Econometrica, Vol. 62, No. 1 (January, 1994), 95–116

A COMPUTATIONALLY PRACTICAL SIMULATION ESTIMATOR FOR PANEL DATA

BY MICHAEL P. KEANE[1]

In this paper I develop a practical extension of McFadden's method of simulated moments estimator for limited dependent variable models to the panel data case. The method is based on a factorization of the MSM first order condition into transition probabilities, along with the development of a new highly accurate method for similating these transition probabilities. A series of Monte-Carlo tests show that this MSM estimator performs quite well relative to quadrature-based ML estimators, even when large numbers of quadrature points are employed. The estimator also performs well relative to simulated ML, even when a highly accurate method is used to simulate the choice probabilities. In terms of computational speed, complex panel data models involving random effects and ARMA errors may be estimated via MSM in times similar to those necessary for estimation of simple random effects models via ML-quadrature.

KEYWORDS: Method of simulated moments, panel data, limited dependent variables, equicorrelation, importance sampling, multinomial probit.

1. INTRODUCTION

AN IMPORTANT PROBLEM in the panel data literature is the estimation of limited dependent variable (LDV) models in the presence of serially correlated errors. Maximum likelihood estimation (ML) of these models generally requires the evaluation of choice probabilities which are multivariate integrals—with the order of integration proportional to the number of time periods in the panel. In order to make estimation practical, it is typically assumed that the covariance matrix of the errors has some simple form which allows the order of integration to be reduced. In particular, it is generally assumed that the errors are i.i.d. or that they are equicorrelated. The latter assumption leads to the popular random effects model (see Heckman (1981)) which is practical to estimate using the Gaussian quadrature procedure described by Butler and Moffitt (1982).

The assumption that errors are either equicorrelated or i.i.d. can be undesirable in many situations. Economic theory often suggests of serially correlated error components. Furthermore, if we wish to use an LDV model for prediction purposes, it is important to determine the error structure which gives the best fit to the data, rather than imposing a particular structure a priori.

[1] I would like to thank the Alfred P. Sloan Foundation, the Federal Reserve Bank of Minneapolis, and the Institute for Empirical Macroeconomics for their support of this research. This paper is taken from the first chapter of my Ph.D. dissertation at Brown University. I thank my dissertation committee of Robert Moffitt, Tony Lancaster, and David Runkle for their advice and comments. The comments of Daniel McFadden, Ariel Pakes, John Geweke, and seminar particpants at the 1988 Winter meetings of the Econometric Society, the University of Chicago, the University of Minnesota, the University of Rochester, and the 1990 Banff Invitational Symposium on Consumer Decision Making and Choice Behavior are also greatly appreciated. All errors are of course the responsibility of the author.

The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

An additional problem with the a priori assumption of a particular error structure in LDV models is that, due to the nonlinearity of these models, the ML estimator is not in general consistent when the error structure is misspecified. For example, if one falsely assumes equicorrelated errors when there is an AR(1) error component, the ML estimator of the regressor coefficients may be inconsistent. If one assumes i.i.d. errors, the ML estimator will be consistent in the presence of any type of serial correlation so long as the errors are normally distributed (see Robinson (1982)). However the resultant estimator is inefficient if serial correlation is present, and the resultant model will not predict as well as one which accounts for serial correlation.

The method of simulated moments (MSM), a new estimation technique developed by McFadden (1989) and Pakes and Pollard (1989), provides a practical method for estimating LDV models in which the choice probabilities are cumbersome multivariate integrals. The idea is to perform Monte-Carlo simulations of these integrals rather than evaluating them numerically. If the simulations are unbiased, the simulation errors tend to cancel out of the first order conditions of a method of moments estimator (MOM) when one sums over observations. Both McFadden and Pakes and Pollard (1989) prove that an MSM estimator based on unbiased simulators is consistent and asymptotically normal, as well as being asymptotically (in simulation size) as efficient as ML, given the optimal choice of the MOM weighting matrix. If an MSM estimator is constructed using simulators that are only asymptotically (in simulation size) unbiased, these consistency and asymptotic normality results continue to hold if simulation size is increased with sample size at a sufficient rate (see also McFadden and Ruud (1991)). When importance sampling or other similar techniques are used to construct simulators which are smooth functions of the model parameters, standard gradient methods can be used to obtain MSM estimates.

Thus, the MSM estimator provides a potential means for estimating LDV models with complex patterns of serial correlation. However, the discussion of LDV models in McFadden is limited largely to the estimation of multinomial probits involving many choices on cross-sectional data. Here the order of integration is equal to $M-1$, where M is the number of possible choices. Application of the MSM to panel data is a difficult problem for the following reason: with T time periods and M possible choices, there are M^T possible choice sequences. The probability of each sequence must be simulated in order to apply MSM directly, and this quickly becomes impractical as T increases.

In this paper, I develop a practical extension of the MSM estimator to the panel data case. This estimator avoids the problem of having to simulate the probabilities of M^T possible choice sequences by expressing the objective function in terms of transition probabilities. A practical method for simulating transition probabilities is developed by using importance sampling techniques. This probability simulator is smooth in the model parameters, so that standard gradient methods of optimization work well.

Unfortunately, for large T, practical simulators of transition probabilities that are smooth functions of the model parameters can be only asymptotically unbiased in the number of random draws used in simulation. As a result, the simulation estimator proposed here can only be made asymptotically unbiased in simulation size when $T \geqslant 4$. To assess the importance of the bias that exists for finite simulation size, I perform a series of Monte-Carlo repeated sampling experiments. These Monte-Carlo experiments are of additional interest, because they are the first repeated sampling experiments for any simulation estimator to be reported in the econometrics literature. Even for large T and small sized simulations, the Monte-Carlo tests indicate that the bias in the MSM estimator is negligible.

The method for simulating probabilities developed in this paper is also of interest in its own right. The method is often referred to as the Geweke-Hajivassiliou-Keane, or GHK, simulator, because simultaneous and independent work by Geweke (1991) and Hajivassiliou led to the development of the same method. Hajivassiliou, McFadden, and Ruud (1991), in an extensive study of alternative methods for the simulation of probabilities, find that the GHK simulator is extremely accurate even for small numbers of Monte-Carlo draws, and that it outperforms all the other methods considered.

In addition to the present paper, other attempts to develop simulation estimators for panel data LDV models have been made by Hajivassiliou and McFadden (1990) and by McFadden and Ruud (1987). These approaches are described in detail in Section 3.2. But basically, the preferred approach of these authors is to use the highly accurate GHK simulator to implement simulated maximum likelihood (SML) or to find the roots of the simulated score (also using GHK). In this paper, I also present the first repeated sampling experiments for the SML estimator based on GHK. The results indicate that SML based on a small number of draws works reasonably well, but that its estimates of serial correlation parameters suffer from substantial biases when serial correlation is strong. These biases are not present with the MSM estimator, and hence SML is clearly dominated by MSM in my experiments.

Besides these alternative simulation estimators, several authors have devised practical non-ML estimators for LDV models with dependent observations which are more efficient than ML estimators that assume i.i.d. errors, but which remain consistent in the presence of serially correlated errors. These include the "orthogonality condition" estimators of Avery, Hansen, and Hotz (1983), the minimum distance method of Chamberlain (1985), the "nonlinear IV" estimators of Bates and White (1987), and the quasi-ML estimator of Poirier and Ruud (1988). However, none of these estimators is as efficient as ML, and their extension to systems of equations or multinomial probit models is computationally burdensome. The simulation estimator for panel data proposed here thus has three important advantages over these procedures. First, it is asymptotically (in simulation size) as efficient as ML; second, being a simulation estimator, it can readily handle multinomial probit situations; and third, it can be

readily extended to estimate nonlinear systems of equations which would be impractical to estimate by ML.

The outline of the paper is as follows. Section 2 describes the probit model. Section 3.1 describes the MSM estimator and Section 3.2 explains the difficulties involved in applying it to panel data. The key section of the paper is 3.3 which presents the construction of the simulation estimator for panel data. Section 3.4 contains the details of the construction of smooth simulators for transition probabilities, and contains a discussion of the asymptotic properties of the MSM estimator based on this simulation procedure. Section 4 presents the Monte-Carlo test results. Section 5 concludes the paper.

