working paper
department
of economics

ETHOD OF SIMULATED MOMENTS FOR ESTIMATION OF DISCRETE
RESPONSE MODELS WITHOUT NUMERICAL INTEGRATION

Daniel McFadden

No. 464 August 1987

massachusetts
institute of
technology
50 memorial drive
cambridge, mass. 02139

A METHOD OF SIMULATED MOMENTS FOR ESTIMATION OF DISCRETE RESPONSE MODELS WITHOUT NUMERICAL INTEGRATION

Daniel McFadden

No. 464 August 1987

A METHOD OF SIMULATED MOMENTS FOR ESTIMATION OF DISCRETE RESPONSE MODELS WITHOUT NUMERICAL INTEGRATION

by

Daniel McFadden
Department of Economics
Massachusetts Institute of Technology
Cambridge, MA 02139

March 1986 (Revised July 1986, August 1987)

ABSTRACT

This paper proposes a simple modification of a conventional method of moments estimator for a discrete response model, replacing response probabilities that require numerical integration with estimators obtained by Monte Carlo simulation. This method of simulated moments (MSM) does not require precise estimates of these probabilities for consistency and asymptotic normality, relying instead on the law of large numbers operating across observations to control simulation error, and hence can use simulations of practical size. The method is useful for models such as high-dimensional multinomial probit (MNP), where computation has restricted applications.

ACKNOWLEDGEMENTS

This research grows out of joint work with Kenneth Train on the estimation of choice models containing variables measured with error. I have particularly benefited from discussions with Ariel Pakes and David Pollard, who pointed out a lacuna in my original analysis of this problem, and whose independent investigation of the asymptotic behavior of simulation experiments motivated several critical steps in my proofs. I have also benefited from suggestions made by Chungrung Ai, Moshe Ben-Akiva, Chris Cavanagh, Vassilis Hajivassiliou, Robert Hall, James Heckman, Hidehiko Ichimura, Charles Manski, Dan Nelson, Peter Phillips, and Paul Ruud. This research was supported in part by National Science Foundation Grant No. SES-8606349.

KEYWORDS: METHOD OF MOMENTS, SIMULATION, MULTINOMIAL PROBIT, DISCRETE RESPONSE

forthcoming, *Econometrica*

1. INTRODUCTION

A classical method of moments estimator θ_{mm} of an unknown parameter vector θ^* minimizes the (generalized) distance from zero of empirical moments

$$(1) \quad \sum_{\text{observations}} \begin{bmatrix}\text{Instrument}\\ \text{Vector}\end{bmatrix}\left(\begin{bmatrix}\text{Observed}\\ \text{Response}\end{bmatrix} - \begin{bmatrix}\text{Expected}\\ \text{Response at } \theta_{mm}\end{bmatrix}\right).$$

For some problems, the expected response function may be difficult to express analytically or compute, but relatively easy to simulate. When this function is replaced by an asymptotically unbiased simulator such that the simulation errors are independent across observations and sufficiently regular in θ, the variance introduced by simulation will be controlled by the law of large numbers operating across observations, making it unnecessary to consistently estimate each expected response. This is the basis for the estimation method developed in this paper, the <u>method of simulated moments</u> (MSM).[1]

This paper focuses on application of MSM to the multinomial probit model. However, the method is more general and can be applied to most moment estimation problems. In a related paper, Pakes and Pollard (1987) have independently proposed minimum distance estimators using simulation, and have established their statistical properties using combinatorial empirical process methods. Most of the statistical results in this paper could be obtained by application of their methods.

Section 2 of this paper gives definitions and notation for discrete response models. Section 3 defines the MSM estimator and gives an informal argument that it is consistent asymptotically normal (CAN). Section 4 discusses issues of computation and statistical efficiency. Sections 5-7 discuss applications of the method to discrete panel data with autoregressive errors, to discrete response models with measurement errors in explanatory variables, and to non-normal discrete response problems. An Appendix contains formal statements of assumptions and results.

2. DEFINITIONS AND NOTATION

Define $C = \{1,\ldots,m\}$ to be a set of mutually exclusive and exhaustive alternatives. A latent variable model for response from C is defined by

(2) $u_i = \alpha x_i$, $i \in C$,

where α is a row vector of individual weights distributed randomly in the population, x_i is a column vector of measured attributes of alternative i, and response i is observed if $u_i \geq u_j$ for $j \in C$ (with zero probability of ties). Let d_i denote a response indicator, equal to one for the observed response, zero otherwise.

Assume $\alpha = a(\theta,\eta)$ is a smooth parametric function of a random vector η, with unknown parameter vector θ taking true value θ^*. Let $g(\eta)$ denote the density of η, and $g_a(\alpha|\theta)$ the induced density of α . Let $\beta(\theta)$ and $\Omega(\theta)$ denote the mean and covariance matrix of α. In applications, it is often convenient to work with a Cholesky factorization of Ω: let $\Gamma(\theta)$ be a upper triangular matrix satisfying $\Gamma'\Gamma = \Omega$.

Define $X_C = (x_1,\ldots,x_m)$ and $u_C = (u_1,\ldots,u_m)$. The response probability for alternative i, $P_C(i|\theta,X_C)$, equals the probability of drawing a latent vector u_C with $u_i \geq u_j$ for $j \in C$, given X_C. Define

(3) $u_{C-i} = (u_1-u_i,\ldots,u_{i-1}-u_i,u_{i+1}-u_i,\ldots,u_m-u_i)$

(4) $X_{C-i} = (x_1-x_i,\ldots,x_{i-1}-x_i,x_{i+1}-x_i,\ldots,x_m-x_i)$.

Then u_{C-i} has a multivariate density $g_U(u_{C-i}|\theta,X_C)$ with mean βX_{C-i} and covariance matrix $X_{C-i}'\Omega X_{C-i}$, and $P_C(i|\theta,X_C)$ equals the nonpositive orthant probability of $u_{C-i} = a(\theta,\eta)X_{C-i}$,

(5) $$P_C(i|\theta,X_C) = \int 1(u_{C-i}\le 0)g_U(u_{C-i}|\theta,X_C)du_{C-i}$$
$$= \int 1(a(\theta,\eta)X_{C-i}\le 0)g(\eta)d\eta,$$

where $1(Q)$ denotes an indicator function for the event Q.

When α is multivariate normal, one obtains the MNP model. For this model, α can be written

(6) $$\alpha = a(\theta,\eta) \equiv \beta(\theta) + \eta\Gamma(\theta),$$

with η a row vector of independent standard normal variates.

In economic applications, the latent variables u_i often have the interpretation of utility or profit, and $P_C(i|\theta,X_C)$ is the choice probability for a population of optimizing agents. The attributes x_i are functions of observed characteristics of the alternatives and of the decision-makers, with αx_i interpreted as an approximation to a general economic function of observed and unobserved characteristics and of the deep parameter θ. Alternative-specific dummy variables may be included in x_i; the associated components of α can be interpreted as alternative-specific additive disturbances.

Let $n = 1,\ldots,N$ index a random sample from the population, yielding observations (d_{Cn},X_{Cn}) with $d_{Cn} = (d_{1n},\ldots,d_{mn})$ and $X_{Cn} = (x_{1n},\ldots,x_{mn})$. The log likelihood of the sample is

(7) $$L(\theta) = \sum_{n=1}^{N} \sum_{i\in C} d_{in} \text{Ln } P_C(i|\theta,X_{Cn}).$$

The associated score is[2]

(8) $$\partial L(\theta)/\partial\theta = \sum_{n=1}^{N} \sum_{i\in C} W_{in}[d_{in} - P_C(i|\theta,X_{Cn})],$$

where

(9) $$W_{in} = \partial \mathrm{Ln}\, P_C(i|\theta,X_{Cn})/\partial\theta.$$

The primary impediment to practical maximum likelihood estimation of θ for the MNP model is computation of the (m-1) dimensional orthant probabilities for u_{C-i} to obtain $P_C(i|\theta,X_C)$. Direct numerical integration is practical for $m \leq 4$ using a method of Owen (1956), modified by Hausman and Wise (1978), or expansions due to Dutt (1976). Otherwise, unless α has a factor-analytic covariance structure with less than four factors, it is usually impractical to carry out the large number of numerical integrations required to iteratively maximize (6). Lerman and Manski (1981) suggest a Monte Carlo procedure for estimating $P(i|\theta,X_C)$ that can be applied to MNP models with large m; but find that it requires an impractical number of Monte Carlo draws to estimate small probabilities and their derivatives with acceptable precision. Daganzo (1980) has developed approximate maximum likelihood estimators for MNP using a normal approximation to maxima of normal variates suggested by Clark (1961). This approach has the drawbacks that the accuracy of the approximation cannot be refined with increasing sample size, and the method can be inaccurate when components have unequal variances; see Horowitz, Sparmonn, and Daganzo (1982).

3. THE METHOD OF SIMULATED MOMENTS

The conventional method of moments estimator of a $k \times 1$ parameter vector θ in the discrete response model $P_C(i|\theta,X_C)$ satisfies

(10) $\theta_{mm} = \text{argmin}_{\theta}\ (d-P(\theta))'W'W(d-P(\theta))$,

where $d-P(\theta)$ denotes the $mN \times 1$ vector of residuals $d_{in} - P_C(i|\theta,X_{Cn})$ stacked by observation and by alternative within observation, and where W is a $K \times mN$ array of instruments of rank $K \geq k$. The instruments may depend on θ, but are evaluated at some fixed θ_o in forming first-order conditions for solution of (10). The instrument array (9), evaluated at θ^* (or at a consistent estimator of θ^*) yields a method of moments estimator asymptotically equivalent to the maximum likelihood estimator for θ, and hence asymptotically efficient. If computation makes exact calculation of the efficient instruments impractical, (9) nevertheless provides a template for instruments that with relatively crude approximations to P and its $mN \times k$ array of derivatives P_θ will yield moderately efficient estimators.[3]

Under mild regularity assumptions, sufficient conditions for classical method of moments estimation to be CAN are (i) and (ii):

(i) The instruments are asymptotically correlated with the score; i.e., the array $\bar{R} = \lim N^{-1}WP_\theta(\theta^*)$ is of maximum rank.

(ii) The conditional expectation of the residuals $d-P(\theta)$, given the instruments, is zero if and only if $\theta = \theta^*$.

In the remainder of this section, I will assume the instruments W are a computationally practical fixed array, defined independently of θ. (Approximation of the optimal instruments (9) is considered in Section 4.) The method of simulated moments (MSM) avoids the computation of $P(\theta)$ required for (10), replacing it with a simulator $f(\theta)$ that is (asymptotically) conditionally unbiased, given W and d, and independent across observations. The MSM estimator is given by any argument θ_{sm} satisfying

(11) $(d-f(\theta_{sm}))'W'W(d-f(\theta_{sm})) \le \inf_{\theta} (d-f(\theta))'W'W(d-f(\theta)) + O(1)$.

Simulators for the Response Probabilities

An unbiased *frequency* simulator $f(\theta)$ is readily calculated from the latent variable model (2): Draw one or more vectors η from the density $g(\eta)$, independently for each observation n, and fix these draws for the remainder of the analysis. Given trial θ, calculate $u_{Cn} = a(\theta,\eta)X_{Cn}$ and calculate the frequency $f_C(i|\theta,X_{Cn})$ with which component i of u_{Cn} is largest. This simulator has discontinuities at values of θ where there are ties for the maximum component of u_{Cn}. For the MNP model, the frequency simulator is computed economically from (6) by drawing standard normal vectors η and calculating $u_{Cn} = (\beta(\theta) + \eta\Gamma(\theta))X_{Cn}$.

