The Econometrics Journal

Econometrics Journal (2010), volume **13**, pp. 1–39.
doi: 10.1111/j.1368-423X.2009.00301.x

Heterogeneity in dynamic discrete choice models

Martin Browning[†] and Jesus M. Carro[‡]

[†] *Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, UK*
E-mail: Martin.Browning@economics.ox.ac.uk

[‡] *Departamento de Economia, Universidad Carlos III de Madrid, Calle Madrid, 126,*
28903 GETAFE, Madrid, Spain
E-mail: jcarro@eco.uc3m.es

First version received: November 2008; final version accepted: September 2009

Summary We consider dynamic discrete choice models with heterogeneity in both the levels parameter and the state dependence parameter. We first present an empirical analysis that motivates the theoretical analysis which follows. The theoretical analysis considers a simple two-state, first-order Markov chain model without covariates in which both transition probabilities are heterogeneous. Using such a model we are able to derive exact small sample results for bias and mean squared error (MSE). We discuss the maximum likelihood approach and derive two novel estimators. The first is a bias corrected version of the Maximum Likelihood Estimator (MLE) although the second, which we term MIMSE, minimizes the integrated mean square error. The MIMSE estimator is always well defined, has a closed-form expression and inherits the desirable large sample properties of the MLE. Our main finding is that in almost all short panel contexts the MIMSE significantly outperforms the other two estimators in terms of MSE. A final section extends the MIMSE estimator to allow for exogenous covariates.

Keywords: *Binary choice*, *Fixed effects*, *Heterogeneous slopes*, *Panel data*, *Unobserved heterogeneity*.

1. INTRODUCTION

Heterogeneity is an important factor to take into account when making inference based on microdata. A significant part of the literature on binary choice models in the recent years has been about estimating dynamic models accounting for permanent unobserved heterogeneity in a robust way. Honoré and Kyriazidou (2000) and Carro (2007) are two examples; surveys of this literature can be found in Arellano and Honoré (2001) and Arellano (2003a). Unobserved heterogeneity in dynamic discrete choice models is usually only allowed through a specific constant individual term, the so-called individual effect. In this paper, we consider that there may be more unobserved heterogeneity than is usually allowed for. In particular, we investigate whether the state dependence parameter in dynamic binary choice models is also individual specific.

In Browning and Carro (2006) we presented two principal objections to allowing for limited heterogeneity. The first is that this rules out, *a priori*, some interesting structural models. The second objection is that whenever we have sufficiently long panels to allow for heterogeneity in slope parameters, we usually find it. In Section 2, we complement the latter analysis with an illustration using consumer milk-type choice from a consumer panel data set. The sample used contains more than 100 periods for each household, so we have a panel with large T. This allows us to overcome the incidental parameters problem and use the standard Maximum Likelihood Estimator (MLE) to test for the presence of permanent unobserved heterogeneity both on the intercept and on the coefficient of the lag of the endogenous variable versus a model where only the intercept is heterogeneous.[1] A likelihood ratio test overwhelmingly rejects the restricted model. Furthermore, the estimates of the parameters of interest are very different when we allow for the more general form of heterogeneity. This illustration serves to further motivate the subsequent theoretical analysis.

Micropanels with a large number of periods is rare. Therefore, we need to find a way to estimate the model with two sources of heterogeneity when the number of periods is small. Furthermore, we want to do that without imposing any restriction on the conditional distribution of the heterogeneous parameters. There are not many examples in the literature where more than one source of heterogeneity is allowed in dynamic models, even for linear models. For example, the surveys of dynamic linear models in Arellano and Honoré (2001), Wooldridge (2002, ch. 11) and (in the statistics literature) Diggle et al. (2002) do not consider the possibility of allowing for heterogeneity other than in the ‘intercept’.

When we consider dynamic discrete choice models, even less is known than for the linear model. Given this relative ignorance we begin by concentrating attention on the simplest possible model and providing a thorough analysis of different estimators in respect to their tractability, bias, mean squared error (MSE) and the power of tests based on them. Thus we consider the model in which a lag of the endogenous variable is the only explanatory variable and both the slope and the intercept are individual specific with an unknown joint distribution. This simple two-state, first-order Markov chain model allows us to make a fully non-parametric analysis and to derive exact analytical expressions for the bias and MSE of the estimators we consider. We show how to use the analytical expression for the bias if T is fixed to correct the MLE estimator and obtain a Non-linear Bias Corrected (NBC) Estimator. We find that both MLE and NBC perform poorly in MSE terms. This leads us to suggest a third alternative estimator which minimizes the integrated MSE; we term this the ‘minimizes the integrated mean square error’ (MIMSE) estimator. This is an attractive estimator since it performs much better than the other two for small values of T but converges to MLE as T becomes large. Moreover, it is computationally very simple. After a thorough examination of the simple case with no covariates, we provide an extension of the MIMSE estimator to the case in which we have exogenous covariates.

The structure of rest of the paper is outlined in the next paragraphs. We regard the positive suggestions below as a first step toward incorporating more heterogeneity in dynamic discrete outcome models than is usually allowed for. Much of the analysis presented is frankly exploratory and leads to inconclusive or even negative results. For example, the exact analytical results for the MLE and NBC with small T indicate that bias reduction techniques are unlikely to lead to useful estimators in these cases.

Section 2 presents the empirical milk analysis that illustrates the need for multiple sources of heterogeneity.

[1] Others have suggested panel data tests for heterogeneous slopes when the time dimension is small; see Pesaran and Yamagata (2008) for a review of these tests and a novel test. Our emphasis in this paper is on allowing for slope heterogeneity rather than simply testing for it.

In Section 3, we study the basic model without covariates and with four observations per unit (including the initial observation). Although taking four observations for one unit may seem excessively parsimonious, this analysis allows us to display almost all of the features of interest in a transparent way. We show that there is no unbiased estimator. Following this we derive the bias for the MLE. An important finding in this respect is that the bias of the MLE estimator of the marginal dynamic effect is always negative; this is the non-linear analogue of the Nickell bias result for linear dynamic models (see Arellano, 2003b). Based on this derivation we define a one-step bias corrected estimator, which we term non-linear biased corrected (NBC). We calculate the *exact* bias and MSE of the MLE and the NBC. We show that whilst NBC reduces the bias it is sometimes worse than MLE for the MSE. The relatively poor performance of the NBC together with the result on the non-existence of an unbiased estimator sets limits on the bias correction route as a solution to the estimation problem for dynamic discrete outcome models.

To take into account that both the MLE and the NBC display high MSE, in Section 4 we present a new estimator that MIMSE. We derive the closed-form expression for the estimator. We also derive the Bayesian posterior assuming a uniform prior over the two transition probabilities and relate our estimators to that.

Section 5 compares the exact finite sample properties of the three estimators (MLE, NBC and MIMSE) with $T > 3$. There are two main conclusions. First, for most of the possible values the NBC is best in terms of bias, both for levels for very small T and for convergence of the bias to zero as T becomes large. Second, MIMSE almost always dominates MLE and NBC on the MSE criterion. We show the exact areas of dominance for MLE; these include most cases that we would ever be interested in.

In Section 6, we shift perspective and consider estimating the *distribution* of parameters of interest in the population of households. In a small-T context this will seem a natural shift given that there are severe limits on how much we can learn about individual parameters with small T. We consider estimators based on the three estimators already considered (MLE, NBC and MIMSE). Using both analytical and simulation analysis, we conclude that MIMSE dominates both other estimators and gives less biased estimators of both the location and dispersion of the distribution. This is the case both as the number of cross-section units becomes large and when it is fixed at the value we have in our empirical application in Section 2. The broad conclusion is that if we are interested in population outcomes then MIMSE performs well relative to the other two estimators.

In Section 7, we extend the MIMSE estimator to allow for exogenous covariates. We propose to use the equivalence between MIMSE and the mean of the posterior distribution with flat priors. This way it can be easily computed using MCMC techniques. Our analysis suggests that MIMSE is a credible and feasible candidate for estimating dynamic discrete choice models. Section 8 concludes and proofs are given in the Appendix.

2. RESULTS FOR A LARGE T PANEL

2.1. *Incorporating heterogeneity*

In this section, we present results for a dynamic discrete choice analysis from a long panel. Specifically, we estimate the patterns of buying full-fat milk (rather than low-fat milk) on a

Danish consumer panel that gives weekly individual purchases by households for more than 100 weeks.[2] Although the results have substantive interest, we present the analysis here mainly to motivate the subsequent econometric theory. A conventional treatment would take

$$y_{it} = 1\{\alpha y_{it-1} + x_{it}'\beta + \eta_i + v_{it} \geq 0\} \quad (t = 0, \ldots, T; i = 1, \ldots, N), \tag{2.1}$$

where y_{it} takes value 1 if household i purchases full-fat milk in week t, and zero otherwise. The parameter η_i reflects unobserved differences in tastes that are constant over time. The parameter α accounts for state dependence on individual choices due to habits. The x_{it} variables are other covariates that affect for the demand for full-fat milk. In our empirical analysis these are the presence of a child aged less than 7, quarterly dummies and a time trend. Since the relative prices of different varieties of milk are very stable across our sample period, it is reasonable to assume that the time trend picks up both price effects and common taste changes. A more flexible specification of model (2.1) that we will also consider is a model with interactions between the lagged dependent variable and the observables.

$$y_{it} = 1\{\alpha y_{it-1} + x_{it}'\beta + (y_{it-1}x_{it})'\gamma + \eta_i + v_{it} \geq 0\} \quad (t = 0, \ldots, T; i = 1, \ldots, N). \tag{2.2}$$

This allows that the state dependence depends on observables but still the only latent factor is the individual specific parameter.

It is conventional to allow for a ‘fixed effect’ η_i as in (2.1). The primary focus of this paper is on whether this makes sufficient allowance for heterogeneity. In particular, we examine whether it is also necessary to allow that the state dependence parameter varies across households and, if it does, how should we estimate if we have a short panel. Thus we take the following extended binary choice model:

$$y_{it} = 1\{\alpha_i y_{it-1} + x_{it}'\beta + \eta_i + v_{it} \geq 0\} \quad (t = 0, \ldots, T; i = 1, \ldots, N). \tag{2.3}$$

In model (2.3), we allow that both the intercept and the state dependence parameter are heterogeneous but the effects of the covariates are assumed to be common across households.[3]

The values of the parameters of (2.3) are not usually of primary interest; rather they can be used to generate other ‘outcomes of interest’. There are several candidates. In this paper, we focus on the dependence of the current probability of y being unity on the lagged value of y; this is the *marginal dynamic effect*:

$$m_i(x) = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1, x) - \Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x). \tag{2.4}$$

Another important outcome of interest is the long-run proportion of time that y_{it} is unity, given a particular fixed x vector. Using standard results from Markov chain theory this is given by

$$\frac{\Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x)}{\Pr(y_{it} = 1 \mid y_{i,t-1} = 0, x) + \Pr(y_{it} = 0 \mid y_{i,t-1} = 1, x)}. \tag{2.5}$$

In this paper, we shall only concern ourselves with the marginal dynamic effect; this is simply to limit what is already a long paper. In this empirical section we assume that the unobserved

[2] In Denmark during the first four years of our sample period there were three levels of fat content in milk: skimmed (0.01%), medium (1.5%) and high (3.5%). In the final year another low-fat (0.5%) milk was introduced. The 3.5% milk is what we call full-fat milk.

[3] We could extend the following empirical analysis to allow for heterogeneous effects of these covariates (and would certainly do so if our main concern was to analyse milk expenditure patterns) but for our purposes here it suffices to consider only heterogeneity in (η, α).

random shock v_{it} is an i.i.d. standard Normal; in the analysis in the following sections we consider the non-parametric case in which the distribution of v_{it} is not known. For the Normal case the marginal dynamic effect is given by

$$m_i(x) = \Phi(\alpha_i + x'\beta + \eta_i) - \Phi(x'\beta + \eta_i), \tag{2.6}$$

where $\Phi(\cdot)$ is the standard Normal cdf.

2.2. The Danish consumer panel

We have a Danish consumer panel that follows the same households for up to five years (with most households exiting the survey before the end of the five-year period) from January 1997 to December 2001. This panel provides data on all grocery purchases during the survey period and some characteristics of the household. Respondents provide detailed information on every item bought. For example, for milk they record the volume and price paid, the store where it is purchased, the fat content and other characteristics of that specific purchase. We aggregate purchases of milk to the weekly level (in Denmark households only consume fresh milk so that taking weekly averages gives positive purchases of milk in every week) and set the full-fat indicator for that week/household to unity if the household buys any full-fat milk in that week; this does not exclude the possibility that they also buy low-fat milk in the same week.

Our strategy in this empirical section is to estimate the parameters of (2.1), (2.2) and (2.3) without imposing any restriction on the joint distribution of α_i and η_i. We thus select a subsample of the data in which the household is observed for at least for 100 weeks so that we are in a large-T context. We assume that this selection is exogenous to the milk-buying decision. We also select on households having the number of changes on their decision with respect to the previous period greater than 10% of the number of periods; without this the parameters for a particular household may not be estimated or may be very imprecisely estimated.[4] We take up this issue in more detail in the next section. This sample selection gives us 371 households who are observed from between 100 and 260 weeks. We then use a standard Probit to estimate; this is a consistent estimator under the assumptions made. If we did not include covariates with common effects ($\beta_i = \beta$) then this estimation strategy would be the same as treating each household as a time series and estimating α_i and η_i (and β_i) for each separately. Given the length of our panel, we invoke standard large-T results.

2.3. Missing observations

Some weeks are missing for some households. This seems to be mainly because households are not disinclined to keep complete records in that week or because of being on holiday.[5] We shall take these missing weeks to be 'missing at random' in the sense that their occurrence is independent of the taste for full-fat milk. There are then two options for dealing with missing

[4] It should be noted that excluding observations that do not change their decision implies no bias on the estimation of models (2.1) and (2.2), because the contribution to the log-likelihood of these observations is zero. To see this, notice that the MLE estimate of η_i will $\pm\infty$, so the likelihood (log-likelihood) of the observations of those households i that never change will be one (zero) at the estimated value of η_i, regardless of the value of the other variables and parameters.

[5] We emphasize again that we are here presenting an illustration. For a substantive study of fat consumption we would need to explicitly model the possibility of some purchases not being recorded.

Table 1. Estimates.

Model	(2.1)	(2.2)	(2.3)
α	0.81	0.71	–
mean(α)	–	–	0.70
SD(α)	–	–	0.76
mean(η)	−0.72	−0.70	−0.73
SD(η)	0.60	0.61	0.70
corr(η, α)	–	–	−0.31
Child present	0.47	0.14	0.38
Quarter 2	−0.04	−0.05	−0.05
Quarter 3	−0.06	−0.08	−0.06
Quarter 4	0.08	0.13	0.09
Trend ($\times$100)	−0.015	−0.014	−0.014
$y_{it-1} *$ Child present	–	0.76	–
$y_{it-1} *$ Quarter 4	–	−0.14	–
Log-likelihood	−27,905	−27,659	−26,376

weeks. Suppose, for example, that week $t-1$ is missing but we observe in weeks $t-2$, t and $t+1$. The first option is to use the probability $\Pr(y_{it} = 1 \mid y_{i,t-2} = 1, x_{it}, x_{i,t-1})$ in the likelihood. This assumes that we can impute $x_{i,t-1}$ which is not problematic in our case (for example, the presence of a child aged less than 7 or the season). The alternative procedure, which we adopt, is to drop observation t and to start again at period $t+1$. When we do this we of course keep (η_i, α_i) constant for each household. Using the latter procedure causes a small loss of efficiency but is much simpler. The proportion of missing observations is about 14% of the total number of observations.