2. THE PROBIT MODEL

Although the simulation estimator which I will describe has more general application, I will follow McFadden (1989) and restrict my exposition to the multinomial probit model. The probit model has the following form. An individual chooses from the set S of M mutually exclusive choices that choice which gives greatest utility. The utility of choice j to person i at time t is given by:

$$(1) \qquad \mu_{ijt} = X_{it}\beta_j + \varepsilon_{ijt} \qquad (j = 1, \ldots, M-1; t = 1, \ldots, T),$$

where X_{it} is a row vector of exogenous regressors and β_j is a vector of coefficients possibly specific to choice j. μ_{iMt} is normalized to 0 for identification. ε_{ijt} is a disturbance which is independent across individuals but has a multivariate normal distribution over choices and time, $\varepsilon_i \sim N(0, \Sigma)$. Here, and below, the suppression of subscripts indicates that a variable becomes a vector including elements for all values of the subscripts. The covariance matrix Σ has the Cholesky decomposition $A'A$, so that $\varepsilon_i = A'\eta_i$ where η_i is i.i.d. We do not observe μ_{ijt} directly, but instead observe only the indicator d_{ijt}, where

$$(2) \qquad d_{ijt} = \begin{cases} 1 & \text{if } \mu_{ijt} \geqslant \mu_{ij't} \ \forall j' \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Let the $K \times 1$ parameter vector θ contain the elements of β and the parameters of the error processes that generate Σ. Thus θ denotes the true parameter vector. Letting j_t be the index of the choice actually made at time t, define $J_{it} = \{j_1, \ldots, j_t\}_i$ as the sequence of choices actually made by person i in periods 1 through t. Then, maximum likelihood estimation of the probit model requires that one calculate the choice probabilities $E(d_{ijt}|J_{i,t-1}, \hat{\theta}, X_i)$ for a large number of trial parameter values $\hat{\theta}$. The problem with the probit model in the panel data case is that, for unrestricted Σ, the time T choice probabilities are $T \cdot (M-1)$ variate integrals. Even in the single period case, estimation is prohibitively expensive for $M \geqslant 4$ (see Kahaner (1991)).

The required order of integration may exceed $T \cdot (M-1)$ if the history of the choice process begins prior to $t = 1$, the first period of observed data. Then, consistent estimation requires that the past history of the (X, ε) process be

"integrated out" of the likelihood function (see Heckman (1981)). This problem is simplified if X is exogenous. Then it is only necessary to integrate over the history of the ε process prior to $t = 1$ to form the appropriate marginal density $f(\varepsilon_{i1}, \ldots, \varepsilon_{iT}|\hat{\theta})$. For many stochastic processes a closed form exists for this marginal density. Examples are when ε_{it} is a sum of an individual random effect plus a stationary invertible ARMA process, as in the probit models I consider in Section 4. In such cases, where a closed form for the marginal density exists, the order of integration remains $T \cdot (M - 1)$ despite the existence of an unobserved past history of the ε_{it} process.

3. THE METHOD OF SIMULATED MOMENTS AND PANEL DATA

3.1. *Background on the Method of Simulated Moments*

The idea of simulation estimation is to simulate integrals like $E(d_{ijt}|J_{i,t-1}, \hat{\theta}, X_i)$ via inexpensive Monte-Carlo methods, rather than evaluate them accurately via expensive numerical methods. For example, consider the integral $P = \int_R f(\mu|\hat{\theta})d\eta$ where $f(\mu|\hat{\theta})$ is the density of the random variable μ given a trial parameter value $\hat{\theta}$ and R is a region in the support of μ. The value of P can be simulated by taking one or more draws from the density $f(\mu|\hat{\theta})$ and calculating the fraction of these draws which fall in the region R. For example, referring to the probit model in equation (1), one could draw values for μ_{ijt}, $j = 1, \ldots, M$, $t = 1, \ldots, T$ from the multivariate normal distribution determined by $X_i\beta$ and Σ and count the percentage that generate a particular choice sequence. This is called a "frequency" simulator. The resulting estimate of P is unbiased even if only one draw from $f(\mu|\hat{\theta})$ is used.[2]

McFadden proposes the use of simulation in the context of method of moments (MOM) estimation to solve the problem of high order multivariate integration in the probit model. In the single period case on which McFadden concentrates, the vector of first order conditions for an MOM estimator is given by:

$$(3) \qquad \sum_{i=1}^{N} W_i\left(d_i - E\left(d_i|\hat{\theta}_{\text{MOM}}, X_i\right)\right) = 0$$

where $\hat{\theta}_{\text{MOM}}$ is the MOM estimator of θ, d_i is an $M \times 1$ column vector of indicator variables for person i, $E(d_i|\hat{\theta}, X_i)$ is an $M \times 1$ column vector of expected values of the elements of d_i conditional on $\hat{\theta}$ and X_i, and W_i is a $K \times M$ weighting matrix. The optimal MOM weights are obtained if the (k, j)th element of W_i is chosen to be $E(d_{ij}|\hat{\theta}_c, X_i)^{-1} \, \partial E(d_{ij}|\hat{\theta}_c, X_i)/\partial\hat{\theta}_{ck}$ for each j and k where $\hat{\theta}_c$ is a consistent (but inefficient) initial estimate of θ. Then the

[2] Frequency simulation has been used to evaluate choice probabilities in the context of ML estimation by Albright, Lerman, and Manski (1977), Lerman and Manski (1981), and Pakes (1986). Note, however, that if one simulates a probability P in a log likelihood function, the resulting estimate of $\ln P$ will be biased because of the nonlinearity of the logarithmic transformation (i.e., $E \ln(P + s) \neq \ln E(P + s) = \ln P$, where s is a mean zero random simulation error). Thus, simulations of P must be very precise in order to keep the bias in simulating $\ln P$ small.

FOC for the MOM estimator are (asymptotically) equivalent to the derivatives of the log-likelihood function, and the estimator is asymptotically efficient.

The important aspect of the MOM estimator is that the $E(d_{ij}|\hat{\theta}, X_i)$, which are the difficult quantities to evaluate, enter *linearly* into the FOC of the estimator. The MSM estimator replaces the difficult to evaluate $E(d_{ij}|\hat{\theta}, X_i)$ by simulators $\hat{E}(d_{ij}|\hat{\theta}, X_i)$. Because these simulators enter the FOC linearly, the simulation errors enter the FOC linearly as well, and simulation error tends to cancel across observations i. Both McFadden and Pakes-Pollard have independently proved that the MSM estimator based on unbiased simulators is consistent and asymptotically normal and that, if frequency simulation is used, it has a covariance matrix which is $(1 + 1/r)$ times that of the MOM estimator, where r is the number of draws used in the simulation. If the optimal MOM weights are consistently simulated, the estimator is asymptotically as efficient as ML. It generally will not be possible to simulate the weights consistently (for fixed r) because of denominator bias, so the MSM estimator will approach the efficiency of ML only as the number of random draws used to simulate the denominators of the weights grows large. McFadden and Ruud (1991) show that these asymptotic results continue to hold for MSM estimators based on biased simulators, so long as the simulation bias converges to zero in probability as sample size increases at a sufficient rate (see Section 3.4).

As a practical matter it may not be desirable to use frequency simulators to evaluate choice probabilities. If $E(d_{ij}|\hat{\theta}, X_i)$ is evaluated using a frequency simulator, $\partial\hat{E}(d_{ij}|\hat{\theta}, X_i)/\partial\hat{\theta}$ is not continuous. In fact, $\hat{E}(d_{ij}|\hat{\theta}, X_i)$ is piecewise constant in $\hat{\theta}$ with discrete jumps at values of $\hat{\theta}$ which produce ties among alternatives. Hence, conventional gradient methods cannot be used to minimize the objective, and random search or pseudo-gradient methods must be used instead.

As discussed in McFadden, it is often more practical to use a techniques such as "importance sampling" to construct "smooth" simulators. Returning to the equation for P we see it can be rewritten as $P = \int_R [f(\mu|\hat{\theta})/\gamma(\mu)]\gamma(\mu)\,d\mu$ where $\gamma(\mu)$ is any density function such that $f(\mu|\hat{\theta})/\gamma(\mu)$ is finite almost everywhere in region R. By definition, this integral is the expectation of the quantity $f(\mu|\hat{\theta})/\gamma(\mu)$. Thus, P can be simulated by evaluating $f(\mu|\hat{\theta})/\gamma(\mu)$ at one or more draws from the "importance sampling" density $\gamma(\mu)$ that fall in the region R. This method is inexpensive because $\gamma(\mu)$ can be chosen in such a way that most or all the μ drawn will fall in the region R. Such a simulator is called "smooth" because, for a given $\mu = \mu^*$, the derivative of $f(\mu^*|\hat{\theta})/\gamma(\mu^*)$ with respect to $\hat{\theta}$ is continuous.[3] Hence standard gradient methods can be used to minimize objectives with respect to $\hat{\theta}$. For stability of the optimization process it is essential that the draws from $\gamma(\mu)$ be held fixed as we iterate on $\hat{\theta}$.

