity we turn to estimation using the covariate model above.

	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	01.6	0.06

Table 7.3: Parameter estimates with five support points

7.3. Model with covariates.

The right panel of Figure 7.1 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; see equation (6.3).[7] We continue to keep 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 7.4 presents the LR statistics against the model with 5 points of support and no covariates.

The $\chi^2(10)$ statistic for the stability test used in the previous subsection is 36; although formally this is a rejection of the use of a linear age term to pick up the time inhomogeneity, it is a considerable improvement on the model without age effects.

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

Table 7.4: Tests for age and cyclical effects

Group			**age**		**cycle**	
	estimates		**t-value**			
	P	θ	**G**	**H**	**G**	**H**
1 2 3						
4						
5						

Table 7.5: t-values for covariates

8. Conclusions.

References

[1] Alessie, R.; Hochguertel, S. and Soest, A. "Ownership of Stocks and Mutual Funds: A Panel Data Analysis." Review of Economics and Statistics, 2004,

[7] 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 the empirical illustration taken here we have taken account of the time invariant factor, cohort.

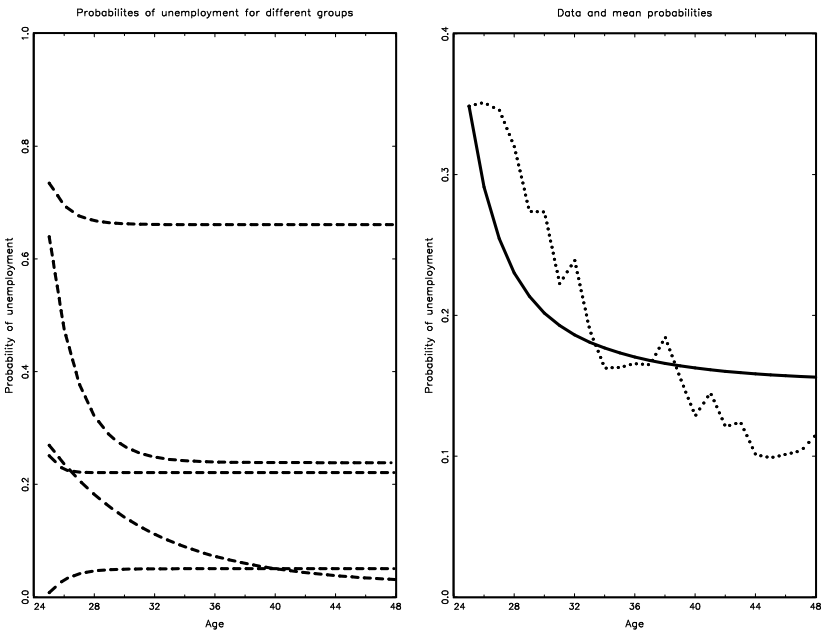

Figure 7.1: Probabilities with 5 points of support.

Figure 7.2: Probabilities with age and cyclical effects.

86(3), pp. 783-96.

[2] Arellano, M. and Honoré, B. "Panel Data Models: Some Recent Developments." Handbook of Econometrics, 2001, 5, pp. 3229-96.

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

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

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

[6] Browning, Martin and Jesus Carro. "Heterogeneity in Dynamic Discrete Choice Models.," Oxford: University of Oxford, 2008.

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

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

[9] Heckman, J. J. "Heterogeneity and State Dependence." Studies in Labor Markets, 1981, 31, pp. 91-140.

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

[11] Hyslop, D. R. "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women." Econometrica, 1999, 67(6), pp. 1255-94.

[12] McLachlan, G. and Peel, D. Finite Mixture Models. Wiley-Interscience, 2004.

A. Proofs.

A.1. Rank of J matrix.