It is also possible to construct *smooth* unbiased simulators $f(\theta)$. This simplifies the iterative computation of the estimator, and its statistical analysis. Let $\gamma(u_{C-i})$ denote a density chosen for the simulation that has the nonpositive orthant as its support. Then (5) can be rewritten

(12) $$P_C(i|\theta,X_C) \equiv \int h(u_{C-i},\theta,X_{C-i})\gamma(u_{C-i})du_{C-i},$$

where $h(u_{C-i},\theta,X_{C-i}) = g_U(u_{C-i}|\theta,X_C)/\gamma(u_{C-i})$. Average $h(u,\theta,X_{C-i})$ for an observation, using one or more Monte Carlo draws from $\gamma(u_{C-i})$ that are taken independently across observations and fixed for different θ. This gives a smooth positive unbiased estimator of $P_C(i|\theta,X_C)$, provided γ has sufficiently thick tails so the expectation of h exists. The density γ can be chosen to facilitate Monte Carlo draws and reduce simulation variance. For example, if γ is independent exponential in each component, then random variates from this distribution can be calculated from logarithms of uniform random numbers from

(0,1). Choices of γ that make h flatter can reduce simulation error, as in Monte Carlo importance sampling. For MNP, $h(u_{C-i},\theta,X_{C-i}) = n(u_{C-i}-\beta(\theta)X_{C-i},X'_{C-i}\Gamma(\theta)'\Gamma(\theta)X_{C-i})/\gamma(u_{C-i})$, where $n(v,A)$ denotes a multivariate normal density centered at zero with covariance matrix A. When γ is exponential, this h is uniformly bounded.

A potential drawback of smooth simulators based on (12) is that they are not constrained to sum up to one for $i \in C$. An alternative class of kernel-smoothed frequency simulators are defined in Section 4 that satisfy summing-up, but are only asymptotically unbiased. Section 4 also defines special unbiased smooth frequency simulators for MNP.

Statistical Properties

I shall argue that MSM estimators are CAN under mild regularity conditions. The main result on the asymptotic properties of these estimators is given in Theorem 1 below. I will assume that the parameter space Θ is a closed convex subset of $[0,1]^k$, that the true θ^* is in the interior of Θ, that the explanatory variables have a distribution with compact support, that the response probabilities are uniformly bounded and twice continuously differentiable with respect to θ, and that the instruments are smooth functions. Define a simulation bias $B(\theta) = N^{-1/2}W(Ef(\theta)-P(\theta))$; this is zero if unbiased simulators are used, and more generally is assumed to satisfy[4]

(13) $\sup_{\Theta} |B(\theta)| = o_p(1)$.

Define a simulation residual process $\zeta(\theta) = N^{-1/2}W(f(\theta)-Ef(\theta))$. These simulation residuals are by construction the normalized sum over observations of independent identically distributed terms, independent of d and uniformly bounded, with $E(\zeta(\theta)|W) = 0$ for each θ. Then $\zeta(\theta)$ is an empirical process in θ that by a standard central limit theorem is pointwise asymptotically normal. I shall need the following critical <u>stochastic boundedness</u> and <u>stochastic equicontinuity</u> assumptions:

(14) $\sup_{\Theta} |\zeta(\theta)| = O_p(1)$,

(15) $\sup_{\theta\in A_N} |\zeta(\theta)-\zeta(\theta^*)| = o_p(1)$,

where $A_N = \{\theta|\ N^{1/2}|\theta-\theta^*|\le O(1)\}$.

I prove these properties for smooth simulators in Proposition 1 at the end of this section; the technically more difficult case of simulators with discontinuities is handled in Appendix Lemma 7.

<u>Theorem 1. Suppose the MSM estimator</u> θ_{sm} <u>defined by</u> (11) <u>satisfies Appendix assumptions</u> [A1] to [A10]. (These are stated informally as the assumptions in the preceding paragraph.) <u>Then</u> θ_{sm} <u>is consistent, with</u> $N^{1/2}(\theta_{sm}-\theta^*)$ <u>converging in distribution to a normal vector with mean zero and covariance matrix</u> $\Sigma_{sm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{sm}\bar{R}(\bar{R}'\bar{R})^{-1}$, <u>with</u> $\bar{R} = \lim N^{-1}WP_\theta(\theta^*)$ <u>and</u> $G_{sm} = \lim N^{-1}EW(d-f(\theta^*))(d-f(\theta^*))'W'$.

Proof: The argument parallels that of Pakes and Pollard (1987). The vector $W(d-f(\theta))$ entering the defining condition (11) for the MSM estimator can be decomposed into four terms,

(16) $$N^{-1/2}W(d-f(\theta)) \equiv [\zeta(\theta^*) - \zeta(\theta)] - B(\theta) + [N^{-1/2}W(d-f(\theta^*)) + B(\theta^*)] - [N^{-1/2}W(P(\theta)-P(\theta^*))].$$

The asymptotic properties of the estimator are argued by applying conditions (13)-(15) to the first two terms in (16), and applying the following arguments to the last two terms.

(a) By construction of the simulator, $N^{-1/2}W(d-f(\theta^*)) + B(\theta^*)$ has expectation zero, given W. Random sampling plus the independence of the simulators across observations implies by a central limit theorem that this term converges in distribution to a multivariate normal vector Z with mean zero and covariance matrix G_{sm}.[5]

(b) The expression $\omega_N(\theta) \equiv N^{-1}W(P(\theta)-P(\theta^*))$ converges uniformly in probability to a smooth function $\omega(\theta)$ with the properties that $\omega(\theta) = 0$ if and only if $\theta = \theta^*$, and that $\bar{R} \equiv \omega_\theta(\theta^*) = \lim \omega_{\theta N}(\theta^*)$ is of rank k.[6]

Consider first the consistency of θ_{sm}. Argument (a) implies $N^{-1/2}W(d-f(\theta^*)) = O_p(1)$. Hence, (11) satisfies

(17) $$N^{-1}(d-f(\theta_{sm}))'W'W(d-f(\theta_{sm}))$$
$$\le N^{-1}\inf_\Theta (d-f(\theta))'W'W(d-f(\theta)) + O(N^{-1})$$
$$\le [N^{-1/2}W(d-f(\theta^*))]'[N^{-1/2}W(d-f(\theta^*))] + O(N^{-1}) = O_p(1),$$

implying $N^{-1/2}W(d-f(\theta_{sm})) = O_p(1)$. Then, multiplying (16) by $N^{-1/2}$ and using (13), (14) and argument (b), one has $\omega_N(\theta_{sm}) = o_p(1)$. But ω_N converges uniformly outside each neighborhood of θ^* to a function bounded away from zero. Hence, the probability that θ_{sm} is contained in any neighborhood of θ^* approaches one.

Next, I argue that $N^{1/2}(\theta_{sm}-\theta^*)$ is stochastically bounded. From (16), the condition $N^{-1/2}W(d-f(\theta^*)) = O_p(1)$ plus (13), (14), and (17) imply

(18) $$O_p(1) = N^{-1/2}W(P(\theta_{sm})-P(\theta^*)).$$

A Taylor's expansion yields

(19) $N^{-1/2}W(P(\theta_{sm})-P(\theta^{*})) = [N^{-1}WP_{\theta}(\theta^{*}) + O(\theta_{sm}-\theta^{*})]N^{1/2}(\theta_{sm}-\theta^{*})$.

Then $N^{-1}WP_{\theta}(\theta^{*}) = \bar{R} + o_p(1)$ and $\theta_{sm} = \theta^{*} + o_p(1)$ imply

(20) $O_p(1) = [\bar{R}+o_p(1)]N^{1/2}(\theta_{sm}-\theta^{*})$.

Since $\bar{R}$ is of rank k, this implies $N^{1/2}(\theta_{sm}-\theta^{*}) = O_p(1)$.

Finally, consider the asymptotic normality of the MSM estimator. An asymptotically normal statistic $\tilde{\theta}$ is defined, and then θ_{sm} is shown to be asymptotically equivalent to it. Let $\tilde{\theta} = \theta^{*} + (\bar{R}'\bar{R})^{-1}\bar{R}'N^{-1}W(d-f(\theta^{*}))$. Argument (a) implies $N^{1/2}(\tilde{\theta}-\theta^{*}) = (\bar{R}'\bar{R})^{-1}\bar{R}'Z + o_p(1) = O_p(1)$. Then $N^{1/2}(\tilde{\theta}-\theta^{*})$ is asymptotically normal with covariance matrix $(\bar{R}'\bar{R})^{-1}\bar{R}'G_{sm}\bar{R}(\bar{R}'\bar{R})^{-1}$. Also, (15) implies $\zeta(\theta^{*}) - \zeta(\tilde{\theta}) = o_p(1)$. Substituting $\tilde{\theta}$ in (16) and applying the Taylor's expansion (19) with $\tilde{\theta}$ in place of θ_{sm} implies

(21) $N^{-1/2}W(d-f(\tilde{\theta})) = Z - [\bar{R} + O(\tilde{\theta}-\theta^{*})]N^{1/2}(\tilde{\theta}-\theta^{*}) + o_p(1)$

$$= [I-\bar{R}(\bar{R}'\bar{R})^{-1}\bar{R}']Z + o_p(1).$$

From (13), (15), and argument (b),

(22) $\Delta \equiv N^{-1/2}W(f(\tilde{\theta})-f(\theta_{sm}))$

$$= N^{-1/2}W(P(\tilde{\theta})-P(\theta_{sm})) + \zeta(\tilde{\theta}) + B(\tilde{\theta}) - \zeta(\theta_{sm}) - B(\theta_{sm})$$

$$= N^{-1/2}W(P(\tilde{\theta})-P(\theta_{sm})) + o_p(1) = \bar{R}N^{1/2}(\tilde{\theta}-\theta_{sm}) + o_p(1)$$

$$= \bar{R}(\bar{R}'\bar{R})^{-1}\bar{R}'Z + o_p(1).$$

Rewrite condition (11) characterizing θ_{sm} as

(23) $N^{-1}(d-f(\tilde{\theta}))'W'W(d-f(\tilde{\theta})) + o_p(1)$

$$\geq N^{-1}(d-f(\theta_{sm}))'W'W(d-f(\theta_{sm}))$$

$$= N^{-1}(d-f(\tilde{\theta}))'W'W(d-f(\tilde{\theta})) + 2N^{-1/2}(d-f(\tilde{\theta}))'W'\Delta + \Delta'\Delta.$$

From (21) and (22), $N^{-1/2}(d-f(\tilde{\theta}))'W'\Delta = o_p(1)$, and (23) implies $\Delta'\Delta = o_p(1)$. But then $\Delta \equiv \bar{R}N^{1/2}(\tilde{\theta}-\theta_{sm}) + o_p(1) = o_p(1)$, implying θ_{sm} and $\tilde{\theta}$ asymptotically equivalent. □

The stochastic boundedness and equicontinuity conditions used in the proof of Theorem 1 can be demonstrated for smooth simulators by the following argument:

Proposition 1. If the simulator $f(\theta)$ is uniformly bounded and twice continuously differentiable, then (14) and (15) hold.

Proof: A second-order Taylor's expansion of ζ about θ^* yields

(24) $$\zeta(\theta)-\zeta(\theta^*) = \zeta_\theta(\theta^*)(\theta-\theta^*) + (1/2)[N^{-1/2}\zeta_{\theta\theta}]\mathrm{vec}([(\theta-\theta^*)]'[N^{1/2}(\theta-\theta^*)]),$$

where $\zeta_{\theta\theta}$ is a $k \times k^2$ array of second derivatives evaluated at points between θ and θ^*. The array ζ_θ satisfies $E(\zeta_\theta(\theta^*)) = 0$, with independence across observations, so a central limit theorem implies $\zeta_\theta(\theta^*) = O_p(1)$. The contribution of each observation to the array $\zeta_{\theta\theta}$ is uniformly bounded, so $N^{-1/2}\zeta_{\theta\theta} = O_p(1)$. Hence, (24) implies, for $A_N = \{\theta |\ N^{1/2}|\theta-\theta^*| \le O(1)\}$,

(25) $$\sup_{\theta\in A_N} |\zeta(\theta)-\zeta(\tilde{\theta})| = O_p(1)\cdot O_p(N^{-1/2}),$$

establishing (15).

