Classical Estimation Methods for LDV Models Using Simulation

Vassilis A. Hajivassiliou
Cowles Foundation for Research in Economics
Yale University

Paul A. Ruud
Department of Economics
University of California at Berkeley

July 1993

Abstract

This paper discusses estimation methods for limited dependent variable (LDV) models that employ Monte Carlo simulation techniques to overcome computational problems in such models. These difficulties take the form of high dimensional integrals that need to be calculated repeatedly but cannot be easily approximated by series expansions. In the past, investigators were forced to restrict attention to special classes of LDV models that are computationally manageable. The simulation estimation methods we discuss here make it possible to estimate LDV models that are computationally intractable using classical estimation methods.

We first review the ways in which LDV models arise, describing the differences and similarities in censored and truncated data generating processes. Censoring and truncation give rise to the troublesome multivariate integrals. Following the LDV models, we described various simulation methods for evaluating such integrals. Naturally, censoring and truncation play roles in simulation as well. Finally, estimation methods that rely on simulation are described. We review three general approaches that combine estimation of LDV models and simulation: simulation of the log-likelihood function (MSL), simulation of moment functions (MSM), and simulation of the score (MSS). The MSS is a combination of ideas from MSL and MSM, treating the efficient score of the log-likelihood function as a moment function.

We use the rank ordered probit model as an illustrative example to investigate the comparative properties of these simulation estimation approaches.

Acknowledgements: We would like to thank John Geweke and Dan McFadden for very helpful comments. John Wald provided expert research assistance. We are grateful to the National Science Foundation for partial financial support, under grants SES-929411913 (Hajivassiliou) and SES-9225111 (Ruud).

Keywords: Multivariate Integration, Limited Dependent Variable Models, Monte Carlo simulation, Maximum Simulated Likelihood, Method of Simulated Moments, Method of Simulated Scores

Handbook of Econometrics:
Classical Estimation Methods for LDV Models Using Simulation

Vassilis A. Hajivassiliou
Cowles Foundation for Research in Economics
Yale University

Paul A. Ruud
Department of Economics
University of California at Berkeley

July 1993

1 Introduction

This Chapter discusses classical estimation methods for limited dependent variable (LDV) models that employ Monte Carlo simulation techniques to overcome computational problems in such models. These difficulties take the form of high dimensional integrals that need to be calculated repeatedly. In the past, investigators were forced to restrict attention to special classes of LDV models that are computationally manageable. The simulation estimation methods we discuss here make it possible to estimate LDV models that are computationally intractable using classical estimation methods.

One of the most familiar LDV models is the binomial probit model, which specifies that the probability that a binomial random variable y is one, conditional on the regression vector x, is $\Phi(x'\beta)$ where $\Phi(\cdot)$ is the univariate standard normal cumulative distribution function (c.d.f.). Although this integral has no analytical expression, Φ has accurate, rapid, numerical approximations. These help make maximum likelihood estimation of the binomial probit model straightforward and most econometric software packages provide such estimation as a feature. However, a simple and common extension of the binomial probit model renders the resulting model too difficult for maximum likelihood computation. Introducing correlation among the observations generally produces a likelihood function containing integrals that cannot be well approximated *and* rapidly computed.

An example places the binomial probit model in the context of panel data in which a cross-section of N experimental units (individuals or households) is observed repeatedly, say in T consecutive time periods. Denote the binomial outcome for the n^{th} experimental unit in the t^{th} time period by $y_{nt} \in \{0,1\}$. In panel data sets, econometricians commonly expect correlation among the y_{nt} for the same n across different t, reflecting the presence of unobservable determinants of y_{nt} that evolve slowly for each experimental unit through time. In order to model such correlation parsimoniously, econometricians have adapted familiar models with correlation to the probit model. One can describe each y_{nt} as the transformation of a latent, normally distributed, y^*_{nt}:

$$y_{nt} = \left\{ \begin{array}{ll} 0 & \text{if } y^*_{nt} < 0 \\ 1 & \text{if } y^*_{nt} \geq 0 \end{array} \right\} \quad \text{where} \quad y^*_{nt} \sim N(x'_{nt}\beta, 1).$$

Then, one can assign the latent y^*_{nt} a nonscalar covariance matrix appropriate to continuously distributed panel data. For example, stacking the y^*_{nt} first by time period and then by experimental unit, a common specification of the covariance matrix is the variance components plus first-order

autoregression model

$$V(y^*) = I_N \otimes \Omega \quad \text{where} \quad \Omega = (1-\sigma_\alpha^2)\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{pmatrix} + \sigma_\alpha^2 J_T, \qquad (1)$$

where J_T is a $T \times T$ matrix of ones, $0 \leq \sigma_\alpha^2 < 1$, and $|\rho| < 1$.

Now consider the impact of such nonscalar covariance matrices on the likelihood for the observed y_{nt}. Although the *marginal* probabilities that y_{nt} is zero or one are unchanged, the likelihood function consists of the *joint* probabilities that the dependent series $\{y_{n1}, y_{n2}, \dots, y_{nT}\}$ are the observed sequences of zeros and ones. These joint probabilities are multivariate normal integrals over T dimensions and there are 2^T possible integrals.[1]

The practical significance of the increased dimensionality of the integrals is that traditional numerical methods generally cannot compute the integrals with sufficient speed and precision to make the computation of the maximum likelihood estimator workable. In this chapter, we review a collection of alternative, feasible, methods based on the ideas of estimation with simulation suggested by McFadden (1989) and Pakes and Pollard (1989).

In Section 2, we describe LDV models and illustrate the computational difficulties classical estimation methods encounter. Section 3 summarizes basic simulation methods, covering censored and truncated sampling methods. Estimation of LDV models and simulation are combined in Section 4 where three general approaches are reviewed: simulation of the log-likelihood function, simulation of moment functions, and simulation of the efficient score. We provide computational examples throughout to illustrate the various methods and their properties. We conclude this chapter with a summary of the main approaches presented.

2 Limited Dependent Variable Models

2.1 The Latent Normal Regression Model

Consider the problem of maximum likelihood estimation given the N observations on the vector of random variables y drawn from a population with cumulative distribution function (c.d.f.) $F(\theta, Y) = \Pr\{y \leq Y\}$. Let the corresponding density function with respect to Lebesgue measure be $f(\theta, y)$. The density f is a parametric function and the parameter vector θ is unknown, finite-dimensional, and $\theta \in \Theta$, where Θ is a compact subset of $\mathbf{R}^K$. Estimation of θ by maximum likelihood (ML) involves the maximization of the log-likelihood function $\ell_N(\theta) \equiv \sum_{n=1}^N \ln f(\theta; y_n)$ over Θ. Often, finding the root of a system of normal equations $\nabla_\theta \ell_N(\theta) = 0$ is equivalent. In the limited dependent variable models that we consider in this chapter, F will be a mixture of discrete and continuous distributions, so that f may consist of nonzero probabilities for discrete values of y and continuous probability densities for intervals of y. These functions are generally difficult to compute because they involve multivariate integrals that do not have closed forms, accurate approximations, or rapid numerical solutions. As a result, estimation of θ by classical methods is effectively infeasible.

[1] A partial list of studies in numerical analysis of such integrals is Clark (1961), Daganzo (1980), Davis and Rabinowitz (1984), Dutt (1973), Dutt (1976), Fishman (1973), Hammersley and Handscomb (1964), Horowitz *et al.* (1981), Moran (1984), Owen (1956), Rubinstein (1981), Stroud (1971), and Thisted (1988).

In this section, we review the various forms of likelihood functions that arise in LDV models. In the first subsection, we discuss models generated as partially observed or censored latent dependent variables. The second subsection describes truncated latent dependent variables. In this case, one views observations in a latent data set as missing entirely from an observed data set. Within these broad categories, we review discrete, mixed discrete/continuous, and mixture likelihood functions. Following our discussion of likelihood functions, subsection 2.6 treats the structure of the score function for LDV models and the last subsection gives a concrete illustration of the intractability of classical estimation methods for the general LDV model.

2.2 Censoring

In general, and particularly in LDV models, one can represent the data generating process for y as an 'incomplete data' or 'partial observability' process in which the observed data vector y is an indirect observation on a latent vector y^*. In such case, y^* cannot be recovered from the *censored* random variable y.

Definition 1 (Censored Random Variables) *Let Y^* be a random variable from a population with c.d.f. $F(Y^*)$ and support $\mathbf{A}$. Let $\mathbf{B}$ be the support of the random variable $Y = \tau(Y^*)$ where $\tau : \mathbf{A} \to \mathbf{B}$ is not invertible. Then Y is a* censored *random variable.*

In LDV models, τ is often called the 'observation rule;' and though it may not be monotonic, τ is generally piece-wise continuous. An important characteristic of censored sampling is that no observations are missing. Observations on y^* are merely abbreviated or summarized, hence the descriptive term 'censored.' Let $\mathbf{A} \subseteq \mathbf{R}^M$ and $\mathbf{B} \subseteq \mathbf{R}^J$.

The latent c.d.f. $F(\theta; Y^*)$ for y^* is related to the observed c.d.f. for y by the integral equation

$$F(\theta; Y) = \int_{\{y^* | \tau(y^*) \leq Y\}} dF(\theta; y^*). \tag{2}$$

In the LDV models that we consider, $F(\theta; y^*)$ is the multivariate normal c.d.f. given by $F(\theta, y^*) = \int \phi(y^* - \mu, \Omega) dy^*$ where Ω is a positive definite matrix, and

$$\phi(y^* - \mu, \Omega) \equiv \{\det[2\pi\Omega]\}^{-1/2} \exp\left[-\frac{1}{2}(y^* - \mu)'\Omega^{-1}(y^* - \mu)\right]. \tag{3}$$

We will refer to this multivariate normal distribution as the $N(\mu, \Omega)$ distribution. The mean vector is often parameterized as a linear function of observed conditioning variables X: $\mu(\beta) = X\beta$, where β is a vector of K_β slope coefficients. The covariance is matrix is usually a function of a vector of K_σ variance parameters σ.

The p.d.f. for y is the function that integrates to $F(\theta; Y)$. In this chapter, integration refers to the Lebesgue-Stieltjes integral and the p.d.f. is a generalized derivative of the c.d.f.[2] This means that the p.d.f. has discrete and continuous components. Everywhere in the support of Y where F is differentiable, the p.d.f. can be obtained by ordinary differentiation:

$$f(\theta; Y) = \frac{\partial^J F(\theta; Y)}{\partial Y_1 \ldots \partial Y_J}. \tag{4}$$

A simple illustration of such p.d.f.'s is given below in Example 2. In the LDV models we consider, F generally has a small number of discontinuities in some dimensions of Y so that F is not

[2]Such densities are formally known as Radon-Nikodym p.d.f.'s. with respect to Lebesgue measure.

differentiable everywhere. At a point of discontinuity Y^d, we can obtain the generalized p.d.f. by partitioning Y into the elements in which F is differentiable, $\{Y_1, \ldots, Y_{J'}\}$ say, and the remaining elements $\{Y_{J'+1}, \ldots, Y_J\}$ in which the discontinuity occurs. The p.d.f. then has the form

$$\begin{aligned} f(\theta; Y) &= \frac{\partial^{J'}}{\partial Y_1 \ldots \partial Y_{J'}} \cdot [F(\theta; Y) - F(\theta; Y - 0)] \\ &= f(\theta; Y_1, \ldots, Y_{J'}) \cdot \Pr\{Y_j = Y_j^d; j > J' | \theta; Y_1, \ldots, Y_{J'}\}, \end{aligned} \tag{5}$$

where the discrete jump $F(\theta; Y) - F(\theta; Y - 0)$ reflects the nontrivial probability of the event $\{Y_j = Y_j^d; j > J'\}$.[3] Examples 1 and 2 illustrate such probabilities.

It is these probabilities, the discrete components of the p.d.f., that pose computational obstacles to classical estimation. One must carry out multivariate integration and differentiation in (2)–(5) to obtain the likelihood for the observed data — see the following example for a clear illustration of this problem. Because accurate numerical approximations are unavailable, this integration is often handled by such general purpose numerical methods as quadrature. But the speed and accuracy of quadrature is inadequate to make the computation of the MLE practical except in special cases.

Example 1 (Multinomial Probit) *The multinomial probit model is a leading illustration of the computational difficulties of classical estimation methods for LDV models, which require the repeated evaluation of (2)–(5). This model is based on the work of Thurstone (1927) and was first analyzed by Bock and Jones (1968). For a multinomial model with $J = M$ possible outcomes, the latent y^* is $N(\mu, \Omega)$ where μ is a $J \times 1$ vector of means and Ω is a $J \times J$ symmetric positive definite covariance matrix. The observed y is often represented as a vector of indicator functions for the maximal element of y^*: $\tau(y^*) = [\mathbf{1}\{y_j^* = \max_i y_i^*\}; j = 1, \ldots, J]$. Therefore, the sampling space $\mathbf{B}$ of y is the set of orthonormal elementary unit vectors, whose elements are all zero except for a unique element that equals one:*

$$\mathbf{B} = \{(1, 0, 0, \ldots, 0),\ (0, 1, 0, 0, \ldots, 0),\ \ldots,\ (0, 0, \ldots, 0, 1)\}.$$

The probability function for y can be written as an integral over $J - 1$ dimensions after noting that the event $\{y_j = 1, y_i = 0, i \neq j\}$ is equivalent to $\{y_j^* - y_i^* \geq 0,\ i = 1, \ldots, J\}$. By creating the first-difference vector $z_j \equiv [y_j^* - y_i^*,\ i = 1, \ldots, J,\ i \neq j] \equiv \Delta_j y^*$ and denoting its mean and covariance by $\mu_j = \Delta_j \mu$ and $\Omega_j = \Delta_j \Omega \Delta_j'$ respectively, $F(\theta; y)$ and $f(\theta; y)$ are both functions of multivariate normal negative orthant integrals of the general form

$$\Phi(\mu, \Omega) \equiv \int_{-\infty}^{0} \ldots \int_{-\infty}^{0} \phi(x + \mu, \Omega)\, dx.$$

We obtain

$$F(\theta; y) = \sum_{j=1}^{J} \mathbf{1}\{y_j \geq 1\} \Phi(-\mu_j, \Omega_j)$$

and

$$f(\theta; y) = \left\{ \begin{array}{cc} \prod_{j=1}^{J} \Phi(-\mu_j, \Omega_j)^{y_j} & \text{if } y \in \mathbf{B} \\ 0 & \text{if otherwise} \end{array} \right\}. \tag{6}$$

[3] The height of the discontinuity is denoted by

$$F(\theta; Y) - F(\theta; Y - 0) \equiv \lim_{\epsilon \downarrow 0} F(\theta; Y) - F(\theta; Y - \epsilon).$$

When $J = 2$, this reduces to the familiar binomial probit likelihood mentioned in the Introduction:

$$\begin{aligned} f(\theta; y) &= \Phi(\mu_2 - \mu_1, 1)^{y_1} \Phi(\mu_1 - \mu_2, 1)^{y_2} \\ &= \Phi(-\mu', 1)^{1-y'} \Phi(\mu', 1)^{y'} \end{aligned} \tag{7}$$

where $\mu' = \mu_1 - \mu_2$ and $y' = y_2$.

If $J > 5$, then the likelihood function 6 is difficult to compute using conventional expansions without special restrictions on the covariance matrix, or without adopting other distributions that imply closed-form expressions. Examples of the former approach are the factor-analytic structures for Ω analyzed in Heckman (1981), Bolduc (1991), and Bolduc and Kaci (1991), and the diagonal Ω discussed in Hausman and Wise (1978), p.310. An example of the latter is the i.i.d. extreme-value distribution which, as McFadden (1973) shows, yields the analytically tractable Multinomial Logit model. See also Lerman and Manski (1981), p.224, McFadden (1981), and McFadden (1986) for further discussions on this issue.

Example 2 (Tobit) *The tobit or censored regression model*[4] *is a simple example of a mixed distribution with discrete and continuous components. This model has a univariate latent structure like probit: $y^* \sim N(\mu, \sigma^2)$. The observation rule is also similar: $\tau(y^*) = \mathbf{1}\{y^* \geq 0\} \cdot y^*$ which leads to the sample space $\mathbf{B} = \{y \in \mathbf{R} \mid y \geq 0\}$ and c.d.f.*

$$F(\theta; Y) = \left\{ \begin{array}{cc} 0 & \text{if } Y < 0 \\ \int_{\{y^* < Y\}} \phi(y^* - \mu, \sigma) dy^* = \Phi(Y - \mu, \sigma^2) & \text{if } Y \geq 0 \end{array} \right\}$$

The p.d.f. is mixed, containing discrete and continuous terms:

$$f(\theta; Y) = \left\{ \begin{array}{cc} 0 & \text{if } Y < 0 \\ \Phi(-\mu, \sigma^2) & \text{if } Y = 0 \\ \phi(Y - \mu, \sigma^2) & \text{if } Y > 0 \end{array} \right\} \tag{8}$$

The discrete jump in F at $Y = 0$ corresponds to the non-zero probability of $\{Y = 0\}$, just as in binomial probit. F is differentiable for $Y > 0$ so that the p.d.f. is obtained by differentiation. Just as in the extension of binomial to multinomial probit, multivariate tobit models present multivariate integrals that are difficult to compute.

Example 3 (Nonrandom Sample Selection) *The nonrandom sample selection model provides a final example of partial observability which generalizes the tobit model.*[5] *In the simplest version, the latent y^* consists of two elements drawn from a bivariate normal distribution where*

$$\Omega(\sigma) = \left[\begin{array}{cc} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right]$$

The observation rule is

$$\tau(y^*) = \left(\begin{array}{c} \tau_1(y^*) \\ \tau_2(y^*) \end{array} \right) = \left(\begin{array}{c} \mathbf{1}\{y_1^* \geq 0\} \\ \mathbf{1}\{y_1^* \geq 0\} \cdot y_2^* \end{array} \right)$$

so that the first element of y is a binomial variable and the second element is an observation on y_2^* when $y_1 = 1$; otherwise, there is no observation of y_2^* because y_2 is identically zero. That is, the sampling space of y is the union of two disjoint sets: $\mathbf{B} = \{(0,0)\} \cup \{(1, y_2),\ y_2 \in \mathbf{R}\}$. Thus,

[4] Tobin (1958).

[5] See Gronau (1974), Heckman (1974), Lewis (1974), Lee (1978), and Lee (1979).

two cases capture the nonzero regions of the c.d.f. of y. First of all, the c.d.f. is constant on $\mathbf{B}_0 = [0,1) \times [0,\infty)$*:*

$$F(\theta; Y) = \int_{\{y_1^* < 0\}} \phi(y^* - \mu, \Omega)\, dy^* = \Phi(-\mu_1, 1),\ Y \in \mathbf{B}_0$$

because y_2^* is unrestricted (and unobserved) in this case. Once Y_1 reaches 1, the entire sampling space for y_1 has been covered and the c.d.f. on $\mathbf{B_1} = [1,\infty) \times \mathbf{R}$ is increased according to

$$\begin{aligned} F(\theta; Y) &= \mathbf{1}\{Y_2 \geq 0\}\, \Phi(-\mu_1, 1) + \textstyle\int_{\{y_1^* \geq 0, y_2^* \leq Y_2\}} \phi(y^* - \mu, \Omega)\, dy^*, \quad Y \in B_1 \\ &= \mathbf{1}\{Y_2 \geq 0\}\, \Phi(-\mu_1, 1) + \Phi\left(\begin{bmatrix} \mu_1 \\ Y_2 - \mu_2 \end{bmatrix}, \begin{bmatrix} 1 & -\sigma_{12} \\ -\sigma_{12} & \sigma_2^2 \end{bmatrix} \right), \quad Y \in B_1 \end{aligned}$$

The p.d.f. will therefore be

$$f(\theta; Y) = \left\{ \begin{array}{cc} \Phi(-\mu_1, 1) & \text{if } Y_1 = 0 \\ \Phi\left(\mu_1 + \sigma_{12}(Y_2 - \mu_2)/\sigma_2^2,\, 1 - \sigma_{12}^2/\sigma_2^2\right) \cdot \phi\left(Y_2 - \mu_2, \sigma_2^2\right) & \text{if } Y_1 = 1 \end{array} \right\}$$

The sample selection process is often more complicated, with several causes of sample selection. In such cases, the latent y_1^* is a vector with each element associated with a different cause of partial observation. The latent y_2^* is observed only if all the elements of y_1^* (suppose there are $J = M - 1$) are positive so that the observation rule is

$$\tau(y^*) = \begin{pmatrix} \tau_1(y^*) \\ \tau_2(y^*) \end{pmatrix} = \begin{pmatrix} \mathbf{1}\{y_1^* \geq 0\} \\ \left(\prod_{j=1}^{J} \mathbf{1}\{y_{1j}^* \geq 0\}\right) \cdot y_2^* \end{pmatrix},$$

where $\mathbf{1}\{y_1^* \geq 0\}$ is a $(M-1) \times 1$ vector of indicator variables. The sampling space is

$$\mathbf{B} = \{y \in \mathbf{R}^M \mid y_M = y_2^*,\ \prod_{j=1}^{M-1} y_j = 1,\ y_j \in \{0,1\},\ j < M\} \cup \{y \in \mathbf{R}^M \mid \prod_{j=1}^{M-1} y_j = 0,\ y_j \in \{0,1\}\},$$

and the likelihood function contains multivariate integrals over the $M-1$ dimensions of y_1^*.