A.1.1. Decomposition of matrix A

From equations (2.5) 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

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

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

$$\mathbf{A}=\mathbf{CE} \tag{A.2}$$

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.1) 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:

$$\mathbf{E}=\begin{bmatrix}(1-P_1)\mathbf{E}_1 & (1-P_2)\mathbf{E}_2 & ... & (1-P_S)\mathbf{E}_S\\ P_1\mathbf{E}_1 & P_2\mathbf{E}_2 & ... & P_S\mathbf{E}_S\end{bmatrix} \tag{A.3}$$

where

$$\mathbf{E}_s'=\begin{bmatrix}1 & G_s & .. & G_s^T & H_s & G_sH_s & .. & G_s^{T-1}H_s & H_s^2 & .. & G_s^{T-2}H_s^2 & . & . & . & H_s^{T-1} & G_sH_s^{T-1} & H_s^T\end{bmatrix} \tag{A.4}$$

is a vector of dimension

$$e_T=\frac{(T+1)(T+2)}{2} \tag{A.5}$$

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.1) 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

$$(Z+n_{11}^j)(T+2)-\frac{(z+n_{11}^j)(z+n_{11}^j+1)}{2}+x+1+n_{01}^j \tag{A.6}$$

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.1) is in column given by (A.6) and row $j-\frac{\Gamma}{2}$. Then,

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_0 & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_1\end{bmatrix} \tag{A.7}$$

The dimension of $\mathbf{C}$ is $\Gamma\times 2e_T$. From (A.1) and (A.6) matrix $\mathbf{C}$ can be easily computed for any given T. For example, with $T=2$

$$\mathbf{C}=\begin{bmatrix}
1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} \tag{A.8}$$

with dimension 8×12.

A.1.2. 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.1) and (A.6). Table 2.2 reports the $rank(\mathbf{C})$, denoted r_T, for $T=2,...,15$. It

turns out that for all those values of T, the rank of $\mathbf{C}$ is given by

$$\begin{aligned} r_T &= T(T+1)+2 \\ &= 2e_{T-1}+2 \\ &= e_T + e_{T-2} + 1 \end{aligned} \tag{A.9}$$

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

$$rank(\mathbf{A}) \leq \min(rank(\mathbf{C}), rank(\mathbf{E})) \leq \min(rank(\mathbf{C}), 2e_T, S) = \min(r_T, S) \tag{A.10}$$

$$rank(\mathbf{A}) \geq rank(\mathbf{C}) + rank(\mathbf{E}) - 2e_T \tag{A.11}$$

where (A.11) 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.1.3. The rank of J

From (3.5) and (A.2), for given S we have a mapping from unobservables to observables given by:

$$\begin{aligned} \pi(\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta_1, ..\theta_S) &= \mathbf{A}(\mathbf{P}, \mathbf{G}, \mathbf{H}) * \theta \\ &= \mathbf{C} * \mathbf{E}(\mathbf{P}, \mathbf{G}, \mathbf{H}) * \theta \end{aligned} \tag{A.12}$$

where the θ S-vector is normalised 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 $J(T,S)$. For local point identification we require that the rank of $J(T,S)$ 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

$$J = \mathbf{C} * D(\mathbf{E} * \theta) \tag{A.13}$$

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:

$$rank(J) \leq \min\left(rank(\mathbf{C}), rank\left(D(\mathbf{E} * \theta)\right)\right) \leq \min\left(rank(\mathbf{C}), 2e_T, 4S-1\right) \tag{A.14}$$

$$rank(J) \geq rank(\mathbf{C}) + rank\left(D(\mathbf{E} * \theta)\right) - 2e_T \tag{A.15}$$

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

$$\begin{bmatrix} ... & -\mathbf{E}_s\theta_s & ... & (1-P_s)\frac{\partial \mathbf{E}_s}{\partial G_s}\theta_s & ... & (1-P_s)\frac{\partial \mathbf{E}_s}{\partial H_s}\theta_s & ... & (1-P_l)\mathbf{E}_l - (1-P_S)\mathbf{E}_S & ... \\ ... & \mathbf{E}_s\theta_s & ... & P_s\frac{\partial \mathbf{E}_s}{\partial G_s}\theta_s & ... & P_s\frac{\partial \mathbf{E}_s}{\partial H_s}\theta_s & ... & P_l\mathbf{E}_l - P_S\mathbf{E}_S & ... \end{bmatrix} \tag{A.16}$$

where $\mathbf{E}_s$ is in equation (A.4), $s = 1, ..., S$ and $l = 1, ..., S-1$.

