Chapter 40

CLASSICAL ESTIMATION METHODS FOR LDV MODELS USING SIMULATION

VASSILIS A. HAJIVASSILIOU
Yale University

PAUL A. RUUD
University of California

Contents

1. Introduction 2384
2. Limited dependent variable models 2386
   2.1. The latent normal regression model 2386
   2.2. Censoring 2386
   2.3. Truncation 2392
   2.4. Mixtures 2394
   2.5. Time series models 2395
   2.6. Score functions 2397
   2.7. The computational intractability of LDV models 2399
3. Simulation methods 2400
   3.1. Overview 2400
   3.2. Censored simulation 2402
   3.3. Truncated simulation 2406
4. Simulation and estimation of LDV models 2408
   4.1. Overview 2408
   4.2. Simulation of the log-likelihood function 2412
   4.3. Simulation of moment functions 2421
   4.4. Simulation of the score function 2428
   4.5. Bias corrections 2435
5. Conclusion 2437
6. Acknowledgements 2438
References 2438

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden
© 1994 Elsevier Science B.V. All rights reserved
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.

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.

One of the most familiar LDV models is the binomial probit model, which specifies that the probability that a binomial random variable $y$ equals 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, $\Phi$ 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, \( y_{ni}^* \):

\[
y_{ni} = \begin{cases} 
0 & \text{if } y_{ni}^* < 0 \\
1 & \text{if } y_{ni}^* > 0 
\end{cases}
\quad \text{where } \ y_{ni}^* \sim N(\chi_m \beta, 1).
\]

Then, one can assign the latent \( y_{ni}^* \) a nonscalar covariance matrix appropriate to continuously distributed panel data. For example, stacking the \( y_{ni}^* \) 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

\[
V(y^*) = I_N \otimes \Omega \quad \text{where} \quad \Omega = \begin{pmatrix} 
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho & 1 & \rho & \ldots & \rho^{T-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \ldots & 1 & \rho \\
\end{pmatrix} + \sigma_s^2 J_T,
\]

(1)

where \( J_T \) is a \( T \times T \) matrix of ones, \( 0 \leq \sigma_s^2 < 1 \), and \( |\rho| < 1 \).

Now consider the impact of such nonscalar covariance matrices on the likelihood for the observed \( y_{ni} \). Although the marginal probabilities that \( y_{ni} \) is zero or one are unchanged, the likelihood function consists of the joint probabilities that the dependent series \( \{ y_{n1}, y_{n2}, \ldots, y_{nT} \} \) 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 \( \theta \) is unknown, finite-dimensional, and \( \theta \in \Theta \), where \( \Theta \) is a compact subset of \( \mathbb{R}^k \). Estimation of \( \theta \) by maximum likelihood (ML) involves the maximization of the log-likelihood function \( l_n(\theta) = \sum_{i=1}^N \ln f(\theta; y_i) \) over \( \Theta \). Often, finding the root of a system of normal equations \( \nabla_{\theta} l_n(\theta) = 0 \) 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 \( \theta \) by classical methods is effectively infeasible.

In this section, we review the various forms of likelihood functions that arise in LDV models. In the second subsection, we discuss models generated as partially observed or censored latent dependent variables. The third 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, Section 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 cases, \( y^* \) 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 \( A \). Let \( B \) be the support of the random variable \( Y = \tau(Y^*) \) where \( \tau : A \to B \) is not invertible. Then \( Y \) is a censored random variable.

In LDV models, \( \tau \) is often called the “observation rule”; and though it may not be monotonic, \( \tau \) 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 \( A \subseteq R^m \) and \( B \subseteq R^l \).

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

\[
F(\theta; Y) = \int_{y^*|\tau(y^*) \leq y} dF(\theta; y^*).
\] (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 \( \Omega \) is a positive definite matrix, and

\[
\phi(y^* - \mu, \Omega) = \left[ \det(2\pi\Omega) \right]^{-1/2} \exp \left[ -\frac{1}{2}(y^* - \mu)'\Omega^{-1}(y^* - \mu) \right].
\] (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 \( \beta \) is a vector of \( K \) slope coefficients. The covariance matrix is usually a function of a vector of \( K \) variance parameters \( \sigma \).

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:

\[
f(\theta; Y) = \frac{\partial^n F(\theta; Y)}{\partial Y_1 \cdots \partial Y_n}.
\] (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.

\(^2\) Such densities are formally known as Radon–Nikodym p.d.f.'s with respect to Lebesgue measure.
of \( Y \) so that \( F \) is not 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

\[
f(\theta; Y) = \sum_{Y_1}^J \cdots Y_J \frac{\partial^J}{\partial Y_1 \cdots \partial Y_J} \left[ F(\theta; Y) - F(\theta; Y - 0) \right]
= f(\theta; Y_1, \ldots, Y_J) \Pr \{ Y_j = Y^d_j, j > J | \theta; Y_1, \ldots, Y_J \},
\]

where the discrete jump \( F(\theta; Y) - F(\theta; Y - 0) \) reflects the nontrivial probability of the event \( \{ Y_j = Y^d_j, j > J \} \).

