Econometrica, Vol. 57, No. 5 (September, 1989), 995–1026

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

BY DANIEL MCFADDEN[1]

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.

KEYWORDS: Method of moments, simulation, multinomial probit, discrete response.

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

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

For some problems, the expected response function may be difficult to express analytically or to compute, but relatively easy to simulate. When this function is replaced by an 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 *method of simulated moments* (MSM).[2]

This paper focuses on application of MSM to discrete response models, particularly the multinomial probit (MNP) model. However, the method is more general and can be applied to most moment estimation problems. In a related paper, Pakes and Pollard (1989) have independently proposed minimum distance

[1] 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. I have shortened the proof of the main theorem by adapting arguments from Pakes and Pollard's independent investigation of the asymptotic behavior of simulation experiments. I have also benefited from suggestions made by Chunrung 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.

[2] 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.

estimators using simulation, and have established their statistical properties using combinatorial empirical process methods. Most of the statistical results in this paper could also 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 previews the conditions under which it is consistent asymptotically normal (CAN). Section 4 discusses issues of computation and statistical efficiency. Sections 5–7, respectively, discuss applications of the method to discrete panel data with autoregressive errors, to discrete response models with measurement errors in explanatory variables, and to nonnormal discrete response problems. Section 8 contains the theorem and lemmas referred to in the text, and their proofs.

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

$$u_i = \alpha x_i, \quad i \in C, \tag{1}$$

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 $u_i \geqslant u_j$ 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. Let $\beta(\theta)$ and $\Omega(\theta)$ denote the mean and covariance matrix of $a(\theta, \eta)$. In applications, it is often convenient to work with a Cholesky factor of Ω: let $\Gamma(\theta)$ be an 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 $u_i \geqslant u_j$ for $j \in C$, given X_C. Define

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

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

Then u_{C-i} has a multivariate density $g_U(u_{C-i}|\theta, X_C)$ with mean βX_{C-i} and covariance matrix $X'_{C-i}\Omega X_{C-i}$, induced by the transformation $u_{C-i} = a(\theta, \eta)X_{C-i}$ of the random vector η. The response probability $P_C(i|\theta, X_C)$ equals the nonpositive orthant probability of u_{C-i},

$$P_C(i|\theta, X_C) = \int 1(u_{C-i} \leqslant 0) g_U(u_{C-i}|\theta, X_C)\, du_{C-i} \tag{2}$$

$$= \int 1(a(\theta, \eta) X_{C-i} \leqslant 0) g(\eta)\, d\eta,$$

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

$$\alpha = a(\theta, \eta) \equiv \beta(\theta) + \eta\Gamma(\theta), \tag{3}$$

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 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

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

The associated score is

$$\partial L(\theta)/\partial\theta = \sum_{n=1}^{N} \sum_{i \in C} W_{in}\left[d_{in} - P_C(i|\theta, X_{Cn})\right], \tag{4}$$

where

$$W_{in} = \partial \ln P_C(i|\theta, X_{Cn})/\partial\theta. \tag{5}$$

Equation (4) is derived using the identity

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

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 $m \leqslant 4$ 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 solve (4). 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 (1981).

3. THE METHOD OF SIMULATED MOMENTS

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

$$\theta_{mm} = \arg\min_\theta (d - P(\theta))'W'W(d - P(\theta)), \tag{6}$$

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 $K \geqslant k$. The instruments may depend on θ, but are evaluated at some fixed θ_0 in forming first-order conditions for solution of (6). The instrument array (5), 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, (5) 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. Computation of instruments is discussed further in Section 4. In the remainder of this section, I will assume W is a given fixed instrument array.

Under mild regularity assumptions, sufficient conditions for classical method of moments estimation to be CAN are the following:

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

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

The *method of simulated moments* (MSM) avoids the computation of $P(\theta)$ required for (6), replacing it with a simulator $f(\theta)$ that is (asymptotically) conditionally unbiased, given W and d, independent across observations, and "well behaved" in θ. An example is the *simple frequency simulator* calculated from the latent variable model (1) by independently drawing, for each observation, one or more vectors η from the density $g(\eta)$, and then for any trial θ calculating $u_{in} = a(\theta, \eta)x_{in}$ and counting the frequency with which the u_{in} for each alternative is maximized. The MSM estimator is given by any argument θ_{sm} satisfying[3]

$$\begin{aligned}(d - f(\theta_{sm}))'W'W(d - f(\theta_{sm})) \\ \leqslant \inf_{\theta \in \Theta} (d - f(\theta))'W'W(d - f(\theta)) + O(1).\end{aligned} \tag{7}$$

This definition assures the existence of a MSM estimator even if the infimum cannot be attained.

[3] A sequence of numbers a_N is $O(1)$ if it is bounded in magnitude by a convergent sequence, and is $o(1)$ if it converges to zero. A sequence of random variables Z_n is $O_p(1)$ if, given $\varepsilon > 0$, there exists δ such that $\text{Prob}(|Z_N| > \delta) < \varepsilon$ for all N. The sequence is $o_p(1)$ if Z_N converges in probability to zero.

Sufficient conditions for the MSM estimator to be CAN are given formally in Theorem 1. They involve the same regularity assumptions and conditions on instruments as classical method of moments estimators, and in addition place two critical restrictions on the simulator $f(\theta)$:

(iii) *The simulation bias* $B(\theta) = N^{-1/2}W(Ef(\theta) - P(\theta))$, *where the expectation is conditioned on* W *and* d, *is either zero, or else satisfies*

$$\sup_{\theta \in \Theta} |B(\theta)| = o(1). \tag{8}$$

(iv) *The simulation residual process* $\zeta(\theta) = N^{-1/2}W(f(\theta) - Ef(\theta))$ *is uniformly stochastically bounded and equicontinuous in* θ; *i.e.*,

$$\sup_{\theta \in \Theta} |\zeta(\theta)| = O_p(1), \tag{9}$$

$$\sup_{\theta \in A_N} |\zeta(\theta) - \zeta(\theta^*)| = o_p(1), \tag{10}$$

where, for a given $\delta > 0$, $A_N = \{\theta | N^{1/2}|\theta - \theta^*| \leqslant \delta\}$.

If $f(\theta)$ is an unbiased simulator of $P(\theta)$ for all θ, and the random numbers used to compute $f(\theta)$ are independent of the observed responses and of any simulation used in the construction of the instrument array, then (iii) is satisfied. The simulation residuals $\zeta(\theta)$ are by construction the normalized sum over observations of independent identically distributed terms that are 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. Elementary arguments (Lemma 4) establish that if $f(\theta)$ is smooth for θ in a compact domain, then (iv) holds. However, (iv) can also be satisfied by simulators that have "well-behaved" discontinuities (Lemma 8), such as the simple frequency simulator. To satisfy (iv), a simulator must avoid "chatter" as θ varies; this will generally require that the Monte Carlo random numbers used to construct $f(\theta)$ *not* be redrawn when θ is changed.

4. COMPUTATIONAL ISSUES AND STATISTICAL EFFICIENCY

Practical use of the MSM estimator requires that the Monte Carlo simulation of the probabilities and their derivatives be practical and "well behaved"; that easily calculated, moderately efficient instruments W be available; 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.

Simulators for the Response Probabilities

I have given the example of a *simple frequency simulator* $f(\theta)$ for the discrete response model generated from the latent variable model (1): Count the frequency $f_C(i|\theta, X_{Cn})$ with which component i of $u_{Cn} = a(\theta, \eta)X_{Cn}$ is largest, where the η are one or more Monte Carlo random vectors from the density $g(\eta)$, independent across observations and fixed for the duration of the analysis. This

simulator is unbiased for all θ, but 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 (3) by drawing standard normal vectors η and calculating $u_{Cn} = (\beta(\theta) + \eta\Gamma(\theta))X_{Cn}$.

It is also possible to construct a *smooth unbiased simulator* $f(\theta)$. 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 (2) can be rewritten as

$$P_C(i|\theta, X_C) \equiv \int h(u_{C-i}, \theta, X_{C-i})\gamma(u_{C-i})\, du_{C-i}, \tag{11}$$

where $h(u_{C-i}, \theta, X_{C-i}) = g_U(u_{C-i}|\theta, X_C)/\gamma(u_{C-i})$. Average $h(u_{C-i}, \theta, X_{C-i})$ 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 is chosen so that h is dominated by a function H independent of θ with $\int H\gamma\, du_{C-i}$ finite. 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 (0, 1) random numbers. Choices of γ that make h flatter can reduce simulation error, as in Monte Carlo importance sampling. For MNP,

$$\begin{aligned} &h(u_{C-i}, \theta, X_{C-i}) \\ &\quad = n(u_{C-i} - \beta(\theta)X_{C-i},\ X'_{C-i}\Gamma(\theta)'\Gamma(\theta)X_{C-i})/\gamma(u_{C-i}), \end{aligned}$$

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.