I next establish (14), using a "chaining" argument. Given an integer i, cover $[0,1]^k$ with 2^{ki} cubes with sides 2^{-i}, and let Θ_i be a set containing one point selected from each cube that intersects Θ. For $\theta \in \Theta$, define $\theta_i = \theta_i(\theta)$ to be the nearest point in Θ_i; then $|\theta-\theta_i(\theta)| < 2^{-i}$ and $|\theta_{i+1}(\theta)-\theta_i(\theta)| < 2^{-i}$. From this construction,

(26) $$|\zeta(\theta)| \le |\zeta(\theta_1)| + \sum_{i=1}^{\infty} |\zeta(\theta_{i+1})-\zeta(\theta_i)|.$$

I shall need a version of Bernstein's inequality, giving an exponential bound for sums of independent random variables: If Y_i are independently identically distributed with $|Y_i| \le c$ and $EY_i = 0$, then for $t > 0$,[7]

(27) $$P\{N^{-1/2}|\sum_{i=1}^{N} Y_i| > t\} \le 2\exp[-t^2/(2c^2+2ct/3N^{1/2})].$$

Let $M \ge 1$ be a uniform bound for $\Sigma_{i\in C}W_{in}(f_C(i|\theta,X_{Cn})-Ef_C(i|\theta,X_{Cn}))$ and for its derivative with respect to θ. Note that $\sum_{i=1}^{\infty} i2^{-i-3} = 1/4$. Then, for any

$C > 48M+8kM \text{ Ln } 2$,

(28) $$P\{\sup_\Theta|\zeta(\theta)| > C\}$$

$$\le P\{|\zeta(\theta_1)|>C/2\} + \sum_{i=1}^{\infty} P\{\sup_\Theta|\zeta(\theta_{i+1}(\theta)-\zeta(\theta_i(\theta))|>i2^{-i-3}C\}$$

(29) $$\le P\{|\zeta(\theta_1)|>C/2\} + \sum_{i=1}^{\infty} 2^{ki}\sup_\Theta P\{|\zeta(\theta_{i+1}(\theta)-\zeta(\theta_i(\theta))|>i2^{-i-3}C\}$$

(30) $$\le 2\exp[-C^2/4(2M^2+MC/3)]$$

$$+ \sum_{i=1}^{\infty} 2^{ki}\ 2\exp[-C^2i^24^{-i-3}/(2M^24^{-i}+2M2^{-i}Ci2^{-i-3}/3)]$$

(31) $$\le 2\exp[-C/4M] + \sum_{i=1}^{\infty} 2\exp[-iC/8M] \le 5\exp[-C/8M].$$

The inequalities (28) and (29) hold since left-hand-side events are contained in the union of the right-hand-side events, while (30) follows by application of the Bernstein inequality, and (31) by use of the bound on C and manipulation of the exponential terms. Given $\varepsilon > 0$, C can then be chosen sufficiently large to make the right-hand-side of (31) less than ε. This proves (14). □

The preceding arguments also apply to the classical method of moments estimator by setting $f(\theta) \equiv P(\theta)$ and $\zeta(\theta) \equiv 0$. Then, the asymptotic covariance matrix of the estimator is $\Sigma_{mm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{mm}\bar{R}(\bar{R}'\bar{R})^{-1}$, with

$$(32) \qquad G_{mm} = \lim N^{-1} \sum_{n=1}^{N} \left\{ \sum_{i\in C} P_C(i|\theta^*, X_{Cn})W_{in}'W_{in} - W_{Cn}'W_{Cn} \right\},$$

$$W_{Cn} = \sum_{i\in C} P_C(i|\theta^*, X_{Cn})W_{in}.$$

Define

$$(33) \qquad G_{ss} = \lim \sum_{n=1}^{N} \sum_{i,j\in C} W_{in}'W_{jn}E(\zeta_{in}\zeta_{jn}),$$

where $\zeta = \zeta(\theta^*)$. Then $G_{sm} = G_{mm} + G_{ss}$, and G_{ss} is the contribution of the simulation to the asymptotic variance. If $f(\theta)$ is the frequency simulator obtained by r independent Monte Carlo draws for each observation, then $G_{ss} = r^{-1}G_{mm}$ and $\Sigma_{sm} = (1+r^{-1})\Sigma_{mm}$. In this case, one draw per observation gives fifty percent of the asymptotic efficiency of the corresponding classical method of moments estimator, and nine draws per observation gives ninety percent relative efficiency. Use of Monte Carlo variance reduction techniques such as antithetic variates, or use of smooth simulators, may improve further the relative efficiency of MSM.

4. COMPUTATIONAL ISSUES AND STATISTICAL EFFICIENCY

Practical use of the MSM estimator requires that easily calculated, moderately efficient instruments W be available, that the Monte Carlo simulation of the probabilities and their derivatives be economical, that iterative algorithms to compute the estimators be fairly stable and efficient, and that estimators for the asymptotic covariance matrix of the estimators be computable.

Choice of Instruments

Consider first the question of suitable instruments. The classical method of moments estimator is asymptotically efficient if and only if, except for stochastically negligible terms, W is proportional to $\partial \mathrm{LnP}(\theta^{*})/\partial\theta$. The asymptotic efficiency of MSM relative to the classical moments estimator approaches one if the expected response in (11) is simulated consistently. The issue is how to construct W to obtain good asymptotic efficiency for MSM without excessive computation.

For the MNP model, the integral (12) defining the response probability can be differentiated with respect to β and Γ, yielding

$$(34)\qquad \partial P_C(i|\theta,X)/\partial\beta = X(X'\Omega X)^{-1}\int_{u\le 0}(u-\beta X)n(u-\beta X,X'\Omega X)du$$

$$\equiv X(X'\Omega X)^{-1}\int_{u\le 0}(u-\beta X)h(u,X,\theta)\gamma(u)du,$$

$$(35)\qquad \partial P_C(i|\theta,X)/\partial\Gamma$$

$$= \Gamma X(X'\Omega X)^{-1}\left\{\int_{u\le 0}[(u-\beta X)'(u-\beta X)-X'\Omega X]\cdot n(u-\beta X,X'\Omega X)du\right\}(X'\Omega X)^{-1}X',$$

$$\equiv \Gamma X(X'\Omega X)^{-1}\left\{\int_{u\le 0}[(u-\beta X)'(u-\beta X)-X'\Omega X]\cdot h(u,X,\theta)\gamma(u)du\right\}(X'\Omega X)^{-1}X',$$

where $\Omega = \Gamma'\Gamma$, $X = X_{C-i}$, and as before $h(u,X,\theta)$ is the ratio of the multivariate normal density to a Monte Carlo sampling distribution $\gamma(u)$ on the nonpositive orthant.[8] Applying the chain rule to $\beta(\theta)$ and $\Gamma(\theta)$ then yields derivatives of the response probabilities with respect to the deep parameters θ. For discrete response models other than MNP, analogues of (34) and (35), involving derivatives of the density g_U with respect to θ, can be defined, and smooth Monte Carlo simulators constructed.

Economical Simulators

Simulation of (34) and (35) based on Monte Carlo draws from the density $\gamma(u)$ yields smooth unbiased estimates of the derivatives. Since the smooth simulator (12) of $P_C(i|\theta)$ is positive, the ratios of simulators of (34) and (35) to the simulator of (12) provides an approximation to the ideal instruments $\partial \text{Ln } P_C/\partial\theta$. The number of draws per observation must go to infinity with sample size if the ideal instruments are to be estimated consistently, permitting MSM to be asymptotically efficient. However, moderately efficient instruments can be obtained with relatively few draws. It is essential for the asymptotic statistics of the MSM estimator that the simulation of (12) and the derivatives used to construct instruments be independent of the draws used to simulate the expected response in (11); however, use of common draws from $\gamma(u)$ to simulate P_C, $\partial P_C/\partial\beta$, and $\partial P_C/\partial\Gamma$ at observation n may improve the approximation of the ideal instruments.

The frequency simulator of $P_C(i|\theta,X_C)$ is economical to compute, as are the smooth simulators (12), (34), and (35) when the Monte Carlo density $\gamma(u)$ permits easy draws. For MNP, a practical choice is γ independent exponential, allowing u to be drawn as a vector of logarithms of uniform random deviates. However, more accurate smooth simulators may be obtained with suitable transformations.

First consider kernel-smoothed frequency simulators that satisfy summing-up. The construction of these simulators starts from a perturbation of the latent variable model (2),

(36) $\tilde{u}_C = u_C + \nu_C b_N$,

where ν_C is a vector whose components are independently distributed with a distribution function Ψ and b_N is a simulation parameter. Assume Ψ has a finite moment generating function $\mu(t)$ for t in a neighborhood of zero. Choose b_N so that for some $\varepsilon > 0$, $N^{\varepsilon+1/2}b_N \to 0$. Associated with (36) is a response probability, obtained by first conditioning on u_{C-i},

(37) $\tilde{P}_C(i|\theta,X_C) = \int \mathbb{K}(u_{C-i}/b_N)G(u_{C-i}|\theta,X_C)du_{C-i} = \int \mathbb{K}(a(\theta,\eta)X_{C-i}/b_N)g(\eta)d\eta$

with $\mathbb{K}(y_1,\ldots,y_{m-1}) = \int \left(\prod_{j=1}^{m-1} \Psi(v-y_j)\right) \Psi(dv)$. As b_N approaches zero, $\mathbb{K}(u_{C-i}/b_N)$ approaches the indicator function $1(u_{C-i}\le 0)$, and $\tilde{P}_C(i|\theta,X_C)$ approaches $P_C(i|\theta,X_C)$ defined by (5). The kernel-smoothed frequency simulator is an average, over Monte Carlo draws from $g(\eta)$, of $\mathbb{K}(a(\theta,\eta)X_{C-i}/b_N)$. This simulator is nonnegative. If the simulators for all $i \in C$ are constructed from common Monte Carlo draws, then they satisfy summing-up. They are strictly positive if the support of Ψ is the real line. Choosing Ψ to be type I extreme value distributed yields a multinomial logit form $\mathbb{K}(v_1,\ldots,v_{m-1}) = 1/(1 + \sum_{j=1}^{m-1} \exp(-v_j))$, and (37) is a multivariate normal mixture of logits.[9] A polynomial kernel such as $\Psi(v) = [6+5v+(2-|v|)v^3]/12$ for $|v| \le 1$ is computationally economical, and for small b_N yields a smoothed simulator that for most draws coincides with the unsmoothed frequency simulator.[10] For MNP, a variant of the kernel-smoothed frequency simulator is unbiased: Write (2) in the form $u_C = \bar{u}_C + \nu b_N$, with ν a standard normal vector and $\bar{u}_C \sim N(\beta X_C, A'A)$, with A upper triangular and $A'A \equiv X_C'\Gamma'\Gamma X_C - b_N I$; this can be done provided b_N is small enough so $X_C'\Gamma'\Gamma X_C - b_N I$ is positive definite. As in (37), conditioning on $\bar{u}_C$,

$$(38)\qquad P_C(i|\theta,X_C) = \int \mathbb{K}((\beta X_{C-i}+\eta A_{C-i})/b_N)g(\eta)d\eta,$$

$$\mathbb{K}(y_1,\ldots,y_{m-1}) = \int \Big(\prod_{j=1}^{m-1} \Phi(v+y_j)\Big) \Phi'(v)dv,$$

with g the multivariate standard normal density, Φ the standard normal distribution, and A_{C-i} the array with columns A_j-A_i for $j \neq i$, where A_j is a column of A. An average of $\mathbb{K}$ in (38) over Monte Carlo draws from g yields an unbiased positive smooth frequency simulator. Adding-up holds if common draws are used for $i \in C$.[11]

For MNP, construction of economical simulators is aided by the use of spherical transformations. Each of the expressions (12), (34), and (35) involves simulation of integrals of the generic form

$$(39)\qquad Q = \int_{u\geq 0} \Big(\prod_{j=1}^{m} u_j^{k_j}\Big)n(u+\mu,\Lambda)du,$$

where $\sum_{j=1}^{m} k_j$ is 0, 1, or 2, $\mu = \beta X_{C-i}$, and $\Lambda = X_{C-i}\Omega X'_{C-i}$. Make the transformation $r = \left[\sum_{j=1}^{m} u_j^2\right]^{1/2}$ and $s_j = u_j/r$. Define

$$(40)\qquad C(n,a,b) = \int_0^\infty r^n\, e^{-(r-b/a)^2 a/2}dr;$$

this is proportional to a parabolic cylinder function (Spanier and Oldham, 1987), and satisfies the recursion

$$(41)\qquad C(0,a,b) = (2\pi/a)^{1/2}\Phi(b/a^{1/2}),$$

$$C(1,a,b) = C(0,a,b)b/a + e^{b^2/2a}/a,$$

$$C(n,a,b) = C(n-1,a,b)b/a + C(n-2,a,b)(n-1)/a \qquad (n \geq 2).$$

Then,

(42) $Q = c_o c_1 E_s C(\sum_{j=1}^{m} k_j + m-1, a, b)(\prod_{j=1}^{m} s_j^{k_j})$,

where s is distributed uniformly on the intersection of the unit sphere and the nonnegative orthant, and

$$a = s'(X'_{C-i}\Omega X_{C-i})^{-1}s,$$

$$b = -\beta X_{C-i}(X'_{C-i}\Omega X_{C-i})^{-1}s,$$

$$c_o = (2\pi)^{1/2} 2^{-3m/2} |\Omega|^{-1/2} \Gamma(m/2)^{-1},$$

$$c_1 = \exp(-[\beta X(X'\Omega X)^{-1}X'\beta' - (\beta X(X'\Omega X)^{-1}s)/s'(X'\Omega X)^{-1}s]/2),$$

with $X = X_{C-i}$, and c_o independent of X and s. To generate uniform draws from the distribution of s, draw a standard normal random vector u, and take

(43) $s_j = |u_j| / \left[\sum_{j=1}^{m} u_j^2 \right]^{1/2}$.