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 are 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 \( \mu \) is a \( J \times 1 \) vector of means and \( \Omega \) 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 \( y^* : z(y^*) = \{1 \{ y^*_j = \max_i y^*_i \} : j = 1, \ldots, J \} \). Therefore, the sampling space \( B \) of \( y \) is the set of orthonormal elementary unit vectors, whose elements are all zero except for a unique element that equals one:

\[
B = \{(1, 0, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)\}.
\]

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, \}

\[
F(\theta; Y) - F(\theta; Y - 0) \equiv \lim_{\epsilon \to 0} \{ F(\theta; Y) - F(\theta; Y - \epsilon) \}.
\]

\footnote{The height of the discontinuity is denoted by}
J}. By creating the first-difference vector $z_j = \{y_i^* - y_i^*, i = 1, \ldots, J, i \neq j\} = \Delta y_i^*$ and denoting its mean and covariance by $\mu_j = \Delta \mu$ and $\Omega_j = \Delta \Omega \Delta_j^\prime$ respectively, $F(\theta; y)$ and $f(\theta; y)$ are both functions of multivariate normal negative orthant integrals of the general form

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

We obtain

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

and

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

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

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

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 $\Omega$ analyzed in Heckman (1981), Bolduc (1991) and Bolduc and Kaci (1991), and the diagonal $\Omega$ 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

---

\(^4\)Tobin (1958).
structure like probit: \( y^* \sim N(\mu, \sigma^2) \). The observation rule is also similar: \( \tau(y^*) = 1\{y^* \geq 0\} \cdot y^* \) which leads to the sample space \( B = \{y \in R| y \geq 0\} \) and c.d.f.

\[
F(\theta; Y) = \begin{cases} 
0 & \text{if } Y < 0, \\
\int_{y^* < Y} \phi(y^* - \mu, \sigma)dy^* = \Phi(Y - \mu, \sigma^2) & \text{if } Y \geq 0.
\end{cases}
\]

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

\[
f(\theta; Y) = \begin{cases} 
0 & \text{if } Y < 0, \\
\Phi(-\mu, \sigma^2) & \text{if } Y = 0, \\
\phi(Y - \mu, \sigma^2) & \text{if } Y > 0.
\end{cases} \tag{8}
\]

The discrete jump in \( F \) at \( Y = 0 \) corresponds to the nonzero 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.\(^3\) In the simplest version, the latent \( y^* \) consists of two elements drawn from a bivariate normal distribution where

\[
\Omega(\sigma) = \begin{bmatrix}
1 & \sigma_{12} \\
\sigma_{12} & \sigma_2^2
\end{bmatrix}
\]

The observation rule is

\[
\tau(y^*) = \begin{pmatrix} \tau_1(y^*) \\ \tau_2(y^*) \end{pmatrix} = \begin{pmatrix} 1\{y^*_1 \geq 0\} \\ 1\{y^*_1 \geq 0\} \cdot y^*_2 \end{pmatrix}
\]

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: \( B = \{(0, 0) \cup \{(1, y_2), y_2 \in R\} \). Thus, two cases capture the nonzero regions of the

c.d.f. of \( y \). First of all, the c.d.f. is constant on \( B_0 = [0, 1) \times (0, \infty) \):

\[
F(\theta; Y) = \Phi(-\mu_1, 1) + \int_{\{y_1^* < 0\}} \phi(y_1^* - \mu, \Omega) \, dy_1^*, \quad Y \in B_0
\]

because \( y_1^* \) 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 \( B_1 = [1, \infty) \times \mathbb{R} \) is increased according to

\[
F(\theta; Y) = \begin{cases} 
1 \{Y_2 \geq 0\} \Phi(-\mu_1, 1) + \int_{y_1^* \geq 0, y_2 \neq y_1^*} \phi(y_1^* - \mu, \Omega) \, dy_1^*, & Y \in B_1, \\
1 \{Y_2 \geq 0\} \Phi(-\mu_1, 1) + \Phi\left(\frac{\mu_1}{Y_2 - \mu_2}, 1 - \frac{\sigma_{12}^2}{\sigma_2^2}\right), & Y \in B_1.
\end{cases}
\]

The p.d.f. will therefore be

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

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

\[
\tau(y^*) = \begin{pmatrix} \tau_1(y^*) \\ \tau_2(y^*) \end{pmatrix} = \begin{pmatrix} 1 \{y_1^* \geq 0\} \\ (\prod_{j=1}^{J} 1 \{y_1^j \geq 0\}) \cdot y_2^* \end{pmatrix}.
\]