Other types of nonrandom sample selection lead to general *discrete/continuous* models and models of *switching regressions with known sample separation*. Such models are discussed extensively in Dubin and McFadden (1984), Hanemann (1984), Lee (1978), Maddala (1983), and Amemiya (1984).

2.3 Truncation

When it is represented as a partial observation, a limited dependent variable is a censored latent variable. Another mechanism for generating limited dependent variables is *truncation*, which refers to dropping observations so that their realization goes unrecorded.

Definition 2 (Truncated Random Variables) *Let $F(Y)$ be the c.d.f. of y^* and let $\mathbf{D}$ be a proper subset of the support of F and $\mathbf{D}^c$ its complement such that $\Pr\{y^* \in \mathbf{D}^c\} > 0$. The function*

$$G(Y) = \left\{ \begin{array}{cc} F(Y)/\Pr\{Y \in \mathbf{D}\} & \text{if } Y \in \mathbf{D} \\ 0 & \text{if } Y \in \mathbf{D}^c \end{array} \right\}$$

is the c.d.f. of a truncated *y^*.*

One can generate a sample of truncated random variables with the c.d.f. G by drawing a random sample of y^* and removing the realizations that are not members of $\mathbf{D}$. This is typically the way truncation arises in practice. To draw a single realization of the truncated random variable, one can draw y^*'s until a realization falls into $\mathbf{D}$. The term 'truncation' derives from the visual effect dropping the set $\mathbf{D}^c$ has on the original distribution when $\mathbf{D}^c$ is a tail region: the tail of the p.d.f. is cut off or truncated.

To incorporate truncation, we expand the observation rule to

$$y = \left\{ \begin{array}{cc} \tau(y^*) & \text{if } y^* \in \mathbf{D} \\ \text{unobserved} & \text{otherwise} \end{array} \right\} \tag{9}$$

where $\mathbf{D}$ is an 'acceptance region.' This situation differs from that of the nonrandom sample selection model in which an observation is still partially observed: At least, every realization is recorded. In the presence of truncation, the observed likelihood requires normalization relative to the latent likelihood:

$$f(\theta; Y) = \frac{\int_{\{y^* \in \mathbf{D} | \tau(y*) = Y\}} dF(\theta; y^*)}{\int_{\mathbf{D}} dF(\theta; y^*)} \tag{10}$$

The normalization by a probability in the denominator makes the c.d.f. proper, with an upper bound of one.

Example 4 (Truncated Normal Regression) *If $y^* \sim N(\mu, \sigma^2)$ and y is an observation of y^* when $y^* > 0$, the model is a truncated normal regression. Setting $\mathbf{D} = \{y \in \mathbf{R} \mid y > 0\}$ makes $\mathbf{B} = \mathbf{D}$ so that the c.d.f. and p.d.f. of y are*

$$F(\theta; Y) = \left\{ \begin{array}{cc} 0 & \text{if } Y \leq 0 \\ \frac{\int_0^Y \phi(y^* - \mu, \sigma) dy^*}{\int_0^\infty \phi(y^* - \mu, \sigma) dy^*} = \frac{\Phi\left(Y - \mu, \sigma^2\right) - \Phi(-\mu, \sigma^2)}{1 - \Phi(-\mu, \sigma^2)} & \text{if } Y > 0 \end{array} \right\}$$

$$f(\theta; Y) = \left\{ \begin{array}{cc} 0 & \text{if } Y \leq 0 \\ \frac{\phi(Y - \mu, \sigma)}{1 - \Phi(-\mu, \sigma^2)} & \text{if } Y > 0 \end{array} \right\}$$

As in the tobit model, a normal integral appears in the likelihood function. However, this integral enters in a nonlinear fashion, in the denominator of a ratio. Clearly, multivariate forms of truncation lead to multivariate integrals in the denominator.

To accommodate both censored and truncated models, in the remainder of this chapter we will often denote the general log-likelihood function for LDV models with a two-part function:

$$\ln f(\theta; y) = \ln f_1(\theta; y) - \ln f_2(\theta; y) \tag{11}$$

where f_2 represents the normalizing probability $\Pr\{y^* \in \mathbf{D}\} = \int_{\mathbf{D}} dF(\theta; y^*)$. In models with only censoring, $f_2 \equiv 1$. But in general, both f_1 and f_2 will require numerical approximation. Note that in this general form, the log-likelihood function can be viewed as the difference between two log-likelihood functions for models with censoring. For example, the log-likelihood of the truncated regression in Example 4 is the difference between the log-likelihoods of the tobit regression in Example 2 and the binomial probit model mentioned in the Introduction and Example 1 (see equations (7) and (8)):[6]

$$\begin{aligned} \mathbf{1}\{Y > 0\} \ln \left[\tfrac{\phi(Y - \mu, \sigma)}{1 - \Phi(-\mu, \sigma^2)} \right] &= \mathbf{1}\{Y > 0\} \ln \left[\phi(Y - \mu, \sigma) \right] + \mathbf{1}\{Y = 0\} \Phi(-\mu, \sigma^2) \\ &\quad - \left[\mathbf{1}\{Y > 0\} \ln \left[1 - \Phi(-\mu, \sigma^2) \right] + \mathbf{1}\{Y = 0\} \Phi(-\mu, \sigma^2) \right] \end{aligned}$$

[6]Note that scale information about y^* is available in the censored and truncated normal regression models than in the case of binary response, so that σ^2 is now identifiable. Hence, the normalization $\sigma^2 = 1$ is not necessary, as it is in the binary probit model where only the discrete information $\mathbf{1}\{Y > 0\}$ is available.

2.4 Mixtures

LDV models have come to include a family of models that do not necessarily have *limited* dependent variables. This family, containing densities called *mixtures*, shares an analytical trait with the LDV models that we have already reviewed: the p.d.f. generally contains discrete probability terms.

Definition 3 (Mixtures) *Let $F(\theta;Y)$ be the c.d.f. of y^* depending on a parameter θ and $H(\theta)$ another c.d.f. Then the c.d.f.*

$$G(Y) = \int F(\theta;Y)\, dH(\theta)$$

is a mixture.

Possible ways in which mixtures arise in econometric models are unobservable heterogeneity in the underlying data generating process (see, for example, Heckman (1981)) and "short-side" rationing rules (Quandt (1972), Goldfeld and Quandt (1975), Laroque and Salanié (1989)). Laroque and Salanié (1989) discuss simulation estimation methods for the analysis of this type of model.

Example 5 (Mixture) *A cousin of the nonrandom sample selection model is the mixture model generated by an underlying trivariate normal distribution, where*

$$\Omega(\sigma) = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

The observation rule maps a three-dimensional vector into a scalar; the rule can be written as

$$y = \mathbf{1}\{y_1^* \geq 0\} \cdot y_2^* + \mathbf{1}\{y_1^* < 0\} \cdot y_3^*$$

An indicator function determines whether y_2^* or y_3^* is observed. An important difference with sample selection is that the indicator itself is not observed. Thus, y is a 'mixture' of y_2^*'s and y_3^*'s. As a result, such mixtures have qualitatively distinct c.d.f.'s, compared to the other LDV models we have discussed. In the present case,

$$\begin{aligned} F(\theta;Y) &= \int_{\{y_1^* \geq 0, y_2^* \leq Y\} \cup \{y_1^* < 0, y_3^* \leq Y\}} \phi(y^* - \mu, \Omega) dy^* \\ &= \int_{\{y_1^* \geq 0, y_2^* \leq Y\}} \phi(y^* - \mu, \Omega) dy^* + \int_{\{y_1^* < 0, y_3^* \leq Y\}} \phi(y^* - \mu, \Omega) dy^* \end{aligned}$$

and

$$\begin{aligned} f(\theta;Y) \equiv \frac{dF(\theta;Y)}{dY} &= \phi(Y - \mu_2, \sigma_2) \int_{\{y_1^* \geq 0\}} \phi(y_1^* - \mu_{1|2}, \Omega_{1|2}) dy_1^* \\ &\quad + \phi(Y - \mu_3, \sigma_3) \int_{\{y_1^* < 0\}} \phi(y_1^* - \mu_{1|3}, \Omega_{1|3}) dy_1^* \\ &= \phi(Y - \mu_2, \sigma_2)\Phi(\mu_{1|2}, \Omega_{1|2}) + \phi(Y - \mu_3, \sigma_3)\Phi(-\mu_{1|3}, \Omega_{1|3}) \end{aligned}$$

where, for $j = \{2, 3\}$,

$$\begin{aligned} \mu_{1|j} \equiv E(y_1^* | y_j^* = Y) &= \mu_1 + \sigma_{1j}(Y - \mu_j)/\sigma_j^2 \\ \Omega_{1|j} \equiv V(y_1^* | y_j^* = Y) &= 1 - \sigma_1^2/\sigma_j^2 \end{aligned}$$

are conditional moments. The p.d.f. particularly demonstrates the weighted nature of the distribution: The marginal distributions of y_2^* and y_3^* are mixed together by probability weights.

2.5 Time Series Models

LDV models are not typically applied to time series datasets, but short time series have played an important role in the analysis of panel or longitudinal data sets. Such time series are another source of high dimensional integrals in likelihood functions. Here we expand our introductory example.

Example 6 (Multiperiod Binary Probit Model) *A random sample of N economic agents is followed over time, with agent n being observed for T periods. The latent variable $y^*_{nt} = \mu_{nt} + \epsilon_{nt}$ measures the net benefit to the agent characterizing an action in period t. Typically, μ_{nt} is a linear index function of a $k \times 1$ vector of exogenous explanatory variables x_{nt}, i.e., $\mu_{nt} \equiv x'_{nt}\beta$. The agent chooses one of two actions in each period, denoted by $y_{nt} \in \{0, 1\}$, depending upon the value of y^*_{nt}:*

$$\tau(y^*) \equiv \begin{cases} y_{nt} = 1 & \text{if } y^*_{nt} > 0 \\ y_{nt} = 0 & \text{if } y^*_{nt} \leq 0 \end{cases} \qquad t = 1, \cdots, T. \tag{12}$$

Hence, the sample space for $\tau(y^*)$ is $\mathbf{B} = \times_{t=1}^{T}\{0, 1\}$, i.e., all possible (2^T) sequences of length T, with 0 and 1 as the possible realizations in each period.

Let the distribution of $y^*_n \equiv (y^*_{n1}, \cdots, y^*_{nT})'$ be the multivariate normal given in equation (3). Then, for individual n the LDV vector $\{y_{nt}\}$, $t = 1, ..., T$, has the discrete p.d.f.

$$f(\beta, \Omega; Y) = \Phi((-1)^{1-y_{n1}} \cdot \mu_{n1}(\beta), \cdots, (-1)^{1-y_{nT}} \cdot \mu_{nT}(\beta), \Omega).$$

This is a special case of the multinomial probit model of example (1), with $J = 2^T$ alternatives and a typically highly restricted Ω, reflecting the assumed serial correlation in the $\{\epsilon_{nt}\}_{t=1}^{T}$ sequence.

By way of illustration, let us consider the specific covariance structure, found very useful in applied work[7]*:*

$$\epsilon_{nt} = \eta_n + \zeta_{nt}, \quad \zeta_{nt} = \rho\zeta_{n,t-1} + \nu_{nt}, \quad |\rho| < 1, \tag{13}$$

and ν, η independent. This implies that

$$\Omega = \sigma^2_\zeta \cdot \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{pmatrix} + \sigma^2_\eta \cdot J_T.$$

The variance parameters σ^2_ζ and σ^2_η cannot both be identified, so the normalization $\sigma^2_\zeta + \sigma^2_\eta = 1$ is used.[8]

The probability of the observed sequence of choices of individual n is

$$\Pr\{y_n; \theta, x_n\} = \int_{a_n(y_n)}^{b_n(y_n)} \phi(y^*_n - \mu_n, \Omega_n)\, dy^*_n,$$

with $\theta \equiv \left(\beta, \sigma^2_\eta, \rho\right)$ and

$$a_{nt} = \begin{cases} 0 & \text{if } y_{nt} = 1 \\ -\infty & \text{if } y_{nt} = 0 \end{cases},$$

$$b_{nt} = \begin{cases} +\infty & \text{if } y_{nt} = 1 \\ 0 & \text{if } y_{nt} = 0 \end{cases}.$$

[7] See Hajivassiliou and McFadden (1990), Börsch-Supan *et al.* (1992), and Hajivassiliou (1993a).

[8] This is the structure assumed in the introductory example — see equation (1) above.

Note that the likelihood of this example is another member of the family of censored models. Time series models like this do not present a new analytical problem. Indeed, such time series models are more tractable for estimation because classical methods do provide consistent, though statistically inefficient, estimators (see Poirier and Ruud (1988), Hajivassiliou (1986), and Avery *et al.* (1983)).[9] Keane (1993) discusses extensively special issues in the estimation by simulation of panel data models and Mühleisen (1991) compares the performance of alternative simulation estimators for such models. Studies of dynamic discrete behavior using simulation techniques are Berkovec and Stern (1991), Bloemen and Kapteyn (1991), Hajivassiliou and Ioannides (1991), Hotz and Miller (1989), Hotz and Sanders (1991), Hotz *et al.* (1991), Pakes (1992), and Rust (1992).

In this chapter, we do not analyze the estimation by simulation of 'long' time series models. We refer the reader to Lee and Ingram (1991), Duffie and Singleton (1993), Laroque and Salanié (1990), and Gourieroux and Monfort (1990) for results on this topic.

2.6 Score Functions

For models with censoring, the score for θ can be written in two ways which we will use to motivate two approaches to approximation of the score by simulation:

$$\nabla_\theta \ln f(\theta; y) = \frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} \tag{14}$$

$$= E\left[\nabla_\theta \ln f(\theta; y^*) \middle| y\right] \tag{15}$$

where ∇_θ is an operator that represents partial differentiation with respect to the elements of θ. The ratio (14) is simply the derivative of the log-likelihood and simulation can be applied to the numerator and denominator separately. The second expression (15), the conditional expectation of the score of the latent log-likelihood, can be simulated as a single expectation if $\nabla_\theta \ln f(\theta; y^*)$ is tractable. Ruud (1986), van Praag and Hop (1987), Hajivassiliou and McFadden (1990), and Hajivassiliou (1992) have noted alternative ways of writing score functions for the purpose of estimation by simulation.

Here is the derivation of (15): Let $F(\theta; y^* \mid y)$ denote the conditional c.d.f. of y^* given that $\tau(y^*) = y$.[10] We let

$$\mathrm{E}[t(y^*) \mid y] \equiv \int t(y^*) dF(\theta; y^* \mid y) \tag{16}$$

denote the expectation of a random variable $t(y^*)$ with respect to the conditional c.d.f. $F(\theta; y^* \mid y)$ of y^* given $\tau(y^*) = y$. Then

$$\frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} = \frac{1}{f(\theta; y)} \int_{\{y^* \mid \tau(y^*) = y\}} \nabla_\theta dF(\theta; y^*)$$

$$= \int_{\{y^* \mid \tau(y^*) = y\}} \frac{\nabla_\theta f(\theta; y^*)}{f(\theta; y^*)} \frac{f(\theta; y^*)}{f(\theta; y)} dy^*$$

[9] Panel data sets in which each agent is observed for the *same* number of time periods T are called *balanced*, while sets with $T_n \neq T$ for some $n = 1, \cdots, N$ are known as *unbalanced*. As long as the determination of T_n is not *endogenous* to the economic model at hand, balanced and unbalanced sets can be analyzed using the same techniques. There exists, however, the interesting case in which T_n is determined endogenously through an economic decision, which leads to a multiperiod sample-selection problem. See Hausman and Wise (1979) for a discussion of this case.

[10] Formally,

$$F(\theta; Y^* \mid \tau(y^*) = y) \equiv \lim_{\epsilon \downarrow 0} \frac{\Pr\{y^* \leq Y^*, y - \epsilon < \tau(y^*) \leq y\}}{\Pr\{y - \epsilon < \tau(y^*) \leq y\}}$$

$$= \quad E\left[\nabla_\theta \ln f(\theta; y^*) | \, \tau(y^*) = y\right]$$

since $\frac{f(\theta;y^*)}{f(\theta;y)} = f(\theta; y^* / \int_{\{y^*|\tau(y^*)=y\}} f(\theta; y^*) dy^*$ is the p.d.f. of the truncated distribution $\{y^*|\tau(y^*) = y\}$.

This formula for the score leads to the following general equations for normal LDV models when y^* has the multivariate normal p.d.f. given in (3):

$$\begin{array}{rcl} \nabla_\mu \ln f(\theta; y) & = & \Omega^{-1}[\mathrm{E}(y^*|y) - \mu] \\ \nabla_\Omega \ln f(\theta; y) & = & \frac{1}{2}\Omega^{-1}\left\{\mathrm{V}(y^*|y) + [\mathrm{E}(y^*|\tau(y^*) = y) - \mu][\mathrm{E}(y^*|\tau(y^*) = y) - \mu]' - \Omega\right\}\Omega^{-1} \end{array} \tag{17}$$

using the standard derivatives for the log-likelihood of a multivariate normal

$$\begin{array}{rcl} \nabla_\mu \ln \phi(y^* - \mu, \Omega) & = & \Omega^{-1}(y^* - \mu) \\ \nabla_\Omega \ln \phi(y^* - \mu, \Omega) & = & \frac{1}{2}\Omega^{-1}[(y^* - \mu)(y^* - \mu)' - \Omega]\Omega^{-1} \end{array} \tag{18}$$

According to (17), the score of a normal LDV model depends only on the first two moments of a truncated multivariate normal random variable z generated by the truncation rule

$$z = \left\{ \begin{array}{cc} y^* & \text{if } \tau(y^*) = y \\ \text{unobserved} & \text{otherwise} \end{array} \right\} \tag{19}$$

The functional form of these moments depends on the specification of the LDV function τ.