The estimation of the asymptotic covariance matrix of the MSM estimator, as well as approximation of efficient instruments, requires simulation of derivatives of the response probabilities with respect to θ. From (11), these derivatives can be written

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

where h_θ denotes the vector of derivatives of h with respect to θ. Then, a smooth unbiased simulator of (12) can be constructed in the same manner as was done for (11). Lemma 10 details this construction for MNP.

A potential drawback of smooth simulators based on (11) is that they are not constrained to sum up to one for $i \in C$. An alternative class of *kernel-smoothed frequency simulators* satisfy summing-up, but are only asymptotically unbiased. The idea underlying these simulators is that (2) can be approximated by an integral

$$\tilde{P}_C(i|\theta, X_C) = \int \mathbb{K}(a(\theta, \eta)X_{C-i}/b_N)g(\eta)\, d\eta, \tag{13}$$

where $\mathbb{K}(u_{C-i}/b_N)$ is a smooth approximation that approaches the indicator function $1(u_{C-i} \leqslant 0)$ as $b_N \to 0$. Let $\mathbb{K}_i(y_1, \ldots, y_m) \equiv \mathbb{K}(y_{C-i})$. I require that $\Sigma_{i \in C} \mathbb{K}_i(u_C/b_N) \equiv 1$. The multinomial logit function $\mathbb{K}_i(y_C) = e^{y_i}/\Sigma_{j \in C} e^{y_j}$ is one example in this class. Smooth simulators satisfying summing-up for $i \in C$ are obtained by averaging $\mathbb{K}_i(a(\theta, \eta)X_C/b_N)$ over a common Monte Carlo sample from $g(\eta)$.

Kernel-smoothed simulators can be derived from a perturbation of the latent variable model (1),

$$\tilde{u}_C = u_C + \nu_C b_N, \tag{14}$$

where ν_C is a vector whose components are independently distributed with a distribution function Ψ, and b_N is a scalar chosen for the simulation. This formulation can be interpreted as a *mixture* of (1) and a contaminating distribution. Assume Ψ has a finite moment generating function $\mu(t)$ for t in a neighborhood of zero. Define the smooth function

$$\mathbb{K}_i(y_1, \ldots, y_m) = \int \left(\prod_{j \neq i} \Psi(y_i - y_j + v) \right) \Psi'(v)\, dv. \tag{15}$$

This kernel will be most practical when Ψ is chosen so that $\mathbb{K}_i$ has an easily calculated closed form. The response probability implied by (14), obtained by first conditioning on u_C and integrating out ν_C, coincides with the approximation (13) when the smoothing function (15) is used. This simulator is nonnegative, and is strictly positive if the support of Ψ is the real line. Lemma 3 shows that a sufficient condition for it to be asymptotically unbiased is that b_N satisfy $N^{\varepsilon + 1/2} b_N \to 0$ for some $\varepsilon > 0$. If the simulators for all $i \in C$ are constructed from common Monte Carlo draws, then they satisfy summing-up. Choosing Ψ to be type I extreme value distributed yields the multinomial logit form, and (13) is a multivariate normal mixture of logits. This model, 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).

A *polynomial kernel* such as $\Psi(v) = [6 + 5v + (2 - |v|)v^3]/12$ for $|v| \leqslant 1$ is computationally economical, yielding a closed form for $\mathbb{K}_i$. One advantage of this kernel is that it 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 $\mathbb{K}_i(u_C/b_N)$ coincides with $1(u_{C-i} \leqslant 0)$ and the kernel-smoothed frequency simulator coincides with the simple frequency simulator. The probability of the converse is of the order $O(b_N)$. 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 $(r+1)N + mrNO(b_N) \leqslant (r+1)N + o(mrN^{-\varepsilon - 1/2})$. This makes the calculation practical even if the number of alternatives is large.

For MNP, a variant due to Stern (1987) of the kernel-smoothed frequency simulator is unbiased. The idea is that the normal latent variable model can be decomposed into a mixture of two normal vectors, one of which is scaled

standard normal. Thus, it is unnecessary to contaminate the model to achieve mixing, permitting an unbiased simulator. Rewrite (1) as $u_C \equiv \bar{u}_C + \nu b_N$, with ν a standard normal vector independent of $\bar{u}_C \sim N(\beta X_C, A'A)$, where A is upper triangular and $A'A \equiv X_C'\Gamma'\Gamma X_C - b_N I$. This can be done if b_N is small enough so that $X_C'\Gamma'\Gamma X_C - b_N I$ is positive definite. It is adequate for the asymptotic properties of MSM using this simulator that b_N be constant in open neighborhoods of Γ. Then, b_N can be set for each observation and estimation iteration by calculating a Cholesky factor of $X_C'\Gamma'\Gamma X_C$ at the trial Γ and choosing b_N less than the smallest diagonal element of this factor. As in (14), conditioning on $\bar{u}_C$,

$$P_C(i|\theta, X_C) = \int \mathbb{K}_i((\beta X_C + \eta A)/b_N) g(\eta)\, d\eta,$$

(16)

$$\mathbb{K}_i(y_1, \ldots, y_m) = \int \left(\prod_{j \neq i} \Phi(y_i - y_j + v) \right) \Phi'(v)\, dv,$$

with g the multivariate standard normal density and Φ the univariate standard normal distribution. An average of $\mathbb{K}_i$ in (16) over Monte Carlo draws from g yields an unbiased positive smooth *Stern frequency simulator*. Adding-up holds for an observation if common draws are used for all $i \in C$.

For MNP, construction of economical simulators is aided by the use of spherical transformations. The expression (11), and the specializations of (12) given in Lemma 10, involve simulation of integrals of the generic form

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

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 $r = (\sum_{j=1}^m u_j^2)^{1/2}$ and $s_j = u_j/r$. Define

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

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

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

(17)

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

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

Then, a *cylinder simulator* is defined from the form

$$Q = c_0 E_s c_1 C\left(\sum_{j=1}^{m} k_j + m - 1, a, b \right) \left(\prod_{j=1}^{m} s_j^{k_j} \right), \tag{18}$$

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

nonnegative orthant, and

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

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

$$c_0 = (2\pi)^{1/2}2^{-3m/2}|\Omega|^{-1/2}\Gamma(m/2)^{-1},$$

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

with $X = X_{C-i}$, and c_0 independent of X and s. To generate uniform draws from the distribution of s, draw a standard normal random vector u, and take

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

Then, (18) is simulated by drawing one or more s, and for each s using the recursion (17) to calculate C. A further refinement is to use control variates for C; Moran (1984) suggests several. Peter Phillips and Vasillis Hajivassiliou suggested the use of spherical transformations for this problem, and Dan Nelson developed many of the details.

The spherical transformation can also be used to calculate a *conditional chi-square frequency simulator* for MNP that is economical, unbiased, and smooth. Let s be a uniform draw from the unit sphere in $\mathbb{R}^K$, and let λ^2 be an independent random variable with a chi-square 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 $P_C(i|\theta, X_C, s) = H_K(\lambda^2_{j+1}) - H_K(\lambda^2_j)$, 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. An advantage of this method is that it simultaneously provides, for all alternatives, simulators satisfying summing-up.

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 unit sphere in $\mathbb{R}^K$, first draw a random basis $s^1, \ldots, s^K$. This can be done 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. To generate a denser set of antithetic points for any integer $T > 1$, take each pair s^i and s^j with $i < j$ and construct the directions $(\pm s^i \cos \pi t/2T \pm s^j \sin \pi t/2T)$ 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.

Numerical experiments on the algorithms discussed in this section suggest that for MNP with 5 to 20 alternatives, the conditional chi-square simulator combined with the Deák construction of antithetic directions is the fastest and most accurate.[4]

Choice of Instruments

Consider 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 $\partial \ln P(\theta^*)/\partial\theta$. Since the asymptotic efficiency of MSM relative to the classical moments estimator can be controlled by the number of Monte Carlo draws used to simulate $P(\theta)$, the issue is how to construct W to obtain good asymptotic efficiency for MSM without excessive computation.

Simulation of (12) based on Monte Carlo draws from the density $\gamma(u)$ yields smooth unbiased estimates of the derivatives. Since the smooth simulator (11) of $P_C(i|\theta, X_C)$ is positive, the ratio of the simulators of (12) and (11) provides an approximation to the ideal instruments $\partial \ln P_C/\partial\theta$. 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, modestly efficient instruments can be obtained with relatively few draws. It is essential for the asymptotic statistics of the MSM estimator that simulators of the response probabilities and their derivatives used to construct instruments be *independent* of the simulator $f(\theta)$ used in the moment condition (7). For instrument construction, use of common draws from γ to simulate the numerator and denominator of $(\partial P_C/\partial\theta)/P_C$ at each observation may improve the efficiency of the instruments.