[3] As discussed by Geweke (1987) and McFadden, smooth simulators are often more efficient than frequency simulators. Thus, an MSM estimator based on smooth simulators may have an asymptotic covariance matrix less than $(1 + 1/r)$ times that of the MOM estimator. The smooth simulators described here are a form of importance sampling because the simulator is a weighted sum of the $f(\mu|\hat{\theta})$ with the weights being $\gamma(\mu)^{-1}$

3.2. *The Problem of Applying the Method of Simulated Moments to Panel Data*

This section describes the difficulties inherent in applying MSM to panel data. In McFadden's original formulation, the FOC for the MSM estimator are constructed in terms of unconditional choice probabilities. In the panel data case, this is only possible if we treat *sequences* of states as the objects of choice. For this purpose, let $j \in \{1, \ldots, M^T\}$ index alternative sequences. Then the d_{ij} become indicators for certain sequences of events, and the $E(d_{ij}|\hat{\theta}, X_i)$ become the probabilities of observing these sequences. Since there are M^T possible sequences of choices, the residual vector $[d_i - E(d_i|\hat{\theta}, X_i)]$ in equation (3) is M^T elements long. To utilize a smooth simulator, it is necessary to calculate the probability of observing each possible sequence for each person. As M and T increase, this will quickly require an impractical number of calculations.

McFadden discusses two ways around this problem, neither of which is adopted here. The first, investigated by McFadden and Ruud, is to construct a frequency simulator based on only r draws from the density $f(\mu|\hat{\theta})$ of random variables μ that determine the sequence of states chosen. This frequency simulator will give nonzero values for the probabilities of at most r of the possible sequences. McFadden shows that an MSM estimator utilizing residuals $[d_{ij} - \hat{E}(d_{ij}|\hat{\theta}, X_i)]$ only for these few sequences, plus that for the observed sequence if it is not among them, has a covariance matrix which is still only $(1 + 1/r)$ times as large as that for the MOM estimator. The drawback of this approach is that the simulator is not smooth, so that costly random search or pseudo-gradient procedures must be employed in estimation. To my knowledge this method has not yet been successfully applied, although advances in non-smooth hill climbing algorithms may make it more appealing.

The second approach, investigated by Hajivassiliou and McFadden, is to simulate directly, via a smooth simulator, the score of the log-likelihood function, which is

$$\sum_{j=1}^{MT} d_{ij}\, \partial \ln E\left(d_{ij}|\hat{\theta}, X_i\right)/\partial\hat{\theta} = \sum_{j=1}^{MT} d_{ij} E\left(d_{ij}|\hat{\theta}, X_i\right)^{-1} \partial E\left(d_{ij}|\hat{\theta}, X_i\right)/\partial\hat{\theta}.$$

Since $d_{ij} = 0$ for all sequences except that which is observed, the score need only be simulated for the particular observed sequence and the problem of having an enormous number of possible sequences is avoided. However, such a simulator is biased because the term $E(d_{ij}|\hat{\theta}, X_i)$ is simulated and the nonlinear transformation $1/E$ taken. Note that this method of simulated scores, or MSS, procedure is essentially identical to SML, so long as the same smooth probability simulator is used in each case, because any trial value $\hat{\theta}$ that gives a local maximum of the simulated log likelihood function is also a root of the simulated score.

Despite the biases involved in these procedures, Hajivassiliou and McFadden and Borsch-Supan and Hajivassiliou (1990) argue that SML or MSS based on the highly accurate GHK simulator performs well using very small numbers of draws. However, extensive repeated sampling studies of such estimators have

not yet been performed.[4] In this paper I present the first such study. I find that SML based on the GHK simulator does perform reasonably well using only small numbers of draws. However, in my Monte-Carlo experiments this SML procedure is dominated by the MSM estimator I describe in the next section.

In addition to these approaches, van Praag and Hop (1987) and Ruud (1992) point out that the score also can be written in terms of the latent vector μ rather than in terms of choice probabilities. Then, it is possible to form unbiased simulators of the score if it is feasible to draw the μ from their conditional distribution determined by agents' observed choices d. This idea forms the basis of "simulated EM algorithm" and "Gibbs sampling" approaches to estimation of LDV models (see Albert and Chib (1993) and McCulloch and Rossi (1992)). Monte-Carlo analysis of the performance of these procedures relative to MSM or SML based on GHK do not yet exist for the panel data case. Geweke, Keane, and Runkle (1992) compare the performance of all these approaches in a cross-section case. An important avenue for future research is to perform Monte-Carlo studies of the simulated EM and Gibbs sampling approaches in the panel data case.

3.3. *Construction of the Simulation Estimator for Panel Data*

The solution to the problem of applying MSM estimation to panel data that is proposed in this section involves two parts. First, to circumvent the problem of having M^T possible sequences of choices, the MOM first order condition is rewritten in terms of transition probabilities. Second, a method for obtaining computationally efficient smooth simulators of these transition probabilities is devised.

Recalling that J_{it} is the sequence of observed choices for person i in periods 1 through t, the MOM first order condition written in terms of transition probabilities has the form:

$$(4) \qquad \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M} W_{ijt}\left[d_{ijt} - E\left(d_{ijt}|J_{i,t-1}, \hat{\theta}_{\mathrm{MOM}}, X_i\right)\right] = 0.$$

Now, rather than M^T, there are only $M \cdot T$ probabilities to be simulated. However, expressing the FOC in this way creates an important new problem—from the second period onward, the transition probabilities $E(d_{ijt}|J_{i,t-1}, \hat{\theta}, X_i)$ must be simulated. In most circumstances, it is either impossible or extremely difficult to construct unbiased smooth simulators of transition probabilities. This is because transition probabilities are generally ratios of two integrals, and simulating a denominator produces bias.

Unbiased frequency simulators of transition probabilities can be constructed using the acceptance-rejection method (henceforth A/R). In the probit model

[4] Borsch-Supan and Hajivassiliou study the performance of SML based on the GHK simulator in a cross-section trinomial probit model using a single Monte-Carlo data set. They do not do a repeated sampling experiment or consider the panel data case.

of Section 2 we see that, for given values of θ and X_i, a draw for the vector of i.i.d. errors η will generate a particular sequence of states. In the simplest A/R method, random vectors η are drawn from a standard normal distribution with density $f(\eta_{i1},\ldots,\eta_{1,t-1})$ until one or more which generate the sequence of states $J_{i,t-1}$ actually chosen by individual i up through period $t-1$ are obtained. I will refer to such sequences of i.i.d. errors η which generate the observed sequence of choices $J_{i,t-1}$ as "accepted conditioning sequences." When such a sequence is obtained by the A/R method, it will have the correct density $f(\eta_{i1},\ldots,\eta_{i,t-1}|J_{i,t-1},\hat{\theta},X_i)$.[5] If we randomly draw from $f(\eta_{it})$ values for the time t elements of η (i.e., the η_{ijt} for $j=1,\ldots,M-1$), then the fraction of these draws which, conditional on a particular accepted conditioning sequence, imply the choice of state j at time t, is an unbiased frequency simulator of $E(d_{ijt}|J_{i,t-1},\hat{\theta},X_i)$.

Unfortunately, the simple A/R method is not practical. An enormous number of draws may be necessary to obtain just one accepted sequence if M^{t-1} is large, making the A/R method extremely expensive.[6] An important innovation in my approach is the use of an alternative method that generates accepted conditioning sequences at low cost. This method involves drawing sequences very inexpensively from an importance sampling distribution, say with density $\phi(\eta_{i1},\ldots,\eta^*_{i,t-1}|J_{i,t-1},\hat{\theta},X_i)$, chosen so that sequences drawn from it are likely to be accepted. Denote by $(\eta^*_{i,1},\ldots,\eta^*_{i,t-1})$ a *particular* sequence drawn from $\phi(\cdot\,|\,\cdot)$ that generates the observed set of choices $J_{i,t-1}$. Of course, since $(\eta^*_{it},\ldots,\eta_{i,t-1})$ is not drawn from $f(\eta_{i,1},\ldots,\eta_{i,t-1}|J_{i,t-1},\hat{\theta},X_i)$ we have, letting E_ϕ denote the expectation over all $(\eta^*_{i,t},\ldots,\eta^*_{i,t-1})$ sequences obtained from ϕ, that $E_\phi\{\hat{E}(d_{ijt}|(\eta^*_{i,1},\ldots,\eta^*_{i,t-1}),\hat{\theta},X_i)\}\neq E(d_{ijt}|J_{i,t-1},\hat{\theta},X_i)$. However, an *unbiased* simulator of the transition probability $E(d_{ijt}|J_{i,t-1},\hat{\theta},X_i)$ may be formed as a weighted sum of transition probabilities conditional on each of several such sequences $(\eta^*_{i,1},\ldots,\eta^*_{i,t-1})_s$ for $s=1,\ldots,S$ as follows:

$$(5)\qquad \hat{E}\left(d_{ijt}|J_{i,t-1},\hat{\theta},X_i\right)=\frac{1}{S}\sum_{s=1}^{S}\omega_s\hat{E}\left(d_{ijt}|(\eta^*_{i,1},\ldots,\eta^*_{i,t-1})_s,\hat{\theta},X_i\right).$$