2.4. *Results for the long panel*

Table 1 contains the estimates of models (2.1), (2.2) and (2.3) by maximum likelihood estimation (MLE). The model with observable variation in the state dependence parameter, (2.2), fits significantly better than the most restricted model (2.1) (a likelihood ratio statistic of 492 with 5 degrees of freedom) but much worse than the general model (2.3). The likelihood ratio test statistic for model (2.1) against (2.3) is 3058 with 370 degrees of freedom and 2566 with 365 degrees of freedom for testing model (2.2) against (2.3). This represents a decisive rejection of the conventional model which only allows for a single ‘fixed effect’. Figure 1 shows the marginal distributions of the two parameters, α_i and η_i; as can be clearly seen the state dependence parameter varies quite widely across households. Restricting the state dependence parameter to be common across households gives significant bias in the mean of the state dependence and in the impact of children. It also gives a value for the variability of the η that is too low. For the general model we find a significant negative correlation between the two parameters; obviously the standard model (2.1) is not able to capture this.

Figure 2 plots the estimated state dependence parameter ($\hat{\alpha}_i$) and its 95% confidence interval for each of our 371 households, sorted from the smallest value of $\hat{\alpha}_i$ to the largest

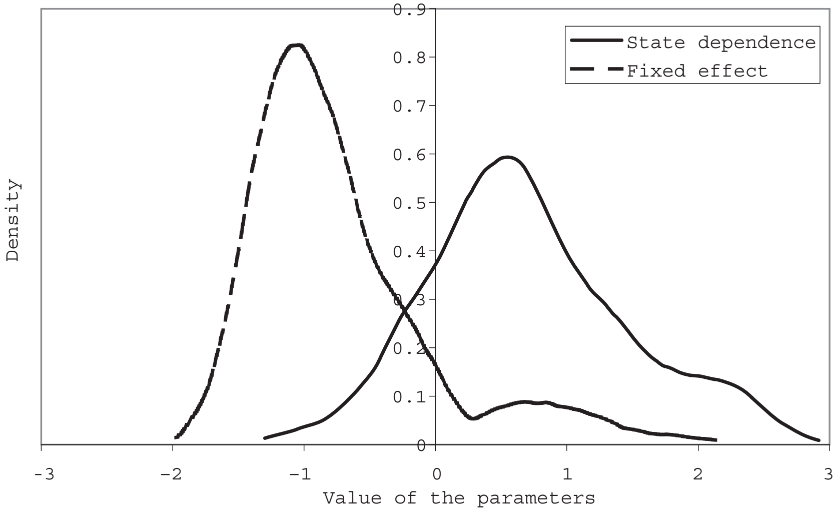

Figure 1. Marginal densities of α_i and η_i.

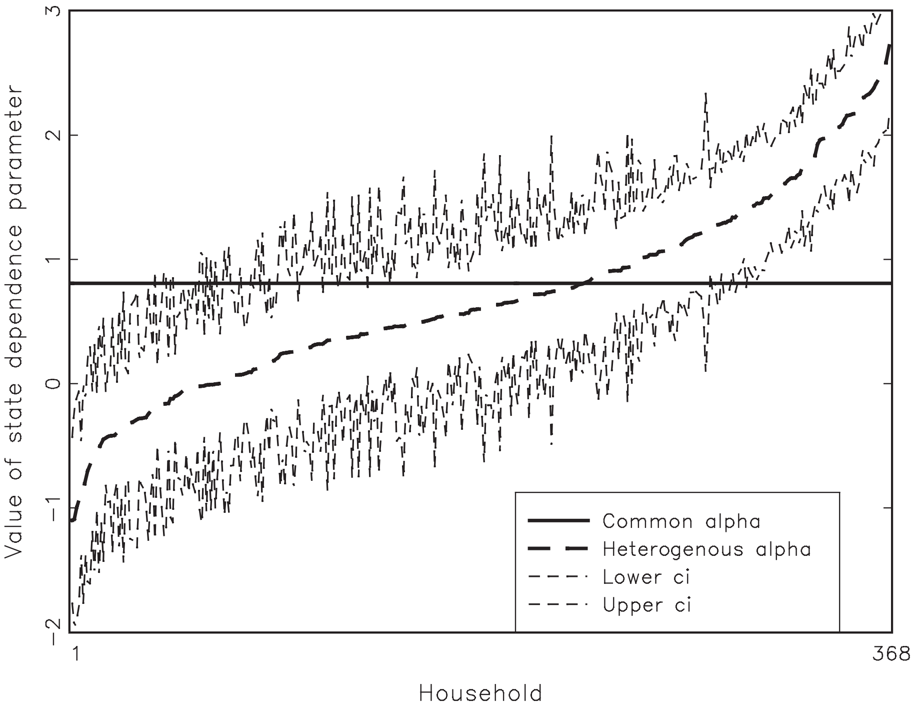

Figure 2. Estimated state dependence parameter, α_i.

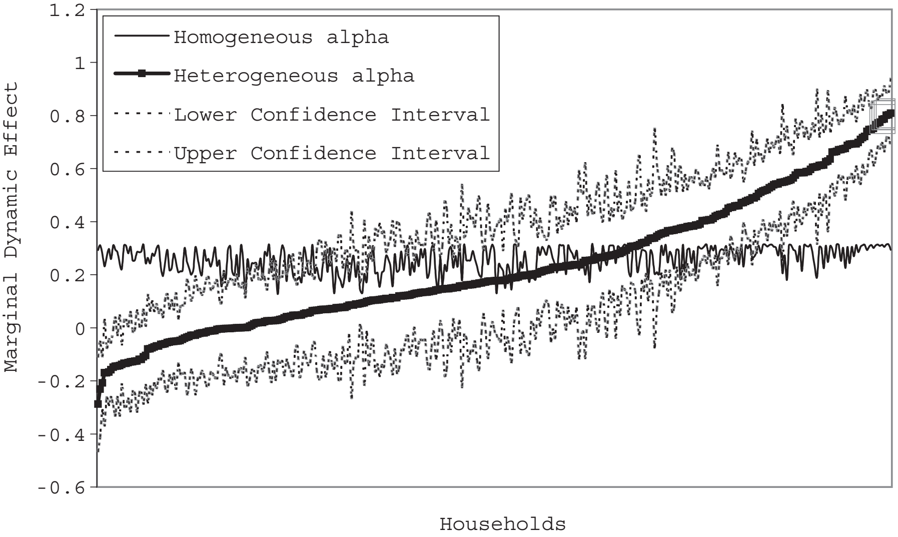

Figure 3. Estimates of marginal dynamic effects.

value.[6] The darker horizontal line is the value of $\alpha = 0.81$ estimated from model (2.1). The proportion of households whose confidence interval of $\hat{\alpha}_i$ contains $\hat{\alpha}$ is 59%. Thus for 41% of our sample the estimated α parameter using a model with more heterogeneity (2.3) is statistically different from the value using model (2.1).

We can also consider the marginal effect, which is of more interest than the parameters that are directly estimated. For both models the marginal effect is different for each household but the variation in the magnitude of the marginal effect among households is greater in model (2.3) than in model (2.1). This is shown in Figure 3; to plot this we set the quarterly dummies and time trend to zero and the child variable to the mode for the household. The x-axis values are sorted according to the values of the marginal effect for the general model (2.3). The flatter (variable) line is for model (2.1) and the increasing curve is the value for model (2.3) (with 95% confidence bands). In this case 46% of households have a marginal effect that is significantly different from that implied by model (2.1) and 52% have a marginal dynamic effect that is not significantly different from zero (at a 5% significance level). The differences between the implications of the two models for the outcome of interest (the marginal dynamic effect) can be seen even more dramatically in Figure 4 and Table 2 which present the estimated distribution of the marginal dynamic effect, for the three estimated models of those households with the child variable equal to zero. We plot a grid on this figure to facilitate comparisons across the three sets of estimates. Once again we see that the extended model gives much more variation across households in the marginal effects. But there are also other strong differences; for example, for the conventional models ((2.1) and (2.2)) all households are estimated to have a positive marginal dynamic effect, whereas for the unrestricted model about 18% have a negative effect (although most are not

[6] In the confidence intervals of Figures 2 and 3 we are ignoring the sampling variability in the estimation of model (2.1) because it is negligible in comparison to the sampling variability in estimating model (2.3).

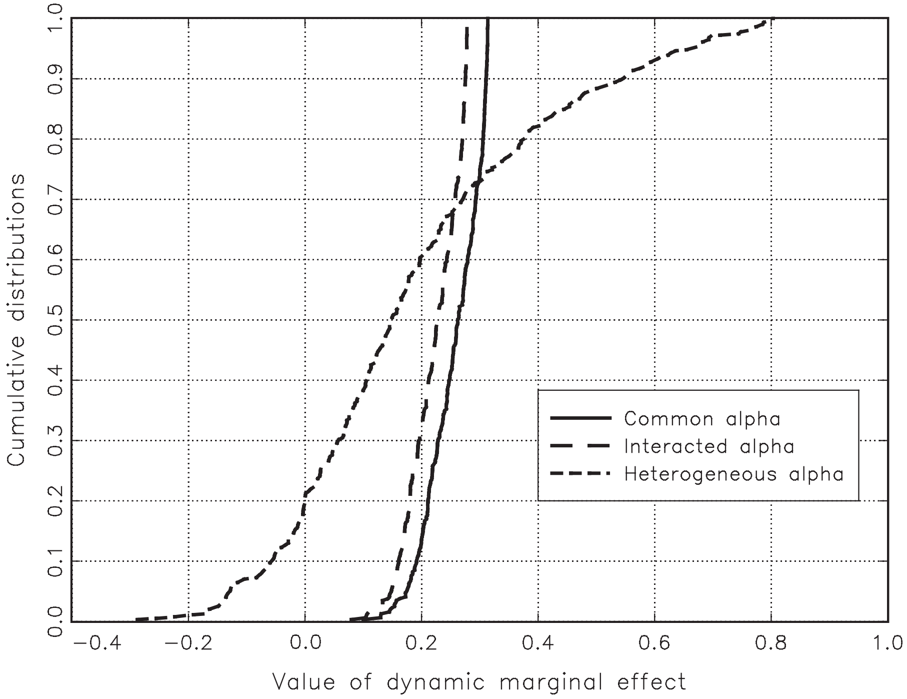

Figure 4. Distribution of the dynamic marginal effect.

Table 2. Distribution of the marginal dynamic effect.

Model	(2.1)	(2.2)	(2.3)
Minimum	0.10	0.08	−0.29
First quartile	0.22	0.19	0.03
Median	0.26	0.23	0.15
Third quartile	0.30	0.26	0.32
Maximum	0.31	0.28	0.80
Mean	0.26	0.22	0.19
SD	0.05	0.04	0.23

‘significantly’ different from zero; see Figure 3). Moreover the mean and median are lower for the extended model.

This empirical analysis serves to illustrate our contention that there is probably more heterogeneity in dynamic models than is allowed for by conventional schemes that only allow ‘intercepts’ to vary across households. We turn now to a consideration of estimation when we do not have the luxury of observing households for very many periods. One option is to formulate a (random effects) parametric model for the conditional joint distribution of $(\alpha, \eta \mid x, y_0)$ and then to estimate the parameters by, say, maximum likelihood. This parametric model would have to accommodate the bimodalities and fat tails displayed by the distributions shown in Figure 1. In this paper, we consider the alternative of estimating non-parametric models which do not restrict the joint distribution of the latent factors.

3. EXACT BIAS AND MSE ANALYSIS, $T = 3$

3.1. A simple model with a lagged dependent variable

The empirical analysis above suggested strongly that we need to allow for heterogeneity in both the intercept and the state dependence parameter when we consider dynamic models. Since relatively little is known about the behaviour of the dynamic non-linear panel data estimators in the simpler case in which we only allow for heterogeneity in the 'intercept' (see, for example, Arellano and Honoré, 2001, sec. 8), we necessarily have to be modest in our aims here. Consequently we restrict attention to the simple model with no covariates, in which case we can dispense with parametric formulations such as (2.3) and focus directly on the two transition parameters:

$$G_i = \Pr(y_{it} = 1 \mid y_{i,t-1} = 0), \tag{3.1}$$

$$H_i = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1). \tag{3.2}$$

This is a two-state, first-order stationary Markov model with a marginal dynamic effect given by:

$$M_i = H_i - G_i. \tag{3.3}$$

There is a large literature on the estimation of Markov models considering such issues as testing for stationarity or the order of the process; the classic reference is Anderson and Goodman (1957) who consider the case in which all agents have the same transition matrix. In general, most investigators assume less heterogeneity than we do here. Exceptions include Billard and Meshkani (1995) and Cole et al. (1995) who both use an empirical Bayes approach,[7] and Albert and Waclawiw (1998) who adopt a quasi-likelihood approach to estimate the first two moments of the joint distribution of the transition probabilities. The distributions plotted in Figure 1 suggest that this may miss important features of the joint distribution.

There are two primary virtues of considering the simplest model of a first-order stationary Markov chain without covariates. The first is that we can derive exact analytical finite sample results and discuss estimation and bias reduction without recourse to simulation. This allows us, for example, to sign the bias for particular estimators for any value of (G, H) and not just for particular values as in Monte Carlo studies. The second advantage is that the analysis here is fully non-parametric and does not require assumptions concerning functional forms. Thus the basic case serves as a general benchmark which we can examine in great and exact detail. We shall only consider estimation conditional on the observed initial value y_{i0}.[8] We start with an exhaustive account of the case in which $T = 3$ and, with no loss of generality, we only consider

[7] This is essentially a random coefficients model.

[8] If we are willing to make assumptions concerning the initially observed value (for example, it is drawn from the long-run distribution) then there may be considerable gain in efficiency when T is small. We do not explore this here to avoid potential biases caused by misspecifications of the distribution of the initial value. For recent results on taking account of the initial conditions problem, see Honoré and Tamer (2006).

paths that start with $y_{it} = 1$. This very simple case is instructive and leads us to reject some possibilities and also suggests general results. In a later section, we consider the general fixed-T case.