For LDV models with truncation, there are no changes to (14)–(16). The only change that (9) requires for (19) is the restriction to the acceptance region $\mathbf{D}$. That is, the score depends on only the first two moments of a truncated multivariate normal random variable z' generated by the truncation rule

$$z' = \left\{ \begin{array}{cc} y^* & \text{if } \tau(y^*) = y,\ y^* \in \mathbf{D} \\ \text{unobserved} & \text{otherwise} \end{array} \right\}$$

As a result, there is a basic change to (17). Because the log-likelihood function of truncated models is the difference between two log-likelihood functions for censored models (see equation (11)), the score is expressed as the difference in the scores for such models:

$$\begin{array}{rcl} \nabla_\theta \ln f(\theta; y) & = & \nabla_\theta \ln f_1(\theta; y) - \nabla_\theta \ln f_2(\theta; y) \\ & = & \mathrm{E}\left[\nabla_\theta \ln f(\theta; y^*) | \, \tau(y^*) = y\right] - \mathrm{E}\left[\nabla_\theta \ln f(\theta; y^*) | \, y^* \in \mathbf{D}\right] \end{array}$$

so that (17) becomes

$$\begin{array}{rcl} \nabla_\mu \ln F(\theta; y) & = & \Omega^{-1}[\mathrm{E}(y^*|\tau(y^*) = y) - \mathrm{E}(y^*|y^* \in \mathbf{D})] \\ \nabla_\Omega \ln F(\theta; y) & = & \frac{1}{2}\Omega^{-1}\left\{\mathrm{E}[(y^* - \mu)(y^* - \mu)'|\tau(y^*) = y]\right. \\ & & \left. - \ \mathrm{E}[(y^* - \mu)(y^* - \mu)'|y^* \in \mathbf{D}]\right\}\Omega^{-1} \end{array}$$

2.7 The Computational Intractability of LDV Models

The likelihood contribution $f(\theta; y_n)$ and the score $\nabla_\theta \ln f(\theta; y_n)$ are functions of at most M-dimensional integrals over the region $\mathbf{D}(y) \equiv \{y|\tau(y^*) = y\}$ in the domain of the $M \times 1$ latent vector y_n^*. The fundamental source of the computational intractability of classical estimation methods for the general LDV model is the repeated evaluation of such integrals. To illustrate, consider a multinomial probit model with $M = 16$ alternatives, with $K = 20$ exogenous variables that vary by alternative. A random sample of $N = 1000$ observations is available. Suppose the $M \times M$ variance-covariance matrix Ω of the unobserved random utilities has $\frac{15 \cdot 16}{2} - 1 = 119$ free elements

(after imposing identification restrictions). Then, the number of parameters to be estimated is $p = 139$. Suppose the analyst uses an iterative Newton-Raphson type of numerical procedure, employing numerical approximations to the first derivatives based on two-sided first differences and that 20 iterations are required to achieve convergence, which is a realistic number.[11] Each iteration requires at least 2p evaluations of the likelihood function for approximating the first derivatives. We thus expect that finding the ML estimator will require about $20 \times 2p$ function evaluations. Since the sample consists of $N = 1000$ individuals, we will have to calculate $N \times 20 \times 2p$ contributions to the likelihood function, each of which, in general, will be 16-dimensional integrals. Let s be the time in seconds a given computer requires to approximate a 16-dimensional integral by numerical quadrature methods. Our hypothetical ML estimation will thus require about $N \times 20 \times 2p \times s$ seconds. On a typical modern supercomputer (say a Cray 1) one could expect $s \approx 2$. Hence, using such a supercomputer, our problem would take about $1000 \times 20 \times 178 \times 2/3600$ hours, which is about 4 *months* of Cray 1 CPU! It is crucial to stress that such numerical quadrature methods offer only poor approximations to integrals of such dimension.[12] The maximum likelihood estimates resulting from 4 months of Cray 1 CPU would be utterly unreliable. The need for alternative estimation methods for these problems is apparent.

3 Simulation Methods

3.1 Overview

Two general approaches to exploiting simulation in parametric estimation are to approximate the likelihood function and to approximate such moment functions as the score. The likelihood function can be simulated by Monte Carlo techniques over the latent marginal distribution $f(\theta; y^*)$ in equation (4) for the mixture case, equation (5) for the discrete/continuous case, and equation (10) for the truncated case. Alternatively, the score can be approximated either by integrating both numerator and denominator in equation (14) or by integrating over the latent conditional p.d.f. $f(\theta; y^*|y)$ as in equation (15). Thus, simulation techniques focus on the simulation from these two distributions, $f(\theta; y^*)$ and $f(\theta; y^*|y)$. The censoring and truncation discussed above for LDV models also appear in simulations and we consider methods for effecting each type of observation rule below. As we will show in Section 4, some simulation estimation methods use *censored simulation* for the estimation of the main types of LDV models discussed in Section 2, (censored, truncated, and mixture models), whereas other estimation methods use truncated simulation for the estimation of these models.

Simulation of standard normal random variables is an old and well-studied problem. Relatively fast algorithms are widely available for generating such random variables on a computer. Thus, consider the simulation of the latent data generating process. We can always write

$$y^* = \mu + \Gamma\eta, \tag{20}$$

where η is a vector of M independent standard normal random variables and Γ is a matrix square root of Ω, so that $\Omega = \Gamma\Gamma'$. It is convenient to set Γ to the (lower triangular) Cholesky factor. Clearly, the latent data generating process can be simulated rapidly with simulations of η for any

[11] See Quandt (1986) for a discussion of issues in numerical optimization methods.

[12] Clark (1961) proposed another numerical approximation method for such integrals — see also Daganzo *et al.* (1977) and Daganzo (1980). The Horowitz *et al.* (1981) study finds serious shortcomings in the numerical accuracy of the Clark method in typical problems with high J and unrestricted Ω.

values of μ and Ω. Such simulations can be used in turn to simulate the likelihood and log-likelihood functions and their derivatives with respect to the parameters.

As in all of the examples given above, the observation rules common in LDV models imply regions of integration that are rectangles: that is, for some matrix A, and vectors b_0, and b_1, possibly with some infinite elements,

$$\{y^* \mid \tau(y^*) = y\} = \{y^* \mid b_0 \leq Ay^* \leq b_1\} \tag{21}$$

where $\text{rank}(A) \leq M$. These are the problems that we will consider. Since Ay^* is also normally distributed, it will often be convenient to simulate Ay^* instead of y^*. In that case, we simply transform the mean vector and covariance matrix to $A\mu$ and $A\Omega A'$, respectively. Without any loss of generality in this section, we set $A = I_M$, the $M \times M$ identity matrix. We denote $\mathbf{D} = \{z \in \mathbf{R}^M \mid b_0 \leq z \leq b_1\}$.

Such regions as (21) have two important analytical properties. First of all, rectangular regions have constant boundaries with respect to the variable of integration, simplifying integration. Secondly, the differentiation in (4) and (5) can be carried out analytically to obtain likelihood functions composed of multivariate normal p.d.f.'s of the form (3) and multivariate normal c.d.f.'s of the form

$$\Pr\{\mathbf{D}; \mu, \Omega\} \equiv \int \mathbf{1}\{y^* \in \mathbf{D}\}\phi(y^* - \mu, \Omega)\, dy^* \tag{22}$$

Thus, simulation of the likelihood can be restricted to terms in $\Pr\{\mathbf{D}; \mu, \Omega\}$. Simulation of the score in (14) involves only the additional terms

$$\begin{array}{rcl} \nabla_\mu \Pr\{\mathbf{D}; \mu, \Omega\} & = & \Omega^{-1} \int \mathbf{1}\{y^* \in \mathbf{D}\}(y^* - \mu)\phi(y^* - \mu, \Omega)\, dy^* \\ \nabla_\Omega \Pr\{\mathbf{D}; \mu, \Omega\} & = & \frac{1}{2}\Omega^{-1} \left\{\int \mathbf{1}\{y^* \in \mathbf{D}\} \left[(y^* - \mu)(y^* - \mu)' - \Omega\right] \phi(y^* - \mu, \Omega)\, dy^*\right\} \Omega^{-1} \end{array} \tag{23}$$

Normalized by $\Pr\{\mathbf{D}; \mu, \Omega\}$, these equations transform to

$$\begin{array}{rcl} \nabla_\mu \ln \Pr\{\mathbf{D}; \mu, \Omega\} & = & \Omega^{-1}[\mathrm{E}(y^* | y^* \in \mathbf{D}) - \mu] \\ \nabla_\Omega \ln \Pr\{\mathbf{D}; \mu, \Omega\} & = & \frac{1}{2}\Omega^{-1} \left\{\mathrm{E}\left[(y^* - \mu)(y^* - \mu)' | y^* \in \mathbf{D}\right] - \Omega\right\} \Omega^{-1} \end{array} \tag{24}$$

which are terms in (17).

In the remainder of this section, we will discuss the simulation of (22)–(24). For this purpose we denote

$$y^*_{-i} \equiv [y^*_m; m = 1, \ldots, M; m \neq i],$$

$$\mu_{-i} \equiv \mathrm{E}(y^*_{-i}), \quad \Omega_{-i,-i} \equiv \mathrm{V}(y^*_{-i}) \quad \text{and} \quad \Omega_{-i,i} \equiv \mathrm{Cov}(y^*_{-i}, y^*_i)$$

and the conditional moments

$$\mu_{-i|i}(y^*_i) \equiv \mathrm{E}(y^*_{-i} | y^*_i) \quad \text{and} \quad \Omega_{-i,-i|i} \equiv \mathrm{V}(y^*_{-i} | y^*_i).$$

These conditional moments have the well-known formulas

$$\mu_{-i|i}(y^*_i) = \mu_{-i} + \Omega_{-i,i}\Omega^{-1}_{i,i}(y^*_i - \mu_i),$$

$$\Omega_{-i,-i|i} \equiv \Omega_{-i,-i} - \Omega_{-i,i}\Omega^{-1}_{ii}\Omega_{i,-i}.$$

The conditional mean and variance of y^*_i given y^*_{-i}, denoted $\mu_{i|-i}$ and $\Omega_{ii|-i}$, are defined analogously. We also define

$$y^*_{<i} \equiv [y^*_m; m = 1, \ldots, i-1]$$

and use a similar notation for the marginal and conditional moments of this subvector of random variables. For example, the conditional mean of y^*_i given $y^*_{<i}$ is

$$\mu_{i|<i}(y^*_{<i}) = \mu_i + \Omega_{i,<i}\Omega^{-1}_{<i,<i}(y^*_{<i} - \mu_{<i}).$$

3.2 Censored Simulation

We begin by focusing on the integrals in (22) and (23) accumulated in the vector

$$\mathrm{E}\left[h(y^*, \mathbf{D})\right] \equiv \mathrm{E}\left[\mathbf{1}\{y^* \in \mathbf{D}\}\begin{pmatrix} 1 \\ y^* \\ \mathrm{vec}(y^* y^{*\prime}) \end{pmatrix}\right] \tag{25}$$

The elements of h are censored random variables. We consider two basic methods of simulation: direct censoring of the multivariate normal random variable and importance sampling.

3.2.1 Multivariate Normal Simulation

A direct method for simulating $\Pr\{\mathbf{D}; \mu, \Omega\}$ and its derivatives is to make repeated Monte Carlo draws for η, use (20) to calculate y^* for each η, and then form an empirical analogue of (25), by working only with the realization that fall in set $\mathbf{D}$. Let $\{\eta_1, \ldots, \eta_R\}$ be R simulated draws from the $N(0, I_M)$ distribution and $\tilde{y}_r = \mu + \Gamma\eta_r$ $(r = 1, \ldots, R)$ so that

$$\bar{h} = \frac{1}{R}\sum_{r=1}^{R} h(\tilde{y}_r, \mathbf{D})$$

is an unbiased simulation of (25). As R gets larger, the sampling variance of $\bar{h}$, $P(1-P)/R$ approaches zero and $\bar{h}$ converges strongly to $\mathrm{E}[h(y^*, \mathbf{D})]$. The simulation of $\Pr\{\mathbf{D}; \mu, \Omega\}$ is simply the observed frequency with which the simulations of y^* fall into $\mathbf{D}$. Its derivatives with respect to μ and Ω are functions of the average simulation of $\mathbf{1}\{y^* \in \mathbf{D}\}y^*$ and $\mathbf{1}\{y^* \in \mathbf{D}\}y^*y^{*\prime}$. We will call this the *crude Monte Carlo* (CMC) simulator. Lerman and Manski (1981) conducted the first extensive application of Monte Carlo integration as a numerical technique to the estimation of LDV models using the CMC simulator.

The CMC is quick to compute and ideal for computers with a "vectorization facility."[13] However, the CMC also has at least two major drawbacks: First, it is *not* continuous in parameters. The simulator jumps at parameter values where a $\tilde{y}_r$ is on the boundary of $\mathbf{D}$. For example, consider parameter values (μ_0, Γ_0) chosen so that the m^{th} element of the r^{th} simulation equals its lower bound in $\mathbf{D}$:

$$\tilde{y}_{rm} = \mu_{0m} + \Gamma_{0m}\eta_r = b_{0m}$$

where Γ_{0m} is the m^{th} row of Γ_0. Decreasing the parameter μ_m from μ_{0m} will cause the indicator $\mathbf{1}\{\tilde{y}_r \in \mathbf{D}\}$ to jump from 1 to 0, and this will result in discrete jumps in the elements of $h(\tilde{y}_r, \mathbf{D})$ and $\bar{h}$. Such discontinuities make computation of estimators and asymptotic distribution theory awkward.[14] Second, the number of computations required by the CMC rises inversely with $\Pr\{\mathbf{D}; \mu, \Omega\}$, which makes it intractable when this probability is small. It should be noted that in principle the accuracy of the CMC can be improved by use of so-called *simulation-variance-reduction techniques*, as, for example, the use of control and antithetic variates. See Hendry (1984) for definitions.

[13] Such a mechanism allows simultaneous operation on adjacent elements of a vector using multiple processors. See Hajivassiliou (1993b) who shows that the CMC exhibits the greatest speed gains from vectorization among 13 alternative simulation methods.

[14] See Quandt (1986) for a discussion of iterative parameter search algorithms and their requirements for differentiability of the function to be optimized.

3.2.2 Importance Sampling

Importance sampling is another general method for reducing the sampling variance of integrals computed by Monte Carlo integration over (censoring) intervals. The CMC involves sampling y^* from the $\phi(y^* - \mu, \Omega)$ p.d.f. and evaluating the function $h(y^*, \mathbf{D})$. A simple generalization of this procedure rewrites $\mathrm{E}[h]$ in terms of another sampling distribution γ:

$$\mathrm{E}[h] = \int h(y^*, \mathbf{D})\phi(y^* - \mu, \Omega)\, dy^* = \int \left[h(\tilde{y}, \mathbf{D})\frac{\phi(\tilde{y} - \mu, \Omega)}{\gamma(\tilde{y}; \mu, \Omega, \delta)}\right] \gamma(\tilde{y}; \mu, \Omega, \delta)\, d\tilde{y}.$$

δ is a vector of parameters characterizing the design of the importance sampler $\gamma(\cdot)$. Note that for $h(\cdot) = 1$, this expression corresponds to $\Pr\{\mathbf{D}; \mu, \Omega\}$. By drawing a random variable $\tilde{y}$ from the importance p.d.f. γ and evaluating the weighted indicator function $h(\tilde{y})w(\tilde{y})$, where

$$w(\tilde{y}) \equiv \frac{\phi(\tilde{y} - \mu, \Omega)}{\gamma(\tilde{y}; \mu, \Omega, \delta)},$$

one obtains an alternative unbiased simulation of $\Pr\{\mathbf{D}; \mu, \Omega\}$. The first advantage offered by importance sampling is the ability to substitute sampling from γ for sampling from ϕ. In some cases, γ may be sampled more quickly, or, in a more general setting, sampling from ϕ may be impractical.

In addition, if γ also has an analytical integral over a truncated sampling region $\mathbf{C}$ such that $\mathbf{D} \subseteq \mathbf{C}$, then this analytical integral can be exploited as an approximation to $\Pr\{\mathbf{D}; \mu, \Omega\}$ as follows:

$$\Pr\{\mathbf{D}; \mu, \Omega\} = \Pr\{\tilde{y} \in \mathbf{C}\} \int_{\mathbf{C}} \mathbf{1}\{\tilde{y} \in \mathbf{D}\} w(\tilde{y}) \frac{\gamma(\tilde{y}; \mu, \Omega, \delta)}{\Pr\{\tilde{y} \in \mathbf{C}\}} d\tilde{y}.$$

By drawing from the truncated p.d.f. $\gamma(\tilde{y}, \mu, \Omega, \delta)/\Pr\{\tilde{y} \in \mathbf{C}\}$, fewer simulations are 'wasted' on outcomes of zero and, in effect, $\Pr\{\tilde{y} \in \mathbf{C}\}w(\tilde{y})$ approximates $\Pr\{\mathbf{D}; \mu, \Omega\}$. When γ is a good approximation to ϕ, so that the ratio of densities $w \equiv \phi/\gamma$ is relatively constant, the sampling variance of the importance-sampling simulator is small. As noted above, the sampling variance of the CMC for a single simulation is $P(1-P)$, while the sampling variance of the importance sampler is

$$\mathrm{V}\left(P_{\mathbf{C}} \cdot \mathbf{1}\{\tilde{y} \in \mathbf{D}\} w(\tilde{y})\right) = P_{\mathbf{C}}^2 \cdot P_{\mathbf{D}} \cdot \left[\mathrm{V}\left(w(\tilde{y}) \mid \tilde{y} \in \mathbf{D}\right) + (1 - P_{\mathbf{D}}) \cdot \mathrm{E}\left(w(\tilde{y}) \mid \tilde{y} \in \mathbf{D}\right)^2\right],$$

where $P_{\mathbf{C}} \equiv \Pr\{\tilde{y} \in \mathbf{C}\}$ and $P_{\mathbf{D}} \equiv \Pr\{\tilde{y} \in \mathbf{D}\}$. In the extreme case that $\gamma = \phi$, $\mathrm{V}\left(w(\tilde{y}) \mid \tilde{y} \in \mathbf{D}\right) = 0$ and $\mathrm{E}\left(w(\tilde{y}) \mid \tilde{y} \in \mathbf{D}\right)^2 = P_{\mathbf{D}}$. Therefore, good approximations to ϕ afford improvements over the CFC. Geweke (1989) introduces importance sampling in Monte-Carlo integration in the context of Bayesian estimation.[15]

Definition 4 (GHK Importance Sampling Simulator) *The GHK importance p.d.f. is the 'recursively truncated' multivariate normal p.d.f.*

$$\gamma(\tilde{y}; \mu, \Omega, \mathbf{D}) = \phi(\tilde{y} - \mu, \Omega) \left[\prod_{m=1}^{M} \left\{\Phi\left(c_{1m}, \sigma^2_{m|<m}\right) - \Phi\left(c_{0m}, \sigma^2_{m|<m}\right)\right\}\right]^{-1} \tag{26}$$

[15] Other investigations of the use of Monte Carlo integration in Bayesian analysis are, *inter alia*, Bauwens (1984), Kloek and van Dijk (1978), and West (1990).

for $\tilde{y} \in \mathbf{D}$ *where* $\sigma_{m|<m} \equiv \sqrt{\Omega_{mm|<m}}$ *and*

$$c_{im} \equiv b_{im} - \mu_{m|<m}(\tilde{y}_{<m}), i = 0, 1.$$

By construction, the support of this p.d.f. is $\mathbf{D}$. *Conditional on* $\tilde{y}_{<m}$, $\tilde{y}_m$ *is univariate truncated normal on* $\mathbf{D}_m$ *with conditional mean determined by* $\tilde{y}_{<m}$. *Draws from* γ *can be made recursively according to the formula*

$$\tilde{y}_m = \mu_{m|<m}(\tilde{y}_{<m}) + \sigma_{m|<m}\Phi^{-1}\left[\omega_m \Phi\left(c_{1m}, \sigma^2_{m|<m}\right) - (1-\omega_m)\Phi\left(c_{0m}, \sigma^2_{m|<m}\right)\right] \tag{27}$$

where the ω *are independently distributed uniform random variables.*[16] *The GHK simulator is the product*

$$h_{GHK}(\tilde{y}) \equiv \prod_{m=1}^{M} \left\{\Phi\left(c_{1m}, \sigma^2_{m|<m}\right) - \Phi\left(c_{0m}, \sigma^2_{m|<m}\right)\right\} \cdot \begin{pmatrix} 1 \\ \tilde{y} \\ vec(\tilde{y}\tilde{y}') \end{pmatrix} \tag{28}$$

is an unbiased simulator of $\mathrm{E}(h)$.