Then, (43) is simulated by drawing one or more s, and for each s using the recursion (41) to calculate *C*. A further refinement is to use control variates for *C*.[12]

The spherical transformation can also be used to calculate an economical unbiased smooth frequency simulator for MNP. Let s be a uniform draw from the unit sphere in $\mathbb{R}^K$, and let λ^2 be a random variable with a Chi-squared distribution with K degrees of freedom, denoted $H_K(\lambda^2)$. Then, the latent variable model for MNP can be written $u_C = (\beta+\lambda s\Gamma)X_C$. Given s, an easy computation yields a partition of $[0,+\infty]$ into intervals $[\lambda_j,\lambda_{j+1}]$, $j = 0,\ldots,m$, on which each of the components of u_C is maximum. (Some of the intervals may be degenerate.) The probability of response i, given s, is $P_C(i|\theta,X_C,s) = H_K(\lambda_{j+1}^2) - H_K(\lambda_j^2)$, where j is the ascending rank of $s\Gamma x_i$ in the vector $s\Gamma X_C$. The λ_j are smooth in θ for almost all X_C, so $P_C(i|\theta,X_C,s)$ is also smooth. The simulator is an average of the $P_C(i|\theta,X_C,s)$ for r random draws of s.

The accuracy of simulators for MNP that are based on spherical transformations can be improved substantially by use of antithetic variates. Deák (1980) gives an effective procedure: For uniform draws from the sphere in $\mathbb{R}^K$, first draw a random basis $s^1,\ldots,s^K$. This can be done for example by drawing K standard normal vectors and applying a Gram-Schmidt orthonormalization. Then use the 2K vectors $\pm s^j$, or the 2K(K-1) vectors $(\pm s^i \pm s^j)$ for $i < j$, as directions for the simulation.[13]

Iterative Estimation methods

A practical estimation procedure is first to use relatively crude instruments, defined independently of θ, to iterate to an initial consistent estimator $\hat{\theta}$, second to simulate the ideal instruments using (12), (34), and (35) at $\hat{\theta}$, and third to carry out one or more iterations using the approximately ideal instruments. Good candidates for crude instruments are low-order polynomials in the explanatory variables; e.g., X_{C-i} for $\partial LnP_C(i|\theta,X_C)/\partial\beta$ and $X_{C-i}X'_{C-i}$ for $\partial LnP_C(i|\theta,X_C)/\partial\Gamma$.[14]

Consider iterative algorithms for calculation of MSM estimators. When smooth simulators are used for f(θ) in (12), and the instruments W are defined independently of θ, then estimates can be computed by Newton-Raphson iteration or a similar second-order method applied to minimize the criterion

(44) $(d-f(\theta))'W'W(d-f(\theta))$.

This criterion may be irregular; in particular, kernel-smoothed frequency simulators may have local flats. Then, optimization methods that use non-local information, such as simulated annealing, may be more reliable; see Press *et al* (1986).

When a frequency simulator is used, (44) is piecewise constant in θ, and non-local methods must be used in iteration. I have tried random search algorithms and pseudo-gradient methods that adaptively approximate slopes using long baselines; the former have performed better. For discrete response models that can be parameterized in terms of mean and variance, such as MNP, convergence can be accelerated using a method due to Manski: Suppose r simulations per observation, and that starting from a trial θ_o, a search direction $\Delta\theta$ has been determined. Consider (44) as a function of $\theta_o+\lambda\Delta\theta$, with λ a step size to be determined. The value λ_{nj} at which there is a jump in (44) from draw j, observation n, is easily calculated. Then, it is practical to enumerate the values of (44) at all the jump points λ_{nj} and choose a global minimum along this search direction.

Generally, iteration using smooth simulators is faster that that using frequency simulators. However, in applications where the number of alternatives is very large, the burden of computing $f(\theta)$ or approximations to the optimal instruments for all alternatives may be excessive. Then, a frequency simulator $f(\theta)$ with r repetitions will be non-zero for at most r alternatives, and the instruments need be computed only for these alternatives plus the observed one. For example, a single Monte Carlo draw for each observation requires calculation of the instruments only for the observed and drawn alternatives, and yields fifty percent of the efficiency of the classical method of moments estimator, no matter how large the set of possible alternatives. Comparable reductions in computation can be achieved using a kernel-smoothed frequency simulator with a kernel of bounded support.

Asymptotic Covariance Matrix

Consider estimation of the asymptotic covariance matrix of the MSM estimator, $\Sigma_{sm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{sm}\bar{R}(\bar{R}'\bar{R})^{-1}$. A consistent estimator of G_{sm} is

$$(45) \quad \hat{G}_{sm} = N^{-1}\sum_{n=1}^{N}\sum_{i,j\in C} W_{in}(d_{in}-f(i|\theta_{sm},X_{Cn}))(d_{jn}-f(i|\theta_{sm},X_{Cn}))W'_{jn}.$$

The matrix $\bar{R} = \lim N^{-1}WP_{\theta}(\theta^{*})$ is consistently estimated by

$$(46) \quad \hat{R} = N^{-1}W\hat{P}_{\theta},$$

where $\hat{P}_{\theta}$ is an unbiased simulator of the array P_{θ}, evaluated at θ_{sm} or any initially consistent estimator $\hat{\theta}$ of θ^{*}, obtained using one or more draws per observation in (34) and (35) or their analogues in models other than MNP, independent of any simulation used to compute W.

To show (45), note first that this expression with θ^{*} in place of θ_{sm} converges to G_{sm} by a law of large numbers; see part (a) of the proof of Theorem 1. Second, by (13) and (15), terms involving the difference of $f(\theta^{*})$ and $f(\theta_{sm})$ are $o_p(1)$. The argument for (46) is the same, but it is necessary to use versions of (13) and (15) for $\hat{P}_{\theta}$. These hold for smooth simulators using the argument of Proposition 1.

5. DISCRETE PANEL DATA WITH AUTOREGRESSIVE ERRORS

Consider longitudinal discrete response data (d_{tn},x_{tn}) for subjects n = 1,...,N observed over t = 1,...,T periods, where $d_{tn} = \pm 1$ indicates a binary response and x_{tn} is a vector of explanatory variables. This problem has 2^T alternative response patterns, large for long panels. A latent variable model that may be appropriate for such data is

(47) $u_{tn} = \beta x_{tn} + \varepsilon_{tn}$,

$d_{tn} = \text{sign}(u_{tn})$,

with $\varepsilon_{tn} = \xi_n + \nu_{tn}$, ξ_n a normal subject-specific disturbance, ν_{tn} a normal first-order autoregressive disturbance, and ξ, ν independent of each other and independent across subjects. If ε_{tn} is stationary, with variance normalized to one, then

$$(48) \qquad \varepsilon_{tn} = (1-\lambda^2)^{1/2}\eta_{0n} + \lambda\left[(1-\rho^2)\sum_{j=0}^{t-2}\rho^j\eta_{t-j,n} + \rho^{t-1}\eta_{1n}\right],$$

with λ^2 the proportion of the variance in the autoregressive error, ρ the serial correlation, and η_{jn} independent standard normal variates. The probability $P(d_n|x_n,\beta,\lambda,\rho)$ of $d_n = (d_{1n},\ldots,d_{Tn})$ given $x_n = (x_{1n},\ldots,x_{Tn})$ equals the probability of a (T+1) dimensional draw $(\eta_{0n},\ldots,\eta_{Tn})$ such that $d_{tn}u_{tn} > 0$ for $t = 1,\ldots,T$.

Full maximum likelihood estimation of this model requires T-dimensional numerical integration to evaluate $P(d_n|x_n,\beta,\lambda,\rho)$, which is computationally impractical for $T > 4$. Ruud (1981) has developed practical consistent estimators using partial likelihoods for small numbers of adjacent periods; see also Chamberlain (1984). The MSM method, starting from initially consistent estimators, offers a computationally practical way to increase efficiency. A frequency simulator or a kernel-smoothed frequency simulator with a finite-support kernel, can be computed using (41), and with a moderate number of repetitions requires simulation of the instruments for a practical number of alternatives per subject, even for large T. Alternately, it may be possible to compute directly an unbiased simulator of the score $\partial \text{Ln}\ P(d_n|\theta,x_n)/\partial\theta$ for the observed response pattern d_n. From the analogues of (34) and (35) for the discrete panel data problem, this requires that one obtain asymptotically unbiased or consistent simulators for the conditional expectation of first and second order polynomials given draws from the nonpositive orthant. Unbiased simulators can be obtained by use of acceptance/rejection methods, or asymptotically unbiased simulators by allowing the number of repetitions in the simulation of the probability in the denominator of $P^{-1}\partial P/\partial\theta$ to go to infinity with sample size. These approaches have been investigated by Ruud and McFadden (1987) and Hajivassiliou and McFadden (1987).

MSM estimation of discrete panel data models extends readily to more general time-series covariance structures, so long as it is practical to Cholesky-factor and invert the covariance matrix to obtain a representation analogous to (48) for the ε_{tn} in terms of independent normal variates, and so long as it is practical to construct instrument arrays for the deep parameters of the problem. The estimator can also be applied to models with general state dependence, provided the initial value problem (Heckman, 1981) can be handled. For example, consider the model

(49) $$u_{tn} = \beta x_{tn} + \psi d_{t-1,n} + \xi_n + \nu_{tn},$$
$$d_{tn} = \mathrm{sign}(u_{tn}),$$

with ξ_n a subject-specific disturbance and ν_{tn} independent across t. If the disturbances are normal, and ν_{tn} has unit variance, then

(50) $$P(d_n|x_n,d_{0n},\beta,\xi_n) = \prod_{t=1}^{T} \Phi(d_{tn}(\beta x_{tn} + \psi d_{t-1,n} + \xi_n)).$$

Suppose the conditional distribution of ξ_n given x_n and d_{0n} can be assumed to depend only on d_{0n}; this is justified if x_n is independent of the past history of the x process. Suppose the inverse distribution of ξ_n given d_{0n} is given a flexible parametric form that spans the true inverse distribution. Then the response probabilities are given by the expectation of (50) with respect to ξ, which can be simulated economically from the inverse distribution. Adding serial correlation to the disturbances ν_{tn} in (49) makes (50) a T-dimensional integral, whose simulation by MSM can be handled jointly with simulation of the expectation with respect to ξ.

6. DISCRETE RESPONSE MODELS WITH MEASUREMENT ERROR

Suppose discrete response for a random sample n = 1,...,N satisfies a latent variable model

(51) $$u_n = \beta z_n + \varepsilon_n,$$
$$d_n = H(u_n),$$

where H maps the row vector u_n into m discrete categories with d_n an indicator for the observed category, and $H^{-1}(d_n)$ the set of u_n yielding the observed category. To simplify notation, assume β is a scalar; generalization merely requires that the construction below be carried out component by component. Suppose z_n is not observed directly, but is related to a vector of observations x_n by

(52) $$x_n = z_n\Lambda + \xi_n,$$

where $\xi_n \sim N(0,\Psi)$, independent of ε. We interpret the x as observations on z measured with error, or as indicators for z. In form, this is a multiple indicator or factor-analytic latent variable model, with Λ giving the factor loadings.