If we take a parametric formulation with an arbitrary cdf $F(\cdot)$ then we have

$$\begin{aligned} G_i &= F(\eta_i), \\ H_i &= F(\alpha_i + \eta_i). \end{aligned} \tag{3.4}$$

Observing this allows us to derive a restriction that is analogous to the usual model (2.1) with a homogeneous state dependence parameter, α_i. Assuming that $F(\cdot)$ is everywhere strictly increasing we can invert both equations to give

$$\alpha_i = F^{-1}(H_i) - F^{-1}(G_i). \tag{3.5}$$

Then the usual homogeneity restriction , $\alpha_i = \alpha$, gives the restriction

$$H_i = F(\alpha + F^{-1}(G_i)). \tag{3.6}$$

It is important to note that this restriction is *parametric* and depends on the chosen cdf. That is, an assumption of a homogeneous state dependence parameter for one distribution is implicitly assuming that the state dependence parameter is heterogeneous for any other distribution, unless α is zero. This emphasizes the arbitrariness in the usual homogeneity assumption since there is no reason why the homogeneity of the state dependence parameter α_i should be linked to the distribution of $F(\cdot)$. Given this arbitrariness, we see as more natural the hypothesis that the marginal dynamic effect is the same for everyone:

$$M_i = M \Rightarrow H_i = M + G_i. \tag{3.7}$$

We shall return to testing for this in Section 3.6 below.

When there are no covariates we can treat each household as an individual (albeit short) time series and drop the i subscript. Table 3 gives the outcomes for the case with $T = 3$ (that is, four observations including period 0) and $y_0 = 1$. The first column gives the name we have given to each case, the second column gives the observed path and the next four columns give the frequencies for observed pairs of outcomes 00, 01, 10 and 11, respectively. The final column gives the probability of observing the path (conditional on $y_0 = 1$) which we denote by

Table 3. Outcomes for $T = 3$.

Case	Path	n_{00}	n_{01}	n_{10}	n_{11}	Probability of case j, p_j
a	1000	2	0	1	0	$(1 - H)(1 - G)(1 - G)$
b	1001	1	1	1	0	$(1 - H)(1 - G)G$
c	1010	0	1	2	0	$(1 - H)G(1 - H)$
d	1011	0	1	1	1	$(1 - H)GH$
e	1100	1	0	1	1	$H(1 - H)(1 - G)$
f	1101	0	1	1	1	$H(1 - H)G$
g	1110	0	0	1	2	$HH(1 - H)$
h	1111	0	0	0	3	HHH

$p_a, p_b, \ldots, p_h$, respectively. This is given by

$$p_j = (G)^{n_{01}^j}(1-G)^{n_{00}^j}(H)^{n_{11}^j}(1-H)^{n_{10}^j}, \tag{3.8}$$

where n_{01}^j is the number of $0 \to 1$ transitions for case j, etc. We now consider the choice of an estimator for this scenario.

3.2. *All estimators are biased*

An *estimator* $(\hat{G}, \hat{H})$ assigns values to G and H for each case $a, b, \ldots, h$:

$$\{\hat{G}, \hat{H}\} : \{a, b, c, d, e, f, g, h\} \to \Im([0, 1]^2), \tag{3.9}$$

where $\Im(X)$ denotes the power set of X. An estimator $(\hat{G}, \hat{H})$ is the correspondence (3.9) evaluated at the random indicator for the paths a to h. For the marginal dynamic effect the correspondence is given by

$$\hat{M} = \hat{H} - \hat{G} : \{a, b, c, d, e, f, g, h\} \to \Im([-1, 1]). \tag{3.10}$$

If the values given by the estimator are unique for each case then the corresponding parameter is point estimated, otherwise the estimator is partially defined. For example, as we shall see in the next subsection, maximum likelihood is point defined for H but only partially defined for G. Before considering particular estimators we show analytically that there is no unbiased estimator for G and H.

PROPOSITION 3.1. *All estimators of* (G, H) *are biased.*

This is a useful result since it shows that there is no point in searching for an unbiased estimator and we consequently have to seek for estimators that have low bias or low MSE. An alternative way to state this result is that for any estimator of (G, H) we can find an alternative estimator and some values of (G, H) which give a lower bias. Thus we will always be in the situation in which we are making trade-offs, even when we restrict attention to bias.

3.3. *Maximum likelihood estimator*

In the current context in which probabilities are given, the most natural estimator is maximum likelihood. The MLE $\{\hat{G}_j^{MLE}, \hat{H}_j^{MLE}\}_{j=a,\ldots,h}$ gives the values of G and H that maximize the probabilities for each case. It is convenient to give the results for any fixed T ($\geq$3) at this point. From (3.8) it is easily seen that the log-likelihood is maximized for values of G and H given by

$$\hat{G}_j^{MLE} = \frac{n_{01}^j}{n_{00}^j + n_{01}^j}, \tag{3.11}$$

$$\hat{H}_j^{MLE} = \frac{n_{11}^j}{n_{10}^j + n_{11}^j}. \tag{3.12}$$

If this mapping exists then the parameter is point estimated. Since we condition on $y_0 = 1$ we always have $(n_{10}^j + n_{11}^j) \neq 0$ so that $\hat{H}_j^{MLE}$ is always defined. The MLE estimator for G does not

Table 4. Outcomes conditioning on point estimation.

		Maximum likelihood		Non-linear bias corrected	
Case	Adjusted probability, $\tilde{p}_j$	$\hat{G}^{MLE}$	$\hat{H}^{MLE}$	$\hat{G}^{BC1}$	$\hat{H}^{BC1}$
a	$\frac{(1-G)(1-G)}{(1+H)}$	0	0	0	0
b	$\frac{(1-G)G}{(1+H)}$	1/2	0	3/8	0
c	$\frac{G(1-H)}{(1+H)}$	1	0	1	0
d	$\frac{GH}{(1+H)}$	1	1/2	1	2/3
e	$\frac{H(1-G)}{(1+H)}$	0	1/2	0	5/6
f	$\frac{HG}{(1+H)}$	1	1/2	1	2/3

exist if we observe $y_t = 1$ for $t = 1, 2, \ldots, T - 1$ ($\Rightarrow n_{00} + n_{01} = 0$). The probability of this is given by:

$$\Pr(\text{non-existence}|y_0 = 1) = H^{T-1}. \tag{3.13}$$

Thus there is always a positive probability of non-existence (so long as $H > 0$) but it goes to zero as T becomes large (so long as $H < 1$). Even for modest T, it is small, unless H is very close to 1. Moreover, for the 'non-existence' case where $n_{10}^j = 0$, i.e. $y_t = 1$ for $t = 1, 2, \ldots, T$ (case h in Table 3), the contribution to the log-likelihood is zero since $\hat{H}_j^{MLE} = 1$ for this case. For the other 'non-existence' case the contribution is not zero, but it is close to zero and it goes to zero as T becomes large, since $\hat{H}_j^{MLE} = \frac{T-1}{T}$. Given these reasons, most investigators ignore the bias introduced by selecting out the non-identifying paths. We thus have two distinct classes of estimator. In the first, we exclude any observation with $n_{00} + n_{01} = 0$. In this case, both G and H are point estimated. When we analyse this case in finite samples, we have to correct the probabilities for sample selection by dividing the given probabilities by $(1 - H^{T-1})$ (and using $\tilde{p}$ to denote adjusted probabilities). The second class of estimator uses all the observed paths but then $\hat{G}^{MLE}$ is only partially defined. We concentrate attention on the former, (point estimated) case and do not consider the partially defined estimator.[9]

Table 4 gives the relevant details for the point-estimated context for $T = 3$ in which we exclude cases g and h. The second column gives the probabilities adjusted for the sample selection and the next two columns give the maximum likelihood estimators for (G, H). These estimators are calculated *without* taking into account that we select out cases g and h (that is, they are based on the unadjusted probabilities given in Table 3). This is largely to conform with current practice which does not adjust probabilities when calculating maximum likelihood estimators for the reasons given in the previous paragraph. The alternative is to use the adjusted probabilities when calculating the MLE; this is perfectly legitimate (and may even be considered better) but it is not the common practice and it leads to estimators which look 'non-standard' so we choose to analyse only the MLE estimator using the unadjusted probabilities. In all the analysis, we always use the adjusted probabilities when calculating biases and MSEs, as previously explained.

[9] The proof that there is no unbiased estimator was given for the no-selection case. It is easy to show by the same methods that there is no unbiased estimator of the pair (G, H) for the class in which we select our cases g and h.

The result of the previous subsection tells us that MLE is biased. Since we have an exact probability model we can go further than this and give the exact bias (using the notation $\hat{G}_j^{MLE}$ to denote the jth element of $\hat{G}^{MLE}$):

$$\begin{aligned} bias(\hat{G}^{MLE}) &= E(\hat{G}^{MLE}) - G = \tilde{p}_a \hat{G}_a^{MLE} + \cdots + \tilde{p}_f \hat{G}_f^{MLE} - G \\ &= \frac{1}{2}\frac{(1-G)G}{(1+H)} \geq 0, \end{aligned} \tag{3.14}$$

$$\begin{aligned} bias(\hat{H}^{MLE}) &= E(\hat{H}^{MLE}) - H = \tilde{p}_a \hat{H}_a^{MLE} + \cdots + \tilde{p}_f \hat{H}_f^{MLE} - H \\ &= \frac{1}{2}\frac{(G-2H-1)H}{(1+H)} \leq 0, \end{aligned} \tag{3.15}$$

$$\begin{aligned} bias(\hat{M}^{MLE}) &= bias(\hat{H}^{MLE}) - bias(\hat{G}^{MLE}) \\ &= \frac{1}{2}\frac{G^2 - G + GH - H - 2H^2}{(1+H)} \leq 0. \end{aligned} \tag{3.16}$$

Although the exact bias depends on the unobserved probabilities G and H, the sign of the bias does not. As can be seen, $\hat{G}^{MLE}$ is always biased upwards and $\hat{H}^{MLE}$ and $\hat{M}^{MLE}$ always have a negative bias. In particular, the bias of the MLE estimate of the marginal dynamic effect, $\hat{M}^{MLE}$, is always negative for interior values of (G, H). This is the analogue of the signable Nickell bias for the linear autoregressive model (see, for example, Arellano, 2003b).[10] We shall return to this in the section in which we consider $T > 3$. The bias of G is maximized at $(G, H) = (0.5, 0)$ and the absolute value of the biases of H and M are both maximized at $(G, H) = (0, 1)$.

Knowing the sign of the bias is sometimes useful since it allows us to put bounds on the possible values of the parameters and the marginal effect. For example, for the marginal effect for case j we have the bounds $[\hat{H}_j^{MLE} - \hat{G}_j^{MLE}, 1]$. Admittedly these are not very tight bounds (particularly for case c), but we should not expect tight bounds if we only observe a household for four periods. One view of the choice of an estimator is then that it reduces to finding an estimator that has the smallest expected bounds. The negative bias result of the previous subsection then states that no estimator gives uniformly tight bounds (that is, smallest bounds independent of the true parameter values).

3.4. Bias corrected estimators

Since we have an exact and explicit form for the bias, one improvement that immediately suggests itself is to use these expressions for the bias with the ML estimates substituted in to define a new (bias corrected) estimator. We define the NBC estimator, which we denote $(\hat{G}^{NBC}, \hat{H}^{NBC})$, as the MLE estimate minus the estimated bias of the latter.[11] We denote the probability of case k using

[10] We have the same pattern of signs for the bias when we consider the case in which $y_0 = 0$, so this is a general result.

[11] The terminology here is to distinguish our correction from linear bias correction estimators as in McKinnon and Smith (1998).

the estimates from observing case j by $p_k(\hat{G}_j^{MLE}, \hat{H}_j^{MLE})$ and define the new estimator by

$$\begin{aligned}\hat{G}_j^{NBC} &= \hat{G}_j^{MLE} - \left[\sum_{k=a}^{f} \tilde{p}_k\left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE}\right)\hat{G}_k^{MLE} - \hat{G}_j^{MLE}\right] \\ &= 2\hat{G}_j^{MLE} - \left[\sum_{k=a}^{f} \tilde{p}_k\left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE}\right)\hat{G}_k^{MLE}\right], \end{aligned} \tag{3.17}$$

$$\begin{aligned}\hat{H}_j^{NBC} &= \hat{H}_j^{MLE} - \left[\sum_{k=a}^{f} \tilde{p}_k\left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE}\right)\hat{H}_k^{MLE} - \hat{H}_j^{MLE}\right] \\ &= 2\hat{H}_j^{MLE} - \left[\sum_{k=a}^{f} \tilde{p}_k\left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE}\right)\hat{H}_k^{MLE}\right]. \end{aligned} \tag{3.18}$$

The values for these are given in the NBC column of Table 4.

We can also derive the biases for the NBC estimator:

$$bias(\hat{G}^{NBC}) = E(\hat{G}^{NBC}) - G = \frac{3}{8}\frac{(1-G)G}{(1+H)} \geq 0, \tag{3.19}$$

$$bias(\hat{H}^{NBC}) = E(\hat{H}^{NBC}) - H = \frac{1}{6}\frac{(3G-6H-1)H}{(1+H)} \lesseqgtr 0, \tag{3.20}$$

$$bias(\hat{M}^{NBC}) = \frac{1}{24}\frac{(9G^2-9G+12GH-4H-24H^2)}{(1+H)} \lesseqgtr 0. \tag{3.21}$$

Note that the bias for H and M is now not necessarily negative. Nevertheless the situation where $bias(\hat{H}^{NBC})$ and $bias(\hat{M}^{NBC})$ are not negative is an extreme case of 'negative autocorrelation' in that it implies that both $\Pr(y_{it} = 1 \mid y_{i,t-1} = 1)$ and $\Pr(y_{it} = 0 \mid y_{i,t-1} = 0)$ are small. The bias for H is positive if the following two conditions are both satisfied: $H < \frac{1}{3}$ and $G > \frac{1}{3} + 2H$. If we restrict attention to values of (G, H) such that $M = H - G > -0.5$ then we can show that the bias of $\hat{M}^{NBC}$ is negative. Comparing the bias for $\hat{G}^{NBC}$ with equation (3.14) we see immediately that the NBC estimator always has a smaller bias for G than MLE. Moreover, if we again restrict attention to $M = H - G > -0.5$ then we can show that the absolute value of biases of H and M are lower for NBC than for MLE. Actually, for M that holds also for $M > -0.8$. Thus, for $T = 3$ and 'reasonable' values of (G, H), the bias correction does indeed lead to a reduction in the bias; although there are some extreme cases for which bias correcting actually increases the bias of the estimator.

The definitions in (3.17) and (3.18) suggest a recursion in which we take the new bias corrected estimator and adjust the bias again. This leads to a second round estimator in which some estimated probabilities exceed unity. If we continue iterating then the estimator does not converge. Formally we can show that there does not exist a limit estimator (see the Appendix) and numerically we have that the iterated estimator does not converge for one case. This may happen when dealing with non-linear transformations as here. Even if a limit estimator had existed, it would still be biased, since we proved there is no unbiased estimator in Proposition 3.1. Given

Table 5. Mean squared errors for estimators.

	Mean squared error	
	$\hat{G}$	$\hat{H}$
MLE	$\frac{1}{4}\frac{(5-4G+4H)(1-G)G}{(1+H)}$	$\frac{1}{4}\frac{(4H^2-4GH+G+1)H}{(1+H)}$
(Mean)	(0.138)	(0.159)
NBC	$\frac{1}{64}\frac{(73-48G+64H)(1-G)G}{(1+H)}$	$\frac{1}{36}\frac{(36H^2-36GH+7G-24H+25)H}{(1+H)}$
(Mean)	(0.140)	(0.158)

Note: Value in parenthesis is mean assuming uniform over (G, H).

this we consider only two candidate estimators: maximum likelihood and the (one-step) nonlinear bias corrected estimator.