In general, starting from any $K \times mN$ array of instruments Z^0, a standard argument from nonlinear generalized least squares shows that the asymptotic minimum variance estimator in the class using linear combinations of instruments in Z^0 is attained by $W = P_\theta(\theta^*)'Z'(Z\hat{P}Z')^{-1}Z$, where $P_\theta(\theta^*)$ is the $mN \times k$ array of derivatives $\partial P_C(i|\theta, X_{Cn})/\partial\theta$, evaluated at θ^*, $\hat{P} = \operatorname{diag} P(\theta^*)$, and

$$Z_{in} = Z^0_{in} - \sum_{j \in C} Z^0_{jn} P_C(j|\theta^*, X_{Cn}).$$

If $Z^0 = \partial \ln P(\theta^*)/\partial\theta$, then $W = \partial \ln P(\theta^*)/\partial\theta$. Approximations to θ^* and to the functions P and P_θ yield approximations to the minimum variance instruments that can be constructed from Z^0.

[4] A program to perform these calculations, in either GAUSS or FORTRAN, is available from the author.

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 (11) and (12) evaluated at $\hat{\theta}$, and third to carry out one or more iterations using these approximately ideal instruments. Good candidates for crude instruments are low-order polynomials in the explanatory variables: If the model is identified, then there are always polynomials in X_{C-i} that have an asymptotic correlation matrix with $\partial P_C(i|\theta, X_C)/\partial\theta$ that is of full rank. For example, in the MNP case, X_{C-i} and $X_{C-i}X'_{C-i}$ will usually be adequate first-round instruments for $\partial \ln P_C(i|\theta, X_C)/\partial\beta$ and $\partial \ln P_C(i|\theta, X_C)/\partial\Gamma$; from Lemma 10, these are a superset of the ideal instruments 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. In the third step when smooth simulators are being used, one Newton-Raphson iteration from $\hat{\theta}$ achieves the maximum asymptotic efficiency attainable from the second-round instruments.

Consider iterative algorithms for calculation of MSM estimators. When smooth simulators are used for $f(\theta)$ in (11), 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

$$(19) \qquad (d - f(\theta))'W'W(d - f(\theta)).$$

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

When a simple frequency simulator is used, (19) is piecewise constant in θ, and nonlocal 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 are made per observation, and that starting from a trial θ_0, a search direction $\Delta\theta$ has been determined. Consider (19) as a function of $\theta_0 + \lambda\,\Delta\theta$, with λ a step size to be determined. The value λ_{nj} at which there is a jump in (19) from draw j and observation n is easily calculated. Then, it is practical to enumerate the values of (19) at all the jump points λ_{nj} and choose a global minimum along this search direction.

Generally, iteration using smooth simulators is faster than 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. In such cases, a frequency simulator $f(\theta)$ with r repetitions will be nonzero 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 gives half 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.

Estimators for the Asymptotic Covariance Matrix

A standard result, obtained by expansion of the first-order condition, is that the asymptotic covariance matrix of a classical method of moments estimator is

$$\Sigma_{mm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{mm}\bar{R}(\bar{R}'\bar{R})^{-1},$$

where

$$\bar{R} = \lim_{N\to\infty} N^{-1}WP_\theta(\theta^*),$$

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

with $W_{\cdot n} = \sum_{i\in C}P_C(i|\theta^*, X_{Cn})W_{in}$. The asymptotic covariance matrix of a CAN method of simulated moments estimator has an almost identical form,

$$\Sigma_{sm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{sm}\bar{R}(\bar{R}'\bar{R})^{-1}, \tag{20}$$

where $G_{sm} = G_{mm} + G_{ss}$, with

$$G_{ss} = \lim_{n\to\infty} N^{-1}\sum_{n=1}^{N}\sum_{i,j\in C} W_{in}W_{jn}'E(Y_{in}Y_{jn}),$$

and $Y_{in} = f_{in}(\theta^*) - P_{Cn}(i|\theta^*, X_{Cn})$. The term G_{ss} is the contribution of the simulation to the asymptotic variance.

If $f(\theta)$ is the simple 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.

Consider estimation of Σ_{sm}. I establish in Lemma 9 that a consistent estimator of G_{sm} is

$$\hat{G}_{sm} = N^{-1}\sum_{n=1}^{N}\sum_{i,j\in C} W_{in}\big(d_{in} - f(i|\theta_{sm}, X_{Cn})\big)\big(d_{jn} - f(j|\theta_{sm}, X_{Cn})\big)W_{jn}'. \tag{21}$$

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

$$\hat{R} = N^{-1}W\hat{P}_\theta, \tag{22}$$

where $\hat{P}_\theta$ is an unbiased simulator of the array P_θ from (12), evaluated at θ_{sm}. The simulator $\hat{P}_\theta$ must be independent of any simulation used in the construction of W. The consistency of $\hat{R}$ continues to hold with any consistent estimator $\hat{\theta}$ in place of θ_{sm}.

5. DISCRETE PANEL DATA WITH AUTOREGRESSIVE ERRORS

Consider longitudinal discrete response data (d_{tn}, x_{tn}) for subjects $n = 1, \ldots, N$ observed over $t = 1, \ldots, T$ 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 could generate such data is

$$u_{tn} = \beta x_{tn} + \varepsilon_{tn},$$
$$d_{tn} = \operatorname{sign}(u_{tn}),$$

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. Define λ^2 to be the proportion of the variance in the autoregressive error, and ρ the serial correlation. The probability of the observed response is

$$P(d_n|X_n, \beta, \lambda, \rho) = \Pr\left(\{\xi_n, \nu_{1n}, \ldots, \nu_{Tn} | d_{tn} \cdot \operatorname{sign}(\beta x_{tn} + \xi_n + \nu_{tn}) > 0\}\right),$$

where $d_n = (d_{1n}, \ldots, d_{Tn})$ and $X_n = (x_{1n}, \ldots, x_{Tn})$. If ε_{tn} is stationary, with variance normalized to one, then

$$\varepsilon_{tn} = (1-\lambda^2)^{1/2}\eta_{0n} + \lambda\left((1-\rho^2)\sum_{j=0}^{t-2}\rho^j\eta_{t-j,n} + \rho^{t-1}\eta_{1n}\right), \tag{23}$$

where the η_{jn} are independent standard normal variates.

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 method that may increase efficiency. A simple frequency simulator or a kernel-smoothed frequency simulator with a polynomial kernel, with a small number of repetitions, requires simulation of the instruments for a small number of alternatives per subject, even for large T.

Hajivassiliou and McFadden (1987) have investigated another approach to this problem which consistently simulates the score $\partial \ln P(d_n|\theta, X_n)/\partial\theta$ for each observation. Estimators with good relative efficiency can then be obtained by iterating to a root of the sample score. A first method for this simulation is to use panel data analogues of (11) and (12), with many Monte Carlo repetitions, to estimate consistently the denominator and numerator of $[\partial P(d_n|\theta, X_n)/\partial\theta]/P(d_n|\theta, X_n)$.

A second *acceptance/rejection* method yields *unbiased* simulators of the score. As in the case of the basic MSM estimator, this permits consistent estimation of θ^* through the operation of the law of large numbers over observations. Consistent simulation of the score for each observation is not required. Let $g_U(u|\theta, X_n)$ denote the density of $(u_{1n}, \ldots, u_{Tn})$ induced by the model, given the density of the disturbances. Then, the response probability and the score satisfy

$$P(d_n|\theta, X_n) = \int_{Q_n} g_U(u|\theta, X_n)\, du,$$

$$\partial \ln P(d_n|\theta, X_n)/\partial\theta = \int_{Q_n} \left[\partial g_U(u|\theta, X_n)/\partial\theta\right] du / P(d_n|\theta, X_n)$$

$$= \int_{Q_n} \left[\partial \ln g_U(u|\theta, X_n)/\partial\theta\right] g_U(u|\theta, X_n, Q_n)\, du$$

where $Q_n = \{u|\text{sign}(u) = d_n\}$, and $g_U(u|\theta, X_n, Q_n)$ is the *conditional* density of u given Q_n. The acceptance-rejection algorithm starts with a density $\gamma(u)$ such that (i) it is easy to draw Monte Carlo samples from $\gamma(u)$, and (ii) there is a known constant M such that $M \geqslant \sup_{Q_n} g_U(u|\theta, X_n)/\gamma(u)$. Draw a pair of points (u, y) with u from $\gamma(u)$ and, independently, y from the uniform density on $[0, M]$. If $u \in Q_n$ and $y \leqslant g_U(u|\theta, X_n)$, accept the variate u; otherwise, reject this pair and draw again. The density of the accepted u is $g_U(u|\theta, X_n, Q_n)$; this is verified by an application of Bayes law. Averaging $\partial \ln g_U(u|\theta, X_n)/\partial\theta$ over the accepted u yields an unbiased simulator of the score. The main drawback of this method is that many draws may be necessary to get accepted u.