Suppose in the population $z_n \sim N(\mu,\Gamma)$, independent of ξ_n. Then the conditional distribution of z given x, suppressing subscripts, is

(53) $$z \sim N(\mu + (x-\mu\Lambda)(\Lambda'\Gamma\Lambda+\Psi)^{-1}\Lambda'\Gamma, \Gamma-\Gamma\Lambda(\Lambda'\Gamma\Lambda+\Psi)^{-1}\Lambda'\Gamma).$$

If the $\varepsilon \sim N(0,\Omega)$ in (51), then

(54) $$u \sim N(\beta\mu+\beta(x-\mu\Lambda)(\Lambda\Gamma\Lambda'+\Psi)^{-1}\Lambda'\Gamma, \beta^2[\Omega+\Gamma\Lambda(\Lambda'\Gamma\Lambda+\Psi)^{-1}\Lambda'\Gamma]).$$

and the response probabilities given x are of MNP form. The MSM estimator for the general MNP model can be adapted directly to this problem, the main practical difficulty being calculation of the derivatives of the Cholesky factor of the covariance matrix with respect to the deep parameters in order to calculate a relatively efficient instrument matrix.

A number of variants of the measurement error model (51) may be encountered in applications, including variables measured with error that are common to several alternatives or interact with alternative-specific dummys, multiple variables measured with possibly correlated errors, and simultaneity between the latent variables and observed indicators. These may alter the details of (52) and (53), but give the same basic structure for the response probabilities and MSM estimator. It is also possible to treat measurement error in discrete response models such as multinomial logit by allowing the ε to have an appropriate distribution. For the logit example, MSM estimation can be used by simulating the expectations of the logit formulas with respect to the conditional distribution of the true explanatory variables. These topics are studied in greater detail in McFadden (1986a,1986b) and Train, McFadden, and Goett (1987).

7. NON-NORMAL DISCRETE RESPONSE MODELS

This paper has focused on estimation of the MNP model. However, the MSM estimator can be applied to any latent variable model in which unbiased estimates of the response probabilities can be obtained economically by Monte

Carlo methods. For example, in the latent variable model (1), it may be reasonable to assume that some components of α are always non-negative, giving monotonicity. This could be modeled by taking the density of α to be multivariate truncated normal, or by giving some components non-negative densities such as gamma. This complicates the analytic representation of response probabilities, but is readily accommodated in Monte Carlo draws from the latent variable model to obtain frequency estimators.

The MSM estimator also permits analysis of discrete response data generated by more complex partial observation functions than the maximum indicator appearing in (1). For example, consider data on ranks of alternatives. With the exception of the multinomial logit model, it is impossible to obtain analytically tractable expressions for probabilities of more than the first few ranks in terms of response probabilities; see Falmange (1978), Barbara and Pattanaik (1985), and McFadden (1986a). However, Monte Carlo drawings from the latent variable model provides unbiased frequency estimators of the ranking probabilities that can be used in MSM estimation.

8. SUMMARY

This paper has proposed a simple modification of a classical method of moments estimator for discrete response models that avoids the necessity for accurate numerical integration to calculate response probabilities, using instead asymptotically unbiased simulators of these probabilities. This method of simulated moments is practical for problems where direct numerical integration is computationally intractable.

9. APPENDIX: THEOREMS AND PROOFS

I use the following notation, mostly collected from Sections 2 and 3 of the paper:

$C = \{1,...,m\}$	the set of possible responses.
$u_i = \alpha x_i$ or $u_C = \alpha X_C$	a latent variable model, $i \in C$, with $X_C = (x_1,...,x_m)$ a $K \times m$ array, $u_C = (u_1,...,u_m)$, $\alpha = a(\theta,\eta)$ a $1 \times K$ vector function, with θ a $k \times 1$ parameter vector with true value θ^*, and η a random vector with density $g(\eta)$ and associated measure g, independent of X_C.
$d_i = 1(u_i \geq u_C)$	indicator for maximum u_i (= observed response).
$P_C(i\|\theta,X_C)$	probability of η such that $a(\theta,\eta)X_C$ is maximized at i, given X_C.
$f_C(i\|\theta,X_C)$	a simulator for $P_C(i\|\theta,X_C)$.
W_i	$k \times 1$ instrument vector, determined as a function $W_i = w_i(\theta,X_C)$.
$n = 1,...,N$	a random sample.
d, $P(\theta)$, $f(\theta)$	$mN \times 1$ vectors formed by stacking d_i, $P_C(i\|\theta,X_C)$, or $f_C(i\|\theta,X_C)$ by alternative, then by observation.
$W = W(\theta)$	$k \times mN$ array formed by stacking W_{in}.
$P_\theta(\theta)$	$mN \times k$ array of derivatives of $P(\theta)$.
θ_{sm}	any vector satisfying $\\|W(\theta_o)(d-f(\theta_{sm}))\\| \leq \inf_\theta \\|W(\theta_o)(d-f(\theta))\\| + O_p(1)$, some θ_o.
$p(X_C)$	density for X_C, with associated measure p.
$R_N(\theta)$, $R(\theta)$, $\bar{R}$	$k \times k$ array of sample covariances, $R_N(\theta) = N^{-1}WP_\theta(\theta)$, $R(\theta) = \lim R_N(\theta)$, $\bar{R} = R(\theta^*)$.[15]

The response probabilities are invariant under monotone transformations of the latent variable model. Hence, without loss of generality, we may normalize $x_1 \equiv 0$, so X_C is contained in a $K(m-1)$ dimensional space. Further, α may be defined without loss of generality to have a compact domain.[16]

The first assumptions made require X_C and θ to have regular domains, and guarantee a zero probability that the latent variables for different alternatives are equal, so the response probabilities are well-defined without additional tie-breaking rules:

[A1] The parameter space Θ is a compact convex subset of $\mathbb{R}^k$, and θ^* is in the interior of Θ.

[A2] The domain $\mathcal{X}$ of the attributes X_C is a compact subset of a K(m-1) dimensional space.

[A3] The random vector η is finite-dimensional with domain $\mathbb{N}$, is independent of X_C, and has a finite mean. The function $\alpha = a(\theta,\eta)$ is continuous on $\Theta \times \mathbb{N}$, and is twice differentiable in θ with these derivatives continuous on $\Theta \times \mathbb{N}$.

[A4] For an open set $\mathcal{X}_o \subseteq \mathcal{X}$ with $p(\mathcal{X}_o) = 1$, the subset $\mathbb{N}(\theta, X_C)$ of $\mathbb{N}$ such that $a(\theta,\eta)X_C$ is distinct in every component has probability one for each $\theta \in \Theta$.

The last assumption is usually imposed in the definition of discrete response models, and can be derived from more basic structural conditions. The following lemma covers common applications, including MNP. When the model contains alternative-specific random effects, the array A_{22} in this result is a (m-1) dimensional identity matrix.

Lemma 1. Suppose [A1] and [A2]. Suppose there is a partition $X_C = \begin{bmatrix} 0 & A_{12} \\ 0 & A_{22} \end{bmatrix}$ such that A_{22} is $(m-1) \times (m-1)$ and almost surely nonsingular. Suppose $\alpha = a(\theta,\eta) \equiv \beta(\theta) + \eta\Gamma(\theta)$ is twice continuously differentiable in θ. Suppose α is partitioned commensurately with X_C, so

$$(55) \qquad [\alpha^1 \;\; \alpha^2] = [\beta^1(\theta) \;\; \beta^2(\theta)] + [\eta^1 \;\; \eta^2]\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \Gamma_{22} \end{bmatrix}.$$

Suppose Γ_{22} is nonsingular for $\theta \in \Theta$. Suppose the density of η^2 conditioned on η^1 is uniformly bounded and continuous, with support $\mathbb{R}^{m-1}$, and suppose η has a finite mean. Then [A3] and [A4] hold.

Proof: $\alpha X_C = \beta X_C + [0, \eta^1(\Gamma_{11}A_{12}+\Gamma_{12}A_{22}) + \eta^2\Gamma_{22}A_{22}]$. With probability one, the term $\eta^2\Gamma_{22}A_{22}$ has a continuous density with support $\mathbb{R}^{m-1}$, given η^1, implying that the probability of a hyperplane where components of αX_C are tied is zero. □

The next assumption guarantees that the response probabilities are well-behaved:

[A5] The probability $P_C(i|\theta,X_C)$ is positive, and twice differentiable in θ, and the probability and its derivatives are continuous, on $\Theta \times \mathcal{X}$.

The following result gives a sufficient condition for [A5] which holds in particular for the MNP model with alternative-specific dummys:

Lemma 2. Suppose the hypotheses of Lemma 1, with A_{22} always nonsingular. Then [A5] holds.

Proof: By the symmetry of the problem in alternatives, it is sufficient to consider $P_C(1|\theta,X_C) = \Pr(\alpha X_C \le 0|X_C)$. Using the decomposition of Lemma 1, $\alpha X_C = [0,s] \le 0$ implies

$$(56) \qquad \eta^2 = [s - \beta^1 A_{12} - \beta^2 A_{22} - \eta^1(\Gamma_{11}A_{12}+\Gamma_{12}A_{22})](\Gamma_{22}A_{22})^{-1}.$$

Then, given η^1, the set $B(\eta^1)$ of η^2 satisfying $\alpha X_C \le 0$ is the intersection of m-1 linearly independent half-spaces, with each bounding hyperplane twice continuously differentiable in (θ,X_C). Hence, $P_C(1|\theta,X_C) = \int_{\eta^1}\int_{B(\eta^1)} g_{2.1}(\eta^2|\eta^1)d\eta^2\, g_1(\eta^1)d\eta^1$ is twice continuously differentiable in (θ,X_C). Since $g_{2.1}$ has support $\mathbb{R}^{m-1}$, the probability is positive. □

The next assumptions concern the instruments and identification of θ^*:

[A6] <u>The function</u> $W_i = w_i(\theta,X_C)$ <u>determining the instruments is continuous, bounded, and twice continuously differentiable on</u> $\Theta \times \mathbb{X}$. (<u>Let</u> M_w <u>denote a bound on</u> w_i <u>and its</u> θ <u>derivative, uniform in</u> i, θ, X_C.)

[A7] <u>The instruments identify</u> θ^*, <u>with</u>

$$(57) \qquad \omega(\theta,\tilde\theta) \equiv \int_{\mathbb{X}} \sum_{i\in C} w_i(\tilde\theta,X_C)[P_C(i|\theta,X_C)-P_C(i|\theta^*,X_C)]dp(X_C)$$

<u>equal to zero if and only if</u> $\theta = \theta^*$, <u>for any</u> $\tilde\theta \in \Theta$.[17]

[A8] $\bar{R}$ <u>is of maximum rank</u>.

To satisfy [A6], instruments constructed by simulation require the use of smooth simulators such as (12), (34), and (35) in the case of MNP. If [A5] holds and the instruments equal the score of the likelihood evaluated at each trial θ, $w_i(\theta,X_C) \equiv \partial \mathrm{Ln}\, P_C(i|\theta,X_C)/\partial\theta$, then $\tilde\theta = \theta$ and ω reduces to

$$\omega(\theta,\tilde{\theta}) = \int_{\mathcal{X}} \sum_{i\in C} P_C(i|\theta^*,X_C)[\partial LnP_C(i|\theta,X_C)/\partial\theta]dp(X_C),$$

the expected score of an observation under maximum likelihood estimation. For this case, [A7] requires that θ^* be the only critical point of the expected log likelihood, a standard identification condition. Also, in this case, $\bar{R}$ equals the information matrix evaluated at θ^*, which is symmetric and nonnegative definite, and by [A7] is definite at some point in every neighborhood of θ^*. Then, [A8] adds only a regularity requirement. In the case of more general instruments, [A7] and [A8] are standard assumptions for the identification and regularity of classical method of moments estimators. Hence, the identification conditions for MSM are the same as for the corresponding classical method of moments estimator.