3.5. Mean squared error of the estimators

The results above have focused on the bias of the maximum likelihood estimators $(\hat{G}^{MLE}, \hat{H}^{MLE})$ and the NBC estimators $(\hat{G}^{NBC}, \hat{H}^{NBC})$. However, the MSE can increase even if the bias is reduced. Thus we also need to consider the MSE of our candidate estimators. The MSE for any estimator is given by

$$\begin{aligned} MSE(\hat{G}) &= E(\hat{G} - G)^2 \\ &= \sum_{j=a}^{f} \tilde{p}_j(G, H)(\hat{G}_j - G)^2, \end{aligned} \tag{3.22}$$

$$\begin{aligned} MSE(\hat{H}) &= E(\hat{H} - H)^2 \\ &= \sum_{j=a}^{f} \tilde{p}_j(G, H)(\hat{H}_j - H)^2. \end{aligned} \tag{3.23}$$

Table 5 gives the exact MSEs for the two estimators; the values given are not symmetric in G and H since we consider only the case with $y_0 = 1$. Given these expressions, it is easy to show neither estimator dominates the other in terms of MSE. For example, if we take $(G, H) = (0.5, 0.5)$ then the MSE of ML estimators of G and H are lower, whereas for $(G, H) = (0.25, 0.75)$ the NBC estimator has the lowest MSE. Given that we have exact expressions for the MSE, we can find the mean for each of our estimators if we assume a distribution for (G, H). The values in parentheses in Table 5 give the means assuming a uniform distribution over $[0, 1]^2$. As can be seen, the two estimators of G and H are quite similar in this regard. We shall return to the MSE analysis in the later sections.

3.6. Inference

The final consideration for the two estimators is their performance for hypothesis testing. In the current context the most important hypothesis we would wish to test is that the marginal dynamic effect is zero: $G = H$. Table 6 gives the probabilities for the six possible paths under $H_0 : G = H$

Table 6. Outcomes for no marginal dynamic effect.

Case	Path	Prob, given $G = H$	$\hat{G}^{MLE}$	$\hat{G}^{NBC}$
a	1000	$\frac{(1-G)^2}{(1+G)}$	0	0
b	1001	$\frac{(1-G)G}{(1+G)}$	1/3	7/18
c	1010	$\frac{(1-G)G}{(1+G)}$	1/3	7/18
d	1011	$\frac{G^2}{(1+G)}$	2/3	38/45
e	1100	$\frac{(1-G)G}{(1+G)}$	1/3	7/18
f	1101	$\frac{G^2}{(1+G)}$	2/3	38/45

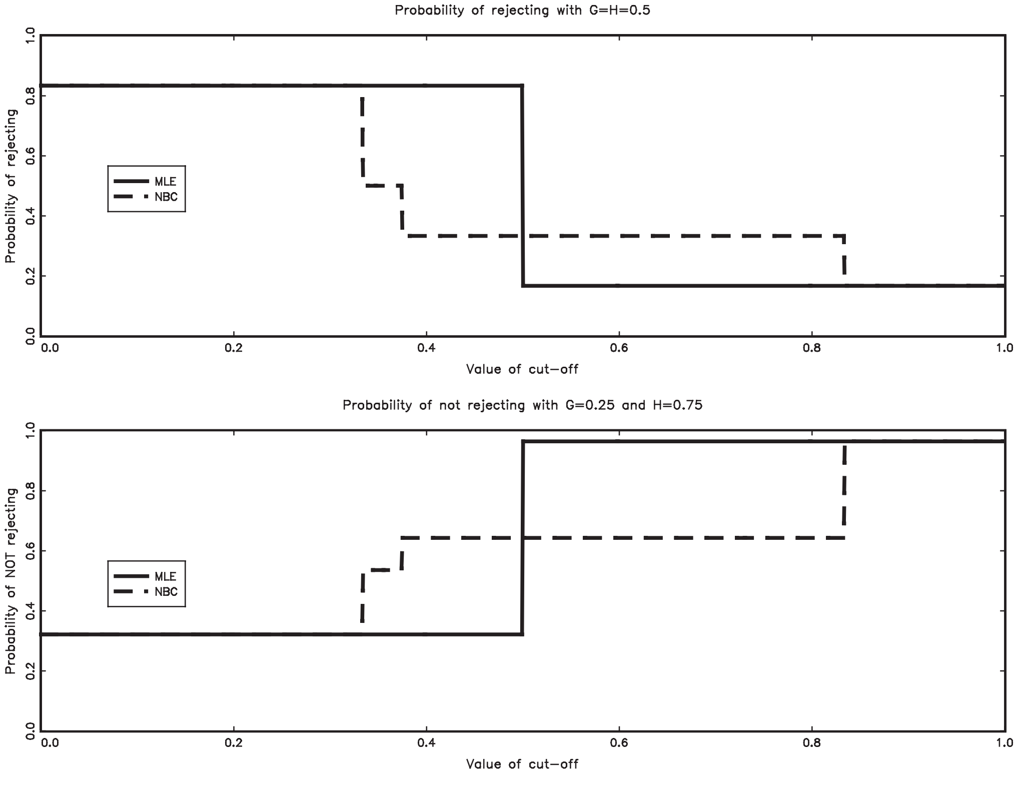

Figure 5. Inference for MLE and NBC.

and the corresponding ML estimator and NBC estimator under the null. To consider inference we have to specify a decision process that leads us to either reject or not reject H_0 consequent on observing one of the cases $a, b, \ldots, f$. We consider symmetric two-sided procedures in which we reject H_0 if $|\hat{M}| > \tau$, where τ is a cut-off value between zero and unity. The top panel of Figure 5 shows the probabilities of rejecting the null when it is true for values of $\tau \in (0, 1)$ and $G = H = 0.5$ when $T = 3$. This shows that neither estimator dominates the other in terms of size. What of the converse: the probability of rejecting H_0 when it is false. The bottom panel of

Figure 5 shows the case with $G = 0.25,\ H = 0.75$ (that is, with reasonably strong positive state dependence). Once again, neither estimator dominates the other.

3.7. *Where does this leave us?*

A number of conclusions arise from a consideration of the simple model with $T = 3$ and $y_0 = 1$:

- There is no unbiased estimator for either the point-identified case nor the partially identified case.
- MLE gives an upwards biased estimator for $G = \Pr(y_{it} = 1 \mid y_{i,t-1} = 0)$ and a downwards biased estimator of $H = \Pr(y_{it} = 1 \mid y_{i,t-1} = 1)$ and the marginal dynamic effect $M = H - G$.
- We can calculate the bias of the MLE and consequently define a one-step bias corrected estimator (NBC).
- The bias corrected estimator makes the absolute value of the bias of G smaller, as compared to MLE. For values of $M > -0.8$ the NBC estimator of M also gives a lower bias in absolute terms than MLE, but not for values of M close to -1.
- NBC does not dominate MLE on an MSE criterion. In fact the mean MSE of the two estimators are very close if we assume that (G, H) are uniformly distributed.
- Neither of the two estimators dominates the other in terms of making inferences.

Most of these conclusions apply in the $T > 3$ case; before considering that explicitly we present a new estimator that is designed to address the relatively poor performance of MLE and NBC for the MSE.

4. MINIMIZING THE INTEGRATED MSE

4.1. *Minimum integrated MSE estimator of M*

The two estimators developed so far are based on MLE but the case for using MLE is not very compelling if we have small samples (see Berkson, 1980, and the discussion following that paper). As we have seen, we can make small sample corrections for the bias to come up with an estimator that is less biased, but our investigations reveal that this is not necessarily better on the MSE criterion. Given that we use the latter as our principal criterion, it is worth investigating alternative estimators that take the MSE into account directly. To focus our discussion we concentrate on the estimator for the marginal effect $M = H - G$. The MSE for an estimator $\hat{M}_j$ (where j refers to an observed path of zeros and ones) is given by:

$$\lambda(\hat{M}; G, H) = \sum_{j=1}^{J} p_j (\hat{M}_j - (H - G))^2. \tag{4.1}$$

As with the bias, we can show that there is no estimator that minimizes the MSE for all values of (G, H) so we have to settle for finding the minimum for some choice of a prior distribution of (G, H). Given that we are looking at the general case in which we have no idea of the context, the obvious choice is the uniform distribution on $[0, 1]^2$. This gives the integrated MSE:

$$\begin{aligned}\psi &= \int_0^1 \int_0^1 \lambda(\hat{M}; G, H)\, dGdH \\ &= \int_0^1 \int_0^1 \sum_{j=1}^{J} p_j (\hat{M}_j - (H - G))^2\, dGdH \\ &= \int_0^1 \int_0^1 \sum_{j=1}^{J} \big(G^{n^j_{01}}(1-G)^{n^j_{00}} H^{n^j_{11}}(1-H)^{n^j_{10}}\big)(\hat{M}_j - (H-G))^2\, dGdH, \qquad (4.2)\end{aligned}$$

where we have substituted for p_j from Table 3. The criterion (4.2) is additive in functions of $\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_J$ so that we can find minimizing values of the estimator considering each case in isolation. Differentiating (4.2) with respect to $\hat{M}_j$, setting the result to zero and solving for $\hat{M}_j$ gives:

$$\hat{M}_j = \frac{\int_0^1 \int_0^1 p_j (H - G) dGdH}{\int_0^1 \int_0^1 p_j dGdH} \qquad (4.3)$$

$$= \frac{\left(\int_0^1 H^{(1+n^j_{11})}(1-H)^{n^j_{10}} dH\right)}{\left(\int_0^1 H^{n^j_{11}}(1-H)^{n^j_{10}} dH\right)} - \frac{\left(\int_0^1 G^{(1+n^j_{01})}(1-G)^{n^j_{00}} dG\right)}{\left(\int_0^1 G^{n^j_{01}}(1-G)^{n^j_{00}} dG\right)}. \qquad (4.4)$$

Using the result that for x and z that are integers we have

$$\int_0^1 Y^x (1-Y)^z dY = \frac{\Gamma(x+1)\Gamma(z+1)}{\Gamma(x+z+2)} = \frac{x!z!}{(x+z+1)!} \qquad (4.5)$$

(where $\Gamma(\cdot)$ is the gamma function) we have the following closed form for the minimum integrated MSE (MIMSE) estimator:

$$\begin{aligned}\hat{M}_j^{MIMSE} &= \frac{\big(n^j_{11}+1\big)!\big(n^j_{10}+n^j_{11}+1\big)!}{\big(n^j_{10}+n^j_{11}+2\big)!n^j_{11}!} - \frac{\big(n^j_{01}+1\big)!\big(n^j_{00}+n^j_{01}+1\big)!}{\big(n^j_{00}+n^j_{01}+2\big)!n^j_{01}!} \\ &= \frac{n^j_{11}+1}{n^j_{10}+n^j_{11}+2} - \frac{n^j_{01}+1}{n^j_{00}+n^j_{01}+2}. \qquad (4.6)\end{aligned}$$

As can be seen, the MIMSE estimator is simply the MLE estimator with $n^j_{st}+1$ replacing n^j_{st} everywhere. It is important to note that the first term on the right-hand side of equation (4.6) is $\hat{H}_j^{MIMSE}$, and the second term is $\hat{G}_j^{MIMSE}$.

The MIMSE point estimates the values of the parameters in cases where the MLE did not exist. Moreover, the MIMSE estimate will always be in the interior of the parameter space (that is, $\hat{M}_j^{MIMSE} \in (-1, 1)$). In terms of computational difficulty, the MIMSE estimator is as easy to compute as the MLE estimator and somewhat easier to compute than the NBC estimator. In particular, we only require observation of the sufficient statistics $\{n^j_{00}, n^j_{01}, n^j_{10}, n^j_{11}\}$ to compute the estimator $\hat{M}_j^{MIMSE}$. Of most importance is that as each $n_{st} \to \infty$ (which would follow from $n \to \infty$ and the transition probabilities being interior) the MIMSE estimator converges to the MLE. Convergence to MLE is a considerable virtue, since then MIMSE inherits all of the desirable asymptotic properties (consistency and asymptotic efficiency) of MLE.

4.2. A Bayesian perspective

The use of a uniform distribution in the derivation of the MIMSE estimator suggests extending to a Bayesian analysis (see Billard and Meshkani, 1995, Cole et al., 1995). Suppose we have a sample Y and parameters $(G, H) \in [0, 1]^2$. The posterior distribution of the parameters is given by

$$P(G, H \mid Y) = \frac{P(Y \mid G, H)P(G, H)}{P(Y)} = \frac{P(Y \mid G, H)P(G, H)}{\iint P(Y \mid G, H)P(G, H)dGdH}, \tag{4.7}$$

where $P(Y \mid G, H)$ is the likelihood of the data and $P(G, H)$ is the prior distribution. In our case:

$$P(Y \mid G, H) = G^{n_{01}}(1-G)^{n_{00}}H^{n_{11}}(1-H)^{n_{10}} \tag{4.8}$$

and we take a uniform prior $P(G, H) = 1$. Then, using the same results used to obtain the closed form for the MIMSE, we have

$$\begin{aligned} P(Y) &= \int_0^1 \int_0^1 G^{n_{01}}(1-G)^{n_{00}}H^{n_{11}}(1-H)^{n_{10}}dGdH \\ &= \frac{n_{11}!n_{10}!}{(n_{10}+n_{11}+1)!}\frac{n_{01}!n_{00}!}{(n_{00}+n_{01}+1)!}. \end{aligned} \tag{4.9}$$

The posterior distribution is given by

$$P(G, H \mid Y) = G^{n_{01}}(1-G)^{n_{00}}H^{n_{11}}(1-H)^{n_{10}}\frac{(n_{00}+n_{01}+1)!(n_{10}+n_{11}+1)!}{n_{11}!n_{10}!n_{01}!n_{00}!}. \tag{4.10}$$

For a Bayesian analysis this provides all that is required from the data for subsequent analysis of, say, the Bayesian risk for the marginal dynamic effect, $M = H - G$. Our interest here is in how this relates to our estimators.

To link to estimators we consider the marginal posterior of G:

$$\begin{aligned} P(G \mid Y) &= \int_0^1 P(G, H \mid Y)dH \\ &= \frac{(n_{00}+n_{01}+1)!(n_{10}+n_{11}+1)!}{n_{11}!n_{10}!n_{01}!n_{00}!}G^{n_{01}}(1-G)^{n_{00}}\int_0^1 H^{n_{11}}(1-H)^{n_{10}}dH \\ &= G^{n_{01}}(1-G)^{n_{00}}\frac{(n_{00}+n_{01}+1)!}{n_{00}!n_{01}!}, \end{aligned}$$

where we have used (4.10). A standard result is that the MLE is the mode of the posterior distribution assuming a flat prior. In the current context, taking the derivative of this expression with respect to G, setting this equal to zero and solving for G gives the maximum likelihood estimator in equation (3.11). To show the link to the MIMSE, we have that the conditional mean

Table 7. Estimates of marginal effect for three estimators.