The GHK simulator was developed by Geweke (1992), Hajivassiliou and McFadden (1990), and Keane (1990). Experience suggests that the sampling variance of $h_{GHK}(\tilde{y})$ is very small so that it approximates $\mathrm{E}(h)$ well in practice. This approximant has the properties of lying in the unit interval, summing to one over all the disjoint rectangular regions surrounding and including $\mathbf{D}$, and being a continuous function of ω, μ, Ω, b_0, and b_1. These properties are discussed in Börsch-Supan and Hajivassiliou (1993). Moreover, Hajivassiliou *et al.* (1992) found conclusive evidence for the superior root-mean-squared-error performance of the GHK method in an extensive Monte-Carlo study comparing the GHK to 12 other simulators for normal rectangle probabilities $\Pr\{\mathbf{D}; \mu, \Omega\}$.

3.3 Truncated Simulation

We now turn to the expectations in (24). These are ratios of the integrals in (25) and cannot be simulated without bias using the censored simulation methods above. Even ignoring the bias, one must ensure that the denominator of the ratio is not zero. For example, the CMC and some importance sampling simulators can yield outcomes of zero for probabilities and thus violate this requirement.[17] In this subsection, we describe two general procedures which draw directly from the truncated distributions associated with the expectations in equation (24).

3.3.1 Acceptance/Rejection Methods

Acceptance/rejection (A/R) methods provide a mechanism for drawing from a conditional density when practical exact transformations from uniform or standard normal variates are not available. The following result is standard; see Devroye (1986), Fishman (1973), or Rubinstein (1981) for proofs.

[16] This method is described extensively in Devroye (1986) and is a simple application of the cumulative probability integral transform result — see Feller (1971). Computationally more efficient methods for generating univariate truncated normal variates exist — for example Geweke (1992). The advantage of the method presented in the preceding equation, however, is that it is continuous in μ, Ω, and ω_m, which, as already mentioned, is a desirable property of simulators for asymptotic theory and for iterative parameter search. The method of constructing γ in this example can also be extended to a bivariate version using a bivariate normal c.d.f. and standardizing adjacent pairs of elements.

[17] It should be noted that one of the attractive properties of the GHK simulator is that it generates simulated probability values that are bounded away from 0 and 1, unlike many other importance sampling simulators. See Börsch-Supan and Hajivassiliou (1993) for details.

Proposition 1 *Suppose $\phi(y^*)$ is a J-dimensional density, and one wishes to sample from the conditional density $\phi(y^*|\mathbf{D}) \equiv \phi(y^*)/\int_{\mathbf{D}} \phi(y^*)\, dy^*$. Suppose $\gamma(\tilde{y})$ is a density with a support $\mathbf{A}$ from which it is practical to sample, with the property that*

$$\sup_{\mathbf{D}} \frac{\phi(\tilde{y})}{\gamma(\tilde{y})} \leq \alpha < +\infty,$$

where $\mathbf{D} \subseteq \mathbf{A}$. Draw $\tilde{y}$ from γ and ω from a uniform density on $[0,1]$, repeat this process until a pair satisfying $\tilde{y} \in \mathbf{D}$ and $\phi(\tilde{y}) \geq \omega\alpha \cdot \gamma(\tilde{y})$ is observed, and accept the associated $\tilde{y}$. Then, the accepted points have density $\phi(\cdot|\mathbf{D})$.

The choice of a suitable comparison density $\gamma(\cdot)$ is important because it determines the expected 'yield' of the acceptance/rejection scheme. The main attractive feature of A/R is that the *accepted* draws have the correct truncated distribution. The practical shortcoming, though, is that the operations necessary until a specific number of draws are accepted may be very large.

The A/R scheme also provides an unbiased simulator of $1/\Pr\{\mathbf{D};\mu,\Omega\}$ if $\int_{\mathbf{D}} \gamma(\tilde{y})\, d\tilde{y} = \Gamma(\mathbf{D})$ is practical to compute. The conditional probability of acceptance, given $\{\tilde{y} \in \mathbf{D}\}$, is $\int_{\mathbf{D}} \phi(\tilde{y})\, d\tilde{y}/\alpha = \Pr\{\mathbf{D}\}/\alpha$, so that the marginal probability of acceptance is $\Gamma(\mathbf{D})\Pr\{\mathbf{D}\}/\alpha$. The distribution of the number of trials to get an acceptance is the geometric and its expectation is $\alpha/[\Gamma(\mathbf{D})\Pr\{\mathbf{D}\}]$. Therefore, if t is the number of draws made until $\tilde{y}$ is accepted, $t \cdot \Gamma(\mathbf{D})/\alpha$ is an unbiased simulator of $1/\Pr\{\mathbf{D}\}$.

Example 7 *The recursively truncated normal p.d.f. in Definition 4 works well in practice as the comparison distribution. A bound on the density ratio is given by*

$$\alpha = \prod_{m=1}^{M} \left\{ \Phi\left(b_{1m} - \mu_{m|<m}(b_{1<m}), \sigma^2_{m|<m}\right) - \Phi\left(b_{0m} - \mu_{m|<m}(b_{0<m}), \sigma^2_{m|<m}\right) \right\}$$

where the conditional moments are conditioned on $\tilde{y}_{<m}$ equal to the boundaries. Since $\mathbf{A} = \mathbf{D}$, $\Gamma(\mathbf{D}) = 1$.

3.3.2 Gibbs Resampling

Gibbs resampling is another way to draw from truncated distributions. An infinite number of calculations are required to generate a finite number of draws with distribution approaching the true one. But convergence to the true distribution is geometric in the number of resamplings, hence the performance of this simulator in practice is generally very satisfactory. In addition, this simulator is continuous and differentiable in the parameters μ and Ω. The Gibbs simulator is based on a Markov chain that utilizes computable univariate truncated normal densities to construct transitions, and has the desired truncated multivariate normal as its limiting distribution.[18] This simulator is defined by the following Markovian updating scheme.

Proposition 2 *Consider the multivariate normal distribution $N(\mu,\Omega)$ truncated on $\mathbf{D}$, which is assumed to be finite. Define a recursive procedure with steps $j = 1,\ldots,J$ in rounds $g = 1,\ldots,G$. Let $\{y^{*(jg)}\}$ be a sequence on $\mathbf{D}$ such that on the j^{th} step of the g^{th} round, the j^{th} element of $y^{*(jg)}$ is computed from $y^{*(j,g-1)}_{-j}$ by*

$$y^{*(jg)}_j = \mu_{j|-j}\left(y^{*(j,g-1)}_{-j}\right) + \sigma_{j|-j} \cdot \Phi^{-1}\left[\omega_{j,g-1}\Phi\left(c^g_{1j}, \sigma_{j|-j}\right) - (1-\omega_{j,g-1})\Phi\left(c^g_{0j}, \sigma_{j|-j}\right)\right]$$

[18] This simulator can be generalized in principle to non-normal distributions, provided the corresponding univariate distributions are easy to sample.

where

$$c_{ij}^{g} \equiv b_{ij} - \mu_{j|-j}\left(y_{-j}^{*(j,g-1)}\right),\ i = 0, 1,$$

and the ω_{jg} are independent uniform [0,1] variates and $\sigma_{j|-j} = \sqrt{\Omega_{jj|-j}}$. Repeat this process G "Gibbs resampling rounds." Then the random draws obtained by this simulator have a distribution that converges in L_1 norm at a geometric rate to the true truncated distribution $\{y^*|y^* \in \mathbf{D}\}$ as the number of Gibbs resampling rounds G grows to infinity.

This result is proved in Hajivassiliou and McFadden (1990). It relies on stochastic relaxation techniques as discussed in Geman and Geman (1984). See also Tierny (1992) for other theoretical results on the Gibbs resampling scheme.[19] We present below Monte Carlo experiments with simulation estimators based on this truncated simulation scheme.

4 Simulation and Estimation of LDV Models

4.1 An Overview

In this Section, we bring together the parametric estimation of the LDV models described in Section 2 with the simulation methods in Section 3. Our focus is the consistent estimation of the parameters of the model; we defer the discussion of limiting distributions to a later section. Our exposition follows the general historical trend of thought in this area. We begin with the application of simulation to approximating the log-likelihood function. Next, we consider the simulation of moment functions. Because of the simulation biases that naturally arise in the log-likelihood approach, the unbiased simulation of moment functions and the method of moments is an alternative approach. Finally, we discuss simulation of the score function. Solving the normal equations of ML estimation is a special case of the method of moments and simulating the score function offers the potential for efficient estimation.

One can organize a description of the methods along the following lines. Figure 1 gives a diagrammatic presentation of a useful taxonomy. In this figure, the various estimation methods are represented as elliptical sets and the properties of the associated simulation methods are represented as rectangular sets. Five families of estimation methods are pictured. All of the methods fall into the class of *generalized method of simulated moments* (GMSM). This is the simulated counterpart to the generalized method of moments (GMM) (see Newey and McFadden (1993)). Within the GMSM, fall the *method of simulated scores* (MSS), the *simulated EM* (SEM), the *method of simulated moments* (MSM), and *maximum simulated likelihood* (MSL). In parallel with the types of LDV models, the simulation methods are divided between censored and truncated sampling. The simulation methods are further separated into those that simulate the efficient score of the LDV models with and without bias.

The MSM is a simulated counterpart to the method of moments (MOM). As the figure shows, the MSM is restricted to simulation methods that generate unbiased simulations using censored simulation methods. The MSL estimation method also rests on censored simulation but, as we will explain, the critical object (the log-likelihood function) is simulated with bias. The SEM algorithm is an extension of the EM algorithm using unbiased simulations from truncated distributions; it falls, therefore, in the upper half of the figure. Of these methods, only the MSS has versions that use both classes of simulation methods, censored and truncated, that we have described above.

[19] The usefulness of Gibbs resampling for Bayesian estimation has been recognized by Geweke (1992), Chib (1993), and by McCulloch and Rossi (1993).

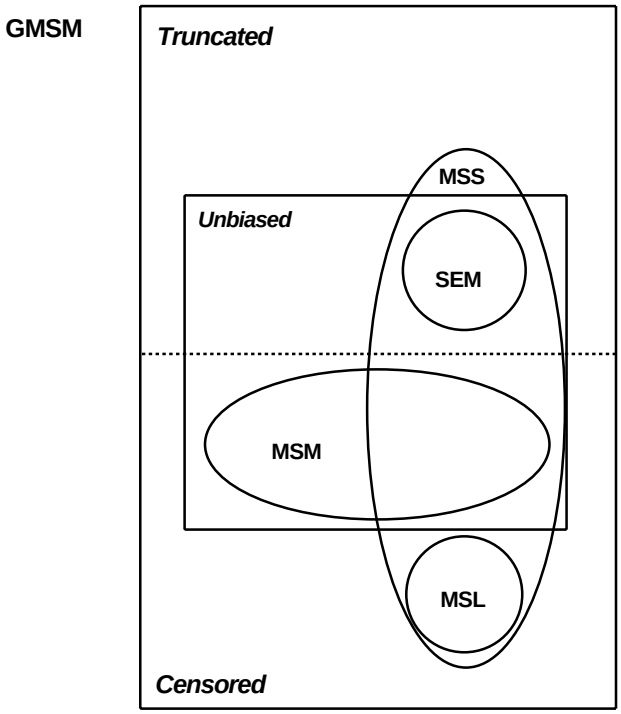

Figure 1: Taxonomy of Simulation Estimators

Throughout this section, we will assume that we are working with models for which the maximum likelihood estimator is well-behaved. In particular, we suppose that the usual regularity conditions are met, ensuring that the ML estimator is the most efficient CUAN estimator. We will illustrate the methods using the *rank ordered probit* model. This LDV model is a natural candidate for most approaches to estimation with simulation and the exact MLE performs well in small samples.

Example 8 (Rank Ordered Probit) *The* rank ordered probit *model is a generalization of the multinomial probit model described in Example 1. Instead of observing only the most preferred (or highest ranked) alternative, each observation records the rank order of the alternatives from most preferred to least preferred. The rank ordering yields considerably more information about the underlying preference parameters than the simpler, highest-ranked-alternative response. Hence, consumer survey designers often prefer to ask for complete rankings.*

We can express the observation rule of rank ordered data algebraically as

$$y = \tau_{ij}(y^*) = \mathbf{1}\left\{y^*_{(i)} = y^*_j\right\}, \quad i, j = 1, \ldots, J$$

where the $\left\{y^*_{(j)}\right\}$ *correspond to the order statistics of* y^**,*

$$y^*_{(1)} \leq y^*_{(2)} \leq \cdots \leq y^*_{(J)},$$

so that the first element of y *is the index of the largest element of* y^* *and so on until the last element is assigned the index of the smallest element of* y^**. The sample space of* y *consists of the*

$J! \equiv J \cdot (J-1) \cdot \ldots \cdot 2$ *different* $J \times J$ *matrices containing zeros and ones such that only a single entry equals one in each row and column:*

$$\mathbf{B} = \left\{ [y_{ij}; i, j = 1, \ldots, J] \mid y_{ij} \in \{0, 1\}, \sum_i y_{ij} = \sum_j y_{ij} = 1 \right\}.$$

Thus, even moderate numbers of alternatives correspond to discrete sampling spaces with many outcomes.

The c.d.f. of y *is not particularly informative; it is simpler to derive the probability of each possible outcome directly: The rank ordering* y *corresponds to values of* y^* *in a set satisfying* $J-1$ *inequalities:*

$$\mathbf{D}(y) \equiv \left\{ y^* \in \mathbf{R}^J \mid y_{1\cdot} y^* \leq y_{2\cdot} y^* \leq \ldots \leq y_{J\cdot} y^* \right\},$$

where $y_{j\cdot}$ *is the row vector* $[y_{j1}, \ldots, y_{jJ}]$*. Such additional inequalities as* $y_{1\cdot} y^* \leq y_{3\cdot} y^*$ *are redundant. As in the multinomial choice model, it is convenient to transform the latent* y^* *into a vector of* $J-1$ *differences:*

$$z_y \equiv \Delta_y y^* = [y_{i\cdot} y^* - y_{i+1,\cdot} y^*; i = 1, \ldots, J-1]$$

where

$$\Delta_y \equiv [y_{ij} - y_{i+1,j}; i = 1, \ldots, J-1; j = 1, \ldots, J]$$

is a $J-1 \times J$ *differencing matrix. According to this definition,* $\mathbf{D}(y) = \{y^* \mid z_y \leq 0\}$*. The transformed random vector* z_y *is also multivariate normal and for all* $Y \in \mathbf{B}$*,*

$$f(\theta; Y) = \Pr\{y = Y; \mu, \Omega\} = \Phi(\Delta_Y \mu, \Delta_Y \Omega \Delta_Y'). \tag{29}$$

One probability term in this p.d.f. is equivalent in computational complexity to the normal orthant integrals of the choice probabilities in Example 1.

We will use the various simulation and estimation methods to estimate this rank ordered probit model in Monte Carlo experiments. Because a natural standard of comparison is the MLE, we present first a Monte Carlo experiment for the MLE in a workable case.

Example 9 *When* $J = 4$*, the MLE is computable using standard approximation methods. In our basic Monte Carlo experiment the population parameters will be*

$$\mu = \begin{bmatrix} -1 \\ -1/3 \\ 1/3 \\ 1 \end{bmatrix} \quad \textit{and} \quad \Omega = \begin{bmatrix} 1 & 1/2 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1/2 & 1 \end{bmatrix}.$$

These values yield a reasonable amount of variation in y *and they induce significant inconsistency in the popular rank ordered logit estimator (Beggs* et al. *(1981)) when it is applied to the data. The block diagonal* Ω *contains covariances among the latent* y^* *that are zero in the latent logit model. The* μ *and* Ω *parameters are not all identifiable and so we normalize by reducing the parameterization to* $\Delta_Y \mu$ *and* $\Delta_Y \Omega \Delta_Y'$ *for* $Y = I_4$*, the* 4×4 *identity matrix. The first variance in* $\Delta_Y \Omega \Delta_Y'$ *is also scaled to 1. In order to restrict* $\Delta_Y \Omega \Delta_Y'$ *to be positive semi-definite, this covariance matrix is also parameterized in terms of its Cholesky square root. Putting the mean parameters first, then stacking the non zero elements of the Cholesky parameters, the identifiable population parameter vector is* $\theta_0 = [-0.6667, -0.6667, -0.6667, 0.5000, 1.3230, 0.0000, -0.3780, 0.9258]$.

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.6667*	*-0.6864*	*0.1317*	*-0.7702*	*-0.6807*	*-0.5921*
θ_2	*-0.6667*	*-0.6910*	*0.2351*	*-0.8354*	*-0.6629*	*-0.5231*
θ_3	*-0.6667*	*-0.7063*	*0.2263*	*-0.8276*	*-0.6648*	*-0.5374*
θ_4	*-0.5000*	*-0.5135*	*0.2265*	*-0.6402*	*-0.5016*	*-0.3645*
θ_5	*1.3230*	*1.3536*	*0.3002*	*1.130*	*1.317*	*1.519*
θ_6	*0.000*	*-0.0127*	*0.1797*	*-0.1241*	*-0.008616*	*0.09545*
θ_7	*-0.3780*	*-0.4081*	*0.1909*	*-0.5158*	*-0.3891*	*-0.2765*
θ_8	*0.9258*	*0.9385*	*0.2461*	*0.7513*	*0.9140*	*1.074*

Table 1: Sample Statistics for Rank Ordered Probit MLE

The basic Monte Carlo experiment will be a random draw from the distribution of each estimator for $N = 100$ *observations on* y. *There will be 500 replications of each estimator. Results of the experiment for the MLE are in Table 1. The MLE has a small bias relative to its sampling variance and the sampling variance is small enough to make hypothesis tests for equal means or zero covariances quite powerful. It appears that the bias in the MLE is largely caused by asymmetry in the sampling distribution: The medians are closer to the population values than the means. Overall, the asymptotic approximation to the distribution of the MLE is good. The inverse information matrix predicts the standard deviations in the fourth column of Table 1 to be 0.1296, 0.1927, 0.1703, 0.2005, 0.2248, 0.1543, 0.1514, 0.1987. Therefore, the actual sampling distribution has more variation than the asymptotic approximaton.*

For the simulation estimators, we will also conduct Monte Carlo experiments for a model with $J = 6$ *alternatives. In that case, the MLE is not easily computed. We will use the population values*

$$\mu = \begin{bmatrix} -1 \\ -3/5 \\ -1/5 \\ 1/5 \\ 3/5 \\ 1 \end{bmatrix} \quad \textit{and} \quad \Omega = \begin{bmatrix} 1 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5/4 & 3/4 & 1/4 & 1/4 \\ 0 & 0 & 3/4 & 5/4 & 1/4 & 1/4 \\ 0 & 0 & 1/4 & 1/4 & 5/4 & 3/4 \\ 0 & 0 & 1/4 & 1/4 & 3/4 & 5/4 \end{bmatrix},$$

which correspond to θ_0 *=[-0.4000, -0.4000, -0.4000, -0.4000, -0.4000, -0.5000, 1.414, 0.000, -0.3536, 0.9354, 0.000, -0.1768, -0.6013, 1.052, 0.000, 0.000, 0.000, -0.4752, 0.8799] when normalizing on* $Y = I_6$.

4.2 Simulation of the Log-Likelihood Function

One of the earliest applications of simulation to estimation was the general computation of multivariate integrals in such likelihoods as that of the multinomial probit by Monte Carlo integration. Crude Monte Carlo simulation can approximate the probabilities of the multinomial probit to any desired degree of accuracy, so that the corresponding *maximum simulated likelihood* (MSL) estimator can approximate the ML estimator.