Here the "sequence weights" ω_s are ratios of the correct joint density

[5] Note, as was mentioned in Section 2, that since the history of ε prior to $t=1$ is assumed unobserved, $f(\cdot\,|\,\cdot)$ must be an appropriate marginal density with respect to ε dated prior to $t=1$. If the ε_{it} process is ergodic Markov and has been operating since the indefinite past, then it is stationary and it is appropriate to draw from the ergodic distribution. For example, if $\varepsilon_{it}=\rho\varepsilon_{i,t-1}+\upsilon_{it}$ where $|\rho|<1$ and $\upsilon_{it}\sim N(0,1-\rho^2)$ then draw ε_{i1} from the $N(0,1)$ distribution. If the ε_{it} process started in the recent past one cannot draw from the ergodic distribution. For example, if ε_{it} is as above, but the process started K periods prior to $t=1$, then draw ε_{i1} from the $N(0,(1-\rho^2)\sum_{q=0}^{K}\rho^{2q})$ distribution. Since $\varepsilon_i=A'\eta_i$, simple mappings exist from these ε distributions to the approximate η distributions. Note that there may be processes for which integration over the past history does not lead to a closed form for the marginal distribution of ε_{it} for $t=1,T$. In such cases, one must simulate the whole history of the ε_{it} process in order to obtain draws from the appropriate marginal distribution.

[6] As discussed in Hajivassiliou and McFadden, more elaborate A/R methods have been devised which reduce the number of draws needed to obtain accepted sequences. However, they find that even the more sophisticated A/R methods are prohibitively expensive in practice.

$f(\eta^*_{i,1}, \ldots, \eta^*_{i,t-1}|J_{i,t-1}, \hat{\theta}, X_i)$ to the importance sampling joint density $\phi(\eta^*_{i,1}, \ldots, \eta^*_{i,t-1}|J_{i,t-1}, \hat{\theta}, X_i)$ of sequence s. Thus the ω_s are importance sampling weights.

Different choices of the importance sampling density ϕ and the method for forming the $\hat{E}$ lead to different estimators. Section 3.4 describes the details of the particular procedure used here to form $\hat{E}$ and to construct accepted conditioning sequences (i.e., the particular importance sampling distribution from which the sequences are drawn), and derives the correct form of the sequence weights corresponding to this procedure. In this procedure, ϕ is specified as the density of sequences constructed by drawing recursively (i.e., one component of η at a time) until a complete η vector implying a person's choice history is obtained, and $\hat{E}$ is also formed using such recursive draws. This amounts to using the GHK method to simulate the transition probabilities. It is important to note, however, that other choices of ϕ and $\hat{E}$ could potentially dominate this procedure. This is an important avenue for future research.

3.4. *Asymptotically Unbiased Simulation of Transition Probabilities via Importance Sampling*

To simulate a time t transition probability we must construct conditioning sequences that generate the observed set of choices up through period $t-1$, denoted $J_{i,t-1}$. Rather than drawing entire vectors $(\eta^*_{i,1}, \ldots, \eta^*_{i,t-1})$ from the unconditional density $f(\cdot)$ and accepting only those vectors for which *all* $\eta_{i\tau}$ for $\tau = 1, t-1$ are accepted (i.e., those that generate the entire set of observed choices $J_{i,t-1}$), the inexpensive method of sequence construction used in this paper is to build up sequences on an element-by-element basis. Specifically, using the $M = 2$ case (where there is only one error per period) to illustrate, the procedure works as follows. A value for η_{i1} is drawn from the truncated standard normal distribution from which *all* draws generate the state actually chosen in period one. This value, call it η^*_{i1}, is retained and a value of η_{i2} is drawn from the truncated normal distribution conditional on η^*_{i1}, all draws from which generate the state actually chosen in period two. The vector $(\eta^*_{i1}, \eta^*_{i2})$ is then an accepted conditioning sequence for periods 1 and 2. Repeating this for $t-1$ periods we obtain one conditioning sequence η for the first $t-1$ time periods.[7]

[7] To draw from the truncated univariate normal, let Y^* be a particular draw from a univariate standard uniform distribution, $Y \in [0,1]$. Let $F(\cdot)$ be the standard normal distribution function. Then, $F(\eta_{it}|a \leqslant \eta_{it} \leqslant b) = (F(\eta_{it}) - F(a))/(F(b) - F(a))$. Letting $F(\eta^*_{it}|a \leqslant \eta_{it} \leqslant b)$ equal Y^*, we have $\eta^*_{it} = F^{-1}\{(F(b) - F(a))Y^* + F(a)\}$, where $F^{-1}(\cdot)$ is the inverse distribution function. In the probit model, the values of a and b may be simply calculated given knowledge of X_i, $\hat{\theta}$ and the previous period draws $(\eta^*_{i,1}, \ldots, \eta^*_{i,t-1})$. As $\hat{\theta}$ is varied in optimization, the draws from the uniform distribution are held fixed. Then, the values of a and b, and hence the draws η^*_{it}, vary smoothly with $\hat{\theta}$. The weights ω_s also vary smoothly with $\hat{\theta}$. It will, therefore, be possible to construct smooth simulators of transition probabilities using formula (5).

Given this method of constructing conditioning sequences, the proper sequence weights are constructed as follows. Begin with the 3-period case. Recalling that j_t denotes the choice actually made at time t, while $J_{i,t} \equiv \{j_1, \ldots, j_t\}$, and suppressing all individual subscripts i, define the sets:

$$A = \{\eta_1 | d_{j1} = 1 \text{ for } j = j_1\},$$

$$B = \{\eta_1, \eta_2 | d_{j1} = 1 \text{ for } j = j_1, d_{j2} = 1 \text{ for } j = j_2\},$$

$$B(\eta_1^*) = \{\eta_2 | d_{j2} = 1 \text{ for } j = j_2 \text{ given } \eta_1 = \eta_1^* \in A\},$$

where η_1^* is the particular value of η_1 drawn from the truncated standard normal distribution. It is clear that $f(\eta_1^*, \eta_2^* | (\eta_1, \eta_2) \in B)$, the joint density of a particular vector (η_1^*, η_2^*) obtained by an unbiased A/R method, does not equal $f(\eta_2^* | \eta_2 \in B(\eta_1^*)) f(\eta_1^* | \eta_1 \in A)$, the importance sampling density obtained when the η^* vector is constructed on an element-by-element basis. Notice, however, that:

$$(6) \qquad f(\eta_1^*, \eta_2^* | (\eta_1, \eta_2) \in B) = \frac{f(\eta_1^*) f(\eta_2^*)}{\int_A \int_{B(\eta_1)} f(\eta_1) f(\eta_2)\, d\eta_2\, d\eta_1},$$

$$f(\eta_1^* | \eta_1 \in A) = \frac{f(\eta_1^*)}{\int_A f(\eta_1)\, d\eta_1},$$

$$f(\eta_2^* | \eta_2 \in B(\eta_1^*)) = \frac{f(\eta_2^*)}{\int_{B(\eta_1^*)} f(\eta_2)\, d\eta_2}.$$