Case	Path	$\hat{M}^{MLE}$	$\hat{M}^{NBC}$	$\hat{M}^{MIMSE}$
a	1000	0	0	1/12
b	1001	−1/2	−3/8	−1/6
c	1010	−1	−1	−5/12
d	1011	−1/2	−1/3	−1/6
e	1100	1/2	5/6	1/6
f	1101	−1/2	−1/3	−1/6
g	1110	–	–	1/10
h	1111	–	–	3/10

of G is given by

$$\begin{aligned} E(G \mid Y) &= \int_0^1 G P(G \mid Y) dG \\ &= \frac{(n_{00} + n_{01} + 1)!}{n_{00}! n_{01}!} \int_0^1 G^{n_{01}+1}(1-G)^{n_{00}} dG \\ &= \frac{(n_{00} + n_{01} + 1)!}{n_{01}! n_{00}!} \frac{(n_{01} + 1)! n_{00}!}{(n_{00} + n_{01} + 1 + 1)!} \\ &= \frac{n_{01} + 1}{n_{00} + n_{01} + 2}, \end{aligned}$$

which is the MIMSE estimator for G (see the second expression on the right-hand side of equation (4.6)).

4.3. Comparing the MIMSE estimator with MLE and NBC, $T = 3$

We now consider how the MIMSE estimator compares to MLE and NBC in terms of finite sample bias and MSE. In the interests of comparability we shall only consider the estimates for cases a to f and exclude the cases for which MLE does not exist. In doing this, we use the adjusted probabilities given in Table 4 that take into account the sample selection. This is consistent with our earlier decision to consider the MLE derived using the uncorrected probabilities but to use the corrected probabilities when considering bias and MSE. Note that the MIMSE estimator does *not* minimize the integrated MSE for the corrected probabilities so that these comparisons are relatively *unfavourable* to MIMSE.

In Table 7 we give the three sets of values for the estimator of M. As can be seen, the estimates of M for MIMSE range from $-5/12$ to 0.3. Figure 6 shows the comparisons of bias and MSE for values of $G = 0.2, 0.5, 0.8$ and $H \in [0, 1]$ when $T = 3$. The left-hand panels give the bias and the right-hand panels give the MSE. As can be seen for the bias, sometimes MIMSE is worse than NBC and sometimes it is better. In particular, since the bias of the MIMSE estimator can be positive or negative, we can have zero bias for some parameter values (for example, at $(G, H) = (0.5, 0.366)$). Turning to the right-hand-side panels for MSE we see that the MIMSE estimator does better than MLE and NBC unless there is strong negative state dependence and

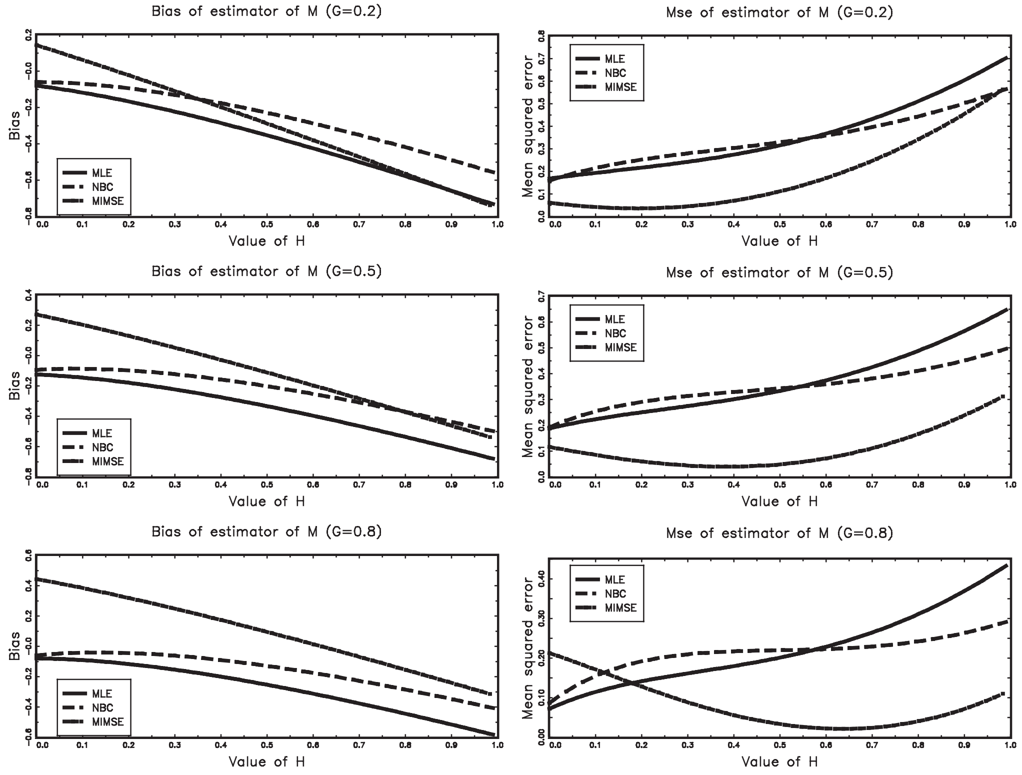

Figure 6. Bias and MSE for three estimators, $T = 3$.

sometimes does very much better. For example, for $(G, H) = (0.5, 0.8)$ (which implies moderate positive state dependence with $M = 0.3$) we have values for the MSE of 0.49, 0.41 and 0.17 for MLE, NBC and MIMSE, respectively.

5. EXACT BIAS AND MSE ANALYSIS FOR FIXED $T > 3$

As before we shall only consider sequences that start with $y_0 = 1$. When considering $T = 3$ we could write down all eight possible cases and show explicit expressions for the bias and MSE. For larger values of T, tables such as Table 3 become impractical. For the observed sequence $\{1, y_1, y_2, \ldots, y_T\}$ there are 2^T possible distinct paths; for convenience we denote 2^T by Γ. An estimator for G and H is given by a mapping from the Γ outcomes to values for $\hat{G}$ and $\hat{H}$. Given (3.11) and (3.12), the bias of the MLE estimators is given by

$$bias(\hat{G}^{MLE}) = \left(\frac{1}{(1 - H^{T-1})} \sum_{j=1}^{\Gamma-2} p_j \left(\frac{n_{01}^j}{n_{00}^j + n_{01}^j} \right) \right) - G, \tag{5.1}$$

$$bias(\hat{H}^{MLE}) = \left(\frac{1}{(1-H^{T-1})} \sum_{j=1}^{\Gamma-2} p_j \left(\frac{n_{11}^j}{n_{10}^j + n_{11}^j} \right) \right) - H. \tag{5.2}$$

Note that the summation is from 1 to $(\Gamma - 2)$ since the last two cases are selected out. The MSEs for the MLE are given by

$$\begin{aligned} MSE(\hat{G}) &= \frac{1}{(1-H^{T-1})} \sum_{j=1}^{\Gamma-2} p_j \left(\left(\frac{n_{01}^j}{n_{00}^j + n_{01}^j} \right) - G \right)^2, \\ MSE(\hat{H}) &= \frac{1}{(1-H^{T-1})} \sum_{j=1}^{\Gamma-2} p_j \left(\left(\frac{n_{11}^j}{n_{10}^j + n_{11}^j} \right) - H \right)^2. \end{aligned} \tag{5.3}$$

These are exact analytical expressions for the bias and MSE. We cannot derive closed form expressions for these (mainly because we cannot display closed form expressions for n_{st}^j) but we can compute the exact values numerically using these formulas. We postpone presenting these until after we define the bias corrected estimator.

As before, we can define a new estimator by taking the bias of the MLE estimator, assuming that the values of G and H are the estimated values and then bias correcting. This gives[12]

$$\begin{aligned} \hat{G}_j^{NBC} &= 2\hat{G}_j^{MLE} - \left[\sum_{k=1}^{\Gamma-2} \tilde{p}_k \left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE} \right) \hat{G}_k^{MLE} \right], \\ \hat{H}_j^{NBC} &= 2\hat{H}_j^{MLE} - \left[\sum_{k=1}^{\Gamma-2} \tilde{p}_k \left(\hat{G}_j^{MLE}, \hat{H}_j^{MLE} \right) \hat{H}_k^{MLE} \right]. \end{aligned} \tag{5.4}$$

Finally, the MIMSE estimator is given by (4.6).

We turn now to the performance of our three estimators as we increase T from 3 to 12. We consider three cases: $(G, H) = (0.75, 0.25)$, $(0.5, 0.5)$ and $(0.25, 0.75)$. The first of these cases is somewhat extreme in that $M = -0.5$ and the y variable has a high probability of changing from period to period. For most contexts (for example, state dependence due to habit formation as in our empirical example) this will never be considered. Nonetheless, there may be circumstances, such as the purchase of a particular small durable, when we see this sort of behaviour. Figure 7 shows the results. The left-hand panels give the bias against T and the right-hand panels give the MSEs against T; note that the y-axis scales vary from panel to panel. We consider first the (absolute) biases. There are two aspects to this. First, how big is the bias for very small T? And, second, how quickly does the bias converge to zero (if it does) as T increases (see, for example, Carro, 2007; Hahn and Newey, 2004). Since we gave an exact analysis of the former for $T = 3$ in the previous section we concentrate here on the second issue. For all three cases shown in Figure 7 the NBC estimator usually has the smallest bias (in absolute value) and appears to be converging to zero faster.[13] Taking the values shown in the figure we can actually be more precise

[12] Since we have to sum over all the $\Gamma - 2$ cases to calculate the NBC, the computation time increases with T. However, for any $T < 24$ a regular PC takes less than a minute in computing the NBC. We have not tried higher T, because most of the micropanels found in practice have fewer than 25 periods.

[13] It is worth noting that the biases for G and H are not so regular and are not even always monotone decreasing in T. Despite this, the difference, M, is well behaved.

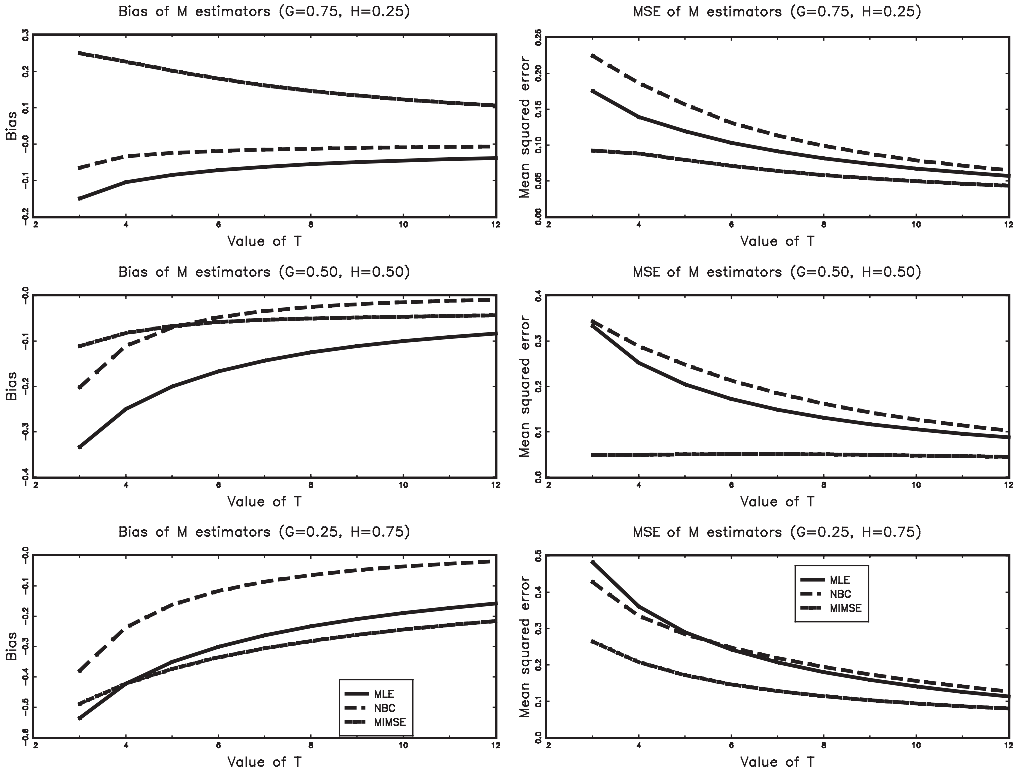

Figure 7. Bias and MSE for estimators of marginal dynamic effect.

than this. To a very high order of approximation we have that the bias of estimator e, which we denote b_e, is polynomial in T with the following dominant term:

$$b_e \simeq \kappa_e T^{\delta_e}, \tag{5.5}$$

where κ_e and δ_e are functions of (G, H). For the case $G = H = 0.5$ and the MLE estimator this is an exact relationship with $\kappa_{MLE} = -1$ and $\delta_{MLE} = -1$ so that the bias is always exactly $-1/T$. Regressions, for the three cases we consider, of the log (absolute) bias on log T gives values of $\delta_{MLE} \simeq -0.9$, $\delta_{NBC} \simeq -2$ and $\delta_{MIMSE} \simeq -0.6$. Thus the bias disappears fastest for NBC and slowest for MIMSE. The exact rates for MLE and NBC are close to the expected orders of $O(T^{-1})$ and $O(T^{-2})$, respectively. Given that the bias of NBC is also usually lowest for $T = 3$ this corroborates what the figure suggests, namely that NBC is superior to MLE and MIMSE in terms of bias.

In addition to the dependence on T, the bias also depends on the values of G and H; that is, on the number of transitions we have. Given that we are analysing paths that start with 1, the bias, for any given T, is higher the closer H is to 1. That is, the bias is higher, the higher the probabilities of having paths with no changes. This effect of H on the bias can be seen in Figure 6, and in Figure 7 where for any given T the bias of the MLE is always higher in the third panel which has the higher H ($H = 0.75$).

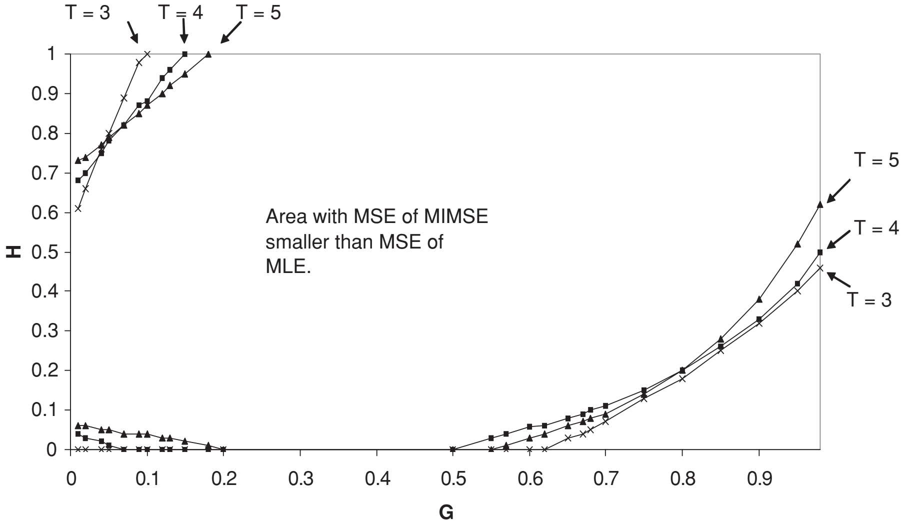

Figure 8. MIMSE and MLE of *M* in terms of MSE.