Definition 5 (Maximum Simulated Likelihood) *Let the log-likelihood function for the unknown parameter vector θ given the sample of observations $(y_n, n = 1, \ldots, N)$ be*

$$\ell_N(\theta) \equiv \sum_{n=1}^{N} [\ln f(\theta; y_n)]$$

and let $\tilde{f}(\theta; y, \omega)$ be an unbiased simulator so that $f(\theta; y) = \mathrm{E}_\omega[\tilde{f}(\theta; y, \omega)|y]$ where ω is a simulated vector of R random variates. The maximum simulated likelihood estimator is

$$\hat{\theta}_{MSL} \equiv \arg\max_\theta \tilde{\ell}_N(\theta)$$

where

$$\tilde{\ell}_N(\theta) \equiv \sum_{n=1}^{N} \ln \tilde{f}(\theta; y_n, \omega_n)$$

for some given simulation sequence $\{\omega_n\}$.

It is important to note that MSL estimator is conditional on the sequence of simulators $\{\omega_n\}$. For both computational stability and asymptotic distribution theory, it is important that the simulations do not change with the parameter values. See McFadden (1989) and Pakes and Pollard (1989) for an explanation of this point.

Example 10 *Börsch-Supan and Hajivassiliou (1993) proposed MSL estimation of the multinomial probit model of Example 1 using the GHK simulator for the choice probabilities. In this example, we make similar calculations for the rank ordered probit model of Example 9. Instead of the normal probality function in (29), we used the probability simulator in the first element of h_{GHK} in (28) to compute the simulated log-likelihood function $\tilde{\ell}_N(\theta)$.*[20] *For the simulations of the probability of each observation, we drew a vector of $J - 1 = 3$ independently distributed uniform random variables for each ω_n. For each replication of $\hat{\theta}_{MSL}$, we drew a new dataset $\{(y_n, \omega_n); n = 1, \ldots, N\}$ before maximizing $\tilde{\ell}_N(\theta)$ over θ. Each $\tilde{f}(\theta; y_n, \omega_n)$ consisted of a single simulation of $f(\theta; y_n)$ ($R = 1$).*

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.6667*	*-0.7230*	*0.1424*	*-0.8219*	*-0.7198*	*-0.6253*
θ_2	*-0.6667*	*-0.6077*	*0.2162*	*-0.7342*	*-0.5934*	*-0.4640*
θ_3	*-0.6667*	*-0.9555*	*0.2520*	*-1.087*	*-0.9256*	*-0.7860*
θ_4	*-0.5000*	*-0.6387*	*0.1430*	*-0.7305*	*-0.6415*	*-0.5379*
θ_5	*1.3230*	*1.2595*	*0.1741*	*1.134*	*1.237*	*1.353*
θ_6	*0.0000*	*0.0131*	*0.1717*	*-0.09063*	*0.01013*	*0.1285*
θ_7	*-0.3780*	*-0.6715*	*0.2088*	*-0.7883*	*-0.6639*	*-0.5292*
θ_8	*0.9258*	*1.3282*	*0.2211*	*1.185*	*1.301*	*1.448*

Table 2: Sample Statistics for Rank Ordered Probit MSLE Using GHK (J=4, R=1)

The results of this Monte Carlo for $J = 4$ are in Table 2. In contrast with the MLE, this MSLE exhibits much larger bias. The median is virtually identical to the mean. The sampling variances are also larger, particularly for the covariance parameters. Nevertheless, this MSLE gives a rough approximation to the population parameters.

[20] The order of integration affects this simulator, but we do not attempt to describe our particular orderings. They were chosen purely on the basis of a convenient algorithm for finding the limits of integration.

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.4000*	*-0.4585*	*0.1504*	*-0.5565*	*-0.4561*	*-0.3664*
θ_2	*-0.4000*	*-0.2489*	*0.2059*	*-0.3898*	*-0.2460*	*-0.0940*
θ_3	*-0.4000*	*-0.5054*	*0.1710*	*-0.6056*	*-0.4957*	*-0.3891*
θ_4	*-0.4000*	*-0.4589*	*0.2013*	*-0.5779*	*-0.4551*	*-0.3216*
θ_5	*-0.4000*	*-0.6108*	*0.1882*	*-0.6934*	*-0.6016*	*-0.5042*

Table 3: Sample Statistics for Rank Ordered Probit MSLE Using GHK (J=6, R=1)

The results of this Monte Carlo for $J = 6$ are in Table 3. For brevity, only the mean parameters are listed. Once again, substantial biases appear in the sample of estimators. Given our experience with $J = 4$, it seems likely that these biases are largely due to simulation. We will confirm this below as we apply other methods to this case.

Note that unbiased simulation of the likelihood function is neither necessary nor sufficient for consistent MSL estimation. Because the estimator is a nonlinear function (through optimization) of the simulator, the MSL estimator will generally be a biased simulation of the MLE even when the criterion function of estimation is simulated without bias because

$$\mathrm{E}\left[\tilde{\ell}(\theta)\right] = \ell(\theta) \not\Rightarrow \mathrm{E}\left[\arg\max_{\theta} \tilde{\ell}(\theta)\right] = \arg\max_{\theta} \ell(\theta).$$

Note also that while unbiased simulation of the likelihood function is often straightforward, unbiased simulation of the *log*-likelihood is generally infeasible. The logarithmic transformation of the intractable function introduces a nonlinearity that cannot be overcome simply. However, to obtain an estimator with the same *probability limit* as the MLE, a sufficient characteristic of a simulator for the log-likelihood is that its sample average converge to the same limit as the sample average log-likelihood. Only by reducing the error of a simulator for the log-likelihood function to zero at a sufficiently rapid rate with sample size can one expect to obtain a consistent estimator. Such results rest on a general proposition that underlies the consistency of many extremum estimators (see Newey and McFadden (1993), Theorem 2.1):

Lemma 1 *Let*

1. $\theta \in \Theta$, *a compact subset of* $\mathbf{R}^K$,
2. $Q_0(\theta)$, $Q_N(\theta)$ *be continuous in* θ,
3. $\theta_0 \equiv \arg\max_{\theta\in\Theta} Q_0(\theta)$ *be unique,*
4. $\hat{\theta}_N \equiv \arg\max_{\theta\in\Theta} Q_N(\theta)$ *and*
5. $Q_N(\theta) \to Q_0(\theta)$ *in probability uniformly in* $\theta \in \Theta$ *as* $N \to \infty$.

Then $\hat{\theta}_N \to \theta_0$ *in probability.*

We will assume from now on that the log-likelihood function is sufficiently regular to exploit this lemma. In particular, we suppose that the y_n are i.i.d., that θ is identifiable, that $f(\theta; y)$ is continuous at each θ in a compact parameter space Θ, and that $\mathrm{E}\left[\sup_{\theta\in\Theta} |\ln f(\theta; y)|\right] < \infty$. We refer

the reader to Newey and McFadden (1993), Theorem 2.5 for further discussion of these conditions and their roles.

For LDV models with censoring, the generic likelihood simulator $\tilde{f}(\theta; y_n, \omega_n)$ is the average of R replications of one of the simulation methods described above:

$$\tilde{f}(\theta; y_n, \omega_n) \equiv \frac{1}{R} \sum_{r=1}^{R} \tilde{f}(\theta; y_n, \omega_{nr}).$$

If the model includes truncation, then the likelihood simulation typically involves a ratio of such averages, because a normalizing probability appears in the denominator, although unbiased simulation of the ratio is possible (see Section 3.3). In any case, the simulation error will generally be $O_P(1/R)$. Thus, a common approach to approximating the log-likelihood function with sufficient accuracy is increasing the number of replications per observation R with the sample size N. This statistical approach is in contrast to a strictly numerical approach of setting R high enough to achieve a specified numerical accuracy independent of sample size.

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.6667*	*-0.6795*	*0.1366*	*-0.7726*	*-0.6774*	*-0.5840*
θ_2	*-0.6667*	*-0.6528*	*0.2267*	*-0.7913*	*-0.6268*	*-0.5029*
θ_3	*-0.6667*	*-0.8327*	*0.2299*	*-0.9686*	*-0.8085*	*-0.6768*
θ_4	*-0.5000*	*-0.5771*	*0.2159*	*-0.7076*	*-0.5641*	*-0.4412*
θ_5	*1.3230*	*1.3582*	*0.2459*	*1.1863*	*1.3184*	*1.5036*
θ_6	*0.0000*	*-0.0121*	*0.2089*	*-0.1380*	*-0.01570*	*0.1275*
θ_7	*-0.3780*	*-0.5034*	*0.2016*	*-0.6256*	*-0.4875*	*-0.3753*
θ_8	*0.9258*	*1.1334*	*0.2454*	*0.9505*	*1.1142*	*1.2814*

Table 4: Sample Statistics for Rank Ordered Probit MSLE Using GHK (J=4, R=5)

Example 11 *For illustration, let us increase the replications in the previous examples from* $R = 1$ *simulation per observation to* 5. *The summary statistics are listed in Tables 4 and 5. In both cases,* $J = 4$ *and* $J = 6$, *the biases are significantly reduced. See Börsch-Supan and Hajivassiliou (1993) for a more extensive Monte Carlo study of the relationship between* R *and bias in the multinomial probit model.*

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.4000*	*-0.4088*	*0.1256*	*-0.4893*	*-0.4053*	*-0.3227*
θ_2	*-0.4000*	*-0.3059*	*0.1776*	*-0.4200*	*-0.2966*	*-0.1846*
θ_3	*-0.4000*	*-0.4554*	*0.1387*	*-0.5373*	*-0.4553*	*-0.3615*
θ_4	*-0.4000*	*-0.4288*	*0.1661*	*-0.5369*	*-0.4219*	*-0.3142*
θ_5	*-0.4000*	*-0.5046*	*0.1773*	*-0.6211*	*-0.4976*	*-0.3872*

Table 5: Sample Statistics for Rank Ordered Probit MSLE Using GHK (J=6, R=5)

In the rank ordered probit model and similar discrete LDV models, all that is necessary for estimator consistency is that $R \to \infty$ as $N \to \infty$. No relative rates are required provided that the likelihood is sufficiently regular. Nor must the simulations ω satisfy any restrictions on dependence across observations. The following proposition, taken from Lee (1993), establishes this situation.

Proposition 3 *Let $f(\theta;y)$ be uniformly bounded away from zero for all $\theta \in \Theta$, a compact set, and all $y \in \mathbf{B}$, the sample space of y. Assume that the set of regularity conditions in the paragraph after Lemma 1 hold. Let $\{\omega_{nr}\}$ be and i.i.d. sequence over the index r. The MSL estimator $\hat{\theta}_{MSL} \equiv \arg\max_\theta \frac{1}{N}\sum_{n=1}^{N} \ln \tilde{f}(\theta; y_n, \omega_n)$ is consistent if $R \to \infty$ as $N \to \infty$.*

Proof. By a uniform law of large numbers and the lower bound of f,

$$\sup_{y,\theta} \left| \frac{\tilde{f}(\theta; y_n, \omega_{nr})}{f(\theta; y_n)} - 1 \right| \xrightarrow{P} 0, \quad \text{as} \quad R \to \infty,$$

so that

$$\sup_{\theta} \frac{1}{N} \left| \tilde{\ell}_N(\theta) - \ell_N(\theta) \right| \xrightarrow{P} 0, \quad \text{as} \quad R \to \infty,\ N \to \infty.$$

Since our regularity assumptions in the paragraph after Lemma 1 guarantee that

$$\sup_{\theta} \left| \frac{1}{N}\ell_N(\theta) - \mathrm{E}[\ln f(\theta; y)] \right| \xrightarrow{P} 0 \quad \text{as} \quad N \to \infty,$$

then $\tilde{\ell}_N(\theta)/N$ also converges uniformly to $\mathrm{E}[\ln f(\theta; y)]$ and consistency follows by Lemma 1. □

Thus, the property of estimator consistency makes modest demands on the simulations of the likelihood function. Strictly speaking, one could employ a common sequence of simulations $\{\omega_r\}$ for all simulated likelihoods which grows at an arbitrarily slow rate with sample size. The differences between simulation designs appear only in the limiting normal distributions of the estimators. It is especially important to note that consistency does not confine such differences to sampling variances. Both the expectations and the variances of the approximate limiting distribution can be affected by the simulation design.

Note that Proposition 3 does not apply to models with elements of y which are continuously distributed and unbounded. Additional work is needed in this area. See Hajivassiliou and McFadden (1990) for the special conditions needed for an example of a multiperiod (panel) autocorrelated tobit model.

From the standpoint of asymptotic distribution theory, the simplest use of simulation makes independent simulations for the contribution of each observation to the likelihood function. If elements of the sequence $\{\omega_{nr}\}$ are independent across the observation index n, as well as the replication index r, then we preserve the independence of the $\tilde{f}(\theta; y_n, \omega_{nr})$ and its derivatives across n, permitting the application of familiar laws of large numbers and central limit theorems. When $\tilde{f}$ is differentiable in θ, we can make a familiar linear approximation for $\hat{\theta}_{MSL}$:

$$0 = \frac{1}{\sqrt{N}}\nabla_\theta \tilde{\ell}(\theta_0) + \left[\frac{1}{N}\nabla^2_\theta \tilde{\ell}(\bar{\theta})\right] \sqrt{N}(\hat{\theta}_{MSL} - \theta_0) \tag{30}$$

where the elements of $\bar{\theta}$ lie on the line segment between $\hat{\theta}_{MSL}$ and θ_0. The consistency of $\hat{\theta}_{MSL}$ implies the consistency of $\bar{\theta}$ which in turn implies that

$$\frac{1}{N}\nabla^2_\theta \tilde{\ell}(\bar{\theta}) \xrightarrow{P} \mathrm{E}\left[\nabla^2_\theta \ln f(\theta_0; y)\right] \equiv \mathcal{I}(\theta_0) \tag{31}$$

using the argument that supports Proposition 3. The leading term is a sum of N i.i.d. terms

$$\frac{1}{\sqrt{N}}\nabla_\theta \tilde{\ell}(\theta_0) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \frac{\nabla_\theta \tilde{f}(\theta_0; y_n, \omega_n)}{\tilde{f}(\theta_0; y_n, \omega_n)}$$

to which we would like to apply a central limit theorem. But we are prevented from this by the fact that the expectation of these terms is not zero. Consider the simple factorization, obtained by adding and subtracting terms,

$$\frac{1}{\sqrt{N}}\nabla_\theta \tilde{\ell}(\theta_0) = \frac{1}{\sqrt{N}}\nabla_\theta \ell(\theta_0) + A_N + B_N \tag{32}$$

where

$$\begin{array}{rcl} A_N & = & \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left\{\nabla_\theta \ln \tilde{f} - \mathrm{E}_\omega\left[\nabla_\theta \ln \tilde{f}\right]\right\} \\ B_N & = & \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left\{\mathrm{E}_\omega\left[\nabla_\theta \ln \tilde{f}\right] - \nabla_\theta \ln f\right\} \end{array} \tag{33}$$

A_N is a sum of i.i.d. terms with zero expectation and can be viewed as the source of pure simulation noise in $\hat{\theta}_{MSL}$. B_N is the potential source of simulation bias. The next result can be used to show that $R/\sqrt{N} \to \infty$ is a sufficient rate of increase to avoid such bias.

Proposition 4 *Let $\tilde{\mu}(\theta; y, \omega)$ be an unbiased simulator for $\mu(\theta; y)$ such that $V(\tilde{\mu}-\mu|y) = O(R^{-1})$. Let $s(\theta; y, \mu)$ be a moment function such that $\mathrm{E}[s(\theta_0; y, \mu)] = 0$. Consider the simulator $\tilde{s}(\theta; y) \equiv s(\theta; y, \tilde{\mu})$ and let $R/\sqrt{N} \to \infty$. If $\tilde{s}$ is Lipschitz in $\tilde{\mu} \in S$ uniformly in θ, then the simulation bias*

$$B_N \equiv \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left\{\mathrm{E}_\omega\left[\tilde{s}(\theta; y)\right] - s(\theta; y, \mu)\right\} \xrightarrow{P} 0.$$

Proof. If $\tilde{s}$ is Lipschitz in $\tilde{\mu}$ uniformly in θ then

$$\tilde{s} - s = \left[\nabla_\mu s(\theta; y, \mu)\right](\tilde{\mu} - \mu) + \left[\nabla_\mu s(\theta; y, \mu^*) - \nabla_\mu s(\theta; y, \mu)\right](\tilde{\mu} - \mu),$$

where μ^* is on the line segment joining $\tilde{\mu}$ and μ. According to the hypothesis of unbiasedness,

$$\mathrm{E}_\omega(\tilde{s} - s) = \mathrm{E}_\omega\left\{\left[\nabla_\mu s(\theta; y, \mu^*) - \nabla_\mu s(\theta; y, \mu)\right](\tilde{\mu} - \mu)\right\}$$

so that

$$\|\mathrm{E}_\omega(\tilde{s} - s)\| \leq M^* \mathrm{E}(\tilde{\mu} - \mu)^2 = O(R^{-1})$$

for some finite M^* according to the Lipschitz hypothesis. Therefore, $B_N = O_P(\sqrt{N}/R)$ and the result follows. □

In the multinomial and rank ordered probit cases, the Lipschitz requirement is generally met by the regularity conditions that bound the discrete probabilities and the smoothness of the probability simulator $\tilde{f}$: $\mu = (f, \nabla_\theta f)$, $\tilde{\mu} = (\tilde{f}, \nabla_\theta \tilde{f})$, and $s = (\nabla_\theta f)/f$. We are not aware of any slower rates for R that avoid bias in the limiting distribution of $\hat{\theta}_{MSL}$.

Proposition 5 *Let f be bounded uniformly away from zero and Lipschitz in θ on a compact space Θ. Let $\tilde{f}(\theta; y, \omega)$ be an unbiased differentiable simulator for $f(\theta; y)$, also bounded uniformly away from zero and Lipschitz in θ on Θ such that $V(\tilde{f}-f) = O(R^{-1})$. Let $R/\sqrt{N} \to \infty$. Then the simulation components*

$$A_N + B_N \equiv \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left\{\nabla_\theta \ln \tilde{f}(\theta; y_n, \omega_n) - \nabla_\theta \ln f(\theta; y_n)\right\} \xrightarrow{P} 0$$

and $\hat{\theta}_{MSL}$ is asymptotically efficient.

Proof. The difference between simulated and exact scores can be written

$$\begin{aligned} A_N + B_N &= \frac{1}{\sqrt{N}} \sum_n \frac{1}{\tilde{f}} \left[\nabla_\theta \tilde{f} - \tilde{f} \cdot \nabla_\theta \ln f \right] \\ &= O(1) \frac{1}{\sqrt{N}} \sum_n \left[\nabla_\theta \tilde{f} - \tilde{f} \cdot \nabla_\theta \ln f \right] \end{aligned}$$

By the Chebychev inequality,

$$\Pr \left\{ \left| \frac{1}{\sqrt{N}} \sum_n \left[\nabla_{\theta_i} \tilde{f} - \tilde{f} \cdot \nabla_{\theta_i} \ln f \right] \right| > \epsilon \right\} \leq \frac{1}{\epsilon \sqrt{N}} \sum_n V \left[\nabla_{\theta_i} \tilde{f} - \tilde{f} \cdot \nabla_{\theta_i} \ln f \right] = O(\sqrt{N}/R)$$

for each component of the gradient. The result follows from this order and equations (30)–(33). □

Propositions 4 and 5 demonstrate that bias is the fundamental hurdle that MSL must overcome. The logarithmic transformation of the likelihood function forces one to increase R with the sample size to obtain a consistent estimator. Given enough simulations to overcome bias, there are enough simulations to make the asymptotic contribution of simulation to the limiting distribution of $\hat{\theta}_{MSL}$ negligible.