The rank of $\mathbf{C}$ has already been calculated on previous subsection. As it happens with the rank of $\mathbf{E}$, the rank of $D(\mathbf{E} * \theta)$ depends on the values of the unknowns $\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta$, so there is no a unique value of $rank(J)$ valid for every case. However, for most of the possible values of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta\}$ the rank of $D(\mathbf{E}*\theta)$ is equal to $\min(2e_T - 1, 4S - 1)$.[8] 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.14) and (A.15) and previous calculations of $rank(\mathbf{C})$ $(= r_T)$ we have that for most of the possible values of $\{\mathbf{P}, \mathbf{G}, \mathbf{H}, \theta\}$

$$\min\left(r_T - 1, 4S-1\right) \leq rank(J) \leq \min\left(r_T, 4S-1\right) \tag{A.17}$$

because $r_T = T(T+1)+2$ is strictly smaller than $2e_T - 1 = (T+1)(T+2)-1$ for any $T \geq 1$.[9] As a matter of fact simulations suggest a general form: $rank\left(J\right) = \min\left(4S-1, r\left(T\right)\right)$ for any (S, T).

A.2. Long run initial probability value.

The probability of the initial observation P is the long run probability given in (4.2). In this situation J is of dimension $\Gamma \times (3S-1)$ since we are not estimating the S probabilities of the initial observation. J is going to be different also

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

[9] A special case where $rank(D(E*\theta))$ is reduced so that $rank(J)$ is smaller than bounds in (3.6) is $H_s = 1 - G_s$ and $P_s = 1 - P_s = 0.5$ for all $s = 1, ..., S$.

because the probabilities of the initial observation are going to be derived with respect to G and H. From equations (4.3), (3.4) and the binomial theorem, any element of a row j of matrix $\mathbf{A}$ is given by

$$\left(\frac{G}{1+G-H}\right)^{y_0^j}\left(\frac{1-H}{1+G-H}\right)^{\left(1-y_0^j\right)} G^{n_{01}^j}(1-G)^{n_{00}^j} H^{n_{11}^j}(1-H)^{n_{10}^j} \quad \text{(A.18)}$$

$$= \sum_{z=0}^{n_{10}^j+(1-y_0^j)} \sum_{x=0}^{n_{00}^j} (-1)^x (-1)^z \binom{n_{00}^j}{x}\binom{n_{10}^j}{z} G^{(x+n_{01}^j+y_0^j)} H^{(z+n_{11}^j)} \frac{1}{1+G-H} \quad \text{(A.19)}$$

Following the same procedure as in previous subsections, we can decompose matrix $\mathbf{A}$ as the product of two matrices:

$$\mathbf{A} = \mathbf{C}^{(lr)}\mathbf{E}^{(lr)} \quad \text{(A.20)}$$

$\mathbf{C}^{(lr)}$ is a $\Gamma \times e_{T+1}$ matrix that contains the coefficients $\left(\sum_{x=0}^{n_{00}^j}(-1)^x(-1)^z\binom{n_{00}^j}{x}\binom{n_{10}^j}{z}\right)$. Each coefficient is in row j and column

$$(Z+n_{11}^j)(T+3) - \frac{(z+n_{11}^j)(z+n_{11}^j+1)}{2} + x + 1 + n_{01}^j + y_0^j \quad \text{(A.21)}$$

and $\mathbf{E}^{(lr)}$ is the following $e_{T+1} \times S$ matrix:

$$\mathbf{E} = \begin{bmatrix} \frac{1}{1+G_1-H_1}\mathbf{E}_1^{(lr)} & \frac{1}{1+G_2-H_2}\mathbf{E}_2^{(lr)} & \ldots & \frac{1}{1+G_S-H_S}\mathbf{E}_S^{(lr)} \end{bmatrix} \quad \text{(A.22)}$$

where

$$\mathbf{E}_s^{(lr)\prime} = \begin{bmatrix} 1 & G_s & .. & G_s^{T+1} & H_s & G_sH_s & .. & G_s^TH_s \\ H_s^2 & .. & G_s^{T-1}H_s^2 & . & . & . & H_s^T & G_sH_s^T & H_s^{T+1} \end{bmatrix}$$

As previously this decomposition is useful to calculate the rank of the Jacobian. This rank will be driven by the rank of $\mathbf{C}^{(lr)}$, which can be exactly calculated for every T. r_T in Table 4.1 gives $rank(\mathbf{J}^{(lr)}) = rank(\mathbf{C}^{(lr)})$.