Turning to the MSEs a radically different pattern emerges. MIMSE has the lowest MSE in all cases and MLE is almost always better than NBC. One feature to note is that although the MLE MSE is clearly converging towards the MIMSE MSE (as we expect theoretically) it is still significantly higher even when we have $T = 10$. The figures for these three cases suggest that MIMSE is usually best in MSE terms. In Figure 8, we display the values of G and H in the unit square for which MIMSE is MSE better than MLE for values of $T = 3, 4, 5$. Note that the sets for the different values of T are not nested The MIMSE estimator performs worse only for extreme values of G and H, particularly those that imply a very negative state dependence.

6. MANY HOUSEHOLDS

6.1. Using MLE, NBC and MIMSE to estimate the distribution of M

In the previous three sections we have considered households in isolation and treated their observed paths as separate time series. However, in most empirical analyses, the interest is not in individual households but in the population. Thus it may be that the distribution of M in the population is of primary interest, rather than the values for particular households. We now consider how the estimators we had before—MLE, NBC and MIMSE—could be used in estimating the distribution of M on the population. We take $T = 9$ (that is, 10 observations per unit, including the initial observation) as being a 'reasonably' long panel in practical terms, but still short enough to give concern over small sample bias. As before we continue with the context in which $y_{i0} = 1$. We present results for three different distributions of (G, H). Firstly we consider a uniform distribution for (G, H) over $[0, 1]^2$. For this distribution we have exact calculations of the properties of the estimators when $T = 9$ and N goes to infinity. The second distribution is the empirical distribution of (G, H) for the 367 households considered in the empirical section above. In this case we simulate a sample with $T = 9$ and large N to display the properties of the estimators when we pool many households. For the final set of simulations we take a uniform distribution for G on $[0.1, 0.9]$ and impose the homogeneous state dependence parameter condition (3.6) for H with a Normal distribution:

$$H_i = \Phi(0.81 + \Phi^{-1}(G_i)). \tag{6.1}$$

The value of $\alpha = 0.81$ is taken from the empirical estimate of (2.1). This can be considered a 'standard' model with homogeneous state dependence.

For the first case, the true distribution of M over the population has the following cdf:

$$F_{M_i}(x) = \begin{cases} \frac{1}{2}(1 + 2x + x^2) & \text{if } x \leq 0, \\ \frac{1}{2}(1 + 2x - x^2) & \text{if } x > 0, \end{cases} \tag{6.2}$$

and pdf

$$f_{M_i}(x) = \begin{cases} (1 + x) & \text{if } x \leq 0, \\ (1 - x) & \text{if } x > 0. \end{cases} \tag{6.3}$$

This implies that the mean and median value of the marginal dynamic effect M are zero. To calculate the estimated distributions, firstly note that $\hat{M}_i$ can only take one of 2^T possible values, since any household sequence observed on the pooled sample will correspond with one of the 2^T combinations of 1's and 0's we can have conditional on the first observation. Then, the distribution of $\hat{M}_i$ when N goes to infinity and T is fixed is given by the probabilities of observing each path j on a pooled sample:

$$\begin{aligned} \Pr(j) = \Pr(j \mid H, G)\Pr(H, G) &= \int_G \int_H p_j f(G, H)\, dGdH \\ &= \int_G \int_H (G)^{n^j_{01}}(1 - G)^{n^j_{00}}(H)^{n^j_{11}}(1 - H)^{n^j_{10}} f(G, H)\, dGdH. \end{aligned} \tag{6.4}$$

In the case of a uniform distribution we are considering,

$$\Pr(j) = \int_0^1 \int_0^1 (G)^{n^j_{01}}(1 - G)^{n^j_{00}}(H)^{n^j_{11}}(1 - H)^{n^j_{10}}\, dGdH \tag{6.5}$$

$$= \frac{n_{11}!n_{10}!}{(n_{10} + n_{11} + 1)!} \frac{n_{01}!n_{00}!}{(n_{00} + n_{01} + 1)!}. \tag{6.6}$$

From this we can derive the distribution of $\hat{M}$ as $N \to \infty$ with a fixed T. The differences in the estimated distribution between the three estimators comes from the different $\hat{M}_i$'s estimated from a given path j (this is what we have studied in previous sections). Figures 9 and 10 give the graphical comparisons of the true distribution and the estimated distributions based on the estimates of M_i for each possible path by MLE, NBC and MIMSE, conditioning on identification of the MLE for $T = 9$ and $N \to \infty$, uniform case. The first, Figure 9, shows the cumulative distributions and the second, Figure 10, shows the $Q - Q$ plot; although the two figures are informationally equivalent, the latter reveals to the eye different detail to the former. Consider first the MLE and NBC estimators. The NBC cdf is always to the right of the MLE estimator, and for many values NBC is closer to the true

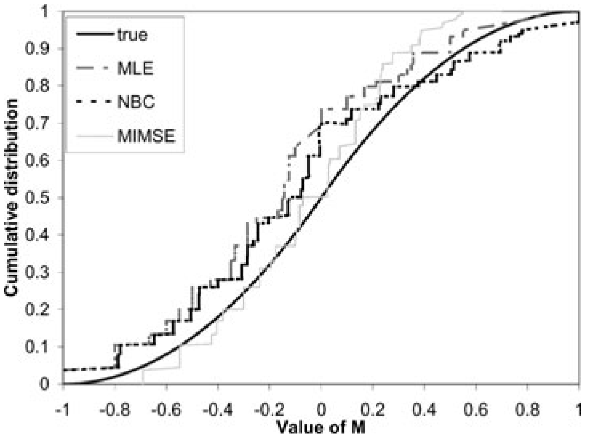

Figure 9. Estimates of the distribution of M.

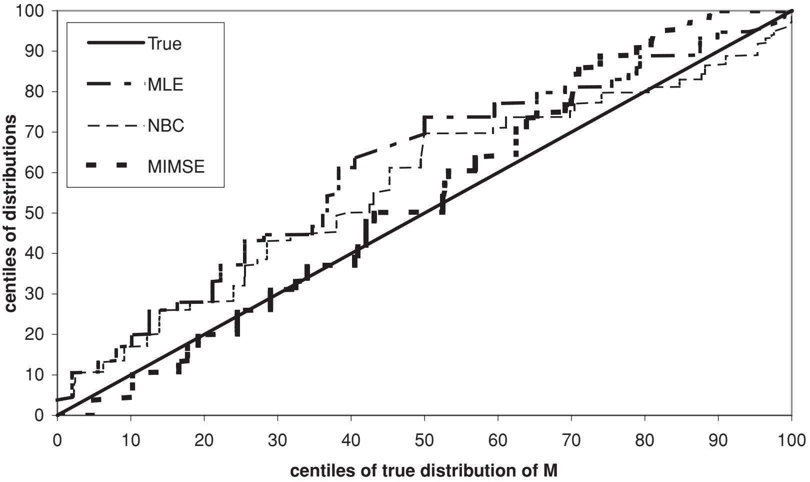

Figure 10. Q–Q plot of estimators of M.

distribution, since the bias is higher in absolute value for MLE as compared to NBC. However, these figures also show that the NBC estimate does worse than MLE for high values of the marginal effect. Thus the lower bias at the lower end for NBC is cancelled out by the higher bias for higher values of M. Hence the MLE usually has a lower variance. A conventional statistic to measure the difference between a true distribution and one for an estimator is the absolute value of the difference between them; that is the Kolmogorov–Smirnov (K–S) statistic:

$$D = \sup_{m \in [-1,1]} |\widehat{F(m)} - F(m)|. \tag{6.7}$$

The NBC estimator dominates the ML estimator on this criterion. Turning to the MIMSE estimator, we see from the $Q - Q$ plot that up to the 6th decile this tracks the true value very closely; the divergences are mainly due to the MIMSE estimator taking on a finite number of values. In particular, the median of the MIMSE estimator and the true distribution are very close. The estimated medians when $N \to \infty$ and $T = 9$ by MLE, NBC and MIMSE converge to −0.14, −0.11 and −0.07, respectively. This close correspondence is to be expected given that the estimator was derived assuming a distribution close to the one used here. At the top of the M distribution, however, the MIMSE tends to underestimate the true value (that is, the cdf is to the left of the true cdf). Despite these differences, the main conclusion from the two figures is that the MIMSE estimator is considerably better than either MLE or NBC in terms of the fit to the true cdf.

The probabilities in (6.6) can also be used in deriving the asymptotic properties of estimates of moments of M_i as $N \to \infty$ and T is held fixed. This can then be used as an approximation to exact finite sample properties in panels with large N and small T. Looking at the mean marginal dynamic effect, the estimated average from our three estimators ($\hat{\hat{M}} = \frac{1}{N}\sum_{i=1}^{N}\hat{M}_i$) converge to the true value as $(T, N) \to \infty$, because $\hat{M}_i \to M_i$ and the sample average converges to the population mean. But for a given T, as $N \to \infty$,

$$\hat{\hat{M}} \to_p E(\hat{M}_i) \neq E(M_i) \tag{6.8}$$

as long as $\hat{M}_i$ is a biased estimator of M_i.[14] Therefore, $\Pr(j)$ in (6.6) will give the probabilities of each possible value of $\hat{M}_i$, allowing us to calculate the asymptotic properties as $N \to \infty$, of the estimators based on moments of M_i:

$$\hat{\hat{M}} \to_p E(\hat{M}_i) = \sum_j \Pr(j)\hat{M}_j, \tag{6.9}$$

$$\sqrt{N}(\hat{\hat{M}} - bias(\hat{M}_i)) \to_d N(0, Var(\hat{M}_i)), \tag{6.10}$$

where $bias(\hat{M}_i) = E(\hat{M}_i) - E(M_i)$. When N goes to infinity and T equals 9, the MLE, NBC and MIMSE estimates of the mean of M converge to −0.16, −0.08 and −0.05, respectively. Thus pooling gives that the MIMSE has lower asymptotic bias than NBC for the mean of M. As for the asymptotic root MSE, we have values of 0.21 for the MLE, 0.25 for the NBC and 0.10 for the MIMSE. Thus MIMSE is best for this criterion and NBC is worst.

The top panels of Figure 11 present results using the empirical distribution (G, H) for the 367 households considered in Section 2. For each pair we simulate 50 paths of length 10 with an initial value of unity (so that we have 18,350 paths in all, before we select out the paths for which MLE is not identified). The mean of M for the data is 0.23 (positive mean state dependence) and the means of the estimates from the simulated data are 0.08, 0.17 and 0.14 for MLE, NBC and MIMSE, respectively. Thus the bias is negative in all three cases and largest in absolute value for MLE and smallest for NBC. This reflects the fact that the NBC estimator usually has a lower bias for any particular path (see Section 5). The median of M for the data is 0.178 and the estimates are 0, 0 and 0.150 for the MLE, NBC and MIMSE, respectively. The latter displays much less bias than the other two estimators. One notable feature of these distributions is that all three display a sharp jump at some point in the distribution; at zero for MLE and NBC (hence the median result) and at about 0.25 for MIMSE. It is this clustering (around zero for MLE and NBC

[14] Also, note that $E(\hat{\hat{M}}) = E(\hat{M}_i)$.

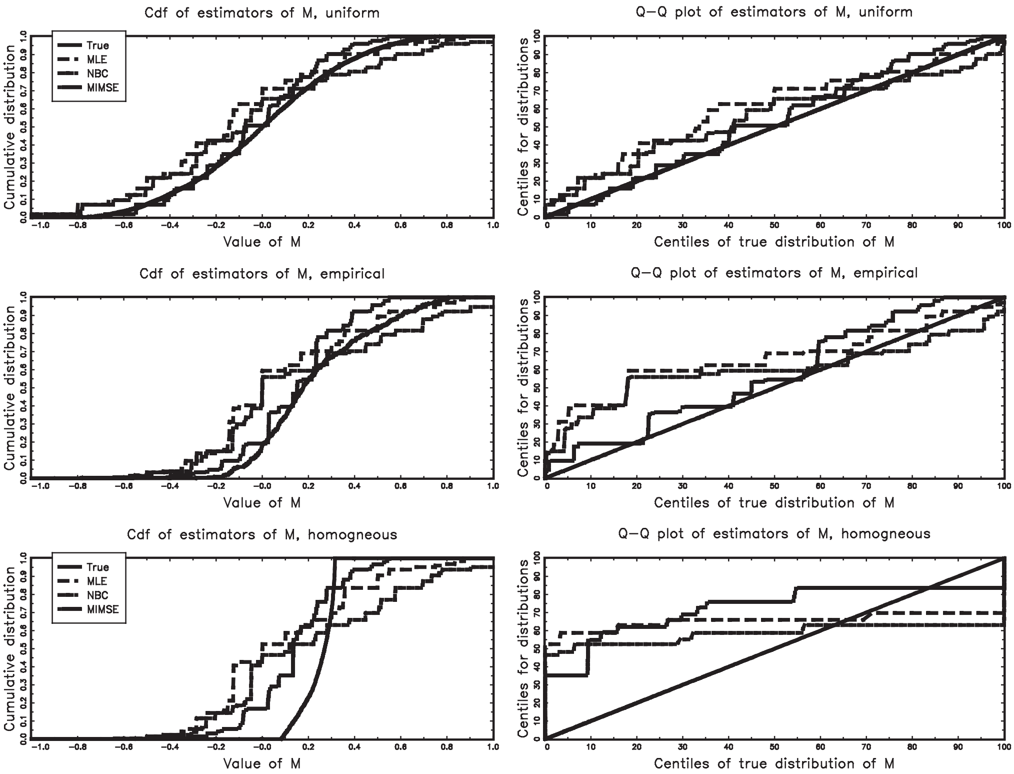

Figure 11. Estimates of the distribution of M.

and close to the true mean for MIMSE) that seems to give the lower mean bias for the MIMSE. Once again, the MIMSE estimator gives a much closer fit to the true distribution.

The final set of simulations assume that the state dependence parameter in a parametric Normal model is constant across the population (note that this does *not* impose that the dynamic marginal effect is the same for everyone). The results for the three estimators are given in the bottom panels of Figure 11. When comparing the three estimators, the conclusion is the same as in the other two simulations: the MIMSE estimator is clearly better than MLE or NBC. And this is true here even for the higher percentiles. Note that the overall fit for all estimators is much worse in the case, mainly due to the efficiency loss caused by not imposing a constant state dependence parameter when estimating. This emphasizes the importance of first testing for slope homogeneity (see Pesaran and Yamagata, 2008).

6.2. Finite sample comparisons

In the previous subsection, we examined the estimated distribution when the number of households becomes large. To end this section, we look at the finite sample performance of the three estimators, in terms of mean bias and root mean squared error (RMSE), when we want to estimate the mean and some quartiles of the distribution of M, with samples where the number of households is large but not unduly so and the number of periods is small. We consider the same three experiments as in the previous subsection. The first simulation experiment consider a uniform distribution for (G, H) over $[0.1, 0.9]^2$. The second distribution is the empirical distribution of (G, H) for the 367 households considered in the empirical section above. For the final set of simulations we take a uniform distribution for G on $[0.1, 0.9]$ and impose the homogeneous state dependence parameter condition (3.5) for H with a Normal distribution, as in equation (6.1), with $\alpha = 0.81$. In all of them, $y_{i0} = 1$ and the number of households N is equal to 367 (the number of households in the empirical illustration). As before, we exclude observations for which MLE is not identified.