There is a simulation design that uses the same total number $(N \cdot R)$ of simulations of ω as the independent design, but applies every simulation of ω to every observation of y. That is, the simulated log-likelihood function is generated according to the double sum $\tilde{\ell}(\theta) = \sum_{n=1}^{N} \sum_{m=1}^{NR} \ln \tilde{f}(\theta; y_n, \omega_m)$. The motivation for this approach is to take advantage of all $N \cdot R$ simulations that must be drawn when R independent simulations are made for each observation. Lee (1993) finds that efficiency requires only that $R \to \infty$ as $N \to \infty$ with this design. This approach appears to gain efficiency without any additional computational cost. However, one simulates each contribution to the likelihood $N \cdot R$ times rather than merely R times, substantially increasing the cost of evaluating the average simulated log-likelihood function. The computational savings gained by pooling simulations in this manner are generally overcome by the added computational cost of calculating $O(N^2)$ likelihoods instead of $O(N^{3/2})$, especially when N is large.

We close our discussion of simulated likelihood functions by noting that the method of simulated pseudo-maximum likelihood (SPML) of Laroque and Salanié (1989) is another early simulation estimation approach for LDV models. This method, originally developed for the mixture models of Subsection 2.4 in the case of the analysis of markets in disequilibrium, uses simulation to overcome the high dimensional integration difficulties that arise in calculating the moments of such models.

Definition 6 (Simulated Pseudo Maximum Likelihood) *Let the observation rule $\tau(y^*) = y$ yield a mixture model with the first two moments $g_1(x_n, \theta) \equiv E(y \mid x_n, \theta)$ and $g_2(x_n, \theta) \equiv E\left((y - Ey)^2 \mid x_n, \theta\right)$. Consider simulating functions $\tilde{g}_j(x_n, \theta, \omega, R)$, $j = 1, 2$, based on auxiliary simulation sequences $\{\omega\}$, such that $\tilde{g}_j(x_n, \theta, \omega, R)$ converge almost surely to $g_j(x_n, \theta)$ as $R \to \infty$, $j = 1, 2$. The simulated pseudo maximum likelihood estimator $\hat{\theta}_{SPML}$ is defined by:*

$$\hat{\theta}_{SPML} \equiv \arg\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \psi\left(y_n, \tilde{g}_1(\cdot), \tilde{g}_2(\cdot)\right), \tag{34}$$

where $\psi(\cdot) \equiv \frac{1}{2}\left[(y_n - g_1(\cdot))^2 / g_2^2(\cdot) + \ln g_2(\cdot)\right]$ corresponds to the log-likelihood contribution assuming $y_n \sim N(g_1(\cdot), g_2(\cdot))$.

Laroque and Salanié (1989) prove that for $x_n \in \mathbf{X} \in \mathbf{R}$, $\theta \in \Theta$ compact, and $\tilde{g}_j(\cdot)$ sufficiently continous on $\mathbf{X} \times \Theta$, then $\hat{\theta}_{SPML} \xrightarrow{p} \hat{\theta}_{PML}$ as $R \to \infty$.[21] It should be noted that for particular

[21] Pseudo maximum likelihood estimation methods, which are special types of the Classical Minimum Distance (CMD) approach, are developed in Gourieroux *et al.* (1984a) and Gourieroux *et al.* (1984b). See Newey and McFadden (1993) for a discussion of CMD and the closely related *generalized method of moments* (GMM).

choices of a pseudo likelihood function $\psi(\cdot)$, the SPML estimator can be shown to be consistent for a *finite* number of simulations R, because it then satisfies the basic linearity property of the MSM approach. Such a choice could be $\psi(\cdot) \equiv (y_n - g_1(\cdot))^2$, which corresponds to the assumption that $y_n \sim N(g_1(\cdot), 1)$.

4.3 Simulation of Moment Functions

The simulation of the log-likelihood is an appealing approach to applying simulation to estimation, but this approach must overcome the inherent simulation bias that forces one to increase R with the sample size. Instead of simulating the log-likelihood function, one can simulate moment functions. When they are linear in the simulations, moment functions can be simulated easily without bias. The direct consequence is that the simulation bias in the limiting distribution of an estimator is also zero, making the need to increase the number of simulations per observation with sample size unnecessary. This was a key insight of McFadden (1989) and Pakes and Pollard (1989).

Method of moments (MOM) estimators have a simple structure. Such estimators are generally constructed from 'residuals' that are differences between observed random variables y and their conditional expectations. These expectations are known functions of the conditioning variables x and the unknown parameter vector θ to be estimated, let $\mathrm{E}(y \mid x, \theta) \equiv \mu(\theta; x)$. Moment equations are built up by multiplying the residuals by various weights or instrumental variable functions and specifying the estimator as the parameter values which equate the sample average of these products with zero: The MOM estimator $\hat{\theta}_{MOM}$ is defined by

$$\frac{1}{N}\sum_{n=1}^{N} w_n(X, \hat{\theta}_{MOM}) \left[y_n - \mu(\hat{\theta}_{MOM};, x_n) \right] = 0. \tag{35}$$

The consistency of such estimators rests on the uniform convergence of the sample averages to their population counterparts for any value of θ as the sample size approaches infinity. When the unique root of the population equations is θ_0, the population value of θ, the root of the sample equations, converges to θ_0. The limiting distribution of $\hat{\theta}_{MOM}$ is derived from the linear expansion

$$0 = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} w_n(\theta_0) e_n(\theta_0) + \left[\frac{1}{N}\sum_{n=1}^{N} w_n(\bar{\theta}) \nabla_\theta e_n(\bar{\theta}) + e_n(\bar{\theta}) \nabla_\theta w_n(\bar{\theta}) \right] \sqrt{N}(\hat{\theta}_{MOM} - \theta_0),$$

where we have denoted the residual by $e_n(\theta) \equiv y_n - \mathrm{E}(y_n \mid x_n, \theta)$ and $\bar{\theta}$ lies between $\hat{\theta}_{MOM}$ and θ_0. Because $\mathrm{E}[e_n(\theta_0)] = 0$, the leading term will generally converge to a limiting normal random variable with zero expectation, implying no asymptotic bias in $\hat{\theta}_{MOM}$:

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} w_n(\theta_0) e_n(\theta_0) \xrightarrow{d} N(0, \Sigma_{MOM}),$$

where

$$\frac{1}{N}\sum_{n=1}^{N} w_n(\theta_0) V[e_n(\theta_0) \mid x_n] w_n(\theta_0)' \xrightarrow{p} \Sigma_{MOM}.$$

One of the matrices in the second term converges to zero:

$$\frac{1}{N}\sum_{n=1}^{N} e_n(\bar{\theta}) \nabla_\theta w_n(\bar{\theta}) \xrightarrow{p} 0.$$

This fact is often exploited by replacing the weights w in (35) with consistent estimates that do not change the limiting distribution of $\hat{\theta}_{MOM}$. Thus under regularity conditions,

$$\sqrt{N}(\hat{\theta}_{MOM} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H'^{-1}),$$

where

$$\frac{1}{N}\sum_{n=1}^{N} w_n(\bar{\theta})\nabla_\theta e_n(\bar{\theta}) \xrightarrow{p} H.$$

Simulation has an affinity with the MOM. Substituting an unbiased, finite-variance simulator for the conditional expectation $\mu(\theta; x_n)$ does not alter the essential convergence properties of these sample moment equations. We therefore consider the class of estimators generated by the *method of simulated moments* (MSM).

Definition 7 (Method of Simulated Moments) *Let $\tilde{\mu}(\theta; x, \omega) = 1/R\sum_{r=1}^{R}\tilde{\mu}(\theta; x, \omega_r)$ be an unbiased simulator so that $\mu(\theta; x) = E[\tilde{\mu}(\theta; x, \omega) \mid x]$ where ω is a simulated random variable. The method of simulated moments estimator is*

$$\hat{\theta}_{MSM} \equiv \arg\min \|\tilde{s}_N(\theta)\|$$

where

$$\tilde{s}_N(\theta) \equiv 1/N \sum_{n=1}^{N} w_n(\theta)\left[y_n - \tilde{\mu}(\theta; x_n, \omega_n)\right] \tag{36}$$

for some sequence $\{\omega_n\}$.

Defining the MSM estimator as a minimizer rather than the root of the simulated moments equation $\tilde{s}(\theta) = 0$ is an important part of making the MSM operational. Newey and McFadden (1993), Sections 1 and 2.2.3, discuss the general difficulties that MOM poses for the construction of consistent estimators. Whereas the structure of ML provides a direct link between parameter identification and estimator consistency, MOM does not. It is often difficult to guarantee that a system of nonlinear equations has a unique solution. MSM inherits these difficulties. Also, the addition of simulation in MSM may introduce problems that were not present in the original MOM formulation. For example, simulated moment equations may not exhibit solutions at all in small samples, leading one to question the reliability of asymptotic approximations. This property may be the greatest practical drawback of this method of estimation using simulations, although it does not greatly affect the asymptotic distribution theory extended from the MOM case.

Example 12 *To construct an MSM estimator for the rank ordered probit model, we construct a set of moment equations corresponding to the elements of y:*

$$\frac{\sum_{n=1}^{N} y_{ijn}}{N} - \Pr\left\{y_{ij} = 1; \hat{\mu}, \hat{\Omega}\right\} = 0,\ i, j = 1, \ldots, J-1.$$

Not all J^2 elements of y are needed because these elements have a singular distribution. As the sampling space of y makes clear, we can focus our attention on the first $J-1$ rows and columns of y.

Because we obtain more moment equations than parameters, we combine the moments of y according to the method of classical minimum distance *(CMD) using the inverse of the sample covariance of the elements of y as the normalizing matrix. Note, however, that one could use more*

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.6667*	*-0.6906*	*0.1481*	*-0.7792*	*-0.6918*	*-0.5948*
θ_2	*-0.6667*	*-0.7887*	*0.3714*	*-0.9496*	*-0.7109*	*-0.5431*
θ_3	*-0.6667*	*-0.6953*	*0.2223*	*-0.8347*	*-0.6594*	*-0.5366*
θ_4	*-0.5000*	*-0.6683*	*0.4271*	*-0.8962*	*-0.5688*	*-0.3679*
θ_5	*1.3230*	*1.4143*	*0.4384*	*1.118*	*1.337*	*1.633*
θ_6	*0.0000*	*0.1764*	*0.5053*	*-0.1957*	*0.08331*	*0.5563*
θ_7	*-0.3780*	*-0.3077*	*0.2765*	*-0.4703*	*-0.3207*	*-0.1747*
θ_8	*0.9258*	*0.7714*	*0.3356*	*0.5955*	*0.7980*	*0.9834*

Table 6: Sample Statistics for Rank Ordered Probit CMD (J=4)

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.6667*	*-0.6976*	*0.1905*	*-0.7915*	*-0.6809*	*-0.5798*
θ_2	*-0.6667*	*-0.9790*	*0.9576*	*-1.099*	*-0.7619*	*-0.5654*
θ_3	*-0.6667*	*-0.8561*	*0.6394*	*-1.008*	*-0.6900*	*-0.4813*
θ_4	*-0.5000*	*-0.7083*	*0.6392*	*-0.8918*	*-0.5559*	*-0.3327*
θ_5	*1.3230*	*1.4733*	*0.8402*	*1.086*	*1.323*	*1.662*
θ_6	*0.0000*	*0.0780*	*0.6268*	*-0.3091*	*0.03749*	*0.4616*
θ_7	*-0.3780*	*-0.3828*	*0.5423*	*-0.5758*	*-0.3099*	*-0.1110*
θ_8	*0.9258*	*0.8560*	*0.6857*	*0.5023*	*0.7341*	*0.9745*

Table 7: Sample Statistics for Rank Ordered Probit CMD, MSM Version, (J=4,R=1)

moments to increase the efficiency of the estimator. For example, the cross-products $y_{ij}y_{kl}$ $(i \neq k, j \neq l)$ contain additional sample information about the population parameters.

The CMD estimation results are described in Table 6 for $J = 4$ ranked alternatives. This classical estimator is much less efficient than the MLE. In addition, it exhibits large bias and skewness in the sampling distribution.

The summary statistics for the MSM version of the CMD estimator are listed in Table 7. There was $R = 1$ simulation of the GHK probability simulator for each observation and each probability. As expected, the sampling variance is larger for the MSM estimator than for the CMD estimator. In addition, the bias and skewness in the CMD estimator for the mean parameters seems to be aggravated by the simulation in the MSM estimator.

We do not present analogous results for $J = 6$ alternatives because the MSM estimator is not practical in this case. With 720 elements in the sampling space, the amount of simulation becomes prohibitive. This illustrates another important drawback in this method: the MSM works best for sample spaces with a small number of elements.

The analogies between MSM and MOM are direct and, as a result, the asymptotic analysis is generally simpler than for MSL. The first difference with MSL appears in the requirements on the simulation design for estimator consistency. Whereas MSL requires that $R \to \infty$ regardless of whether simulations are independent across observations, MSM yields consistent estimators with fixed R provided that the simulations vary enough to make a law of large numbers work. Because the simulated moments are linear in the simulations, one has the option of applying the law of large numbers to large numbers of observations alone, or in combination with large numbers of

simulations.

Proposition 6 *Let $\tilde{\mu}(\theta; x, \omega)$ be an unbiased, finite variance, simulator for $\mu(\theta; x)$ and let either*

1. *$\{\omega_{nr}; n = 1, \ldots, N, r = 1, \ldots, R\}$ be i.i.d. random variables for fixed R, or*
2. *$\{\omega_r; r = 1, \ldots, N\}$ be an i.i.d. sequence for $R = N$ and let $\omega_{nr} = \omega_r$, $n = 1, \ldots, N$.*

Then $\hat{\theta}_{MSM} \xrightarrow{p} \theta_0$ under the regularity conditions

1. *$s_N(\theta) \equiv 1/N \sum_{n=1}^{N} w_n(\theta) [y_n - \mu(\theta; x_n)]$ is continuous in θ,*
2. *$s_N(\theta) \to s_0(\theta) \equiv \text{plim} 1/N \sum_{n=1}^{N} w_n(\theta) [\mu(\theta_0; x_n) - \mu(\theta; x_n)]$ in probability uniformly in $\theta \in \Theta$, a compact parameter space,*
3. *$s_0(\theta)$ is continuous in θ and $s_0(\theta)$ equals zero only at θ_0.*

Proof. The average difference between the classical moment functions and their simulated counterparts is

$$s_N(\theta) - \tilde{s}_N(\theta) = \frac{1}{N} \sum_{n=1}^{N} w_n(\theta) [\tilde{\mu}(\theta; x_n, \omega_n) - \mu(\theta; x_n)] \tag{37}$$

$$= \frac{1}{NR} \sum_{n=1}^{N} \sum_{r=1}^{R} w_n(\theta) [\tilde{\mu}(\theta; x_n, \omega_{nr}) - \mu(\theta; x_n)], \tag{38}$$

where $s_N(\theta) \equiv 1/N \sum_{n=1}^{N} w_n(\theta) [y_n - \mu(\theta; x_n)]$. Under design (1), the $\{\tilde{\mu}_n - \mu_n\}$ are an i.n.i.d. sequence so that a uniform law of large numbers applied to (37) implies $s_N(\theta) - \tilde{s}_N(\theta) \xrightarrow{p} 0$ as $N \to \infty$. Under design (2), $s_N(\theta) - \tilde{s}_N(\theta)$ is written in (38) as a U-statistic and a uniform law of large numbers for U-statistics (Lee (1993)) implies $s_N(\theta) - \tilde{s}_N(\theta) \xrightarrow{p} 0$ as $N \to \infty$. Therefore, in either case, by continuity, $\|s_N(\theta) - \tilde{s}_N(\theta)\| \xrightarrow{p} 0$ uniformly in θ and Lemma (1) implies the result. □

The opportunity to fix R for all sample sizes offers significant computational savings that are a key motivation for interest in the MSM. As we shall see below, the benefits of the dependent design are generally modest. Thus, while the theoretical applicability of U-statistics to MSM is interesting in itself, we will not consider it further in this section.[22] We continue with the analogy between the MOM and the MSM. Note first of all that an analogous linear expansion for $\hat{\theta}_{MSM}$ exists:

$$0 = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} w_n(\theta_0) \tilde{e}_n(\theta_0) + \left[\frac{1}{N} \sum_{n=1}^{N} w_n(\bar{\theta}) \nabla_\theta \tilde{e}_n(\bar{\theta}) + \tilde{e}_n(\bar{\theta}) \nabla_\theta w_n(\bar{\theta}) \right] \sqrt{N} (\hat{\theta}_{MSM} - \theta_0),$$

where we have denoted the simulated residual by $\tilde{e}_n(\theta) \equiv y_n - \tilde{\mu}(\theta; x_n)$ and $\bar{\theta}$ lies between $\hat{\theta}_{MSM}$ and θ_0. Because $E[\tilde{e}_n(\theta_0)] = 0$, the leading term will generally converge to a limiting normal random variable with zero expectation, implying no asymptotic bias in $\hat{\theta}_{MSM}$:

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} w_n(\theta_0) e_n(\theta_0) \xrightarrow{d} N(0, \Sigma_{MSM}),$$

[22] See Lee (1993).

where

$$\frac{1}{N}\sum_{n=1}^{N} w_n(\theta_0)V[\tilde{e}_n(\theta_0) \mid x_n]w_n(\theta_0)' \xrightarrow{p} \Sigma_{MSM}.$$

Also, as before,

$$\frac{1}{N}\sum_{n=1}^{N} \tilde{e}_n(\bar{\theta})\nabla_\theta w_n(\bar{\theta}) \xrightarrow{p} 0,$$

so that under regularity conditions,

$$\sqrt{N}(\hat{\theta}_{MSM} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma_{MSM}H'^{-1}),$$

where

$$\frac{1}{N}\sum_{n=1}^{N} w_n(\bar{\theta})\nabla_\theta \tilde{e}_n(\bar{\theta}) \xrightarrow{p} H.$$

The equivalence of the H matrices also rests on the unbiased simulation of μ: If $\mu(\theta; x) = E[\tilde{\mu}(\theta; x, \omega) \mid x]$, then $\nabla_\theta \mu(\theta; x) = \nabla_\theta E[\tilde{\mu}(\theta; x, \omega) \mid x] = E[\nabla_\theta \tilde{\mu}(\theta; x, \omega) \mid x]$ for the smooth simulators described in Section 3.

While the first moment of the MSM estimator does not depend on R, the limiting covariance matrix, and hence relative efficiency, does. Simulation noise introduces a generic difference between the covariance matrices of $\hat{\theta}_{MOM}$ and $\hat{\theta}_{MSM}$. Intuition suggests, and theory confirms, that the larger R is, the more efficient the MSM estimator will be as the simulation noise is diminished. The extra variation in $\hat{\theta}_{MSM}$ is contained in the object (37). This term is generated conditional on the realizations of y and is, by definition, independently distributed of the classical moment function. Inflating the simulation noise by $\sqrt{N}$ and evaluating it at θ_0, we can apply a central limit theorem to it to obtain the following result.

Proposition 7 *$\Sigma_{MSM} = \Sigma_{MOM} + 1/R \cdot \Sigma_S$ where*

$$\frac{1}{N}\sum_{n=1}^{N} w_n(\theta_0) \cdot V\left[\tilde{\mu}(\theta_0; x_n, \omega_{nr}) - \mu(\theta_0; x_n)\right] \cdot w_n(\theta_0)' \xrightarrow{p} \Sigma_S.$$

If it were not for the simulation noise, the MSM estimator would be as efficient as its MOM counterpart. McFadden (1989) noted that in the special case where $\tilde{\mu}$ is obtained by averaging simulations of the data generating process itself, $\Sigma_S = \Sigma_{MOM}$ and $\Sigma_{MSM} = (1 + 1/R)\Sigma_{MOM}$. In this case, the inefficiency of simulation is easy to measure and one observes that 10 replications are sufficient to reduce the inefficiency to 10% compared to classical MOM.

The proposition suggests that full efficiency would be obtained if we simply increased R without bound as N grows. That intuition is formalized in the next proposition, which is analogous to Proposition 5 (see McFadden and Ruud (1992)).