The next assumption concerns the simulator $f_C(i|\theta,X_C)$:

[A9] Vectors $(\eta_{1n},\ldots,\eta_{rn})$ are drawn, by simple random sampling or otherwise, independently of W and d, and independently for different n, so that each η has marginal density $g(\eta)$. Define $\varphi_C(i|\theta,X_{Cn})$ to be the frequency in the r draws for observation n of the event that $a(\theta,\eta_{jn})X_{Cn}$ is maximized at component i. Define $f_C(i|\theta,X_{Cn})$ to be any uniformly bounded function of θ, X_{Cn} and $(\eta_{1n},\ldots,\eta_{rn})$ satisfying

(58) $$E\{f_C(i|\theta,X_{Cn})|W,d\} = P_C(i|\theta,X_{Cn}) + O(N^{-\varepsilon-1/2})$$

for some $\varepsilon > 0$; and for some M_φ, M_f, $\lambda > 0$, and all $\theta,\tilde{\theta} \in \Theta$ and $X_{Cn} \in \mathcal{X}$,

(59) $$|f_C(i|\theta,X_{Cn})-f_C(i|\tilde{\theta},X_{Cn})| \le M_\varphi|\varphi_C(i|\theta,X_{Cn})-\varphi_C(i|\tilde{\theta},X_{Cn})| + M_f|\theta-\tilde{\theta}|^\lambda.$$

Condition (58) requires that the simulator be asymptotically unbiased, while (59) requires that it be at least as smooth in θ as the frequency simulator. Condition (59) is satisfied trivially by either the frequency simulator, or by a smooth simulator such as that based on (12). If the simulator is differentiable, then $\lambda = 1$; the assumption also allows $0 < \lambda < 1$, corresponding to "polynomial" non-differentiability. A simulator satisfying (59) will be termed λ-Lipschitz in neighborhoods where φ_C is constant. The following lemma establishes sufficient conditions for a kernel-smoothed frequency simulator to satisfy (58).

Lemma 3. Suppose the assumptions of Lemma 1 hold, with A_{22} always nonsingular. Suppose a kernel-smoothed frequency simulator (37) with a distribution function Ψ having a finite moment generating function in a neighborhood of the origin, and with $N^{\varepsilon+1/2}b_N \to 0$. Then, (58) holds.

Proof: Let $\mu(t)$ be the moment generating function for Ψ. There exists $\tau > 0$ such that $\Psi(v) \le e^{\tau v}\mu(-\tau)$ for all $v < 0$ and $1-\Psi(v) \le e^{-\tau v}\mu(\tau)$ for all $v > 0$. Then, for $c \equiv \max_j y_j/2 > 0$, $\mathbb{K}(y_1,\ldots,y_{m-1}) = \int_{v\le c} (\prod_{j=1}^{m-1} \Psi(v-y_j))\, \Psi(dv) + \int_{v>c} (\prod_{j=1}^{m-1} \Psi(v-y_j))\, \Psi(dv) \le e^{-\tau c}\mu(-\tau) + e^{-\tau c}\mu(\tau)$, with the first term the result of bounding the product at negative arguments, and the second the result of bounding the measure at positive arguments. A similar argument for $c < 0$ yields $\mathbb{K}(y_1,\ldots,y_{m-1}) \ge (1 - e^{-\tau|c|}\mu(\tau))^m \ge 1 - me^{-\tau|c|}\mu(\tau)$. Define $I(A) = \int_A |\mathbb{K}(u_{C-i}/b_N) - 1(u_{C-i}\le 0)|G(u_{C-i}|\theta,X_C)du_{C-i}$. Define A_1 to be the set of u_{C-i} less than $-Mb_N$ in every component, A_2 to be the set of u_{C-i} greater than Mb_N in at least one component, with M a positive constant, and $A_3 = \mathbb{R}^{m-1}-A_1-A_2$. Then, the bounds on $\mathbb{K}$ imply $I(A_1) \le e^{-\tau M}\mu(\tau)m$ and $I(A_2) \le e^{-\tau M}(\mu(\tau)+\mu(-\tau))$. Further, $I(A_3) \le \sum_{j\ne i} \text{Prob}(|u_j-u_i| \le Mb_N)$. But (56) in the proof of Lemma 2 holds when the second partition is of dimension one and s is the value of a single component u_j-u_i of u_{C-i}. Then, letting M_γ be a uniform bound on the conditional density of η^2 given η^1, $\text{Prob}(|u_j-u_i| \le Mb_N) \le 2Mb_NM_\gamma$. Therefore,

$I(A_3) \le 2mMb_N M_\gamma$. Then, $N^{1/2}|\tilde{P}_C(i|\theta,X_C) - P_C(i|\theta,X_C)| \le N^{1/2}(I(A_1) + I(A_2) + I(A_3)) \le N^{1/2}(e^{-\tau M}\mu(\tau)m + e^{-\tau M}(\mu(\tau)+\mu(-\tau)) + 2mMb_N M_\gamma)$. Choose $M = \tau^{-1}\text{Ln } N$. Then, the right-hand-side of the last inequality goes to zero if $N^{1/2}(\text{Ln } N)b_N \to 0$. The condition $N^{1/2+\varepsilon}b_N \to 0$ implies the required limit. □

The next result characterizes the regularity in θ of the simulated moments, and guarantees that with probability one, the condition defining θ_{sm} has a solution with $|W(d-f(\theta_{sm}))| \le mM_W M_\varphi/r$:

<u>Lemma</u> 4. <u>Suppose</u> [A1]-[A9]. <u>Then</u>, <u>almost</u> <u>surely</u>, $W(d-f(\theta))$ <u>is</u> <u>uniformly</u> λ-<u>Lipschitz</u> <u>in</u> θ <u>except</u> <u>for</u> <u>a</u> <u>closed</u> <u>subset</u> Θ_o <u>of</u> Θ <u>with</u> <u>Lebesgue</u> <u>measure</u> <u>zero</u>, <u>and</u> <u>the</u> <u>jumps</u> <u>in</u> <u>this</u> <u>function</u> <u>on</u> Θ_o <u>are</u> <u>bounded</u> <u>by</u> $mM_W M_\varphi/r$.

Proof: Define $I(\theta,X_C,\eta) = 0$ if the components of $a(\theta,\eta)X_C$ are all distinct, and $I(\theta,X_C,\eta) = 1$ otherwise. For each $\theta \in \Theta$, [A4] implies

(60) $$0 = \int_{\mathbb{N}} \int_{\mathcal{X}} I(\theta,X_C,\eta)dg(\eta)\, dp(X_C),$$

and hence

(61) $$0 = \int_{\Theta} \int_{\mathbb{N}} \int_{\mathcal{X}} I(\theta,X_C,\eta)dg(\eta)\, dp(X_C)\, d\theta.$$

Applying Fubini's theorem to (54), there exists a set $\mathcal{X}_1 \subseteq \mathcal{X}$ with probability measure one, for $X_C \in \mathcal{X}_1$ a set $\mathbb{N}(X_C) \subseteq \mathbb{N}$ with probability measure one, and for $(X_C,\eta) \in \mathcal{X}_1 \times \mathbb{N}(X_C)$ a set $\Theta_1(X_C,\eta) \subseteq \Theta$ of full Lebesgue measure on which $I(\theta,X_C,\eta) = 0$. The continuity of $a(\theta,\eta)$ in θ implies that if $I(\theta,X_C,\eta) = 0$, then this is also true in a neighborhood of θ, so $\Theta_1(X_C,\eta)$ is open.

The function $W(d-f(\theta))$ is defined by N independent draws X_{Cn} with density $p(X_C)$, and for each n, r draws $(\eta_{1n},\ldots,\eta_{rn})$, each with marginal density $g(\eta)$. Hence, with probability one, $X_{Cn} \in \mathcal{X}_1$ and $\eta_{jn} \in \mathbb{N}(X_{Cn})$ for

$j = 1,\ldots,r$ and $n = 1,\ldots,N$, implying $\Theta_N = \bigcap_{n=1}^{N} \bigcap_{j=1}^{r} \Theta_1(X_{Cn},\eta_{jn})$ is an open set of full measure. But, by [A9], $f_C(i|\theta,X_C)$ is uniformly λ-Lipschitz with constant M_f on Θ_N.

Suppose $\theta \notin \Theta_N$, so $\theta \notin \Theta_1(X_{Cn},\eta_{jn})$ for some (n,j). With probability one, θ is contained in $\Theta_1(X_{Cn'},\eta_{j'n'})$ for $(n,'j') \neq (n,j)$. Hence, using (52), the discontinuity in $|W(d-f(\theta))|$ is at most $mM_W M_\varphi/r$ with probability one. □

Assumption [A4] implied that the set of η for which there are ties in the components of $a(\theta,\eta)X_C$ has probability zero for all θ and almost all X_C. The next assumption requires that the geometry of $a(\theta,\eta)$ be such that the exceptional set $\mathbb{N}(\theta,X_C)^c$ of η where ties occur varies smoothly in (θ,X_C).

[A10] <u>There exists</u> M_g <u>and</u> $\lambda > 0$ <u>such that for</u> $X_C \in \mathcal{X}_o$ <u>and almost all</u> $\theta \in \Theta$, <u>the set</u> $B_\delta(\theta,X_C) = \{\eta|\eta\in\mathbb{N}(\tilde{\theta},X_C)^c$ <u>for some</u> $|\tilde{\theta}-\theta|\le\delta\}$ <u>has</u> $g(B_\delta(\theta,X_C)) \le M_g\delta^\lambda$.[18]

(INSERT FIGURE 1 ABOUT HERE)

Figure 1 illustrates the construction of $B_\delta(\theta,X_C)$. The assumption holds if the set-valued function $\mathbb{N}(\theta,X_C)^c$ is transversal at θ or if there is at most a polynomial singularity. The next result shows that with regularity conditions, the case $a(\theta,X_C) = \beta(\theta) + \eta\Gamma(\theta)$ for the MNP model satisfies [A10].

<u>Lemma</u> 5. <u>Suppose the hypotheses of Lemma</u> 1, <u>with</u> A_{22} <u>always non-singular, and</u> [A6]-[A9]. <u>Then</u> [A10] <u>holds</u>.

Proof: Suppose a tie between alternatives 1 and 2, so $\alpha x_2 = 0$. Using the notation of (55) and (56), partition $\alpha_1 = (\alpha_1, \alpha_3, \ldots, \alpha_m)$ and let α_2 denote the second component. Then,

$$\text{(62)} \qquad \eta^2 = [-\beta^1 A_{12} - \beta^2 A_{22} - \eta^1(\Gamma_{11}A_{12}+\Gamma_{12}A_{22})](\Gamma_{22}A_{22})^{-1}.$$

This function $\eta^2 = \psi(\theta, X_C, \eta^1)$ is continuously differentiable in (θ, X_C), and hence has a Taylor's expansion

$$\text{(63)} \qquad \psi(\tilde{\theta}, \tilde{X}_C, \eta^1) - \psi(\theta, X_C, \eta^1) = [\lambda_1 + \eta^1\lambda_2]\begin{bmatrix} \tilde{\theta}-\theta \\ \tilde{X}_C - X_C \end{bmatrix},$$

where λ_1 and λ_2 are vectors of continuous derivatives of $\psi(\theta, X_C, \eta^1)$ evaluated between (θ, X_C) and $(\tilde{\theta}, \tilde{X}_C)$. Then uniform continuity on compact $\Theta \times \mathcal{X}$ implies there exists a constant M_ψ such that for $|(\theta, X_C)-(\tilde{\theta}, \tilde{X}_C)| \le \delta$,

$$\text{(64)} \qquad |\psi(\tilde{\theta}, \tilde{X}_C, \eta^1) - \psi(\theta, X_C, \eta^1)| \le M_\psi(1+|\eta^1|)\delta.$$

Then the set $N_2(\tilde{\theta}, \tilde{X}_C, \eta^1) = \{\eta^2 \mid |\eta^2 - \psi(\tilde{\theta}, \tilde{X}_C, \eta^1)| \le M_\psi(1+|\eta^1|)\delta\}$ contains all η^2 solving (55) for $|(\theta, X_C)-(\tilde{\theta}, \tilde{X}_C)| \le \delta$, and satisfies

$$\text{(65)} \qquad \int_{\eta^1} g_1(\eta^1)d\eta^1 \int_{N_2(\theta, X_C, \eta^1)} g_{2.1}(\eta^2|\eta^1)d\eta^2$$

$$\le M_\psi(1+E|\eta^1|)\delta M_\gamma \equiv 2M_g\delta/m(m-1),$$

where M_γ bounds $g_{2.1}$. There are m(m-1) possible combinations of tied alternatives, each of which can with permutations of components of X_C, α, and η and relocation of X_C be put in the form above. The sum of the bounds for each combination gives [A10]. □

Given $\varepsilon > 0$, a finite family of random functions F_ε is said to <u>bracket</u> a family of random functions F if for each $Y \in F$ there exist $\underline{Y}, \overline{Y} \in F_\varepsilon$ such that $\underline{Y} \le Y \le \overline{Y}$ and $E(\overline{Y}-\underline{Y}) < \varepsilon$. The logarithm of of the number of elements in the smallest set F_ε that brackets F, denoted $H(\varepsilon)$, is called <u>metric</u>

entropy with bracketing. The following result establishes stochastic equicontinuity conditions for families whose metric entropy does not rise too rapidly as ε falls.