Therefore, defining $\omega(\eta_1^*)$ as the ratio of the correct to the importance sampling density, we have:

$$(7) \qquad \omega(\eta_1^*) = \frac{f(\eta_1^*, \eta_2^* | (\eta_1, \eta_2) \in B)}{f(\eta_2^* | \eta_2 \in B(\eta_1^*)) f(\eta_1^* | \eta_1 \in A)} = \frac{\text{prob}(\eta_2 \in B(\eta_1) | \eta_1 = \eta_1^*)}{\text{prob}(\eta_2 \in B(\eta_1) | \eta_1 \in A)}$$

where $\text{prob}(\cdot | \cdot)$ denotes integration over a conditional normal density. Hence, $\omega(\eta_1^*)$ is the proper weight to apply to simulated probabilities conditional on sequences (η_1^*, η_2^*) in order for the simulations to be unbiased. To see this, consider simulating the probability of a certain event in period 3, $d_{ij3} = 1$ for some $j \in \{1, \ldots, M\}$, conditional on $(\eta_1^*, \eta_2^*) \in B$. Describe this event as $(\eta_1, \eta_2, \eta_3) \in C_j$ where C_j is a subset of B. Define

$$C_j(\eta_1^*, \eta_2^*) = \{\eta_3 | d_{ij3} = 1 \text{ for } j \in \{1, \ldots, M\} \text{ given } (\eta_1^*, \eta_2^*) \in B\},$$

$$\chi_{C_j(\eta_1^*, \eta_2^*)}(\eta_3) = \begin{cases} 1 & \text{if } \eta_3 \in C_j(\eta_1^*, \eta_2^*), \\ 0 & \text{otherwise;} \end{cases}$$

then:

$$
\begin{aligned}
(8)\qquad & \operatorname{prob}\left[(\eta_1,\eta_2,\eta_3)\in C_j \mid (\eta_1,\eta_2)\in B\right] \\
&= \int_A \int_{B(\eta_1)} \int_{-\infty}^{\infty} \chi_{C_j(\eta_1,\eta_2)}(\eta_3) \\
&\quad \times f(\eta_3) f(\eta_1,\eta_2 \mid (\eta_1,\eta_2)\in B)\, d\eta_3\, d\eta_2\, d\eta_1 \\
&= \int_A \int_{B(\eta_1)} \int_{-\infty}^{\infty} \chi_{C_j(\eta_1,\eta_2)}(\eta_3) f(\eta_3)\omega(\eta_1) \\
&\quad \times f(\eta_2 \mid \eta_2 \in B(\eta_1)) f(\eta_1 \mid \eta_1 \in A)\, d\eta_3\, d\eta_2\, d\eta_1.
\end{aligned}
$$

Thus, the quantity $\omega(\eta_1^*)\chi_{C_j(\eta_1^*,\eta_2^*)}(\eta_3^*)$ evaluated at a particular draw $(\eta_1^*, \eta_2^*, \eta_3^*)$ from the importance sampling density $f(\eta_3)f(\eta_2 \mid \eta_2 \in B(\eta_1))f(\eta_1 \mid \eta_2 \in A)$, which is what we have when we obtain draws for η_1 and η_2 on an element-by-element basis and draw η_3 randomly, is an unbiased simulator of the desired transition probability. Here $\omega(\eta_{1s}^*)$ is the importance sampling weight.

Since the sequence weights and the sequences themselves are smooth functions of the model parameters (see footnote 7), the construction of a smooth simulator by a further application of importance sampling to $f(\eta_3)$ is straightforward. Simply rewrite equation (8) as:

$$
\begin{aligned}
(9)\qquad & \int_A \int_{B(\eta_1)} \int_{C_j(\eta_1,\eta_2)} \left[\frac{f(\eta_3)}{\gamma(\eta_3)}\omega(\eta_1)\right]\gamma(\eta_3) \\
&\quad \times f(\eta_2 \mid \eta_2 \in B(\eta_1)) f(\eta_1 \mid \eta_1 \in A)\, d\eta_3\, d\eta_2\, d\eta_1.
\end{aligned}
$$

An unbiased smooth simulator is obtained by evaluating the quantity $f(\eta_3)\omega(\eta_1)/\gamma(\eta_3)$ at a draw from the importance sampling density $\gamma(\eta_3)f(\eta_2 \mid \eta_2 \in B(\eta_1))f(\eta_1 \mid \eta_1 \in A)$. $\gamma(\eta_3)$ is chosen so that most or all random draws for η_3 fall in the region $C_j(\eta_1, \eta_2)$.

To form a more accurate smooth simulator, we may take an average of simulated transition probabilities over several conditioning sequences (as in equation (5)). Additionally, we may also average over several different draws from the importance sampling density $\gamma(\eta_3)$. Letting S be the number of conditioning sequences and R be the number of importance sampling draws, we obtain:

$$
\begin{aligned}
(10)\qquad \hat{E}\left(d_{ij3} \mid J_{i2}, \hat{\theta}, X_i\right) &= \frac{1}{S}\sum_{s=1}^{S} \omega_s(\eta_{1s}^*) \\
&\quad \times \frac{1}{R}\sum_{r=1}^{R} \hat{E}\left(d_{ij3} \mid \eta_{i3sr}, (\eta_{i1s}^*, \eta_{i2s}^*), \hat{\theta}, X_i\right) \\
&= \frac{1}{S}\sum_{s=1}^{S} \omega_s(\eta_{1s}^*) \\
&\quad \times \frac{1}{R}\sum_{r=1}^{R} \left[f\left(\eta_{i3sr} \mid \eta_{i3sr} \in C_j(\eta_{i1s}^*, \eta_{i2s}^*)\right)/\gamma(\eta_{i3sr})\right]
\end{aligned}
$$

where η_{i3sr} is the value of η_{i3} generated by the sth conditioning sequence and the rth importance sampling draw.

Turning from the three-period example to the general case, the proper weights for t element conditioning sequences can be constructed recursively. Define $\eta(t) = (\eta_1, \ldots, \eta_t)$ and define $\eta(J_t)$ as the set of sequences that generate the observed set of choices through the first t periods. Then $\omega(\eta_1^*, \ldots, \eta_1^*)$, the weight that corresponds to any $t+1$ element accepted conditioning sequence $\eta^*(t+1) = (\eta_1^*, \ldots, \eta_t^*, \eta_{t+1}^*)$ that begins with the elements $(\eta_1^*, \ldots, \eta_t^*)$, is defined recursively as:

$$(11) \qquad \omega(\eta^*(t)) = \frac{\omega(\eta^*(t-1))\,\text{prob}\left(\eta_{t+1} \in \eta(J_{t+1}) | \eta(t) = \eta^*(t)\right)}{\text{prob}\left(\eta_{t+1} \in \eta(J_{t+1}) | \eta(t) \in \eta(J_t)\right)}$$

which is obtained by the same type of algebraic manipulations used to obtain equation (7).

Finally, it is important to note that Monte-Carlo simulation of choice probabilities conditional on accepted conditioning sequences, as in equation (10), may be unnecessary. Choice probabilities conditional on accepted conditioning sequences are integrals over only the time t values of the η_{ijt}. Thus, in multinomial probit with $M \leqslant 3$, direct numerical evaluation is feasible, since the order of integration involved is only $M-1$. In the case where $M \geqslant 3$, one possible choice for forming $\hat{E}$ is to use a recursive simulation procedure like that used to construct the conditioning sequences. Note that a time t choice probability may be written as:

$$\begin{aligned}
&\hat{E}\left(d_{ijt} | \eta^*(t-1), \hat{\theta}, X_i\right) \\
&\quad = \text{prob}\left(\mu_{ikt} < \mu_{ijt}\ \forall k \neq j | \eta^*(t-1), \hat{\theta}, X_i\right) \\
&\quad = \text{prob}\left(\{\eta_{ikt}\}_{k=1,M-1} \in \eta(J_t) | \eta^*(t-1), \hat{\theta}, X_i\right).
\end{aligned}$$

This probability may be simulated recursively using a procedure that has the same mathematical form as the procedure for generating accepted conditioning sequences. First, draw from a truncated normal an η_{i1t} that belongs to the set $\eta(J_t)$ conditional on $\eta^*(t-1)$. Then this value, call it η_{i1t}^*, is retained and a value of η_{i2t} is drawn from a truncated normal distribution such that $(\eta_{i1t}^*, \eta_{i2t}) \in \eta(J_t)$ conditional on $\eta^*(t-1)$. This is repeated until the whole η_{it} vector is constructed. Then, an unbiased simulator of the choice j probability is given by:

$$\begin{aligned}
(12) \quad &\hat{E}\left(d_{ijt} | \eta^*(t-1), \hat{\theta}, X_i\right) \\
&\quad = \text{prob}\left(\eta_{i1t} \in \eta(J_t) | \eta^*(t-1), \hat{\theta}, X_i\right) \\
&\qquad \cdot \text{prob}\left(\eta_{i2t} \in \eta(J_t) | \eta_{i1t}^*, \eta^*(t-1), \hat{\theta}, X_i\right) \\
&\qquad \cdot \cdots \cdot \text{prob}\left(\eta_{i,M-1,t} \in \eta(J_t) | \eta_{i1t}^*, \ldots, \eta_{i,M-2,t}^*, \eta^*(t-1), \hat{\theta}, X_t\right).
\end{aligned}$$

This procedure for simulating choice probabilities is referred to as the Geweke-Hajivassiliou-Keane (or GHK) simulator in the paper by Hajivassiliou, McFadden, and Ruud (1991), who compare its performance to many alternative probability simulators and find that it is the most accurate of those considered. It is important to note that the GHK simulator has an importance sampling interpretation, where the importance sampling density $\gamma(\cdot)$ for η_{it} is chosen to be of the same form $\phi(\cdot)$ that was used to draw accepted conditioning sequences, and the importance sampling weights are of a form analogous to that given in equation (11). Thus, the GHK procedure for simulating within period choice probabilities may be integrated with the recursive procedure for constructing conditioning sequences. Specifically, the conditioning sequences may be constructed element-by-element *within* each period, as in (12), and the sequence weights for transition probability simulation will continue to have a structure similar to (11), except that they will be a product of within period weights.