Proposition 8 *If $R = O(N^\alpha)$, $\alpha > 0$, then $\sqrt{N}(\hat{\theta}_{MOM} - \hat{\theta}_{MSM}) \xrightarrow{d} 0$.*

For any given residual and instrumental variables, there generally exist optimal weights among MOM estimators and the same holds for MSM as well. In what is essentially an asymptotic counterpart to the Gauss-Markov theorem, if $H = \Sigma_{MSM}$ then the MSM estimator is optimal Hansen (1982). To construct an MSM estimator that satisfies this restriction, one normalizes the

simulated residual by its variance and makes the instrumental variables the partial derivatives of the conditional expectation of the simulated moment with respect to the unknown parameters:

$$w_n = [\nabla_\theta \mu(\theta; x_n)] \{\mathrm{V}[y_n - \tilde{\mu}(\theta; \omega_n, x_n)]\}^{-1}$$

One can approximate these functions using simulations that are independent of the moment simulations with R fixed, but efficiency will require increasing R with sample size. If $\tilde{\mu}$ is differentiable in θ, then independent simulations of the $\nabla_\theta \tilde{\mu}$ are unbiased simulators of the instruments. Otherwise, discrete numerical derivatives can be employed. The covariance matrix can be estimated using the sample variance of $\tilde{\mu}$ and the simulated variance of y. Inefficiency in simulated instruments constructed in this way has two sources: the simulation noise and the bias in the inverse of an estimated variance. Both sources disappear asymptotically if R approaches infinity with N. While it is critical that the simulations of w be independent of the simulations of $\tilde{\mu}$, there is no obvious advantage to simulating the individual components of w independently. In some cases, for example simulating a ratio, it appears that independent simulation may be inferior.[23]

4.4 Simulation of the Score Function

Interest in the efficiency of estimators naturally leads to attempts to construct an efficient MSM estimator. The obvious way to do this is to simulate the score function as a set of simulated moment equations. Within the LDV framework however, unbiased simulation of the score with a finite number of operations is not possible with simple censored simulators; the efficient weights are nonlinear functions of the objects that require simulation. Nevertheless, it may be possible with the aid of simulation to construct good approximations that offer improvements in efficiency over simpler MSM estimators.

There is an alternative approach based on truncated simulation. We showed in Section 2 that every score function can be expressed as the expectation of the score of a latent data generating process taken conditional on the observed data. In the particular case of normal LDV models, this conditional expectation is taken over a truncated multivariate normal distribution and the latent score is the score of an untruncated multivariate normal distribution. Simulations from the truncated normal distribution can replace the expectation operator to obtain unbiased simulators of the score function.

In order to include both the censored and truncated approaches to simulating the score function, we define the *method of simulated scores* as follows.[24]

Definition 8 (Method of Simulated Scores) *Let the log-likelihood function for the unknown parameter vector θ given the sample of observations $(y_n, n = 1, \ldots, N)$ be $\ell_N(\theta) \equiv \sum_{n=1}^{N} \ln f(\theta; y_n)$. Let $\tilde{\mu}(\theta; y_n, \omega_n) = 1/R \sum_{r=1}^{R} \tilde{\mu}(\theta; y_n, \omega_{nr})$ be an asymptotically (in R) unbiased simulator of the score function $\mu(\theta; y) = \nabla \ln f(\theta; y)$ where ω is a simulated random variable. The* method of simulated scores *estimator is $\hat{\theta}_{MSS} \equiv \arg\min_{\theta \in \Theta} \|\tilde{s}_N(\theta)\|$ where $\tilde{s}_N(\theta) \equiv 1/N \sum_{n=1}^{N} \tilde{\mu}(\theta; y_n, \omega_n)$ for some sequence $\{\omega_n\}$.*

Our definition includes all MSL estimators as MSS estimators, because they implicitly simulate the score with a bias that disappears asymptotically with the number of replications R. But there are also MSS estimators without simulation bias for fixed R. These estimators rely on simulation from the truncated conditional distribution of the latent y^* given y. We turn to such estimators first.

[23] A Taylor series expansion suggests that positive correlation between the numerator and denominator of a ratio can yield a smaller variance than independent simulation.

[24] The term was coined by Hajivassiliou and McFadden (1990).

4.4.1 Truncated Simulation of the Score

The truncated simulation methods described in Section 3.3 provide unbiased simulators of the LDV score (17), which is composed of elements of the form (24). Such simulation would be ideal, because R can be held fixed, thus leading to fast estimation procedures. The problem is that these truncated simulation methods pose new problems for the MSS estimators that use them.

The first truncated simulation scheme, discussed in subsection 3.3.1 above, is the A/R method. This provides simulations that are discontinuous in the parameters, a property shared with the CMC. A/R simulation delivers the first element in a simulated sequence that falls into a region which depends on the parameters under estimation. As a result, changes in the parameter values cause discrete changes in which element in the sequence is accepted. An example of this phenomenon is to suppose that one is drawing a sequence of normal random variables $\{\eta_r\} \sim N(0, I_M)$ in order to obtain truncated multivariate normal random variables for rank ordered probit estimation. Given the observation y, one seeks a simulation from $\mathbf{D}(y)$, as defined in Example 8. Let the simulation of y^* be $\tilde{y}_t(\mu_1, \Gamma_1) \equiv \mu_1 + \Gamma_1 \eta_t$ at the parameter values (μ_1, Γ_1). At neighboring parameter values where two elements of the vector $\tilde{y}_t(\mu, \Gamma)$ are equal, the A/R simulation is at the point of jumping from the value $\tilde{y}_t(\mu, \Gamma)$ to another point in the sequence $\{\tilde{y}_s(\mu, \Gamma)\}$. See Hajivassiliou and McFadden (1990) and McFadden and Ruud (1992) for treatments of the special asymptotic distribution theory for such simulation estimators. Briefly described, this distribution theory requires a degree of smoothness in the estimator with respect to the parameters that permits such discontinuties but allows familiar linear approximations in the limit. See Ruud (1991) for an illustrative application.

The second truncated simulation scheme we discussed above was the Gibbs resampling simulation method. See subsection 3.3.2. This method is continuous in the parameters provided that one uses a continuous univariate truncated normal simulation scheme. But this simulation method also has a drawback: Strictly applied, each simulation requires an infinite number of resampling rounds. In practice, Gibbs resampling is truncated and applied as an approximation. The limited Monte Carlo evidence that we have seen suggests that such approximation is reliable.

Simulation of the efficient score fits naturally with the EM algorithm for computing the MLE derived by Dempster *et al.* (1977). The EM algorithm includes a step in which one computes an expectation with respect to the truncated distribution of y^* conditional on y. Ruud (1991) suggested that a *simulated* EM (SEM) algorithm could be based on simulation of the required expectation.[25] This substitution provides a computational algorithm for solving the simulated score of MSS estimators.

Definition 9 (EM Algorithm) *The EM algorithm is an iterative process for computing the MLE of a censored data model. On the i^{th} iteration, the EM algorithm solves*

$$\theta^{i+1} = \arg\max_{\theta} Q(\theta, \theta^i; y) \tag{39}$$

where the function Q is

$$Q(\theta^1, \theta^0; y) \equiv \mathrm{E}_{\theta^0}\left[\ln f(\theta^1; y^*) \middle| y\right] \tag{40}$$

where $\mathrm{E}_{\theta^0}[\cdot \mid y]$ indicates an expectation measured with respect to $f(\theta^0; y^* | y)$.

If Q is continuous in both θ arguments, then (39) is a contraction mapping that converges to a root of the normal equations; as Ruud (1991) points out,

$$\theta = \theta^1 = \theta^0 \Rightarrow \nabla_{\theta^1} Q(\theta^1, \theta^0; y) = \nabla_\theta \ln F(\theta; y) \tag{41}$$

[25] van Praag *et al.* (1989) and van Praag *et al.* (1991) also investigated this approach and applied it in a study of the Dutch labor market.

so that the first-order conditions for an iteration of (39) and the normal equations for ML are intimately related.

Unlike the log-likelihood function, this Q can be simulated without bias for LDV models because the latent likelihood $f(\theta; y^*)$ is tractable and Q is linear in $\ln f(\theta; y^*)$ (see equation (40)). According to (41), unbiased simulation of Q implies a means for unbiased simulation of the score. Although it is not guaranteed, an unbiased simulator of Q usually yields a contraction mapping to a stationary point.

For LDV models based on a latent multivariate normal distribution, the iteration in (39) is quite simple to compute, given Q or a simulation of Q. If $f(\theta; y^*) = \phi(y^* - \mu; \Omega)$, then

$$\mu^1 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{\theta^0}(y_n^* \mid y_n) \quad \text{and} \quad \Omega^1 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}_{\theta^0}\left[(y_n^* - \mu^1)(y_n^* - \mu^1)' \mid y_n\right], \tag{42}$$

which are analogous to the equations for the MLE using the latent data. This algorithm is often quite slow, however, in a neighborhood of the stationary point of (39). Any normalizations necessary for identification of θ can be imposed at convergence. See Ruud (1991) for a discussion of these points.

Example 13 (SEM Estimation) *In this example, we will apply the SEM procedure to the rank ordered probit model of our previous examples. We simulated an (approximately) unbiased $\tilde{Q}$ of Q by drawing simulations of y_n^* from its truncated normal distribution conditional on y_n using the Gibbs resampling method truncated to 10 rounds. The support of this truncated distribution is specified as $\mathbf{D}(y)$ in Example 8. The simulated estimators were computed according to (42), after replacing the expectations with the averages of independent simulations.*

Parameter	Population Value	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
θ_1	*-0.4000*	*-0.3827*	*0.1558*	*-0.4907*	*-0.3848*	*-0.2757*
θ_2	*-0.4000*	*-0.4570*	*0.3271*	*-0.5992*	*-0.4089*	*-0.2455*
θ_3	*-0.4000*	*-0.4237*	*0.2262*	*-0.5351*	*-0.3756*	*-0.2766*
θ_4	*-0.4000*	*-0.4268*	*0.2710*	*-0.5319*	*-0.3891*	*-0.2580*
θ_5	*-0.4000*	*-0.4300*	*0.2622*	*-0.5535*	*-0.3794*	*-0.2521*

Table 8: Sample Statistics for Rank Ordered Probit SEM Using Gibbs Simulation (J=6,R=5)

The usual Monte Carlo results for 500 experiments with $J = 6$ ranked alternatives are reported in Table 8 for data sets containing 100 observations and $R = 5$ simulations per observation. These statistics are comparable to those in Table 5 for the MSL estimator of the same model with the same number of simulation replications. The biases for the true parameter values appear to be appreciably smaller in the SEM estimator, while the sampling variances are larger. We cannot judge either estimator as an approximation to the MLE, because the latter is prohibitively difficult to compute.

Although truncated simulation is generally more costly, the SEM estimator remains a promising general approach to combining simulation with relatively efficient estimation. It is the only method that combines unbiased simulation of the score with optimization of an objective function and the latter property appears to offer substantial computational advantages.

4.4.2 Censored Simulation of Ratios

The censored simulation methods in Section 3.2 can also be applied to approximating the efficient score. These simulation methods tend to be much faster computationally than the truncated simulation methods, but censored simulations introduce simulation bias in much the same way as in the MSL. Censored simulation can be applied to discrete LDV models by noting that the score function of an LDV model with observation rule $y \equiv \tau(y^*)$ can generally be written in the ratio form:

$$\begin{aligned} \frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} &= \frac{\int_{\{y^* | \tau(y^*)=y\}} \nabla_\theta dF(\theta; y^*)}{\int_{\{y^* | \tau(y^*)=y\}} dF(\theta; y^*)} \\ &= \frac{E(\nabla_\theta dF(\theta; y^*) \mid \tau(y^*) = y)}{\Pr\{y^* \mid \tau(y^*) = y\}}, \end{aligned}$$

where $F(\theta; y^*|y)$ is the conditional c.d.f. of y given $\tau(y^*) = y$. See subsection 2.6 for more details. van Praag and Hop (1987), McFadden (1989), and Hajivassiliou and McFadden (1990) note that this form of the score function offers the potential of estimation by simulation.[26] A MSS estimator can be constructed by simulating separately the numerator and denominator of the score expressions:

$$\tilde{s}_N(\theta) = \frac{1}{N} \sum_{n=1}^{N} \frac{\tilde{d}(\theta; y_n, \omega_{1n})}{\tilde{p}(\theta; y_n, \omega_{2n})}, \tag{43}$$

where $\tilde{d}(\theta; y_n, \omega_{1n}) = 1/R_1 \sum_{r=1}^{R_1} \tilde{d}(\theta; y_n, \omega_{1nr})$ is an unbiased simulator of the derivative function $\nabla_\theta f(\theta, y)$ and $\tilde{p}(\theta; y_n, \omega_{2n}) = 1/R_2 \sum_{r=1}^{R_2} \tilde{p}(\theta; y_n, \omega_{2nr})$ is an unbiased function of the probability expression $f(\theta; y_n)$. Hajivassiliou and McFadden (1990) prove that when the approximation of the scores in ratio form is carried out using the GHK simulator, the resulting MSS estimator is consistent and asymptotically normal when $N \to \infty$ and $R_2/\sqrt{N} \to \infty$. The number of simulations for the numerator expression, R_1, affects the efficiency of the resulting MSS estimator. Because the unbiased simulator $\tilde{p}(\theta; y, \omega_2)$ of $f(\theta; y)$ does not yield an unbiased simulator of the reciprocal $1/f(\theta; y)$ in the simulator $1/\tilde{p}(\theta; y, \omega_2)$, R_2 must increase with sample size to obtain a consistent estimator. This is analogous to simulation in MSL. In fact, this simulation scheme is equivalent to MSL when $\omega_1 = \omega_2$ and $\tilde{d} = \nabla_\theta \tilde{p}$.

McFadden and Ruud (1992) note that MSM techniques can also be used generally to remove the simulation bias in such MSS estimators. In discrete LDV models, where y has a sampling space $\mathbf{B}$ that is countable and finite, we can always write y as a vector of dummy variables for each of the possible outcomes so that

$$\mathrm{E}_\theta(y_i) = \Pr\{y_i = 1; \theta\} = f(\theta; Y) \quad \text{if} \quad Y_i = 1,\ Y_j = 0,\ j \neq i.$$

Thus,

$$\mathrm{E}_\theta \left[\frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} \right] = 0 = \sum_{Y \in \mathbf{B}} f(\theta; Y) \cdot \frac{\nabla_\theta f(\theta; Y)}{f(\theta; Y)}$$

and the score can be written

$$\begin{aligned} \frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} &= \sum_{Y \in \mathbf{B}} \mathbf{1}\{y = Y\} \frac{\nabla_\theta f(\theta; Y)}{f(\theta; Y)} \\ &= \sum_{Y \in \mathbf{B}} [\mathbf{1}\{y = Y\} - f(\theta; Y)] \frac{\nabla_\theta f(\theta; Y)}{f(\theta; Y)}. \end{aligned} \tag{44}$$

[26] See Hajivassiliou (1993c) for a survey of the development of simulation estimation methods for LDV models.

Provided that the "residual" $\mathbf{1}\{y = Y\} - f(\theta; Y)$ and the "instrumental variables" $\nabla_\theta f(\theta; Y)/f(\theta; Y)$ are simulated independently, equation (44) provides a moment function for the MSM. In this form, the instrumental variables ratio can be simulated with bias as in (43) because the residual term is independently distributed and possesses a marginal expectation equal to zero at the population parameter value. For example, we can alter (43) to

$$\tilde{s}_N(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{Y\in\mathbf{B}}\left[\mathbf{1}\{y_n = Y\} - \tilde{p}(\theta; Y, \omega_{1n})\right]\frac{\tilde{d}(\theta; Y, \omega_{2n})}{\tilde{p}(\theta; Y, \omega_{2n})}, \tag{45}$$

where ω_1 and ω_2 are independent pseudo-random variables. While such bias does not introduce inconsistency into the MSM estimator, the simulation bias does introduce inefficiency because the moment function is not an unbiased simulator of the score function. This general approach underlies the estimation method for multinomial probit originally proposed by McFadden (1989).

4.4.3 MSM versus MSS

MSM and MSS are natural competitors in estimation with simulation because each has a comparative advantage. MSM uses censored simulations that are cheap to compute, but it cannot simulate the score without bias within a finite number of calculations. MSS uses truncated simulations that are expensive to compute (and introduce jumps in the objective function with A/R simulations), but simulates the score (virtually) without bias. McFadden and Ruud (1992) make a general comparison of the asymptotic covariance matrices that suggests when one method is preferable to another.

Consider the special MSS case in which the simulations $\tilde{Y}^*(\theta; Y, \omega)$ are drawn from the latent conditional distribution and the exact latent score $\nabla_\theta \ell^*$ is available so that

$$\tilde{s}_{MSS}(\theta) = R_1^{-1}\sum_{r=1}^{R_1}\nabla_\theta \ell^*[\theta; \tilde{Y}^*(\theta; Y, \omega)].$$

Then Σ_S, the contribution of simulation to the covariance matrix of the estimator, has a useful interpretation:

$$\Sigma_S = (\Sigma_* - \Sigma_0)/R$$

where $\Sigma_* = \mathrm{E}_\omega\{\nabla_\theta \ell^*(\theta; Y^*)[\nabla_\theta \ell^*(\theta; Y^*)]'\}$ is the information matrix of the latent log-likelihood. The simulation noise is proportional to the information loss due to partial observability.

In the simplest applications of censored simulation to the MSM, the simulations are independent of sample outcomes and their contribution to the moment function is additively separable from the contribution of the data: Thus we can write $\tilde{s}_{MSM}(\theta) = g(\theta; Y, \omega_2) - \tilde{g}(\theta; \omega_1, \omega_2)$ (see (45)). In that case, Σ_S simplifies to $\mathrm{V}\left\{\sqrt{N}[\tilde{g}(\theta_0; \omega_1, \omega_2)]\right\}$. In general, the simulation process makes R independent replications of the simulations $\{\omega_r; r = 1, \ldots, R\}$, so that

$$\tilde{g}(\theta; \omega_1, \omega_2) = R^{-1}\sum_{r=1}^{R}\tilde{g}(\theta; \omega_{1r}, \omega_2)$$

and $\Sigma_S = R^{-1}\mathrm{V}_{\theta_0}[\tilde{g}(\theta_0; \omega_1, \omega_2)]$. In an important special case of censored simulation, the simulation process makes R independent replications of the modeled data generating process, $\{\tilde{Y}(\theta; \omega_{1r}); r = 1, \ldots, R\}$, so that

$$\tilde{g}(\theta; \omega_1, \omega_2) = R_2^{-1}\sum_{r=1}^{R_2} g[\theta; \tilde{Y}(\theta; \omega_{1r}), \omega_2] \tag{46}$$

and $\Sigma_S = R^{-1}\mathrm{V}[g(\theta_0; Y, \omega')] = \Sigma_M/R$. Then the MSM covariance matrix equals $1 + 1/R$ times the classical MOM covariance matrix without simulation $G^{-1}\Sigma_M(G')^{-1}$. Now let us specialize to simulation of the score. For simplicity, suppose that the simulated moment functions are unbiased simulations of the score: $\mathrm{E}[\tilde{s}_{MSM}(\theta) \mid Y] = \nabla_\theta \ell(\theta; Y)$. Of course in most cases, the MSM estimator will have a simulation bias for the score. The asymptotic variance of the MSM estimator is

$$\begin{aligned} \Sigma_M &= \lim_{N\to\infty} \mathrm{V}[\tilde{s}_{MSM}(\theta_0) - \nabla_\theta \ell(\theta_0; Y)] \\ &= \lim_{N\to\infty} \mathrm{V}[\tilde{s}_{MSM}(\theta_0)] + \Sigma_0 \\ &= \Sigma_W + \Sigma_0 \end{aligned}$$

where $\Sigma_S = \Sigma_M/R$ and Σ_W holds additional variation attributable to the simulation of the score. If the MSS and MSM estimators use the same number of simulation replications, we can make a simple comparison of the relative efficiency of the two methods. The difference between the asymptotic covariance matrices is

$$R^{-1}\Sigma_0^{-1}[\Sigma_0 + (R+1)\Sigma_W - (\Sigma_* - \Sigma_0)]\Sigma_0^{-1}.$$

This expression gives guidance about the conditions under which censored simulation is likely to dominate truncated. It is already obvious that if Σ_W high, so that censored simulation is inefficient due to a poor approximation of the score, then truncated simulation is likely to dominate. On the other hand, if Σ_0 is low, because partial observability causes a large loss in information, then estimation with censored simulation is likely to dominate truncated.