Lemma 6. Assume F is a uniformly bounded family of measurable random functions satisfying $\int_0^1 H(\varepsilon^2)^{1/2}d\varepsilon$ finite, where H is the metric entropy with bracketing. Suppose $y_1, y_2, \ldots$ are independent identically distributed copies of $Y-EY$ for $Y \in F$. Define $\|Y\| = E|Y-EY|$. Then for every $\lambda > 0$,

(66) $$\lim_{\delta\to 0} \limsup_{N\to 0} \Pr\{\sup_{F} \sup_{\|y-\tilde{y}\|\le\delta} N^{-1/2}|\sum_{n=1}^{N}(y_n-\tilde{y}_n)| > \lambda\} = 0.$$

Proof: Dudley (1984), Theorem 6.2.1, establishes (59). I use a restatement from Alexander (1987), Theorem 2.1. □

The next result establishes that the simulation residuals satisfy stochastic equicontinuity and boundedness conditions sufficient for the MSM estimator to be CAN. The critical step is to show that these residuals satisfy the assumption on metric entropy required by Lemma 6.[19]

Lemma 7. Suppose [A1]-[A10]. Define $\zeta(\theta) = N^{-1/2}W(f(\theta)-Ef(\theta))$ and $B(\theta) = N^{-1/2}W(Ef(\theta)-P(\theta))$. Then given $\lambda, \delta > 0$, there exists M such that

(67) $$\limsup_N \sup_{\Theta} |B(\theta)| = 0,$$

(68) $$\sup_N \Pr\{\sup_{\Theta} |\zeta(\theta)| > M\} < \lambda,$$

and for $A_N = \{\theta |\ N^{1/2}|\theta-\theta^*| \le \delta\}$,

(69) $$\limsup_N \Pr\{\sup_{\theta\in A_N} |\zeta(\theta)-\zeta(\theta^*)| > \lambda\} = 0.$$

Proof: Condition (67) is immediate from [A8]. Assume $\Theta \in [0,1]^k$. For any integer j, cover $[0,1]^k$ with 2^{kj} cubes with sides 2^{-j}, and for each $X_C \in \mathcal{X}_o$, let $\Theta_j(X_C)$ be a set containing one point selected from each cube that intersects Θ. By [A10], the selection can be made so that $g(B_\delta(\theta,X_C)) \le M_g\delta^\lambda$ for $\theta \in \Theta_j(X_C)$. Define $\theta_j(\theta,X_C)$ to be point in $\Theta_j(X_C)$ nearest to θ; then $|\theta-\theta_j(\theta)| \le 2^{-j} \equiv \delta_j$.

Let $Y(\theta,X_C) = \sum_{i\in C} W_i f_C(i|\theta,X_C)$. Define $q_j(\theta,X_C)$ to be the number of draws η_s for $s = 1,\ldots,r$ with $\eta_s \in B_{\delta_j}(\theta,X_C)$. Using the notation of (59), and λ satisfying [A9] and [A10], define

$$(70)\qquad Y_j^o(\theta,X_C) = mM_W(M_f|\theta-\theta_j(\theta)|^\lambda+M_\varphi q_j(\theta,X_C)/r).$$

Then, by [A10], $Eq_j(\theta,X_C) \le rg(B_{\delta_j}(\theta,X_C))$, implying

$$(71)\qquad EY_j^o(\theta,X_C) \le mM_WM_f|\theta-\theta_j(\theta)|^\lambda + mM_WM_\varphi g(B_{\delta_j}(\theta,X_C))$$

$$\le mM_W(M_f+M_\varphi M_g)\delta_j{}^\lambda \equiv M_o\delta_j{}^\lambda.$$

From (59),

$$(72)\qquad |Y(\theta,X_C)-Y(\theta_j(\theta),X_C)| \le mM_W(M_f|\theta-\theta_j(\theta)|^\lambda)$$

$$+ mM_WM_\varphi \max_{i\in C}|\varphi_C(i|\theta,X_C)-\varphi_C(i|\theta_j(\theta),X_C)|$$

$$\le mM_W(M_f|\theta-\theta_j(\theta)|^\lambda+M_\varphi q_j(\theta,X_C)/r)$$

Hence, $\underline{Y}_j(\theta,X_C) \equiv Y(\theta_j(\theta),X_C) - Y_j^o(\theta,X_C) \le Y(\theta,X_C) \le \overline{Y}_j(\theta,X_C) \equiv Y(\theta_j(\theta),X_C) + Y_j^o(\theta,X_C)$. Given $\varepsilon > 0$, choose j to be the smallest integer such that $2^{-\lambda j} < \varepsilon$. Then the 2^{kj+1} functions $\underline{Y}_j$ and $\overline{Y}_j$ bracket $Y(\theta,X_C)$, $\theta \in \Theta$. This implies that the metric entropy with bracketing for $F = \{Y(\theta,X_C)|\theta\in\Theta\}$ satisfies $H(\varepsilon) \le (kj+1)\mathrm{Ln}2 \le (-\mathrm{Ln}\varepsilon)k/\lambda + (k+1)\mathrm{Ln}2$, and hence $\int_0^1 H(\varepsilon^2)^{1/2}d\varepsilon \le \int_0^1 H(\varepsilon^2)d\varepsilon \le (k+1)\mathrm{Ln}2 - 2(k/\lambda)\int_0^1 \mathrm{Ln}\varepsilon\, d\varepsilon < \infty$. This establishes the assumptions of Lemma 5, so (66) holds.

For any $\delta > 0$, forming the expectation of (59) and using [A10], $|\theta-\tilde{\theta}| < \delta$ implies $E|Y(\theta,X_C)-Y(\tilde{\theta},X_C)| < mM_w(M_f+M_\varphi M_g)\delta \equiv M_o\delta$. Hence, (66) can be written in the form

$$\text{(73)} \quad \lim_{\delta\to0} \limsup_{N\to0} \Pr\{\sup_{\Theta} \sup_{|\theta-\tilde{\theta}|\le\delta} |\zeta(\theta)-\zeta(\tilde{\theta})| > \lambda\} = 0.$$

Taking $\delta = 2^{-j}$, $j = 1,2,\ldots$, one has $|\theta-\tilde{\theta}| < \delta$ for $N \ge 4^j$ and $N^{1/2}|\theta-\tilde{\theta}| < 1$. Then (73) implies (69).

Next prove (68). From (73), given $\lambda > 0$, there exists $\delta > 0$ such that

$$\text{(73)} \quad \limsup_{N\to0} \Pr\{\sup_{\Theta} \sup_{|\theta-\tilde{\theta}|\le\delta} |\zeta(\theta)-\zeta(\tilde{\theta})| > \lambda\} < \lambda.$$

Choose any $\theta_o \in \Theta$. A central limit theorem implies there exists M_o such that $\sup_N \Pr\{|\zeta(\theta_o)| > M_o\} < \lambda$. But any $\theta \in \Theta$ can be written $\theta = \theta_o + \sum_{j=0}^{J} (\theta_{j+1}-\theta_j)$ with $\theta_j = (j/J)\theta + (1-j/J)\theta_o$ and J the smallest integer exceeding $1/\delta$. Then,

$$\text{(74)} \quad \Pr\{|\zeta(\theta)| > M_o+\lambda J\}$$

$$\le \Pr\{|\zeta(\theta_o)| > M_o \;\&\; |\zeta(\theta_{j+1})-\zeta(\theta_j)| < \lambda,\ j=0,\ldots,J\} < \lambda.$$

Then (67) holds with $M = M_o + \lambda J$. □

In this lemma, the construction in (70) and the following arguments hold even if the number of repetitions r is a random function of θ, X_C, and N. Then, in particular, the lemma holds for simulators formed by acceptance/rejection methods with random stopping rules, and for consistent simulators where r increases with sample size.

Let $\hat{\theta}_N$ be a sequence in Θ and assume that the instruments are evaluated at $\hat{\theta}_N$ for each N. The $\hat{\theta}_N$ might be non-stochastic, or a sequence of initially consistent estimators, or might equal the MSM estimator θ_{sm}. In the last case, θ_{sm} solves $\|W(\theta_{sm})(d-f(\theta_{sm}))\| \leq \inf_{\theta}\|W(\theta_{sm})(d-f(\theta))\| + o_p(1)$. Lemma 7 holds in all these cases.

FOOTNOTES

1. The idea of simulating response probabilities from an underlying latent variable model, generating the response probabilities by stochastic integration, is standard in the area of computer simulation; see Hammersley and Handscomb (1964), Fishman (1973), and Lerman and Manski (1981). This literature has concentrated on simulating the response probabilities to a level of accuracy that enables their use in standard maximum likelihood procedures.

2. Use the identity

$$0 \equiv \sum_{i\in C} \partial P_C(i|\theta,X_C)/\partial\theta \equiv \sum_{i\in C} [\partial LnP_C(i|\theta,X_C)/\partial\theta]P_C(i|\theta,X_C).$$

3. Starting from any K × mN array of instruments Z^o, and taking account of the structure of the covariance matrix of the residuals, one can show by a standard argument from non-linear least squares that the asymptotic minimum variance estimator in the class using linear combinations of instruments in Z^o is attained by $W = P_\theta(\theta^*)'Z'(Z\hat{P}Z')^{-1}Z$, where $P_\theta(\theta^*)$ is the mN × k array of derivatives $\partial P_C(i|\theta,X_{Cn})/\partial\theta$, evaluated at θ^*, $\hat{P}$ = diag $P(\theta^*)$, and

$$Z_{in} = Z^o_{in} - \sum_{j\in C} Z^o_{jn}P_C(j|\theta^*,X_{Cn}).$$

If $Z^o = \partial LnP(\theta^*)/\partial\theta$, then $W = \partial LnP(\theta^*)/\partial\theta$. Approximations to θ^* and to the functions P and P_θ yield approximations to the minimum variance instruments that can be constructed from Z^o.

4. Appendix Lemma 6 and assumption [A9] give sufficient conditions for simulators to satisfy (13).

5. The independence of the simulators across observations can be relaxed to any process that is sufficient to give the term in (a) asymptotically

normal and to give stochastic equicontinuity of the simulation residuals.

6. Continuous differentiability and compactness imply $|\omega_N(\theta)-\omega_N(\tilde{\theta})| \le M|\theta-\tilde{\theta}|$, where $M \ge \max_\Theta|R(\theta)|$. Given $\varepsilon > 0$, extract a finite covering of Θ with neighborhoods of radius less than $\varepsilon/3M$; let Θ_ε denote the finite set of centers of these neighborhoods. Then choose N_ε sufficiently large so that for $N > N_\varepsilon$, $P\{\max_{\Theta_\varepsilon} |\omega_N(\theta)-\omega(\theta)| > \varepsilon/3\} < \varepsilon$. By construction of Θ_ε, for each $\theta \in \Theta$, there is a $\tilde{\theta} \in \Theta_\varepsilon$ such that $|\omega_N(\theta)-\omega(\theta)| \le |\omega_N(\tilde{\theta})-\omega(\tilde{\theta})| + 2\varepsilon/3$. Hence, $P\{\max_\Theta|\omega_N(\theta)-\omega(\theta)| > \varepsilon\} < \varepsilon$. Regularity condition [A8] implies $R(\theta^*)$ nonsingular.