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 (23) 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 MSM 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

$$\begin{aligned} u_{tn} &= \beta x_{tn} + \psi d_{t-1,n} + \xi_n + \nu_{tn}, \\ d_{tn} &= \text{sign}(u_{tn}), \end{aligned} \tag{24}$$

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

$$P(d_n|x_n, d_{0n}, \beta, \xi_n) = \prod_{t=1}^{T} \Phi\left(d_{tn}(\beta x_{tn} + \psi d_{t-1,n} + \xi_n)\right). \tag{25}$$

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* $G^{-1}(p|d_{0n})$ of the cumulative distribution function $p = G(\xi_n|d_{0n})$ is placed in a flexible parametric family that spans the true inverse distribution. Then the response probabilities are given by the expectation of (25) with respect to ξ, which can be simulated economically from G^{-1}. Adding serial correlation to the disturbances ν_{tn} in (24) makes (25) 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 $n = 1, \ldots, N$ satisfies a latent variable model

$$\begin{aligned} u_n &= \beta z_n + \varepsilon_n, \\ d_n &= H(u_n), \end{aligned} \tag{26}$$

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

$$x_n = z_n \Lambda + \xi_n,$$

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

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

If the $\varepsilon \sim N(0, \Omega)$ in (26), then

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

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 (26) may be encountered in applications, including variables measured with error that are common to several alternatives or interact with alternative-specific dummies, multiple variables measured with possibly correlated errors, and simultaneity between the latent variables and observed indicators. These may alter the details of the density of z given x and the density of u, 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. NONNORMAL 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 nonnegative, giving monotonicity. This could be modeled by taking the density of α to be multivariate truncated normal, or by giving some components nonnegative 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. STATISTICAL PROPERTIES OF MSM ESTIMATORS

The conditions under which MSM estimators are consistent asymptotically normal (CAN) are stated formally and proved in this section. I use the following notation, mostly collected from Sections 2 and 3 of the paper:

$C = \{1, \ldots, m\}$: the set of possible responses.

$u_i = \alpha x_i$ or $u_C = \alpha X_C$: a latent variable model, $i \in C$, with $X_C = (x_1, \ldots, x_m)$ a $\bar{K} \times m$ array, $u_C = (u_1, \ldots, u_m)$, $\alpha = a(\theta, \eta)$ a $1 \times \bar{K}$ 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. Let $X_{C-i} = (x_1 - x_i, \ldots, x_{i-1} - x_i, x_{i+1} - x_i, \ldots, x_m - x_i)$, and $u_{C-i} = (u_1 - u_i, \ldots, u_{i-1} - u_i, u_{i+1} - u_i, \ldots, u_m - u_i)$.

$p(X_C)$: density for X_C, with associated measure p.

$g_U(u_{C-i} | \theta, X_C)$: the density of u_{C-i} induced by the transformation $u_{C-i} = a(\theta, \eta) X_{C-i}$ of the random vector η.

$d_i = 1(u_{C-i} \leqslant 0)$: observed response (= indicator for maximum u_i).

$P_C(i|\theta, X_C)$: probability of η such that $a(\theta, \eta)X_C$ is maximized at i, given X_C; the *discrete response probability*.
$f_C(i|\theta, X_C)$: a simulator for $P_C(i|\theta, X_C)$.
W_i: $K \times 1$ instrument vector, determined as a function $W_i = w_i(\theta, X_C)$, with $K \geqslant k$.
$n = 1, \ldots, N$: 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 $\|W(\theta_0)(d - f(\theta_{sm}))\| \leqslant \inf_\theta \|W(\theta_0)(d - f(\theta))\| + O(1)$, and θ_0 either prespecified or estimated.
$R_N(\theta)$, $R(\theta)$, $\bar{R}$: $K \times k$ array of sample covariances, $R_N(\theta) = WP_\theta(\theta)/N$, $R(\theta) = \lim R_N(\theta)$, $\bar{R} = R(\theta^*)$.

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 $\bar{K}(m-1)$ dimensional space. Further, α may be defined without loss of generality to have a compact range: Given a latent variable model $u_C = \alpha X_C$, the transform $\tilde{\alpha} = \alpha/(1 + \|\alpha\|)$ has a range contained in $[-1, 1]^{\bar{K}}$, and the model $\tilde{u}_C = \tilde{\alpha} X_C$ yields the same response probabilities.

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:

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

ASSUMPTION A2: *The domain $\mathbb{X}$ of the attributes X_C is a compact subset of a $\bar{K}(m-1)$ dimensional space.*

ASSUMPTION 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 $\Theta \times \mathbb{N}$.*

ASSUMPTION A4: *For an open set $\mathbb{X}_0 \subseteq \mathbb{X}$ with $p(\mathbb{X}_0) = 1$, 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*

$$X_C = \begin{pmatrix} 0 & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

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*

$$(27) \qquad [\alpha^1 \quad \alpha^2] = [\beta^1(\theta) \quad \beta^2(\theta)] + [\eta^1 \quad \eta^2]\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \Gamma_{22} \end{bmatrix}.$$

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. *Q.E.D.*

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

ASSUMPTION 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* $\Theta \times \mathbb{X}$.

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

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 $P_C(1|\theta, X_C) = \Pr(\alpha X_C \leqslant 0 | X_C)$. Using the decomposition of Lemma 1, $\alpha X_C = [0, s] \leqslant 0$ implies

$$(28) \qquad \eta^2 = [s - \beta^1 A_{12} - \beta^2 A_{22} - \eta^1(\Gamma_{11}A_{12} + \Gamma_{12}A_{22})](\Gamma_{22}A_{22})^{-1}.$$

Then, given η^1, the set $B(\eta^1)$ of η^2 satisfying $\alpha X_C \leqslant 0$ 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. *Q.E.D.*

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

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

ASSUMPTION A7: *The instruments identify θ^*, with*

$$\omega(\theta, \tilde{\theta}) \equiv \int_{\mathbb{X}} \sum_{i \in C} w_i(\tilde{\theta}, X_C)\left[P_C(i|\theta, X_C) - P_C(i|\theta^*, X_C)\right] dp(X_C) \tag{29}$$

equal to zero if and only if $\theta = \theta^*$, for any $\tilde{\theta} \in \Theta$.

ASSUMPTION A8: *$\bar{R}$ is of maximum rank.*

To satisfy A6, instruments constructed by simulation require the use of smooth simulators such as (11) and (12). If A5 holds and the instruments equal the score of the log likelihood evaluated at each trial θ, $w_i(\theta, X_C) \equiv \partial \ln P_C(i|\theta, X_C)/\partial\theta$, then $\tilde{\theta} = \theta$ and ω reduces to

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

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. If crude instruments independent of θ are used to obtain an initially consistent estimator, then ω in (29) 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^*.

The next assumptions concern the simulator $f_C(i|\theta, X_C)$:

ASSUMPTION 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 $g(\eta)$. The simulator $f_C(i|\theta, X_{Cn})$ is a uniformly bounded function of θ, X_{Cn} and $(\eta_{1n}, \ldots, \eta_{rn})$.*

ASSUMPTION A10: *For the simulator $f_C(i|\theta, X_{Cn})$, the simulation bias converges to zero uniformly in θ as $N \to \infty$; i.e., $B(\theta) = N^{-1/2}W(Ef(\theta) - P(\theta))$, where the*

expectation is conditioned on W and d, satisfies

$$\sup_{\theta\in\Theta} |B(\theta)| = o(1). \tag{8}$$

ASSUMPTION A11: *For the simulator* $f_C(i|\theta, X_{Cn})$, *the simulation residual process* $\zeta(\theta) = N^{-1/2}W(f(\theta) - Ef(\theta))$ *is uniformly stochastically bounded and equicontinuous in* θ; *i.e.*,

$$\sup_{\theta\in\Theta} |\zeta(\theta)| = O_p(1); \tag{9}$$

and for each $\delta > 0$, *defining* $A_N = \{\theta | N^{1/2}|\theta - \theta^*| \leq \delta\}$,

$$\sup_{\theta\in A_N} |\zeta(\theta) - \zeta(\theta^*)| = o_p(1). \tag{10}$$

The main result on the asymptotic properties of MSM estimators is given in the theorem below. Following the proof of this theorem, I examine the conditions under which various frequency simulators, including the simple frequency simulator and smooth simulators, will satisfy A11.

THEOREM 1: *Suppose the MSM estimator* θ_{sm} *defined by* (7) *satisfies Assumptions* A1 *to* A11. *Then* θ_{sm} *is consistent, with* $N^{1/2}(\theta_{sm} - \theta^*)$ *converging in distribution to a normal vector with mean zero and covariance matrix* $\Sigma_{sm} = (\bar{R}'\bar{R})^{-1}\bar{R}'G_{sm}\bar{R}(\bar{R}'\bar{R})^{-1}$, *with* $\bar{R} = \lim N^{-1}WP_\theta(\theta^*)$ *and* $G_{sm} = \lim N^{-1}EW(d - f(\theta^*))(d - f(\theta^*))'W'$.