The drawback of my importance sampling technique for simulating transition probabilities is that the denominators of the sequence weights involve multiple integrals. Observe that in (12) the denominator of the weight for a t period sequence is a t variate integral. Simulation of tth period transition probabilities thus requires evaluation of sequence weights that involve $T-1$ variate integrals. Thus, the denominators of the weights must be simulated when $T \geqslant 4$ in order to avoid cumbersome 3-variate or higher numerical integration.

Note that the denominators of the sequence weights equal the expected value of the numerators over all possible sequences. Thus, given S sequences, an unbiased simulator of their sequence weight denominators is given by the mean of their sequence weight numerators. Such simulation of the sequence weight denominators induces denominator bias. This leads to biased simulators of transition probabilities when simulated sequence weights $\hat{\omega}(\eta_s^*(t-1))$ are substituted for the $\omega(\eta_s^*(t-1))$ in equation (10). In the case of biased simulation, McFadden and Ruud (1991) show that if the difference between the simulated MOM first order condition $W(d-\hat{E}(d|J,\hat{\theta},X))$ and the exact MOM first order condition $W(d-E(d|J,\hat{\theta},X))$ is $O_p(S^{-1/2})$ uniformly in $\hat{\theta}$, and if $S/\sqrt{N}\to\infty$, then the asymptotic bias of the MSM estimator converges to zero in probability uniformly in $\hat{\theta}$, and McFadden's (1989) consistency and asymptotic normality results continue to hold. This $O_p(S^{-1/2})$ condition is simple to verify in the present case. Define:

$$(13)\qquad \omega_{Ast} \equiv \omega_A\left(\eta_s^*(t)\right) = \operatorname{prob}\left(\eta_{t+1} \in \eta(J_{t+1}) | \eta(t) = \eta^*(t)\right) \cdot \operatorname{prob}\left(\eta_t \in \eta(J_t) | \eta(t-1) = \eta^*(t-1)\right) \cdot \cdots \cdot \operatorname{prob}\left(\eta_1 \in \eta(J_1)\right).$$

Then, defining $\hat{\omega}_{BSt} = S^{-1}\sum_{s=1}^{S}\omega_{Ast}$, the simulated sequence weights may be written as $\hat{\omega}(\eta_s^*(t)) = \omega_{Ast}/\hat{\omega}_{BSt}$. For person i, alternative j and time t, the

difference between the simulated and exact FOC contributions is:

$$(14)\qquad W_{ijt}\left[E\left(d_{ijt}|J_{i,t-1},\hat{\theta},X_i\right)-S^{-1}\sum_{s=1}^{S}\frac{\omega_{As,t-2}}{\hat{\omega}_{BS,t-2}}\hat{E}\left(d_{ijt}|\eta_s^*(t-2),\hat{\theta},X_i\right)\right]$$

$$=W_{ijt}\left[E\left(d_{ijt}|J_{i,t-1},\hat{\theta},X_i\right)\right.$$

$$\left.-S^{-1}\sum_{s=1}^{S}\frac{\omega_{As,t-2}}{\omega_{B,t-2}}\hat{E}\left(d_{ijt}|\eta_s^*(t-2),\hat{\theta},X_i\right)\right]$$

$$+W_{ijt}\left[S^{-1}\sum_{s=1}^{S}\left[\frac{\omega_{As,t-2}}{\omega_{B,t-2}}-\frac{\omega_{As,t-2}}{\hat{\omega}_{BS,t-2}}\right]\hat{E}\left(d_{ijt}|\eta_s^*(t-2),\hat{\theta},X_i\right)\right].$$

The first term on the right has expectation zero. The second term is the source of bias. Since W_{ijt} and $\hat{E}(d_{ijt}|\eta_s^*(t-2),\hat{\theta},X_i)$ are bounded, it is sufficient to show that $S^{-1}\Sigma_{s=1}^{S}(\omega_{As,t-2}/\omega_{B,t-2}-\omega_{As,t-2}/\hat{\omega}_{BS,t-2})$ is $O_p(S^{-1/2})$. Suppressing the time subscript, the first order Taylor expansion with remainder of $S^{-1}\Sigma_{s=1}^{S}\omega_{As}/\hat{\omega}_{BS}$ around $S^{-1}\Sigma_{s=1}^{S}\omega_{As}/\omega_B$ is (after some algebraic manipulation):

$$(15)\qquad S^{-1}\sum_{s=1}^{S}\frac{\omega_{As}}{\hat{\omega}_{BS}}=S^{-1}\sum_{s=1}^{S}\frac{\omega_{As}}{\omega_B}-S^{-1}\sum_{s=1}^{S}\frac{\omega_{As}}{\hat{\omega}_{BS}}$$

$$\times\left[\frac{\hat{\omega}_{BS}-\omega_B}{\omega_B}\frac{\hat{\omega}_{BS}}{\omega_B}+\frac{\omega_B^2-\hat{\omega}_{BS}^2}{\omega_B^2}-2\frac{\omega_B-\hat{\omega}_{BS}}{\omega_B}\right].$$

The terms $(\hat{\omega}_B-\omega_B)\hat{\omega}_{BS}/\omega_B^2$ and $(\omega_B^2-\hat{\omega}_{BS}^2)/\omega_B^2$ and $(\omega_B-\hat{\omega}_{BS})/\omega_B$ can all be shown to be $O_p(S^{-1/2})$ by simple applications of the Chebyshev inequality. The quantity $S^{-1}\Sigma_{s=1}^{S}(\omega_{As}/\hat{\omega}_{BS})$ is one by construction. Thus the $O_p(S^{-1/2})$ condition holds.

Despite the fact that $S/\sqrt{N}$ must go to infinity as N increases for the simulation estimator proposed here to be consistent and asymptotically normal, the Monte-Carlo results in Section 4 show little evidence of bias in the proposed estimator, even for small S. I believe the reason that simulation bias is relatively unimportant in the present case is that all time t transition probabilities for person i are simulated conditional on the same set of conditioning sequences, all of which have the same simulated denominator. Thus, all are biased by the same proportion. It is simple to show that any proportional bias of all time t transition probabilities will have no affect asymptotically (in N) on the MOM first order conditions if the optimal MOM weights are used. Of course the optimal weights are only simulated, but if the simulated optimal weights are highly correlated with the true optimal weights there should be a strong tendency for this proportional bias to cancel out of the MSM first order conditions.

Finally, a referee has observed that, using a technique due to McFadden and Ruud (1991), one may, after substituting the biased simulators (equation (5)) into the MOM first order condition (equation (4)), multiply through by the simulated sequence weight denominators $\hat{\omega}_{BS}$ to obtain the expression:

$$(19)\qquad \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{M}\frac{W_{ijt}}{\hat{\omega}_{BSi,t-2}} \times\left[d_{ijt}\hat{\omega}_{BSi,t-2}-\frac{1}{S}\sum_{s=1}^{S}\omega_{Asi,t-2}\hat{E}\left(d_{ijt}|\eta_s^*(t-2),\hat{\theta}_{\mathrm{MSM}},X_i\right)\right]\approx 0.$$

Here the term in brackets is a mean zero residual. Thus, the usual consistency and asymptotic normality results hold for this MSM estimator with *fixed* simulation size. Since this alternative estimator involves different simulated moments and weights than are given by equation (5), it will in general have different small sample properties. An important avenue for future research is to compare the small sample properties of these two estimators.

4. MONTE-CARLO TESTS

This section contains three Monte-Carlo tests of the proposed simulation estimator for panel data. Three types of Monte-Carlo data sets are created using different values for the true model parameters. For each parameter vector, 1,000 Monte-Carlo data sets are created, and a repeated sampling experiment is performed in which MSM estimates are obtained for each data set. For each estimation, a different seed for the probability simulator is used. The Monte-Carlo data for the tests are constructed using the latent variable model of Section 2, with the number of alternatives $M = 2$. The independent variables in all three models are a constant and a variable X that is normally distributed with a mean of 6, an across individual variance of 3, and a within individual variance of 2. Each individual is observed over 8 time periods. Thus, the simulation estimator is required to simulate 8-variate integrals. The number of people N is set to 500. Two types of models are considered in the experiments: a random effects model, so that comparisons can be made with ML, and a random effects plus AR(1) error model that is not feasible to estimate by ML. For comparison purposes, I have also included the SML estimator based on the GHK simulator in all three experiments.