As before, we take the number of observed periods equal to 10 ($T = 9$). Table 8 contains the true values and mean estimates of the mean marginal dynamic effect (mean M), and of the median and the other quartiles of the distribution of M. Mean bias and RMSE over 1000 simulations are also reported. The results are in accordance with the conclusions from the previous subsection. In terms of RMSE, the MIMSE estimator is significantly better than other two, except for the highest quartile, where MLE has a better RMSE in two of the three experiments. For the mean marginal dynamic effect, NBC has slightly smaller RMSE than MIMSE in the last two experiments. However, the NBC estimator of the median M, performs significantly worse than MIMSE, both in terms of mean bias and RMSE.

7. EXTENSION TO THE CASE WITH COVARIATES

In the previous sections, we considered the case without covariates. This allowed us to derive exact results for the MLE (including the sign of the bias in the dynamic marginal effect) and to consider exact bias corrections instead of corrections based on the leading terms of an asymptotic approximation. We now extend the application of the alternative estimators proposed in this paper to the semiparametric case with covariates, allowing for full heterogeneity. That is,

$$y_{it} = 1\left\{\alpha_i y_{it-1} + x_{it}'\beta_i + \eta_i + v_{it} \geq 0\right\} \quad t = 0, \ldots, T;\ i = 1, \ldots, N, \tag{7.1}$$

where x_{it} is a vector of exogenous variables. We assume that identification conditions are satisfied. These include, for instance, the condition that x_{it} covariates vary over time for person i.

7.1. *Discrete covariates*

If x_{it} contains only discrete variables, it is conceptually simple to extend our estimators. For a single binomial covariate we have:

$$\begin{aligned} H_{i0} &= \Pr(y_{it} = 1 \mid y_{it-1} = 1, x_{it} = 0), \\ G_{i0} &= \Pr(y_{it} = 1 \mid y_{it-1} = 0, x_{it} = 0), \\ H_{i1} &= \Pr(y_{it} = 1 \mid y_{it-1} = 1, x_{it} = 1), \\ G_{i1} &= \Pr(y_{it} = 1 \mid y_{it-1} = 0, x_{it} = 1), \end{aligned}$$

as parameters to be estimated. The estimators are analogous to the case without covariates. The only difference is that now we have to look not only at the 0 → 1 transition in the y_{it} but also at the possible values of x_{it}. That is, the likelihood of an observed path is

$$G_{0s}^{n^j_{01|0}}(1 - G_{0s})^{n^j_{00|0}} H_{0s}^{n^j_{11|0}}(1 - H_{0s})^{n^j_{10|0}} G_{1s}^{n^j_{01|1}}(1 - G_{1s})^{n^j_{00|1}} H_{1s}^{n^j_{11|1}}(1 - H_{1s})^{n^j_{10|1}}, \tag{7.2}$$

Table 8. Estimation of the quartiles and the mean of the distribution of M on the population.

		Parameters of interest			
		1st quartile	Median	3rd quartile	Mean
	1. (G, H) from a Uniform distribution $(0.1, 0.9)^2$				
	True value	−0.233	0	0.237	0
Mean estimate	MLE	−0.384	−0.144	0.121	−0.129
	NBC	−0.344	−0.081	0.245	−0.046
	MIMSE	−0.240	−0.030	0.156	−0.041
Mean bias	MLE	−0.150	−0.143	−0.116	−0.130
	NBC	−0.110	−0.080	0.008	−0.047
	MIMSE	−0.006	−0.030	−0.081	−0.041
RMSE	MLE	0.158	0.143	0.125	0.132
	NBC	0.124	0.083	0.046	0.053
	MIMSE	0.023	0.058	0.084	0.044
	2. (G, H) from the empirical distribution for the 367 households				
	True value	0.047	0.177	0.381	0.227
Mean estimate	MLE	−0.143	0.000	0.327	0.080
	NBC	−0.124	0.002	0.488	0.169
	MIMSE	0.026	0.153	0.240	0.139
Mean bias	MLE	−0.190	−0.176	−0.054	−0.148
	NBC	−0.171	−0.175	0.107	−0.058
	MIMSE	−0.021	−0.024	−0.141	−0.088
RMSE	MLE	0.190	0.176	0.063	0.149
	NBC	0.172	0.175	0.115	0.062
	MIMSE	0.023	0.029	0.142	0.089
	3. Homogeneous state dependence parameter				
	True value	0.1597	0.2516	0.2974	0.2202
Mean estimate	MLE	−0.1259	0.0013	0.3397	0.0926
	NBC	−0.0520	0.0430	0.5084	0.1977
	MIMSE	0.0359	0.1440	0.2441	0.1529
Mean bias	MLE	−0.2856	−0.2503	0.0422	−0.1276
	NBC	−0.2117	−0.2087	0.2109	−0.0225
	MIMSE	−0.1239	−0.1076	−0.0533	−0.0674
RMSE	MLE	0.2856	0.2506	0.0475	0.1291
	NBC	0.2118	0.2162	0.2137	0.0323
	MIMSE	0.1248	0.1085	0.0560	0.0683

where $n^j_{01\,|\,0}$ is the number of $y_{t-1} = 0 \to y_t = 1$ transitions for path j given $x_t = 0$, $n^j_{01\,|\,1}$ 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. For example, the MLE for H_{i0} is given by $\hat{H}^{MLE}_{i0} = n^j_{11\,|\,0}/(n^j_{10\,|\,0} + n^j_{11\,|\,0})$ (and similarly for the other parameters). The NBC and MIMSE estimators are obtained similarly by following the same procedure as in previous sections. In fact, the MIMSE estimator has the same simple form as in (4.6), replacing n^j_{01} by $n^j_{01\,|\,0}$ or $n^j_{01\,|\,1}$ depending on the whether we want to obtain a transition probability given $x_{it} = 0$ or given $x_{it} = 1$.

7.2. *Continuous covariates*

The approach in the previous subsection has the virtue of being non-parametric but it quickly becomes infeasible as the number of values of the discrete covariate increases and/or as the number of covariates increases. Moreover, this non-parametric approach is not feasible if we have a continuous covariate. Therefore we go back to the parametric assumption about v_{it} in (7.1) that we made in Section 2. In this subsection, for simplicity in the notation we consider only one x covariate. We assume that $-v_{it}$ follows a distribution with cdf F. Then, the log-likelihood for each i is

$$lk_i(\gamma_i) = \sum_{t=1}^{T} \log[F(\alpha_i y_{it-1} + \beta_i x_{it} + \eta_i)(2y_{it} - 1) + 1 - y_{it}] \tag{7.3}$$

and the MLE of $\gamma_i = (\alpha_i, \eta_i, \beta_i)'$ is the value that maximizes lk_i. The first-order conditions are a set of non-linear equations that do not have a closed form solution.

Here it is not possible to repeat the analysis in previous sections to derive the exact bias of the MLE, even numerically. It is only possible to get information about the bias by simulation. Although simulation is a very good way of finding the bias of an estimator of the parameters in (7.1), a correction made based on the bias from simulations will include the simulation error on top of the problems inherent in non-linear bias corrections. Moreover, as we have seen for the case without covariates, the MIMSE estimator outperforms both the MLE and NBC in terms of MSE. As a result of all this, we consider only the MIMSE estimator as an alternative to the MLE.

In what follows, we omit the subscript i since, as before, we are considering each individual in isolation. The MSE for an estimator $\hat{\gamma}_j$ (where j refer to an observed path of zeros and ones), conditional on x, is given by

$$\lambda(\hat{\gamma}_j; \gamma) = E[(\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma)\,|\,x; \gamma] = \sum_j p_j(\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma), \tag{7.4}$$

where p_j is the likelihood of the observations of path j conditional on x:

$$p_j = \prod_{t=1}^{T} (F(\alpha y_{jt-1} + \beta x_t + \eta)(2y_{jt} - 1) + 1 - y_{jt}). \tag{7.5}$$

The non-informative or flat prior for parameters between $-\infty$ and $+\infty$ is the Jeffrey's prior:

$$p(\gamma)d\gamma = d\gamma. \tag{7.6}$$

This is the equivalent to the uniform prior when the parameter is in the [0, 1] interval that we used in Section 4.[15] This gives the *integrated MSE*:

$$\psi = \int \lambda(\hat{\gamma}_j; \gamma) d\gamma = \int \sum_j p_j (\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma) d\gamma, \tag{7.7}$$

where the integrals are between $-\infty$ and $+\infty$. The criterion (7.7) is additive in functions of j so that we can find minimizing values of the estimator considering each case in isolation. MIMSE is the value of $\hat{\gamma}_j$ that minimize ψ. Differentiating (7.7) with respect to $\hat{\gamma}_j$, setting the result to zero and solving for $\hat{\gamma}_j$ gives

$$\hat{\gamma}_j^{MIMSE} = \frac{1}{\int p_j d\gamma} \int p_j \gamma d\gamma. \tag{7.8}$$

Since (7.8) do not have an analytical closed form solution, these integrals have to be solved numerically. Alternatively, we can take advantage of the relation between MIMSE and the mean of the posterior of γ when using a flat prior. The posterior distribution of γ if the priors are those in (7.6) is

$$P(\gamma \mid Y) = \frac{1}{\int p_j d\gamma} p_j, \tag{7.9}$$

where p_j is the likelihood of the data (given γ and x) written in (7.5). Since the denominator is a constant that does not depend on $\hat{\gamma}$ nor γ, we have that

$$\begin{aligned}
\min_{\hat{\gamma}_j} \psi &= \min_{\hat{\gamma}_j} \int (\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma) p_j d\gamma \\
&= \min_{\hat{\gamma}_j} \frac{1}{\int p_j d\gamma} \int (\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma) p_j d\gamma \\
&= \min_{\hat{\gamma}_j} \int \left((\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma) \frac{1}{\int p_j d\gamma} p_j \right) d\gamma \\
&= \min_{\hat{\gamma}_j} \int (\hat{\gamma}_j - \gamma)'(\hat{\gamma}_j - \gamma) P(\gamma \mid Y) d\gamma.
\end{aligned} \tag{7.10}$$

Therefore, minimizing the *integrated MSE* is equal to minimizing the expected posterior loss function with a quadratic loss function. As it is proved, for instance, in page 24 of Zellner (1971), this minimum in (7.10) is equal to the mean of the posterior function. Hence, we can obtain the MIMSE estimates of γ by computing the mean of the posterior function (7.9).

The only difficult part to compute in (7.9) is $\int p_j d\gamma$, since the likelihood has a simple analytical form and does not require any integral (given a known cdf). The Metropolis–Hastings Algorithm can be used to obtain draws from the posterior density without computing $\int p_j d\gamma$. Since the likelihood can be calculated very easily, we can make very many iterations in this MCMC algorithm, both to guarantee convergence to the posterior and to obtain a good number of valid draws. Once we have many draws we simple compute the average to obtain the MIMSE estimator. One important advantage of this procedure is that we can automatically accommodate a large number of covariates (subject to the identification conditions).

[15] See Zellner (1971) for further discussion on non-informative priors.

In order to illustrate the usefulness of MIMSE when having covariates, we simulate (7.1), obtain the marginal dynamic effect M from the MLE and the MIMSE estimates for different T and values of the parameters. In particular, we choose values of γ so that G, H and M evaluated at the mean value of x are equal to the values used in Figure 6 for the case without covariates. This will allow to have graphs as comparable as possible. The specific details of the simulations are: $\beta = 1$, $x_{it} \underset{\text{i.i.d.}}{\sim} N(0, 1)$, $v_{it} \underset{\text{i.i.d.}}{\sim}$ logistic, we make 10,000 simulations and 20,000 iterations in the Metropolis–Hastings Algorithm of which the first 10,000 are for burn-in and of the 10,000 made after convergence every fifth is retained as a draw from the posterior. This is made for $T = 4, \ldots, 13$ and for the following three values of the α and η parameters. Firstly with $\alpha = -2.2$ and $\eta = 1.1$, which imply $G = 0.75$, $H = 0.25$ and $M = -0.5$ when computed at the mean of x. Secondly, with $\alpha = 0$ and $\eta = 0$, which imply $G = 0.5$, $H = 0.5$ and $M = 0$ when computed at the mean of x. Thirdly, with $\alpha = 2.2$ and $\eta = -1.1$, which imply $G = 0.25$, $H = 0.75$ and $M = 0.5$ when computed at the mean of x.

As we did in Figure 7, Figure 12 shows the mean bias and MSE of the MLE and MIMSE in estimating the marginal dynamic effect at the mean value of x for those three sets of values of the parameters as we increase T. Note that the results as T increases are not as smooth as in

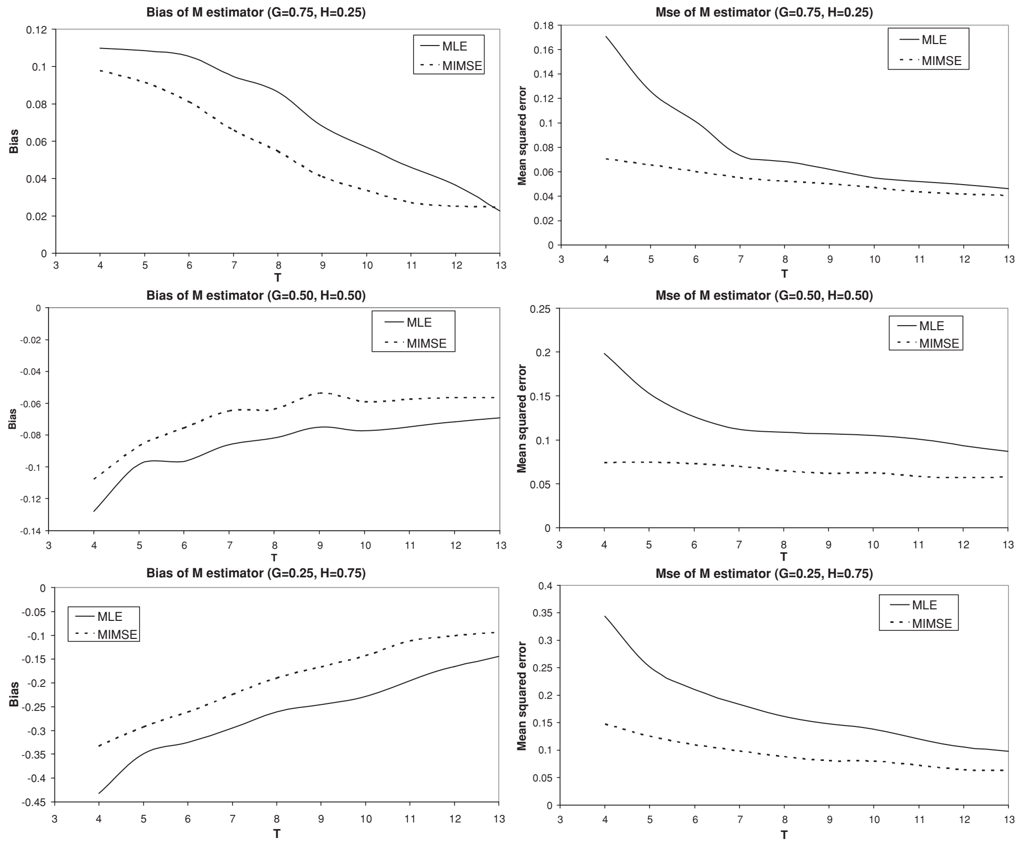

Figure 12. Bias and MSE for estimators of marginal dynamic effect in a model with covariates.