PROOF: The argument parallels that of Pakes and Pollard (1989). The vector $W(d - f(\theta))$ entering the defining condition (7) for the MSM estimator can be decomposed into four terms,

$$\begin{aligned} N^{-1/2}W(d-f(\theta)) \equiv\ & [\zeta(\theta^*) - \zeta(\theta)] - B(\theta) \\ & + \left[N^{-1/2}W(d-f(\theta^*)) + B(\theta^*)\right] \\ & - \left[N^{-1/2}W(P(\theta) - P(\theta^*))\right]. \end{aligned} \tag{30}$$

The asymptotic properties of the estimator are argued by applying conditions (8) to (10) to the first two terms in (30), 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 and the independence of the simulators across observations, plus the implication from A6 that $W(d - f(\theta))$ is uniformly bounded, imply by the Lindeberg-Feller central limit theorem that this expression converges in distribution to a multivariate normal vector Z with mean zero and covariance matrix G_{sm}.

[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. To establish this result, note that continuous differentiability and compactness imply $|\omega_N(\theta) - \omega_N(\tilde{\theta})| \leq M|\theta - \tilde{\theta}|$, where $M \geq \max_\Theta |R(\theta)|$. 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 $N > N_\varepsilon$, $\Pr\{\max_{\Theta_\varepsilon} |\omega_N(\theta) - \omega(\theta)| > \varepsilon/3\} < \varepsilon$. By construction of Θ_ε, for each $\theta \in \Theta$, there is a $\tilde{\theta} \in \Theta_\varepsilon$ such that $|\omega_N(\theta) - \omega(\theta)| \leq |\omega_N(\tilde{\theta}) - \omega(\tilde{\theta})| + 2\varepsilon/3$. Hence, $\Pr\{\max_\Theta |\omega_N(\theta) - \omega(\theta)| > \varepsilon\} < \varepsilon$. Regularity condition A8 implies $R(\theta^*)$ nonsingular.

Consider first the consistency of θ_{sm}. Argument [a] implies $N^{-1/2}W(d - f(\theta^*)) = O_p(1)$. Hence, (7) satisfies

$$
\begin{aligned}
(31) \qquad & N^{-1}(d - f(\theta_{sm}))'W'W(d - f(\theta_{sm})) \\
& \leq N^{-1} \inf_\Theta (d - f(\theta))'W'W(d - f(\theta)) + O(N^{-1}) \\
& \leq \left[N^{-1/2}W(d - f(\theta^*))\right]'\left[N^{-1/2}W(d - f(\theta^*))\right] + O(N^{-1}) \\
& = O_p(1),
\end{aligned}
$$

implying $N^{-1/2}W(d - f(\theta_{sm})) = O_p(1)$. Then, multiplying (30) by $N^{-1/2}$ and using (8), (9), 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, proving consistency.

Next, I argue that $N^{1/2}(\theta_{sm} - \theta^*)$ is stochastically bounded. From (30), the condition $N^{-1/2}W(d - f(\theta^*)) = O_p(1)$ from argument [a], (8), (9), and (31) implies

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

A Taylor's expansion[5] yields

$$
\begin{aligned}
& N^{-1/2}W\big(P(\theta_{sm}) - P(\theta^*)\big) \\
& \qquad = \left[N^{-1}WP_\theta(\theta^*) + O(\theta_{sm} - \theta^*)\right] N^{1/2}(\theta_{sm} - \theta^*).
\end{aligned}
$$

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

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

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, (10) implies $\zeta(\theta^*) - \zeta(\tilde{\theta}) = o_p(1)$. Substituting $\tilde{\theta}$ in (30) and applying the argument above

[5] In this and following proofs, a Taylor's expansion of a vector of functions will mean a component by component expansion, with remainders representable as expansion terms evaluated at intermediate arguments that differ across components.

with $\tilde{\theta}$ in place of θ_{sm} implies

$$(32)\qquad N^{-1/2}W\big(d-f(\tilde{\theta})\big) = Z - \big[\overline{R} + O(\tilde{\theta}-\theta^*)\big]N^{1/2}(\tilde{\theta}-\theta^*) + o_p(1)$$
$$= \Big[I - \overline{R}(\overline{R}'\overline{R})^{-1}\overline{R}'\Big]Z + o_p(1).$$

From (32), (10), argument [b], a Taylor's expansion of $P(\tilde{\theta}) - P(\theta_{sm})$, and the result that $N^{-1}WP_\theta(\theta_{sm}) = \overline{R} + o_p(1)$,

$$(33)\qquad \Delta \equiv N^{-1/2}W\big(f(\tilde{\theta}) - f(\theta_{sm})\big)$$
$$= N^{-1/2}W\big(P(\tilde{\theta}) - P(\theta_{sm})\big) + \zeta(\tilde{\theta}) + B(\tilde{\theta}) - \zeta(\theta_{sm}) - B(\theta_{sm})$$
$$= N^{-1/2}W\big(P(\tilde{\theta}) - P(\theta_{sm})\big) + o_p(1) = \overline{R}N^{1/2}\big(\tilde{\theta}-\theta_{sm}\big) + o_p(1)$$
$$= \overline{R}(\overline{R}'\overline{R})^{-1}\overline{R}'Z + o_p(1).$$

Rewrite condition (7) characterizing θ_{sm} as

$$(34)\qquad N^{-1}\big(d-f(\tilde{\theta})\big)'W'W\big(d-f(\tilde{\theta})\big) + o_p(1)$$
$$\geqslant N^{-1}\big(d-f(\theta_{sm})\big)'W'W\big(d-f(\theta_{sm})\big)$$
$$= N^{-1}\big(d-f(\tilde{\theta})\big)'W'W\big(d-f(\tilde{\theta})\big)$$
$$+ 2N^{-1/2}\big(d-f(\tilde{\theta})\big)'W'\Delta + \Delta'\Delta.$$

From (32) and (33), $N^{-1/2}(d-f(\tilde{\theta}))'W'\Delta = o_p(1)$, and (34) implies $\Delta'\Delta = o_p(1)$. But then $\Delta \equiv \overline{R}N^{1/2}(\tilde{\theta}-\theta_{sm}) + o_p(1) = o_p(1)$, implying θ_{sm} and $\tilde{\theta}$ are asymptotically equivalent. *Q.E.D.*

The next result establishes sufficient conditions for a kernel-smoothed frequency simulator to satisfy the asymptotic unbiasedness condition (8):

LEMMA 3: *Suppose the assumptions of Lemma* 1 *hold, with* A_{22} *always nonsingular. Suppose a kernel-smoothed frequency simulator with a kernel of the form* (15). *Assume the distribution function* Ψ *has a finite moment generating function in a neighborhood of the origin. Assume* $N^{\varepsilon+1/2}b_N \to 0$ *for some* $\varepsilon > 0$. *Then* A10 *holds.*

PROOF: Let $\mu(t)$ be the moment generating function of Ψ. There exists $\tau > 0$ such that $\Psi(v) \leqslant e^{\tau v}\mu(-\tau)$ for all $v < 0$ and $1 - \Psi(v) \leqslant e^{-\tau v}\mu(\tau)$ for all $v > 0$. If $c \equiv \max_{j\neq i}(y_j - y_i)/2 > 0$, then $\mathbb{K}_i(y_1, \ldots, y_m) = \int_{v\leqslant c}(\prod_{j\neq i}\Psi(v + y_i - y_j))\Psi'(v)\,dv + \int_{v>c}(\prod_{j\neq i}\Psi(v + y_i - y_j))\Psi'(v)\,dv \leqslant e^{-\tau c}\mu(-\tau) + e^{-\tau c}\mu(\tau)$, with the first term the result of bounding Ψ at negative arguments, and the second the result of bounding the probability at positive arguments. If $c < 0$, a similar argument yields $\mathbb{K}_i(y_1, \ldots, y_m) \geqslant (1 - e^{-\tau|c|}\mu(\tau))^m \geqslant 1 - me^{-\tau|c|}\mu(\tau)$. Let $\mathbb{K}(y_{C-i}) \equiv \mathbb{K}_i(y_C)$. Define $I(A) = \int_A |\mathbb{K}(u_{C-i}/b_N) - 1(u_{C-i} \leqslant 0)|g_U(u_{C-i}|\theta, X_C)\,du_{C-i}$. 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 $I(A_1) \leqslant e^{-\tau M}\mu(\tau)m$ and $I(A_2) \leqslant e^{-\tau M}(\mu(\tau)+\mu(-\tau))$. Further, $I(A_3) \leqslant \sum_{j\neq i} \Pr(|u_j - u_i| \leqslant Mb_N)$. But (28) 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, $\Pr(|u_j - u_i| \leqslant Mb_N) \leqslant 2Mb_N M_\gamma$. Therefore, $I(A_3) \leqslant 2mMb_N M_\gamma$. Then, $N^{1/2}|\tilde{P}_C(i|\theta, X_C) - P_C(i|\theta, X_C)| \leqslant N^{1/2}(I(A_1) + I(A_2) + I(A_3)) \leqslant N^{1/2}(e^{-\tau M}\mu(\tau)m + e^{-\tau M}(\mu(\tau)+\mu(-\tau)) + 2mMb_N M_\gamma)$. Choose $M = \tau^{-1}\ln N$. Then, the right-hand side of the last inequality goes to zero if $N^{1/2}(\ln N)b_N \to 0$. The condition $N^{1/2+\epsilon}b_N \to 0$ implies the required limit. *Q.E.D.*