The computational speed of both simulation estimators is such that repeated sampling experiments are feasible. For MSM, I have set the number of conditioning sequences S at 10. Since $M = 2$, the choice probabilities conditional on these sequences are evaluated numerically. For SML the number of draws used to form the GHK simulator is also set to 10. For the random effects model, ML-quadrature estimates are obtained using 16 quadrature points. For MSM, a single estimation of the random effects plus AR(1) error model on the 500 individual, 4,000 observation size data sets typically requires approximately 1.5 cpu minutes on an IBM 3090 computer, while runs on an IBM 486 PC

TABLE I

REPEATED SAMPLING EXPERIMENT—1,000 REPLICATIONS—RANDOM EFFECTS MODEL

		MSM—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std($\hat{\beta}$)	Mean Sîd($\hat{\beta}$)	t-Stat Bias
ρ	.600	.60603	.02903	.03304	6.55
Constant	−.900	−.89991	.08861	.08750	.03
X	.250	.24936	.01386	.01443	−1.45
		ML-Quadrature—16 Points			
Parameter	True Value	Mean $\hat{\beta}$	Std($\hat{\beta}$)	Mean Sîd($\hat{\beta}$)	t-Stat Bias
ρ	.600	.59628	.02635	.02668	−4.46
Constant	−.900	−.91017	.08149	.08293	−3.95
X	.250	.25152	.01310	.01332	3.67
		Simulated ML-GHK—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std($\hat{\beta}$)	Mean Sîd($\hat{\beta}$)	t-Stat Bias
ρ	.600	.57954	.02826	.02518	−22.99
Constant	−.900	−.90126	.08267	.07987	−.48
X	.250	.24952	.01331	.01289	−1.14

Note: 1,000 Monte-Carlo data sets were created using the true parameter values shown. Each data set consisted of 500 individuals and 4,000 person year observations. Mean $\hat{\beta}$ refers to mean of the estimated parameters over all 1,000 data sets. Std($\hat{\beta}$) refers to the sample standard deviation of the estimated parameters. Mean Sîd($\hat{\beta}$) refers to the mean of the estimated parameter standard errors over all 1,000 data sets. Typical times for a single estimation of the random effects model on an IBM 3090 are 1.2 cpu minutes for MSM and 0.8 cpu minutes for ML-quadrature.

(without math coprocessor) require an average of only 4.5 minutes. The SML runs typically require about 30% more time than the MSM runs. MSM requires that the probabilities of choosing all alternatives be simulated, while SML requires only that the probabilities of the chosen alternatives be simulated. However, this computational advantage for SML is counterbalanced by the fact that SML requires derivative calculations on each iteration, while MSM only requires that derivatives be calculated once to form the initial weighting matrix.

The results from the first repeated sampling experiment are reported in Table I. In this experiment, a random effects model is used. The true parameter values are $\rho = 0.60$, constant $= -0.90$, and X coefficient $= 0.25$, where ρ denotes the fraction of variance due to the individual effect. The mean and empirical standard deviation of the MSM, SML, and ML-quadrature estimates, along with the mean of the estimated standard errors, are reported. Also reported are t statistics for the statistical significance of the biases, based on the empirical standard errors. The mean MSM estimates of the constant and X coefficient are extremely close to the true values (e.g., 0.24936 vs. 0.250 for the X coefficient) and the estimated biases in these coefficients are insignificant. The mean MSM estimate of the fraction of variance due to the random effect is significantly larger than the true value (0.60603 vs. 0.6000 with a t statistic of 6.55 for the bias) but obviously the magnitude of the bias is small. The ML-quadrature estimates all suffer from biases that are significant but negligible in magnitude. (In an earlier version of this paper, I found that ML-quadrature

estimates suffer from serious bias in these types of models when serial correlation is of this magnitude and conventional numbers of quadrature points—3 to 9—are used.) The sample standard deviations of the MSM estimates range from 5.3 percent higher than that of ML-quadrature for the X coefficient (0.01386 vs. 0.01310) to 10.2 percent higher for ρ. Given the use here of a smooth simulator with simulation size 10 and of (simulated) optimal weights, these results conform well with the MSM to MOM asymptotic covariance matrix ratio of $(1 + 1/R)$ that obtains when frequency simulation is used (more efficient smooth simulators being expected to improve on this ratio). The only drawback in these results is that the mean estimated standard deviation of the ρ estimates is 13.8 percent larger than the empirical standard deviation (0.03304 vs. 0.02903), but the empirical and mean estimated standard deviations are quite close for the constant and X coefficient.

The SML estimates in Table I also appear acceptable. The biases in the estimates of the intercept and slope coefficient are insignificant. The estimate of ρ is very significantly biased downward, and the magnitude of this bias is more than three times greater than the magnitude of the bias in $\hat{\rho}$ for MSM. However, the magnitude of the bias still appears negligible. The empirical standard errors for the SML estimates are very close to those for MSM, and in all cases slightly smaller. The estimated SML standard errors are slightly downward biased.

The second repeated sampling experiment involves 1,000 replications of data constructed to have random effects plus an AR(1) error component ($\rho = 0.20$, AR(1) = 0.60, giving a first lagged autocorrelation of 0.68). The results are reported in Table II. Only MSM and SML results are reported, because ML by quadrature is not feasible for this model. The MSM results are again very

TABLE II

REPEATED SAMPLING EXPERIMENT—1,000 REPLICATIONS—RANDOM EFFECTS + AR(1) ERROR MODEL

		MSM—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std($\hat{\beta}$)	Mean $\widehat{\text{Std}}(\hat{\beta})$	t-Stat Bias
ρ	.200	.19401	.06970	.07065	−2.72
AR(1)	.600	.60421	.04713	.04816	2.83
Constant	−.900	−.90041	.08449	.08431	−.15
X	.250	.24998	.01348	.01404	−.05
		Simulated ML-GHK—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std($\hat{\beta}$)	Mean $\widehat{\text{Std}}(\hat{\beta})$	t-Stat Bias
ρ	.200	.21915	.05368	.04838	11.26**
AR(1)	.600	.53747	.04183	.03922	−47.37**
Constant	−.900	−.90298	.08123	.07703	1.16
X	.250	.24984	.01329	.01251	−.38

Note: 1,000 Monte-Carlo data sets were created using the true parameter values shown. Each data set consisted of 500 individuals and 4,000 person year observations. Mean $\hat{\beta}$ refers to mean of the estimated parameters over all 1,000 data sets. Std($\hat{\beta}$) refers to the sample standard deviation of the estimated parameters. Mean $\widehat{\text{Std}}(\hat{\beta})$ refers to the mean of the estimated parameter standard errors over all 1,000 data sets. Typical times for a single MSM estimation of the random effects plus AR(1) error model are 1.5 cpu minutes on an IBM 3090 and 4.5 minutes on an IBM 486 PC.

impressive. Bias is insignificant for the estimates of the intercept and X coefficient. Bias in the estimates of ρ and the AR(1) coefficient are significant but very small in magnitude. In fact, the mean estimated parameter values are all within 0.09 empirical standard errors of the true values, and the mean estimated standard errors are all within 4.2 percent of the empirical standard errors. The SML estimates of the intercept and X coefficient also exhibit no significant bias. However, the SML estimator has some difficulty in uncovering the true correlation structure in this example. In particular, the estimates of the AR(1) parameter are significantly biased downward (0.53747 vs. 0.60 with a t statistic of -47.37 for the bias), and the magnitude of this bias exceeds 10% of the true value.

In the third experiment, I have set $\rho = 0.20$, AR(1) = 0.90, constant = -0.600, and X coefficient = 0.10. Thus, the first lagged autocorrelation is 0.92. Despite this high degree of serial correlation, the MSM estimator continues to perform well. All four parameter estimates are significantly biased, but in each case the magnitude of the bias is negligible. The SML estimator, on the other hand, has severe difficulty in uncovering the true error structure in this example. It greatly overestimates the fraction of variance due to the random effect (43.8% vs. a true value of 20%) and substantially understates the AR(1) parameter (0.8007 vs. a true value of 0.90). Note that although the bias in $\hat{\rho}$ may appear small in percentage terms (only 11%), an AR(1) parameter of 0.8 implies serial correlation that dies out much more quickly than if the AR(1) parameter were 0.9. Thus, the SML estimates imply less short run persistence and more very long run persistence in choice behavior than does the true model. The SML estimates of the intercept and slope also show significant bias, but again, as in experiments 1 and 2, this bias is negligible in magnitude. Finally, note that in experiment 3 both the MSM and SML estimators have difficulty in estimating the standard errors of the correlation structure parameters. For SML the means of the estimated standard errors for $\hat{\rho}$ and $\hat{\text{AR}}(1)$ are 30% and 25% below the empirical standard errors, respectively. For MSM, they are both over 3 times larger than the empirical standard errors. This may be because with such strong serial correlation the effective sample size is small, causing the asymptotic distribution of the estimators to be a poor approximation to the finite sample distribution.