Figure 7, because this is based on simulations and seven are exact calculations. As can be seen, the MIMSE is performing better than the MLE in terms of MSE for all T and the three values of (G, H) as it did in the case without covariates. However, here MIMSE is better than the MLE also in terms of bias. In the comparison between Figures 7 and 12, it is important to note that, for most of the cases, both the MLE and the MIMSE have smaller biases (in absolute value) and MSEs in the case with covariates than in the case without covariates. This is not surprising since we have added exogenous variations to the model. This is an indication that the detail and exact results for models without covariates in previous sections could be taken as a worse case reference when adding exogenous covariates. For a similar reason, the marginal dynamic effect in model (7.1) that is considered in Figure 12, is more problematic than looking at the marginal effect of x.

8. CONCLUSIONS

We have considered in detail the dynamic choice model with heterogeneity in both the intercept (the ‘fixed effect’) and in the autoregressive parameter. We motivated this analysis by considering the estimates from a long panel in which we could effectively treat each household as a single time series. This analysis suggested strongly that both the parameters vary systematically across households. Moreover, the results of this analysis gave us a joint distribution over the two latent variables that may be difficult to pick up with the usual fully parametric random coefficients model. Consequently, we examined the finite sample properties of non-parametric estimators. In the case without covariates we present exact analytical results for the bias and MSE.

We found the following for a simple two-state first-order Markov chain model:

(1) There is no unbiased estimator for the transition probabilities.

(2) Conditioning on identification, we found that the MLE estimate of the marginal dynamic effect:

$$\Pr(y_{it} = 1 \mid y_{i,t-1} = 1) - \Pr(y_{it} = 1 \mid y_{i,t-1} = 0) \tag{8.1}$$

has a negative bias. This is the non-linear analogue of the Nickell finding that in the linear autoregressive model panel data estimates of the autoregressive parameter are biased toward zero but note that our results are exact finite sample calculations. The degree of bias depends on the parameter values and the length of the panel, T. The bias of the MLE estimator of the marginal dynamic effect does diminish as we increase the length of the panel, but even for $T = 16$ it can be high.

(3) Based on the analysis of bias, we constructed an NBC estimator as a two-step estimator with the MLE as the first step. We find that this estimator does indeed reduce the bias for most cases (as compared to MLE) but in MSE terms it is similar or even worse than MLE. For all but extreme values of negative state dependence, the NBC estimator also has a negative bias for the marginal dynamic effect. A detailed examination of the distribution of the MLE and NBC estimators for $T = 3$ and $T = 10$ suggested that neither can be preferred to the other.

(4) Given the relatively poor performance of the MLE and NBC in terms of MSE, we constructed an estimator that MIMSE and that has a simple closed form. This estimator coincides with the mean of the posterior distribution assuming a uniform prior. The MIMSE

estimator is sometimes better than MLE and NBC in terms of bias but usually it is worse. In terms of MSE, however, it is much better than either of the first two estimators, particularly when there is some positive state dependence.

(5) Turning to the many-person context, we considered a joint distribution of $\Pr(y_{it} = 1 \mid y_{i,t-1} = 1)$ and $\Pr(y_{it} = 1 \mid y_{i,t-1} = 0)$ over the population and use our non-parametric estimators to estimate the empirical distribution of the parameters. Exact calculations and simulations with $T = 9$ and large N suggest that the MIMSE-based estimator significantly outperforms the MLE and NBC estimators in recovering the distribution of the marginal dynamic effect.

The conclusion from our exact analyses on a single observed path and from simulations in a many-unit context is that the MIMSE estimator is superior to MLE or a particular bias corrected version of MLE.

As emphasized in Section 3, we deemed it necessary to examine the no-covariate case in great detail given that we know very little about the performance of alternative dynamic choice estimators which allow for a great deal of heterogeneity. However, for most analyses, we would also want to condition on covariates. The results in Section 7 suggest that MIMSE is a credible and feasible candidate for estimating dynamic discrete choice models with exogenous covariates.

ACKNOWLEDGMENTS

The authors thank two referees, the editor, Manuel Arellano, Bo Honoré, Thierry Magnac, Enrique Sentana and participants in the CAM workshop on ‘Limited dependent variable models’ in July 2004, the 12th Conference on Panel Data in Copenhagen, a CEMFI seminar, and a seminar at Universidad de Alicante for helpful comments and discussion. This work was supported by the EU AGE RTN, HPRN-CT-2002-00235 and by the Danish National Research Foundation through its grant to CAM.

REFERENCES

Albert, P. and M. Waclawiw (1998). A two state Markov chain for heterogeneous transitional data: a quasi-likelihood approach. *Statistics in Medicine 17*, 1481–93.

Anderson and Goodman (1957). Statistical inference about Markov chains. *Annals of Statistics 28*, 89–100.

Arellano, M. (2003a). Discrete choice with panel data. *Investigaciones Económicas XXVII*, 423–58.

Arellano, M. (2003b). *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M. and B. Honoré (2001). Panel data models: some recent developments. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, *Volume 5*, 3229–96. Amsterdam: Elsevier Science.

Berkson, J. (1980). Minimum chi-squared, not maximum likelihood! *Annals of Statistics 8*, 457–87.

Billard, L. and M. Meshkani (1995). Estimation of a stationary Markov chain. *Journal of the American Statistical Association 90*, 307–15.

Browning, M. and J. M. Carro (2006). Heterogeneity and Microeconometrics Modelling. In R. Blundell, W. K. Newey and T. Persson (Eds.), *Advances in Economics and Econometrics*, *Theory and Applications: Ninth World Congress of the Econometric Society*, *Volume 3*, 45–74. New York: Cambridge University Press.

Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics 140*, 503–28.

Cole, B. F., M. T. Lee, G. A. Whitmore and A. M. Zaslavsky (1995). An empirical Bayes model for Markov-dependent binary sequences with randomly missing observations. *Journal of the American Statistical Association 90*, 1364–72.

Diggle, P., P. Heagerty, K.-Y. Liang and S. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford: Oxford University Press.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel data models. *Econometrica 72*, 1295–319.

Honoré, B. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica 68*, 839–74.

Honoré, B. and E. Tamer (2006). Bounds on parameters in dynamic discrete choice models. *Econometrica 74*, 611–29.

McKinnon, J. G. and A. A. Smith (1998). Approximate bias correction in econometrics. *Journal of Econometrics 85*, 205–30.

Pesaran, H. and T. Yamagata (2008). Testing slope homogeneity in large panels. *Journal of Econometrics 142*, 50–93.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley.

APPENDIX

A.1. Proof of Proposition 3.1

Take any estimator $\{\hat{G}, \hat{H}\} = \{(G_a, G_b, \ldots, G_h),\ (H_a, H_b, \ldots, H_h)\}$. If $\hat{G}$ is unbiased then we have

$$\begin{aligned} G = E(\hat{G}) &= \sum_{j=a}^{h} p_j G_j \\ &= (G_h - G_g)H^3 + (G_c + G_e - G_d - G_f)GH^2 + (G_g - G_e)H^2 \\ &\quad + (G_b - G_a)G^2H + (2G_a + G_d + G_f - G_b - 2G_c - G_e)GH + (G_e - G_a)H \\ &\quad + (G_a - G_b)G^2 + (G_b + G_c - 2G_a)G + G_a. \end{aligned} \tag{A.1}$$

Equating the last four terms on the right-hand side with the left-hand side in order to obtain the values of the coefficients that make the right- and left-hand-side polynomials of G and H equal, that is,

$$\begin{aligned} G_a &= 0, \\ G_b + G_c - 2G_a &= 1, \\ G_a - G_b &= 0, \\ G_e - G_a &= 0, \end{aligned}$$

gives

$$G_a = G_b = G_e = 0, \qquad G_c = 1. \tag{A.2}$$

Substituting into the first three terms and equating gives

$$G_e = G_g = G_h, \qquad 1 + G_e = G_d + G_f. \tag{A.3}$$

Substituting this into the term for GH gives the contradiction

$$\begin{aligned} 0 &= 2G_a + G_d + G_f - G_b - 2G_c - G_e \\ &= 0 + 1 + G_e - 0 - 2 - G_e = -1. \end{aligned} \tag{A.4}$$

If $\hat{H}$ is unbiased, then

$$\begin{aligned} H &= E(\hat{H}) = \sum_{j=a}^{h} p_j H_j \\ &= (H_h - H_g)H^3 + (H_c + H_e - H_d - H_f)GH^2 + (H_g - H_e)H^2 \\ &\quad + (H_b - H_a)G^2H + (2H_a + H_d + H_f - H_b - 2H_c - H_e)GH + (H_e - H_a)H \\ &\quad + (H_a - H_b)G^2 + (H_b + H_c - 2H_a)G + H_a \end{aligned} \tag{A.5}$$

and calculations similar to those for $\hat{G}$ also lead to a contradiction. □

A.2. The recursive biased corrected estimator

If we iterate on (3.17) and (3.18) and the process converges (so that $|G_j^{(k+1)} - G_j^{(k)}| \to 0$ as $k \to \infty$ and similarly for H) then we have the limit estimators when $G_j^{(k+1)} = G_j^{(k)}$:

$$\begin{aligned} \hat{G}_j &= E^{(\infty)}(\hat{G}) \\ &= p_a\left(G_j^{(\infty)}, H_j^{(\infty)}\right)\hat{G}_a + \cdots + p_f\left(G_j^{(\infty)}, H_j^{(\infty)}\right)\hat{G}_f, \end{aligned} \tag{A.6}$$

$$\begin{aligned} \hat{H}_j &= E^{(\infty)}(\hat{H}) \\ &= p_a\left(G_j^{(\infty)}, H_j^{(\infty)}\right)\hat{H}_a + \cdots + p_f\left(G_j^{(\infty)}, H_j^{(\infty)}\right)\hat{H}_f. \end{aligned} \tag{A.7}$$

This gives two equations in two unknowns for each case $a, \ldots, f$. The first issue in this iteration is whether there is a solution to these two equations, i.e. whether there is a fixed point on the iterative process. Ideally we would like to have a unique solution for each case that satisfies $G_j^{(\infty)} \in [0, 1]$ and $H_j^{(\infty)} \in [0, 1]$. We can do this for cases a, b, c, d, f but not for case e. To see this, note that the equations for case e are:

$$0 = 0.5p_b\left(G_e^{(\infty)}, H_e^{(\infty)}\right) + \left(p_c\left(G_e^{(\infty)}, H_e^{(\infty)}\right) + p_d\left(G_e^{(\infty)}, H_e^{(\infty)}\right) + p_f\left(G_e^{(\infty)}, H_e^{(\infty)}\right)\right) \tag{A.8}$$

$$= \frac{1}{2}\frac{\left(3 + 2H_e^{(\infty)} - G_e^{(\infty)}\right)G_e^{(\infty)}}{\left(1 + H_e^{(\infty)}\right)}, \tag{A.9}$$

$$\begin{aligned} 0.5 &= 0.5\left(p_d\left(G_e^{(\infty)}, H_e^{(\infty)}\right) + p_e\left(G_e^{(\infty)}, H_e^{(\infty)}\right) + p_f\left(G_e^{(\infty)}, H_e^{(\infty)}\right)\right) \\ &= \frac{1}{2}\frac{\left(1 + G_e^{(\infty)}\right)H_e^{(\infty)}}{\left(1 + H_e^{(\infty)}\right)}. \end{aligned} \tag{A.10}$$

This set of equations has no solution that satisfies the constraints. To see this, if $G_e^{(\infty)} = 0$ then the second equation implies a contradiction. Thus we must have $3 + 2H_e^{(\infty)} - G_e^{(\infty)} = 0$. Substituting this into the second equation gives

$$2\left(H_e^{(\infty)}\right)^2 + 3H_e^{(\infty)} - 1 = 0, \tag{A.11}$$

which does not have any roots between zero and unity.

Table A.1. Outcomes conditioning on point identification.

		MLE		NBC		Limit estimator	
Case	Prob	$\hat{G}$	$\hat{H}$	$G^{(1)}$	$H^{(1)}$	$G^{(\infty)}$	$H^{(\infty)}$
a	$\frac{(1-H)(1-G)(1-G)}{(1-H^2)}$	0	0	0	0	0	0
b	$\frac{(1-H)(1-G)G}{(1-H^2)}$	1/2	0	3/8	0	0.382	0
c	$\frac{(1-H)G(1-H)}{(1-H^2)}$	1	0	1	0	1	0
d	$\frac{(1-H)GH}{(1-H^2)}$	1	1/2	1	2/3	1	1
e	$\frac{H(1-H)(1-G)}{(1-H^2)}$	0	1/2	0	5/6	(0)	(1)
f	$\frac{H(1-H)G}{(1-H^2)}$	1	1/2	1	2/3	1	1

A second issue is where the iterated estimators converge to. For case *e* the recursion goes outside the interval [0, 1] and never reach a fixed point for H. Nonetheless, the other five cases converge to their fixed points; these are given in Table 4. We can also take the estimates for case *e* that minimize the sum of the differences between the expected values of the ML estimators and the values of the latter:

$$\left\{\frac{1}{2}\frac{\left(3+2H_e^{(\infty)}-G_e^{(\infty)}\right)G_e^{(\infty)}}{\left(1+H_e^{(\infty)}\right)}\right\}^2+\left\{\frac{1}{2}\frac{\left(1+G_e^{(\infty)}\right)H_e^{(\infty)}}{\left(1+H_e^{(\infty)}\right)}-0.5\right\}^2. \tag{A.12}$$

The minimizing values are $G_e^{(\infty)}=0$ and $H_e^{(\infty)}=1$ (shown in parentheses in Table A.1 to indicate that they are biased). In fact, these values are a solution of the equation for G but not for H.

The biases for the limit estimator are given by:

$$\phi\left(G^{(\infty)}\right)=E\left(G^{(\infty)}\right)-G=0.382\frac{(1-G)G}{(1+H)}\geq 0, \tag{A.13}$$

$$\varphi\left(H^{(\infty)}\right)=E\left(H^{(\infty)}\right)-H=\frac{(G-H)H}{(1+H)}\gtrless 0. \tag{A.14}$$

For H we can now have a positive bias if $M=H-G<0$ and it is unbiased if $H=G$, i.e. if $M=0$. It could be seen that the bias for G is smaller than for ML but larger than for the one-step estimator. The bias for H is smaller than for the MLE or the NBC for some values of (G,H), but not for all.