Thus, we might expect that the censored simulation method will dominate the truncated one for the multinomial probit model, particularly if $\Sigma_W = 0$. That, however, is a special case in which a more efficient truncated simulation estimator can be constructed from the censored simulation estimator. Because $\mathrm{E}[\tilde{s}(\theta) \mid Y] = \nabla_\theta \ell(\theta; Y)$,

$$\mathrm{E}[g(\theta; Y, \omega_2) - \tilde{g}(\theta; \omega_1, \omega_2)] = \nabla_\theta \ell(\theta; Y)] \quad \Leftrightarrow \quad \mathrm{E}[\tilde{g}(\theta; \omega_1, \omega_2)] = \mathrm{E}\left\{g[\theta; \tilde{Y}(\theta; \omega), \omega']\right\} = 0 \ \ \forall\theta.$$

The bias correction is obviously unnecessary and only increases the variance of the MSM estimator. But an MSM estimator based on $g(\theta; Y, \omega)$ is a truncated simulation MSM estimator; only simulation for the particular Y observed is required. We conclude that the censored method can outperform the truncated method only by choosing $\mathrm{E}_\omega[e(\theta)] \neq \nabla_\theta \ell(\theta; Y)$ in such a way that the loss in efficiency in Σ_M is offset by low Σ_0 and low Σ_W.[27]

4.5 Bias Corrections

In this section, we interpret estimation with simulation as a general method for removing bias from approximate parametric moment functions, following McFadden and Ruud (1992). The approximation of the efficient score is the leading problem in estimation with simulation. In a comparison of the MSM and MSS approximations, we have just described a simple trade-off. On the one hand, the simulated term in the residual of (45) that replaces the expectation in (44) is clearly redundant when the instrumental variables are $\nabla_\theta f(\theta; Y)/f(\theta; Y)$: The expectation of the simulated terms multiplied by the instruments is identically zero for *all* parameter values so that the simulation merely adds noise to the score and the resulting estimator. On the other hand, the simulated residual is clearly necessary when the instruments are not ideal. Without the simulation, the moment equation is invalid and the resultant estimators are inconsistent.

[27] The actual difference in asymptotic covariance matrices is more complicated than the formula above however, because $G \neq \Sigma_M \neq \Sigma_0$.

This trade-off motivates a general structure of simulated moments estimators. We can interpret the extra simulation term as a bias correction to an approximation of the score. For example, one can view the substitution of non-ideal weights into the original score function as an approximation to the score, chosen for its computational feasibility. Because the approximation introduces bias, the bias is removed by simulating the (generally) unknown expectation of the approximate score. Suppose the moment restrictions have a general form

$$\mathrm{E}[s(\theta_0; y, X) \mid X] = 0.$$

When the moment function s is computationally burdensome, an approximation $g(\theta; y, X, \omega)$ becomes a feasible alternative. The additional argument ω represents an ancillary statistic containing the "coefficients" of the approximation. In general, such approximation will introduce inefficiency and bias into MOM estimators constructed from g. Simulation of g over the distribution of y produces an approximate bias correction $\tilde{g}(\theta; X, \omega', \omega)$, where ω' represents the simulated component. Thus, we consider estimators $\hat{\theta}$ that satisfy

$$g(\hat{\theta}; y, X, \omega) - \tilde{g}(\hat{\theta}; X, \omega', \omega) = 0 \tag{47}$$

MSM estimators have this general form; and feasible MSS estimators generally do, too.

4.5.1 A Score Test for Estimator Bias

The appeal of simulation estimators without bias correction is substantial. Although, the simulation of moments or scores overcomes a substantial computational difficulty in the estimation of LDV models, there may remain practical difficulties in solving the simulated moment functions for the estimators. Whereas maximum likelihood possesses a powerful relationship between the normal equations and the likelihood function, moment equations generally do not satisfy such 'integrability' conditions. As a result, there is not even a guarantee that a root of the estimating equations exists. Bias correction can introduce a significant amount of simulation noise to estimators. For these reasons, the approximation of the log-likelihood function itself through simulation still offers an important opportunity to construct feasible and relatively efficient estimators.

MSS, and particularly MSL, estimators can be used without bias correction if the bias is negligible relative to the sampling error of the estimator and the magnitude of the true parameter. A simple score test for significant bias can be developed and implemented easily.

Conditional on the MSS estimator, the expectation of the simulated bias in the approximate score should be zero. The conditional distribution of the elements of the bias correction are i.n.i.d. random variables to which a central limit theorem can be applied. In addition, the White-Eicker estimator of the covariance matrix of the bias elements is consistent so that the usual quadratic statistic, measuring the statistical significance of the bias term, can be computed. As an alternative to testing the significance of this statistic, the bias correction term can be used to compute a local approximate confidence region for the biases in the moment function or the estimated parameters. This has the advantage of providing a way to assess whether the biases are important for the purposes of inference.

5 Conclusion

In this chapter, we have described the use of simulation methods to overcome the difficulties in computing the likelihood and moment functions of LDV models. These functions contain multivariate

integrals that cannot be easily approximated by series expansions. However, unbiased simulators of these integrals can be computed easily.

We began by reviewing the ways in which LDV models arise, describing the differences and similarities in censored and truncated data generating processes. Censoring and truncation give rise to the troublesome multivariate integrals. Following the LDV models, we described various simulation methods for evaluating such integrals. Naturally, censoring and truncation play roles in simulation as well. Finally, estimation methods that rely on simulation were described in the final section of this chapter. We organized these methods into three broad groups: MSL, MSM, and MSS. These are not mutually exclusive groups. But each group has a different motivation: MSL focuses on the log-likelihood function, the MSM on moment functions, and the MSS on the score function. The MSS is a combination of ideas from MSL and MSM, treating the efficient score of the log-likelihood function as a moment function.

Software for implementing these methods is not yet widely available. But as such tools spread, and as improvements in the simulators themselves are developed, simulation methods will surely become a familiar tool in the applied econometrician's workshop.

6 Acknowledgements

We would like to thank John Geweke and Dan McFadden for very helpful comments. John Wald provided expert research assistance. We are grateful to the National Science Foundation for partial financial support, under grants SES-929411913 (Hajivassiliou) and SES-9122283 (Ruud).

References

Amemiya, T. 1984. Tobit Models: A Survey. *Journal of Econometrics*, **24**, 3–61.

Avery, R., Hansen, L., and Hotz, V. 1983. Multiperiod Probit Models and Orthogonality Condition Estimation. *International Economic Review*, **24**, 21–35.

Bauwens, L. 1984. *Bayesian Full Information Analysis of Simultaneous Equation Models using Integration by Monte Carlo.* Berlin: Springer-Verlag.

Beggs, S., Cardell, S., and Hausman, J. 1981. Assessing the potential demand for electric cars. *Journal of Econometrics*, **17**, 1–20.

Berkovec, J., and Stern, S. 1991. Job Exit Behavior of Older Men. *Econometrica*, **59**, 189–210.

Bloemen, H., and Kapteyn, A. 1991. *The Joint Estimation of a Non-linear Labour Supply Function and a Wage Equation Using Simulated Response Probabilities.* mimeo, Tilburg University.

Bock, R.D., and Jones, L.V. 1968. *The measurement and prediction of judgement and choice.* San Francisco: Holden-Day.

Bolduc, D. 1991. Generalized Autoregressive Errors in the Multinomial Probit Model. *Transportation Research B — Methodological.* forthcoming.

Bolduc, D., and Kaci, M. 1991. *Multinomial Probit Models with Factor-Based Autoregressive Errors: A Computationally Efficient Estimation Approach.* mimeo, Université Laval.

Börsch-Supan, A., and Hajivassiliou, V. 1993. Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models. *Journal of Econometrics*, **58(3)**, 347–368.

Börsch-Supan, A., Hajivassiliou, V., Kotlikoff, L., and Morris, J. 1992. Health, Children, and Elderly Living Arrangements: A Multi-Period Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors. *Pages 79–108 of:* Wise, D. (ed), *Topics in the Economics of Aging*. Chicago: University of Chicago Press.

Chib, S. 1993. Bayes Regression with Autoregressive Errors: A Gibbs Sampling Approach. *Journal of Econometrics*, **58(3)**, 275–294.

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.

Daganzo, C., Bouthelier, F., and Sheffi, Y. 1977. Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm. *Transportation Science*, **11**, 338–358.

Davis, P., and Rabinowitz, P. 1984. *Methods of Numerical Integration*. New York: Academic Press.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

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

Dubin, J., and McFadden, D. 1984. An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. *Econometrica*, **52(2)**, 345–362.

Duffie, D., and Singleton, K. 1993. Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica*, **61(4)**, 929–952.

Dutt, J. 1973. A Representation of Multivariate Normal Probability Integrals by Integral Transforms. *Biometrika*, **60**, 637–645.

Dutt, J. 1976. Numerical Aspects of Multivariate Normal Probabilities in Econometric Models. *Annals of Economic and Social Measurement*, **5**, 547–562.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*. New York: Wiley.

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

Geman, S., and Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. 1989. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, **57**.

Geweke, J. 1992. Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints. *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium*, 571–578.

Goldfeld, S., and Quandt, R. 1975. Estimation in a Disequilibrium Model and the Value of Information. *Journal of Econometrics*, **3(3)**, 325–348.

Gourieroux, C., and Monfort, A. 1990. *Simulation Based Inference in Models with Heterogeneity.* mimeo, INSEE.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. 1984b. Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica*, **52**, 701–720.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. 1984a. Pseudo Maximum Likelihood Methods: Theory. *Econometrica*, **52**, 681–700.

Gronau, R. 1974. The Effect of Children on the Housewife's Value of Time. *Journal of Political Economy*, **81**, 168–199.

Hajivassiliou, V. 1986. *Serial Correlation in Limited Dependent Variable Models: Theoretical and Monte Carlo Results.* Cowles Foundation Discussion Paper No.803.

Hajivassiliou, V. 1990. *The Method of Simulated Scores: A Presentation and Comparative Evaluation.* mimeo, Cowles Foundation for Research in Economics, Yale University.

Hajivassiliou, V. 1992. *The Method of Simulated Scores: A Presentation and Comparative Evaluation.* Cowles Foundation Discussion Paper, Yale University.

Hajivassiliou, V. 1993a. *Estimation by Simulation of the External Debt Repayment Problems.* Cowles Foundation Discussion Paper, Yale University.

Hajivassiliou, V. 1993b. Simulating Normal Rectangle Probabilities and Their Derivatives: The effects of Vectorization. *International Journal of Supercomputer Applications*, **?**, ?

Hajivassiliou, V. 1993c. Simulation Estimation Methods for Limited Dependent Variable Models. *Pages 519–543 of:* Maddala, G.S., Rao, C.R., and Vinod, H.D. (eds), *Handbook of Statistics (Econometrics)*, vol. 11. Amsterdam: North-Holland.

Hajivassiliou, V., and McFadden, D. 1990. *The Method of Simulated Scores, with Application to Models of External Debt Crises.* Cowles Foundation Discussion Paper No. 967.

Hajivassiliou, V., , and Ioannides, Y. 1991. *Switching Regressions Models of the Euler Equation: Consumption Labor Supply, and Liquidity Constraints.* mimeo, Cowles Foundation for Research in Economics, Yale University.

Hajivassiliou, V., McFadden, D., and Ruud, P. 1992. *Simulation of Multivariate Normal Orthant Probabilities: Methods and Programs.* mimeo, Cowles Foundation for Research in Economics, Yale University.

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

Hanemann, M. 1984. Discrete/Continuous Models of Consumer Demand. *Econometrica*, **52(3)**, 541–562.

Hansen, L.P. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**, 1029–1054.

Hausman, J., and Wise, D. 1978. A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*, **46**, 403–426.

Hausman, J., and Wise, D. 1979. Attrition Bias in Experimental and Panel Data: The Gary Negative Income Maintenance Experiment. *Econometrica*, **47(2)**, 445–473.

Heckman, J. 1974. Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, **42**, 679–694.

Heckman, J. 1979. Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153–161.

Heckman, J. 1981. Dynamic Discrete Models. *Pages 179–195 of:* Manski, C., and McFadden, D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.

Hendry, D. 1984. Monte Carlo Experimentation in Econometrics. *Pages 937–976 of:* Griliches, Z., and Intriligator, M. (eds), *Handbook of Econometrics*, vol. 2. Amsterdam: North Holland.

Horowitz, J., Sparmonn, J., and Daganzo, C. 1981. An Investigation of the Accuracy of the Clark Approximation for the Multinomial Probit Model. *Transportation Science*, **16**, 382–401.

Hotz, V.J., and Miller, R. 1989. *Conditional Choice Probabilities and the Estimation of Dynamic Programming Models*. GSIA Working Paper 88-89-10.

Hotz, V.J., and Sanders, S. 1991. *The Estimation of Dynamic Discrete Choice Models by the Method of Simulated Moments*. NORC, University of Chicago.

Hotz, V.J., Miller, R., Sanders, S., and Smith, J. 1991. *A Simulation Estimator for Dynamic Discrete Choice Models*. mimeo, NORC, University of Chicago.

Keane, M. 1990. *A Computationally Efficient Practical Simulation Estimator for Panel Data, with Applications to Estimating Temporal Dependence in Employment and Wages*. mimeo, University of Minnesota.

Keane, M. 1993. Simulation Estimation Methods for Panel Data Limited Dependent Variable Models. *Page ? of:* Maddala, G.S., Rao, C.R., and Vinod, H.D. (eds), *Handbook of Statistics (Econometrics)*, vol. 11. Amsterdam: North-Holland.

Kloek, T., and van Dijk, H. 1978. Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo. *Econometrica*, **46**, 1–20.

Laroque, G., and Salanié, B. 1989. Estimation of Multi-Market Disequilibrium Fix-Price Models: An Application of Pseudo Maximum Likelihood Methods. *Econometrica*, 831–860.

Laroque, G., and Salanié, B. 1990. *The Properties of Simulated Pseudo-Maximum Likelihood Methods: The Case of the Canonical Disequilibrium Model*. Working Paper No. 9005, CREST-Departement de la Recherche, INSEE.

Lee, B.-S., and Ingram, B. 1991. Simulation Estimation of Time-Series Models. *Journal of Econometrics*, **47**, 197–205.

Lee, L.-F. 1978. Unionism and Wage Rates: A Simultaneous Equation Model with Qualitative and Limited Dependent Variables. *International Economic Review*, **19**, 415–433.

Lee, L.-F. 1979. Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables. *Econometrica*, **47**, 977–996.

Lee, L.-F. 1993. On the Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models. *Econometric Theory*. forthcoming.

Lerman, S., and Manski, C. 1981. On the Use of Simulated Frequencies to Approximate Choice Probabilities. *Pages 305–319 of:* Manski, C., and McFadden, D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.

Lewis, H.G. 1974. Comments on Selectivity Biases in Wage Comparisons. *Journal of Political Economy*, **82(6)**, 1145–1155.

Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

McCulloch, R., and Rossi, P.E. 1993. *An Exact Likelihood Analysis of the Multinomial Probit Model*. Working Paper 91-102, Graduate School of Business, University of Chicago.

McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. *Pages 105–142 of:* Zarembka, P. (ed), *Frontiers in Econometrics*. New York: Academic Press.

McFadden, D. 1981. Econometric Models of Probabilistic Choice. *Pages 198–272 of:* Manski, C., and McFadden, D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.

McFadden, D. 1986. Econometric Analysis of Qualitative Response Models. *Pages 1395–1457 of:* Griliches, Z., and Intriligator, M. (eds), *Handbook of Econometrics*, vol. 2. Amsterdam: North Holland.

McFadden, D. 1989. A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica*, **57**, 995–1026.

McFadden, D., and Ruud, P. 1992. *Estimation by Simulation*. University of California at Berkeley Working Paper.

Moran, P. 1984. The Monte Carlo Evaluation of Orthant Probabilities for Multivariate Normal Distributions. *Australian Journal of Statistics*, **26**, 39–44.

Mühleisen, M. 1991. *On the Use of Simulated Estimators for Panel Models with Limited-Dependent Variables*. mimeo, University of Munich.

Newey, W.K., and McFadden, D.L. 1993. Estimation in Large Samples. *Page ? of:* Engle, Rob, and McFadden, Daniel (eds), *Handbook of Econometrics*, vol. Vol. 4. North Holland.

Owen, D. 1956. Tables for Computing Bivariate Normal Probabilities. *Annals of Mathematical Statistics*, **27**, 1075–1090.

Pakes, A. 1992. *Estimation of Dynamic Structural Models: Problems and Prospects Part II: Mixed Continuous-Discrete Controls and Market Interactions*. mimeo, Yale University.

Pakes, A., and Pollard, D. 1989. Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, **57**, 1027–1057.

Poirier, D., and Ruud, P.A. 1988. Probit with Dependent Observations. *Review of Economic Studies*, **55**, 593–614.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. 1986. *Numerical Recipes.* Cambridge: Cambridge University Press.

Quandt, R. 1972. A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, **67**, 306–310.

Quandt, R. 1986. Computational Problems in Econometrics. *Pages 1395–1457 of:* Griliches, Z., and Intriligator, M. (eds), *Handbook of Econometrics*, vol. 1. Amsterdam: North Holland.

Rubinstein, R. 1981. *Simulation and the Monte Carlo Method.* New York: Wiley.

Rust, J. 1992. *Estimation of Dynamic Structural Models: Problems and Prospects Part II: Discrete Decision Processes.* SSRI Working Paper #9106, University of Wisconsin at Madison.

Ruud, P. 1986. *On the Method of Simulated Moments for the Estimation of Limited Dependent Variable Models.* mimeo, University of California at Berkeley.

Ruud, P. 1990. *A Note on Computing Multinomial Probit Estimators by Simulation.* Department of Economics, University of California at Berkeley.

Ruud, P. 1991. Extensions of Estimation Methods Using the EM Algorithm. *Journal of Econometrics*, **49**, 305–341.

Stern, S. 1992. A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models. *Econometrica*, **60**, 943–952.

Stroud, A. 1971. *Approximate Calculation of Multiple Integrals.* New York: Prentice Hall.

Thisted, R. 1988. *Elements of Statistical Computing.* New-York: Chapman-Hall.

Thurstone, L. 1927. A Law of Comparative Judgement. *Psychological Review*, **34**, 273–286.

Tierny, L. 1992. *Markov Chains for Exploring Posterior Distributions.* University of Minnesota Working Paper.

Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26**, 24–36.

van Dijk, H.K. 1987. Some Advances in Bayesian Estimation Methods Using Monte Carlo Integration. *Pages 205–261 of:* Fomby, T.B., and Rhodes, G.F. (eds), *Advances in Econometrics*, vol. 6. Greenwich, Connecticut: JAI Press.

van Praag, B.M.S., and Hop, J.P. 1987. *Estimation of Continuous Models on the Basis of Set-Valued Observations.* Erasmus University Working Paper, presented at the ESEM Copenhagen.

van Praag, B.M.S., Hop, J.P., and Eggink, E. 1989. *A Symmetric Approach to the Labor Market by Means of the Simulated Moments Method with an Application to Married Females.* Erasmus University Working Paper, presented at the EEA Augsburg.

van Praag, B.M.S., Hop, J.P., and Eggink, E. 1991. *A Symmetric Approach to the Labor Market by Means of the Simulated EM-Algorithm with an Application to Married Females.* Erasmus University Working Paper, presented at the ESEM Cambridge.

West, M. 1990. *Bayesian Computations: Monte-Carlo Density Estimation.* Duke University, Discussion Paper #90-A10.