7. The inequality, from Gine and Zinn (1986; Lemma 3.2), is

$$P\{\sum_{i=1}^{N} Y_i > t\} \le \exp[-t^2/(2N\sigma^2 + 2ct/3)],$$

where $\sigma^2 = EY_i^2$; see also Pollard (1984) or Shorack and Wellner (1986). Replace t by $N^{1/2}t$ and use $\sigma^2 \le c^2$ to obtain (24).

8. Consider the normal density

$$n(u-\mu,\Lambda) = (2\pi)^{-1/2}|\Lambda|^{-1/2}e^{(u-\mu)'\Lambda^{-1}(u-\mu)/2},$$

with $\Lambda = X'\Gamma'\Gamma X$ and $\mu = \beta X$. The following matrix differentiation formulas are derived by writing out terms from the familiar expressions $\partial \mathrm{Ln}|A|/\partial A = A^{-1}$ and $\partial A^{-1}/\partial A = -A^{-1}\otimes A^{-1}$ (which hold when A is symmetric, but identity of cross-terms is not imposed in the differentiation):

$$\partial \mathrm{Ln}|X'\Gamma'\Gamma X|/\partial\Gamma = 2\Gamma X(X'\Gamma'\Gamma X)^{-1}X',$$

$$\partial(z'(X'\Gamma'\Gamma X)^{-1}z)/\partial\Gamma = -2\Gamma X(X'\Gamma'\Gamma X)^{-1}zz'(X'\Gamma'\Gamma X)^{-1}X'.$$

The derivatives of Ln $n(u-\beta X, X'\Gamma'\Gamma X)$ are then

$$\partial \mathrm{Ln}\ n/\partial\beta = X(X'\Gamma'\Gamma X)^{-1}(u-\beta X)',$$

$$\partial \mathrm{Ln}\ n/\partial\Gamma = -\Gamma X(X'\Gamma'\Gamma X)^{-1}X' + \Gamma X(X'\Gamma'\Gamma X)^{-1}(u-\beta X)'(u-\beta X)(X'\Gamma'\Gamma X)^{-1}X'.$$

This construction was suggested by Paul Ruud.

9. A mixture of multinomial logit models, with the mixture interpreted as the result of taste variations in the population, has been of independent interest as a discrete choice model; see Westin (1974) and McFadden (1984).

10. The polynomial kernel limits the number of alternatives for which calculations must be done. If an observation has every component of u_{C-i} greater than $2b_N$ in magnitude, then $\mathbb{K}(u_{C-i})$ coincides with $1(u_{C-i}\le 0)$; the probability of the converse is $O(b_N)$. If for a draw of u_C for an observation, all component differences exceed $2b_N$ in magnitude, the kernel-smoothed frequency simulator coincides with the simple frequency simulator. Then, in a sample of size N with r Monte Carlo draws per observation, the expected number of alternatives for which further calculation is required to obtain the simulator and corresponding instruments is bounded by $(r+1)N + mrNO(b_N) \le (r+1)N + O(mrN^{-\varepsilon+1/2})$. This makes the calculation practical even if the number of alternatives is large.

11. Discussions with Jim Heckman contributed to the formulation of kernel-smoothed frequency simulators. The unbiased kernel-smoothed frequency simulator for the MNP model was suggested by Steve Stern.

12. Moran (1984) suggests several control variates. Peter Phillips and Vasillis Hajivassiliou suggested the use of spherical transformations for this problem, and Dan Nelson developed many of the details.

13. To generate a denser set of antithetic points, for any integer $T > 1$, and each pair s^i and s^j with $i < j$, construct the directions $(\pm ts^i \pm (T-t)s^j)$ for $t = 1,\ldots,T-1$. Combined with the points $\pm s^i$, this gives $4(T+1)$ evenly spaced points on each great circle, for a total of $2K + 2TK(K-1)$ directions.

14. When $\beta = 0$ and Γ consists of an identity submatrix corresponding to

alternative-specific dummy variables and a zero submatrix corresponding to the remaining variables, then X_{C-i} and $X_{C-i}X'_{C-i}$ are, except for proportional constants, a superset of the ideal instruments. If the model is identified, then it will always be possible to find low-order polynomials in X_{C-i} that have an asymptotic correlation matrix with $\partial P_C(i|\theta)/\partial\theta$ that is of full rank. Thus, the crude instruments proposed may not be grossly inefficient. In the third step when smooth simulators are being used, one iteration from $\hat{\theta}$ using Newton's method achieves the maximum asymptotic efficiency attainable from better instruments simulated at $\hat{\theta}$.

15. With random sampling, $R_N(\theta)$ converges almost surely to a limit $R(\theta)$, for each θ, by application of a strong law of large numbers.

16. For any continuous function $\alpha = a(\theta,\eta)$, the transformed latent variable model $u_C = (1+\|\alpha\|)^{-1}\alpha X_C$ yields the same response probabilities, and is uniformly Lipschitz in X_C.

17. If crude instruments independent of θ are used to obtain an initially consistent estimator, then ω in (50) is independent of $\tilde{\theta}$. If approximations to the ideal instruments are calculated, starting from an initially consistent estimator, then it is sufficient that [A7] hold for $\tilde{\theta}$ in a neighborhood of θ^*.

18. The following argument establishes that $B_\delta(\theta,X_C)$ is closed, and hence measurable, for $(\theta,X_C) \in \Theta \times X_o$: If $\eta^j \in B_\delta$ and $\eta^j \to \eta^0$, then $\eta^j \in N(\theta^j,X_C^j)^c$, a closed set, for some (θ^j,X_C^j) in a closed δ neighborhood of (θ,X_C), by [A3]. Hence, using the continuity of $a(\theta,\eta)X_C$ in (θ,X_C), $\eta^0 \in N(\theta^0,X_C^0)^c$ for each limit point (θ^0,X_C^0) of (θ^j,X_C^j).

19. I am indebted to Ariel Pakes and David Pollard for discussions that led to the formulation of this lemma.

REFERENCES

Alexander, K. (1987) "The central limit theorem for empirical processes on Vapnic-Cervonenkis classes" *The Annals of Probability* 15, 178-203.

Barbara, S. and Pattanaik (1985) "Falmange and the rationalizability of stochastic choices in terms of random orderings" *Econometrica* 54, 707-715.

Chamberlain, G. (1984) "Panel Data" in Z. Griliches and M. Intriligator (ed) *Handbook of Econometrics*, v.2, North Holland, Amsterdam, 1247-1320.

Clark, C. (1961) "The greatest of a finite set of random variables" *Operations Research* 9, 145-162.

Daganzo, C. (1980) *Multinomial Probit*, Academic, New York.

Deák, I. (1980) "Three digit accurate multiple normal probabilities" *Numerische Mathematik* 35, 369-380.

Devroye, L. (1986) *Non-Uniform Random Variate Generation* Springer, New York.

Dudley, R. (1984) "A course on empirical processes" Ecole d'Été de Probabilités de Saint-Flour, XII-1982, *Lecture Notes in Mathematics* 1097, Springer, New York, 1-142.

Dutt, J. (1976) "Numerical aspects of multivariate normal probabilities in econometric models" *Annals of Economic and Social Measurement* 5, 547-562.

Falmange, J. (1978) "A representation theorem for finite random scale systems" *Journal of Mathematical Psychology* 18, .

Fishman, G. (1973) *Concepts and Methods of Digital Simulation*, Wiley, New York.

Hajivassiliou, V. and D. McFadden (1987) "The debt repayment crises of LDC's: estimation by the method of simulated moments" Yale Univ., Working Paper.

Hammersley, J. and D. Handscomb (1964) *Monte Carlo Methods*, Methuen, London.

Hausman, J. and D. Wise (1978) "A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences" *Econometrica* 46, 403-426.

Heckman, J. (1981) "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process " in C. Manski and D. McFadden (ed) *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, 179-195.

Hendry, D. (1984) "Monte Carlo experimentation in econometrics" in Z. Griliches and M. Intriligator (ed) *Handbook of Econometrics*, v.2, North Holland, Amsterdam, 937-976.

Horowitz, J., J. Sparmonn, and C. Daganzo (1981) "An investigation of the accuracy of the Clark approximation for the multinomial probit model" *Transportation Science* 16, 382-401.

Giné, E. and J. Zinn (1986) "Lectures on the central limit theorem for empirical processes" Probability and Banach Spaces *Lecture Notes in Mathematics* 1221, Springer, New York, 50-113.

Lerman, S. and C. Manski (1981) "On the use of simulated frequencies to approximate choice probabilities" in C. Manski and D. McFadden (ed) *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, 305-319.

McFadden, D. (1984) "Econometric analysis of qualitative response models" in Z. Griliches and M. Intriligator (ed) *Handbook of Econometrics*, v.2, North Holland, Amsterdam, 1395-1457.

McFadden, D. (1986a) "The choice theory approach to marketing problems" *Marketing Science*, 5, 275-297.

McFadden, D. (1986b) "Discrete response to unobserved variables for which there are multiple indicators", MIT, Working Paper.

Moran, P. (1984) "The Monte Carlo evaluation of orthant probabilities for multivariate normal distributions" *Australian Journal of Statistics* 26, 39-44.

Owen, D. (1956) "Tables for computing bivariate normal probabilities" *Annals of Mathematical Statistics* 27, 1075-1090.

Pakes, A. and D. Pollard (1985) "The asymptotic distribution of simulation experiments" *Econometrica*, forthcoming.

Pollard, D. (1984) *Convergence of Stochastic Processes*, Springer, New York.

Pollard. D. (1985) "New ways to prove central limit theorems" *Econometric Theory* 1, 295-314.

Press, W., B. Flannery, S. Teukolsky, W. Vetterling (1986) *Numerical Recipes*, Cambridge University Press, Cambridge.

Ruud, P. (1981) *Misspecification Errors in Limited Dependent Variable Models*, MIT, PhD Dissertation.

Ruud, P. and D. McFadden (1987) "Estimation of limited dependent variable models from the regular exponential family by the method of simulated moments", Univ. of California, Berkeley, Working Paper.

Shorack, G. and J. Wellner (1986) *Empirical Processes with Applications to Statistics*, Wiley, New York.

Spanier, J. and K. Oldham (1987) *An Atlas of Functions*, Hemisphere, Washington.

Train, K., D. McFadden, and A. Goett (1987) "Consumer attitudes and voluntary rate schedules for public utilities" *Review of Economics and Statistics*, forthcoming.

Westin, R. (1974) "Predictions from binary choice models" *Journal of Econometrics* 2, 1-16.

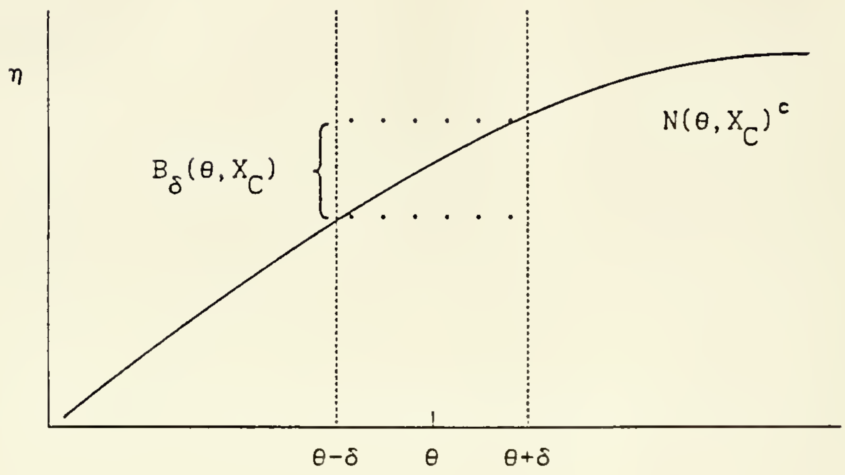

FIGURE 1