A note of caution is in order with regard to using empirical means and standard deviations to characterize the small sample distributions of these estimators. Although the MSM, SML, and ML-quadrature estimators are all aysmptotically normal, they may depart substantially from normality in small samples—in which case the empirical mean and standard deviation do not adequately characterize their distributions. Furthermore, the first and second moments of the estimators for the AR(1) parameter, constant, and X coefficient are not finite in finite samples. In practice, this may cause empirical means and standard deviations to be dominated by extreme outlier estimates.

To address these problems, I also calculated the 1st, 2nd, and 3rd quantiles of the parameter estimates for all the Monte-Carlo experiments in Tables I–III. Looking at quantiles does not affect the ranking of performance of the estima-

TABLE III

REPEATED SAMPLING EXPERIMENT—1,000 REPLICATIONS—RANDOM EFFECTS + AR(1) ERROR MODEL

		MSM—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std$(\hat{\beta})$	Mean $\hat{\text{Std}}(\hat{\beta})$	t-Stat Bias
ρ	.200	.19248	.12484	.42393	−1.90
AR(1)	.900	.90473	.01949	.05903	7.63
Constant	−.600	−.59245	.07496	.07572	3.19
X	.100	.09882	.00923	.01036	−4.07
		Simulated ML-GHK—10 Draws			
Parameter	True Value	Mean $\hat{\beta}$	Std$(\hat{\beta})$	Mean $\hat{\text{Std}}(\hat{\beta})$	t-Stat Bias
ρ	.200	.43922	.13932	.10036	54.24
AR(1)	.900	.80070	.05027	.03752	−64.45
Constant	−.600	−.61785	.07034	.06648	−8.41
x	.100	.10292	.00885	.00899	10.43

Note: 1,000 Monte-Carlo data sets were created using the true parameter values shown. Each data set consisted of 500 individuals and 4,000 person year observations. Mean $\hat{\beta}$ refers to mean of the estimated parameters over all 1,000 data sets. Std$(\hat{\beta})$ refers to the sample standard deviation of the estimated parameters. Mean $\hat{\text{Std}}(\hat{\beta})$ refers to the mean of the estimated parameter standard errors over all 1,000 data sets. Typical times for a single MSM estimation of the random effects plus AR(1) error model are 1.5 cpu minutes on an IBM 3090 and 4.5 minutes on an IBM 486 PC.

tors. In all cases, except for the parameter ρ in Table III, the actual quantiles were negligibly different from those that would be predicted based on normality (a table of these results is available on request). For the MSM estimates of ρ in Table III, the 1st, 2nd, and 3rd quantiles are 0.08989, 0.14948, and 0.30745. Given the mean and empirical standard deviation of the estimates and assuming normality, one would predict these quantiles to be 0.10822, 0.19248, and 0.27675 respectively. For the SML estimates of ρ the quantiles are 0.36737, 0.46609, and 0.54078 respectively, while given normality one would predict 0.34518, 0.43922, and 0.53326. Although the distributions of both estimators depart substantially from normality, there were no extreme outliers influencing the first two moments in either case. Since the true value of ρ is 0.200, while the mean MSM estimate is 0.19248 and the median MSM estimate is 0.14948, MSM does not perform as well under a median criterion. However, it continues to dominate SML, for which the median is 0.46609 and all but 72 of the 1,000 estimates exceed the true value.

5. CONCLUSION

In this paper I have used a factorization of the MSM first order condition into transition probabilities in order to construct a practical MSM estimator for panel data. The key problem in constructing such an estimator is the inexpensive simulation of transition probabilities without the introduction of substantial bias. I develop a particular very inexpensive importance sampling technique for simulation of transition probabilities which is asymptotically unbiased in simulation size. This simulation method is of interest in its own right, since Hajivassiliou, McFadden, and Ruud, who call it the GHK simulator, find that

this method outperforms all other probability simulators from a large set of alternative methods they consider. A battery of Monte-Carlo tests reveal no significant bias in an MSM estimator that utilizes this method to simulate transition probabilities, even for small simulation sizes. For models in which ML is feasible, the MSM estimator performs well relative to quadrature-based methods using many more quadrature points than are typically used in practice. For models in which ML is not feasible, the MSM estimator outperforms SML based on the highly accurate GHK probability simulator. Specifically, when serial correlation is very strong I find that SML produces severely biased estimates of the serial correlation structure, while the MSM estimator does not. The computational speed of the MSM estimator is such that it is possible to estimate panel data models with complex error structures involving random effects and ARMA errors in times similar to those necessary for estimation of simple random effects models by quadrature.

I motivated this paper by arguing that the ability to deal with complex patterns of serial correlation in LDV models may be important for out-of-sample prediction purposes. As an example, Elrod and Keane (1992) have applied the method developed in this paper to estimate detergent choice models with eight alternatives and 30 time periods. We find that a probit model that allows for a rich pattern of serial correlation outperforms all competing choice models in terms of accuracy in forecasting agents' future purchases.

Industrial Relations Center and Dept. of Economics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Manuscript received August, 1990; final revision received April, 1993.

REFERENCES

ALBERT, J., AND S. CHIB (1993): "Bayesian Analysis of Binary and Polychotomous Data," forthcoming, *Journal of the American Statistical Association*.

ALBRIGHT, R., S. LERMAN, AND C. MANSKI (1977): "Report on the Development of an Estimation Program for the Multinomial Probit Model," Report prepared by Cambridge Systematics for the Federal Highway Administration.

AVERY, R., L. HANSEN, AND V. J. HOTZ (1983): "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24, 21–35.

BATES, C., AND H. WHITE (1987): "Efficient Instrumental Variable Estimation of Systems of Equations with Nonspherical Errors," UCSD Discussion Paper #87–14.

BORSCH-SUPAN, A., AND V. HAJIVASSILIOU (1990): "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," Cowles Foundation Discussion Paper 960.

BUTLER, J. S., AND R. MOFFITT (1982): "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50, 761–764.

CHAMBERLAIN, G. (1985): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intrilligator. Amsterdam: North-Holland Publishing Co.

ELROD, T., AND M. KEANE (1992): "A Factor-Analytic Probit Model for Estimating Market Structure in Panel Data," forthcoming, *Journal of Marketing Research*.

GEWEKE, J. (1987): "Bayesian Inference in Econometric Models using Monte-Carlo Integration," *Econometrica*, 57, 1317–1340.

——— : "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," *Computing Science and Statistics: Proceedings of the Twenty-third Symposium on the Interface*, forthcoming.

GEWEKE, J., M. KEANE, AND D. RUNKLE (1992): "Alternative Computational Approaches to Statistical Inference in the Multinomial Probit Model," University of Minnesota manuscript.

HAJIVASSILIOU, V., AND D. MCFADDEN (1990): "The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crisis," Cowles Foundation Discussion Paper 967.

HAJIVASSILIOU, V., D. MCFADDEN, AND P. RUUD (1991): "Simulation of Multivariate Normal Orthant Probabilities: Methods and Programs," Working Paper, Yale University.

HECKMAN, J. (1981): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden. Cambridge: MIT Press.

KAHANER, D. K. (1991): "A Survey of Existing Multinomial Quadrature Routines," in *Statistical Multiple Integration: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference*, ed. by N. Flournoy and R. K. Tsutakawa. Providence, RI: American Mathematical Society.

LERMAN, S., AND C. MANSKI (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden. Cambridge: MIT Press.

MCCULLOCH, R., AND P. ROSSI (1992): "An Exact Likelihood Analysis of the Multinomial Probit Model," Working Paper, University of Chicago.

MCFADDEN, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57, 995–1026.

MCFADDEN, D., AND P. RUUD (1987): "Estimation of Limited Dependent Variable Models from the Regular Exponential Family by the Method of Simulated Moments," Working Paper, University of California, Berkeley.

——— : "Estimation by Simulation," Working Paper, University of California, Berkeley.

PAKES, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–784.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1058.

POIRIER, D., AND P. RUUD (1988): "Probit with Dependent Observations," *Review of Economic Studies*, 55, 593–614.

ROBINSON, P. (1982): "On the Asymptotic Properties of Models Containing Limited Dependent Variables," *Econometrica*, 50, 27–41.

RUUD, P. (1992): "Extensions of Estimation Methods using the EM Algorithm," forthcoming in *Journal of Econometrics*.

VAN PRAAG, B. M. S., AND J. P. HOP (1987): "Estimation of Continuous Models on the Basis of Set-valued Observations," manuscript, Erasmus University, Rotterdam.