The stochastic boundedness and equicontinuity conditions in A11 can be demonstrated for *smooth* simulators by the following argument:

LEMMA 4: *If A1 to A9 hold, and the simulator $f(\theta)$ is uniformly bounded and twice continuously differentiable, then* (9) *and* (10) *hold.*

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

$$\zeta(\theta) - \zeta(\theta^*) = \zeta_\theta(\theta^*)(\theta - \theta^*) + (1/2)\left[N^{-1/2}\zeta_{\theta\theta}\right]\operatorname{vec}\left(\left[(\theta-\theta^*)\right]\left[N^{1/2}(\theta-\theta^*)\right]'\right), \tag{35}$$

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, (35) implies, for $A_N = \{\theta | N^{1/2}|\theta - \theta^*| \leqslant \delta\}$,

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

establishing (10).

I next establish (9), 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,

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

I shall need Bernstein's inequality, which states that independent identically distributed random variables Y_i with $EY_i = 0$ and $|Y_i| \leqslant c$ satisfy

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

for $t > 0$, where $\sigma^2 = EY_i^2$; see Giné and Zinn (1986), Lemma 3.2, and also Pollard (1984) and Shorack and Wellner (1986). Replace t by $N^{1/2}t$ and use

$\sigma^2 \leqslant c^2$ to obtain

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

Let $M \geqslant 1$ 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 constant C satisfying the bound $C > 48M + 8kM \ln 2$,

$$(36)\qquad P\left\{ \sup_{\Theta} |\zeta(\theta)| > C \right\}$$
$$\leqslant P\left\{ |\zeta(\theta_1)| > C/2 \right\}$$
$$+ \sum_{i=1}^{\infty} P\left\{ \sup_{\Theta} |\zeta(\theta_{i+1}(\theta)) - \zeta(\theta_i(\theta))| > i2^{-i-3}C \right\}$$

$$(37)\qquad \leqslant P\left\{ |\zeta(\theta_1)| > C/2 \right\}$$
$$+ \sum_{i=1}^{\infty} 2^{ki} \sup_{\Theta} P\left\{ |\zeta(\theta_{i+1}(\theta)) - \zeta(\theta_i(\theta))| > i2^{-i-3}C \right\}$$

$$(38)\qquad \leqslant 2\exp\left[-C^2/4(2M^2 + MC/3)\right]$$
$$+ \sum_{i=1}^{\infty} 2^{ki} 2\exp\left[-C^2 i^2 4^{-i-3}/\left(2M^2 4^{-i} + 2M2^{-i}Ci2^{-i-3}/3\right)\right]$$

$$(39)\qquad \leqslant 2\exp\left[-C/4M \right] + \sum_{i=1}^{\infty} 2\exp\left[-iC/8M \right] \leqslant 5\exp\left[-C/8M \right].$$

The inequalities (36) and (37) hold since left-hand-side events are contained in the union of the right-hand-side events, while (38) follows by application of the Bernstein inequality, and (39) 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 (39) less than ε. This proves (9). *Q.E.D.*

The following results establish conditions under which simulators can have jumps, but these jumps are sufficiently "well-behaved" so (9) and (10) are satisfied. The simple frequency simulator in the MNP model, in particular, is shown to satisfy these sufficient conditions.

ASSUMPTION A12: *Define $\varphi_C(i|\theta, X_{Cn})$ to be the simple frequency in the r draws for observation n of the event that $a(\theta, \eta_{jn})X_{Cn}$ is maximized at component i. Assume the simulator $f_C(i|\theta, X_{Cn})$ is a uniformly bounded function of θ, X_{Cn}, and $(\eta_{1n}, \ldots, \eta_{rn})$ satisfying*

$$(40)\qquad \left| f_C(i|\theta, X_{Cn}) - f_C\left(i|\tilde{\theta}, X_{Cn}\right)\right|$$
$$\leqslant M_\varphi \left|\varphi_C(i|\theta, X_{Cn}) - \varphi_C\left(i|\tilde{\theta}, X_{Cn}\right)\right| + M_f|\theta - \tilde{\theta}|^\lambda$$

for some M_φ, M_f, $\lambda > 0$, and all θ, $\tilde{\theta} \in \Theta$ and $X_{Cn} \in \mathbb{X}$.

Condition (40) requires that the simulator be at least as smooth in θ as the simple frequency simulator. Condition (40) is satisfied trivially by either the frequency simulator, or by a smooth simulator. If the simulator is differentiable at non-jump points, then $\lambda = 1$; the assumption also allows $0 < \lambda < 1$, corresponding to "polynomial" nondifferentiability. A simulator satisfying (40) will be termed λ-Lipschitz in neighborhoods without jumps.

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 $|W(d - f(\theta_{sm}))| \leq mM_w M_\varphi / r$. Then, any bounded sequence $a_N = O(1)$ on the right-hand side of the definition (7) of a MSM estimator that satisfies $a_N \geq (mM_w M_\varphi / r)^2$ implies existence of the estimator.

LEMMA 5: *Suppose* A1–A9 *and* A12. *Then, almost surely,* $W(d - f(\theta))$ *is uniformly* λ*-Lipschitz in* θ *except for a closed subset* Θ_0 *of* Θ *with Lebesgue measure zero, and the jumps in this function on* Θ_0 *are bounded by* $mM_w M_\varphi / r$.

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

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

and hence

$$0 = \int_{\Theta} \int_{\mathbb{N}} \int_{\mathbb{X}} I(\theta, X_C, \eta)\, dg(\eta)\, dp(X_C)\, d\theta. \tag{41}$$

Applying Fubini's Theorem to (41), there exists a set $\mathbb{X}_1 \subseteq \mathbb{X}$ with probability measure one; for $X_C \in \mathbb{X}_1$, a set $\mathbb{N}(X_C) \subseteq \mathbb{N}$ with probability measure one; and, for $(X_C, \eta) \in \mathbb{X}_1 \times \mathbb{N}(X_C)$, 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, $X_{Cn} \in \mathbb{X}_1$ 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 A12, $f_C(i|\theta, X_C)$ 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 (40), the discontinuity in $|W(d - f(\theta))|$ is at most $mM_w M_\varphi / r$ with probability one. *Q.E.D.*

Assumption A4 implies 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). Define the

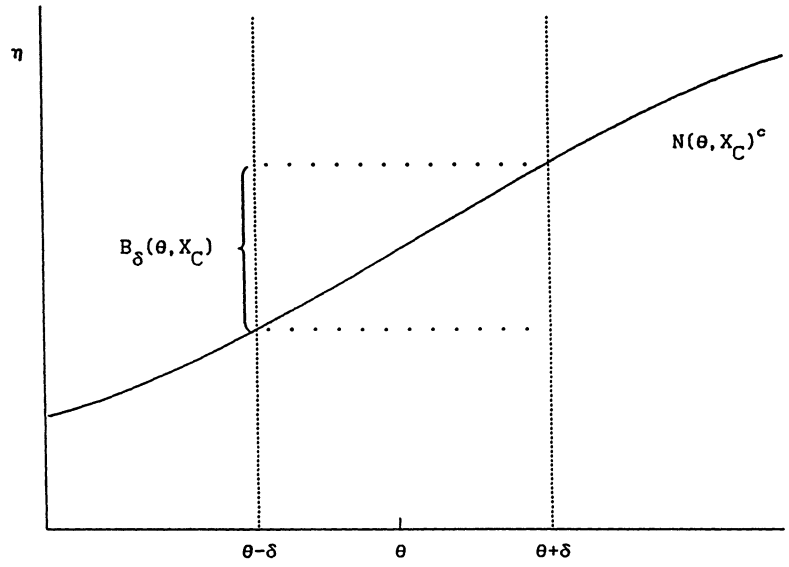

FIGURE 1

set

$$(42) \qquad B_\delta(\theta, X_C) = \left\{ \eta \mid \eta \in \mathbb{N}\left(\tilde{\theta}, X_C\right)^c \text{ for some } |\tilde{\theta} - \theta| \leqslant \delta \right\}.$$

Figure 1 illustrates the construction of $B_\delta(\theta, X_C)$. The following argument establishes that $B_\delta(\theta, X_C)$ is closed, and hence measurable, for $(\theta, X_C) \in \Theta \times \mathbb{X}_0$: If $\eta^j \in B_\delta$ and $\eta^j \to \eta^0$, then $\eta^j \in \mathbb{N}(\theta^j, X_C^j)^c$, 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), $\eta^0 \in \mathbb{N}(\theta^0, X_C^0)^c$ for each limit point (θ^0, X_C^0) of (θ^j, X_C^j).

ASSUMPTION A13: *There exists* M_g *and* $\lambda > 0$ *such that for* $X_C \in \mathbb{X}_0$ *and almost all* $\theta \in \Theta$, *the set* $B_\delta(\theta, X_C)$ *has* $g(B_\delta(\theta, X_C)) \leqslant M_g \delta^\lambda$.

The assumption requires that the measure of the set of η yielding ties for θ in a δ-neighborhood shrink toward zero at a "polynomial rate" as δ goes to zero. This condition 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 $a(\theta, \eta) = \beta(\theta) + \eta\Gamma(\theta)$ satisfies A13; this assumption then holds in particular for the MNP model.

LEMMA 6: *Suppose the hypotheses of Lemma* 1, *with* A_{22} *always nonsingular, and* A6–A9. *Then* A13 *holds.*

PROOF: Suppose a tie between alternatives 1 and 2, so $\alpha x_2 = 0$. Using the notation of (27) and (28), partition $\alpha_1 = (\alpha_1, \alpha_3, \ldots, \alpha_m)$ and let α_2 denote the

second component. Then,

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

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

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

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 $\Theta \times \mathbb{X}$ implies there exists a constant M_ψ such that for $|(\theta, X_C) - (\tilde{\theta}, \tilde{X}_C)| \leqslant \delta$,

$$\left|\psi(\tilde{\theta}, \tilde{X}_C, \eta^1) - \psi(\theta, X_C, \eta^1)\right| \leqslant M_\psi(1 + |\eta^1|)\,\delta.$$

Then the set $\mathbb{N}_2(\tilde{\theta}, \tilde{X}_C, \eta^1) = \{\eta^2 | |\eta^2 - \psi(\tilde{\theta}, \tilde{X}_C, \eta^1)| \leqslant M_\psi(1 + |\eta^1|)\,\delta\}$ contains all η^2 solving $\eta^2 = \psi(\tilde{\theta}, \tilde{X}_C, \eta^1)$ for $|(\theta, X_C) - (\tilde{\theta}, \tilde{X}_C)| \leqslant \delta$, and satisfies

$$\int_{\eta^1} g_1(\eta^1)\,d\eta^1 \int_{\mathbb{N}_2(\theta, X_C, \eta^1)} g_{2.1}(\eta^2|\eta^1)\,d\eta^2$$
$$\leqslant M_\psi(1 + E|\eta^1|)\,\delta M_\gamma \equiv 2M_g\delta/m(m-1),$$

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 A13. *Q.E.D.*

Given $\varepsilon > 0$, a finite family of random functions F_ε is said to *bracket* a family of random functions F if for each $Y \in F$ there exist $\underline{Y}, \overline{Y} \in F_\varepsilon$ such that $\underline{Y} \leqslant Y \leqslant \overline{Y}$ and $E(\overline{Y} - \underline{Y}) < \varepsilon$. The logarithm of the number of elements in the smallest set F_ε that brackets F, denoted $H(\varepsilon)$, is called *metric entropy with bracketing*. The following result establishes stochastic equicontinuity conditions for families whose metric entropy does not rise too rapidly as ε falls.

LEMMA 7: *Assume F is a uniformly bounded family of measurable random functions. Assume F satisfies $\int_0^1 H(\varepsilon^2)^{1/2}\,d\varepsilon$ finite, where H is the metric entropy with bracketing. Suppose $Y \Rightarrow y_1$, $Y \Rightarrow y_2, \ldots$ denote independent identically distributed realizations of deviations from the mean, $Y - EY$, for $Y \in F$. Define $\|Y\| = E|Y - EY|$. Then for every $\lambda > 0$,*

$$\lim_{\delta \to 0} \limsup_{N \to \infty} \Pr\left\{ \sup_{\substack{Y \in F \\ Y \Rightarrow y_n}} \sup_{\substack{Y' \in F \\ \|Y' - Y\| \leqslant \delta \\ Y' \Rightarrow y'_n}} N^{-1/2} \left| \sum_{n=1}^{N} (y_n - y'_n) \right| > \lambda \right\} = 0. \tag{43}$$

PROOF: Dudley (1984), Theorem 6.2.1, establishes (43). I use a restatement from Alexander (1987), Theorem 2.1. *Q.E.D.*

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 7.

LEMMA 8: *Suppose A1–A9, A12–A13. Then A11 holds.*

PROOF: 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 $X_C \in \mathbb{X}_0$, let $\Theta_j(X_C)$ be a set containing one point selected from each cube that intersects Θ. By A13, the selection can be made so that $g(B_\delta(\theta, X_C)) \leqslant M_g \delta^\lambda$ for $\theta \in \Theta_j(X_C)$. Define $\theta_j(\theta) \equiv \theta_j(\theta, X_C)$ to be a point in $\Theta_j(X_C)$ nearest to θ; then $|\theta - \theta_j(\theta)| \leqslant 2^{-j} \equiv \delta_j$.

Let $Y(\theta, X_C) = \Sigma_{i \in C} W_i f_C(i|\theta, X_C)$. 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 (40), and λ satisfying A12 and A13, define

$$Y_j^0(\theta, X_C) = mM_w\Big(M_f\big|\theta - \theta_j(\theta)\big|^\lambda + M_\varphi q_j(\theta, X_C)/r\Big).$$

Then, by A13, $\mathrm{Eq}_j(\theta, X_C) \leqslant rg(B_{\delta_j}(\theta, X_C))$, implying

$$\begin{aligned} EY_j^0(\theta, X_C) &\leqslant mM_w M_f\big|\theta - \theta_j(\theta)\big|^\lambda + mM_w M_\varphi g\big(B_{\delta_j}(\theta, X_C)\big) \\ &\leqslant mM_w\big(M_f + M_\varphi M_g\big)\,\delta_j^\lambda \equiv M_0 \delta_j^\lambda. \end{aligned}$$

From (40),

$$\begin{aligned} &\big|Y(\theta, X_C) - Y\big(\theta_j(\theta), X_C\big)\big| \\ &\quad\leqslant mM_w\Big(M_f\big|\theta - \theta_j(\theta)\big|^\lambda\Big) + mM_w M_\varphi \max_{i \in C}\big|\varphi_C(i|\theta, X_C) \\ &\qquad\qquad - \varphi_C\big(i|\theta_j(\theta), X_C\big)\big| \\ &\quad\leqslant mM_w\Big(M_f\big|\theta - \theta_j(\theta)\big|^\lambda + M_\varphi q_j(\theta, X_C)/r\Big). \end{aligned}$$

Hence, $\underline{Y}_j(\theta, X_C) \equiv Y(\theta_j(\theta), X_C) - Y_j^0(\theta, X_C) \leqslant Y(\theta, X_C) \leqslant \overline{Y}_j(\theta, X_C) \equiv Y(\theta_j(\theta), X_C) + Y_j^0(\theta, X_C)$. 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 $H(\varepsilon) \leqslant (kj+1)\ln 2 \leqslant (-\ln \varepsilon)k/\lambda + (k+1)\ln 2$, and hence $\int_0^1 H(\varepsilon^2)^{1/2}\,d\varepsilon \leqslant \int_0^1 H(\varepsilon^2)\,d\varepsilon \leqslant (k+1)\ln 2 - 2(k/\lambda)\int_0^1 \ln \varepsilon\, d\varepsilon < \infty$. This establishes the assumptions of Lemma 7, so (43) holds.

For any $\delta > 0$, forming the expectation of (40) and using A12, $|\theta - \tilde{\theta}| < \delta$ implies $E|Y(\theta, X_C) - Y(\tilde{\theta}, X_C)| < mM_w(M_f + M_\varphi M_g)\,\delta \equiv M_0 \delta$. Hence, (43) can be written in the form

$$\lim_{\delta \to 0} \limsup_{N \to \infty} \Pr\big\{\big|\zeta(\theta) - \zeta(\tilde{\theta})\big| > \lambda\big\} = 0. \tag{44}$$

Taking $\delta = 2^{-j}$, $j = 1, 2, \ldots$, one has $|\theta - \tilde{\theta}| < \delta$ for $N \geq 4^j$ and $N^{1/2}|\theta - \tilde{\theta}| < 1$. Then (44) implies (10).

Next prove (9). From (44), given $\lambda > 0$, there exists $\delta > 0$ such that

$$\limsup_{N \to \infty} \Pr\left\{ \sup_{\theta \in \Theta} \sup_{\substack{\tilde{\theta} \in \Theta \ \& \\ |\theta - \tilde{\theta}| \leq \delta}} |\zeta(\theta) - \zeta(\tilde{\theta})| > \lambda \right\} < \lambda.$$

Choose any $\theta_0 \in \Theta$. A central limit theorem implies there exists M_0 such that $\sup_N \Pr\{|\zeta(\theta_0)| > M_0\} < \lambda$. But any $\theta \in \Theta$ can be written $\theta = \theta_0 + \sum_{j=0}^{J}(\theta_{j+1} - \theta_j)$ with $\theta_j = (j/J)\theta + (1 - j/J)\theta_0$ and J the smallest integer exceeding $1/\delta$. Then,

$$\Pr\{|\zeta(\theta)| > M_0 + \lambda J\}$$

$$\leq \Pr\{|\zeta(\theta_0)| > M_0 \ \& \ |\zeta(\theta_{j+1}) - \zeta(\theta_j)| < \lambda, \ j = 0, \ldots, J\} < \lambda.$$

Then (9) holds. *Q.E.D.*

In this lemma, the construction holds 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 nonstochastic, or a sequence of initially consistent estimators, or might equal the MSM estimator θ_{sm}. In the last case, θ_{sm} solves

$$\|W(\theta_{sm})(d - f(\theta_{sm}))\| \leq \inf_{\theta} \|W(\theta_{sm})(d - f(\theta))\| + O(1).$$

Lemma 8 holds in all these cases.

The next result establishes the consistency of the estimators (21) and (22) of the components of the asymptotic covariance matrix of the MSM estimator.

LEMMA 9: *Assume A1–A11. Assume $\hat{P}_\theta(\theta)$ is a simulator of $P_\theta(\theta)$ that is twice continuously differentiable in θ, such as the smooth simulator* (12) *with the density γ chosen so that the function $h_\theta(u_{C-i}, \theta, X_{C-i})$ is dominated by an integrable function (i.e., some $H(u_{C-i}, X_{C-i})$ satisfies $|h_\theta| \leq H$ and $\int H(u_{C-i}, X_{C-i})\gamma(u_{C-i})\, du_{C-i}$ finite). Then the estimator $\hat{\Sigma}_{sm}$ for Σ_{sm} given in* (21) *is consistent, as is the estimator $\hat{R} = N^{-1}W\hat{P}_\theta$ for $\bar{R}$ given in* (22), *with $\hat{P}_\theta$ evaluated at θ_{sm} or any consistent estimator of θ^*.*

PROOF: To show (21), 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 (8)–(10), terms involving the difference of $f(\theta^*)$ and $f(\theta_{sm})$ are $o_p(1)$. The argument for (22) is the same, but it is necessary to use versions of (8)–(10) for $\hat{P}_\theta$. These hold for smooth simulators by Lemma 4. *Q.E.D.*

Expressions for the derivatives of the response probabilities with respect to the parameters θ, in a form suitable for application of simulation methods, are needed for the construction of instruments, and consistent estimation of the MSM covariance matrix. They are also needed for Newton-Raphson type iterative search for the estimators. For multinomial probit, Paul Ruud has suggested a characterization of the derivatives of the response probability with respect to β and Γ. Applying the chain rule to $\beta(\theta)$ and $\Gamma(\theta)$ yields derivatives of the response probabilities with respect to the deep parameters θ. These equations provide a template for construction of good crude instruments for MNP. Note that these equations require simulation of only the first and second order censored moments of the multivariate distribution, which can be done efficiently using cylinder function simulators.

LEMMA 10: *Assume the MNP model, generated by the latent variable model* (1) *with* α *satisfying* (3). *Then, the derivatives of the response probabilities with respect to the parameters* β, Γ *are*

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

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

and

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

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

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

where $\Omega = \Gamma'\Gamma$, $X = X_{C-i}$, *and* $h(u, X, \theta)$ *is the ratio of the multivariate normal density to a Monte Carlo sampling distribution* $\gamma(u)$ *on the nonpositive orthant.*

PROOF: 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 $\partial \ln|A|/\partial A = A^{-1}$ 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):

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

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

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

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

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

Q.E.D.

Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Manuscript received October, 1987; final revision received September, 1988.

REFERENCES

ALEXANDER, K. (1987): "The Central Limit Theorem for Empirical Processes on Vapnic-Cervonenkis Classes," *The Annals of Probability*, 15, 178–203.

BARBERA, S., AND P. PATTANAIK (1986): "Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings," *Econometrica*, 54, 707–715.

CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North Holland, 1247–1320.

CLARK, C. (1961): "The Greatest of a Finite Set of Random Variables," *Operations Research*, 9, 145–162.

DAGANZO, C. (1980): *Multinomial Probit*. New York: Academic Press.

DEÁK, I. (1980): "Three Digit Accurate Multiple Normal Probabilities," *Numerische Mathematik*, 35, 369–380.

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

DUDLEY, R. (1984): "A Course on Empirical Processes," Ecole d'Été de Probabilités de Saint-Flour, XII-1982, *Lecture Notes in Mathematics* 1097. New York: Springer, 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, 52–72.

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

GINÉ, E., AND J. ZINN (1986): "Lectures on the Central Limit Theorem for Empirical Processes," Probability and Banach Spaces, *Lecture Notes in Mathematics*, 1221. New York: Springer, 50–113.

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*. London: Methuen.

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 *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden. Cambridge: MIT Press, 179–195.

HENDRY, D. (1984): "Monte Carlo Experimentation in Econometrics" in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North Holland, 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.

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, 305–319.

MCFADDEN, D. (1984): "Econometric Analysis of Qualitative Response Models," in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North Holland, 1395–1457.

——— (1986a): "The Choice Theory Approach to Marketing Problems," *Marketing Science*, 5, 275–297.

——— (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 (1989): "The Asymptotic Distribution of Simulation Experiments," *Econometrica*, 57, 1027–1057.

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

——— (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295–314.

PRESS, W., B. FLANNERY, S. TEUKOLSKY, AND W. VETTERLING (1986): *Numerical Recipes*. Cambridge: Cambridge University Press.

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*. New York: Wiley.

SPANIER, J., AND K. OLDHAM (1987): *An Atlas of Functions*. Washington: Hemisphere.

STERN, S. (1987): "A Method for Smoothing Simulated Moments of Probabilities in Multinomial Probit Models," University of Virginia, Working Paper.

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.

References

The Central Limit Theorem for Empirical Processes on Vapnik-Cervonenkis Classes

Kenneth S. Alexander

The Annals of Probability, Vol. 15, No. 1. (Jan., 1987), pp. 178-203.

Stable URL:

http://links.jstor.org/sici?sici=0091-1798%28198701%2915%3A1%3C178%3ATCLTFE%3E2.0.CO%3B2-D

Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings

Salvador Barberá; Prasanta K. Pattanaik

Econometrica, Vol. 54, No. 3. (May, 1986), pp. 707-715.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28198605%2954%3A3%3C707%3AFATROS%3E2.0.CO%3B2-2

The Greatest of a Finite Set of Random Variables

Charles E. Clark

Operations Research, Vol. 9, No. 2. (Mar. - Apr., 1961), pp. 145-162.

Stable URL:

http://links.jstor.org/sici?sici=0030-364X%28196103%2F04%299%3A2%3C145%3ATGOAFS%3E2.0.CO%3B2-U

LINKED CITATIONS

A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences

Jerry A. Hausman; David A. Wise

Econometrica, Vol. 46, No. 2. (Mar., 1978), pp. 403-426.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28197803%2946%3A2%3C403%3AACPMFQ%3E2.0.CO%3B2-8

Tables for Computing Bivariate Normal Probabilities

Donald B. Owen

The Annals of Mathematical Statistics, Vol. 27, No. 4. (Dec., 1956), pp. 1075-1090.

Stable URL:

http://links.jstor.org/sici?sici=0003-4851%28195612%2927%3A4%3C1075%3ATFCBNP%3E2.0.CO%3B2-2

Simulation and the Asymptotics of Optimization Estimators

Ariel Pakes; David Pollard

Econometrica, Vol. 57, No. 5. (Sep., 1989), pp. 1027-1057.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28198909%2957%3A5%3C1027%3ASATAOO%3E2.0.CO%3B2-R

Consumer Attitudes and Voluntary Rate Schedules for Public Utilities

Kenneth E. Train; Daniel L. McFadden; Andrew A. Goett

The Review of Economics and Statistics, Vol. 69, No. 3. (Aug., 1987), pp. 383-391.

Stable URL:

http://links.jstor.org/sici?sici=0034-6535%28198708%2969%3A3%3C383%3ACAAVRS%3E2.0.CO%3B2-O