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

\[
B = \left\{ y \in \mathbb{R}^M \mid y_M = 0, \prod_{j=1}^{M-1} y_j = 1, y_j \in \{0, 1\}, \ j < M \right\}
\]

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

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 \( D \) be a proper subset of the support of \( F \) and \( D^c \) its complement such that \( \Pr \{ y^* \in D^c \} > 0 \). The function

\[
G(Y) = \begin{cases} 
F(Y)/\Pr \{ Y \in D \} & \text{if } Y \in D, \\
0 & \text{if } Y \in D^c.
\end{cases}
\]

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 \( D \). 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 \( D \). The term "truncation" derives from the visual effect dropping the set \( D^c \) has on the original distribution when \( D^c \) 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

\[
y = \begin{cases} 
\tau(y^*) & \text{if } y^* \in D, \\
\text{unobserved} & \text{otherwise},
\end{cases}
\]  

(9)

where \( D \) 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:

\[
f(\theta; Y) = \frac{\int_{\{y^* \in D \mid \tau(y^*) = Y\}} dF(\theta; y^*)}{\int_{D} dF(\theta; y^*)}.
\]  

(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 \( D = \{ \ y \in R \ | \ y > 0 \ \} \) makes \( B = D \) so that the c.d.f. and p.d.f. of \( y \) are

\[
F(\theta; Y) = \begin{cases} 
0 & \text{if } Y \leq 0, \\
\int_{0}^{Y} \frac{\phi(y^* - \mu, \sigma) \ dy^*}{\int_{0}^{\infty} \phi(y^* - \mu, \sigma) \ dy^*} = \frac{\Phi(Y - \mu, \sigma^2) - \Phi(-\mu, \sigma^2)}{1 - \Phi(-\mu, \sigma^2)} & \text{if } Y > 0,
\end{cases}
\]

\[
f(\theta; Y) = \begin{cases} 
0 & \text{if } Y \leq 0, \\
\phi(Y - \mu, \sigma) & \text{if } Y > 0.
\end{cases}
\]

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:

\[
\ln f(\theta; y) = \ln f_1(\theta; y) - \ln f_2(\theta; y)
\]

(11)

where \( f_2 \) represents the normalizing probability \( \Pr \{ y^* \in D \} = \int_D dF(\theta; y^*) \). In models with only censoring, \( f_2 = 1 \). 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)).

\*Note that scale information about \( y^* \) is available in the censored and truncated normal regression models which is not in the case of binary response, so that \( \sigma^2 \) 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 \( I \{ Y > 0 \} \) is available.
\[ 1 \{ Y > 0 \} \ln \left( \frac{\phi(Y - \mu, \sigma)}{1 - \Phi(-\mu, \sigma^2)} \right) = 1 \{ Y > 0 \} \ln [\phi(Y - \mu, \sigma)] + 1 \{ Y = 0 \} \Phi(-\mu, \sigma^2) \]

- \left[ 1 \{ Y > 0 \} \ln [1 - \Phi(-\mu, \sigma^2)] + 1 \{ Y = 0 \} \Phi(-\mu, \sigma^2) \right].

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 \( \theta \) 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), Larroque and Salanié (1989)). Larroque 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

\[ y = 1 \{ y_1^* \geq 0 \} \cdot y_2^* + 1 \{ y_1^* < 0 \} \cdot y_3^* \]
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,

\[
F(\theta; Y) = \left\{ \begin{array}{ll}
\int_{\Omega^*_1 \cap \Omega^*_2 \cap \Omega^*_3 \subseteq \Omega} \phi(y^*- \mu, \Omega) \, dy^*, \\
\int_{\Omega^*_1 \cap \Omega^*_2 \cap \Omega^*_3 \subseteq \Omega} \phi(y^*- \mu, \Omega) \, dy^* + \int_{\Omega^*_1 \cap \Omega^*_2 \cap \Omega^*_3 \subseteq \Omega} \phi(y^*- \mu, \Omega) \, dy^*,
\end{array} \right.
\]

and

\[
f(\theta; Y) \equiv \frac{dF(\theta; Y)}{dY} = \phi(Y - \mu_2, \sigma_2^2) \int_{\Omega^*_1 \cap \Omega^*_2 \subseteq \Omega} \phi(y^*_1 - \mu_{112}, \Omega_{112}) \, dy^*_1 \\
+ \phi(Y - \mu_3, \sigma_3^2) \int_{\Omega^*_1 \cap \Omega^*_3 \subseteq \Omega} \phi(y^*_1 - \mu_{113}, \Omega_{113}) \, dy^*_1 \\
= \phi(Y - \mu_2, \sigma_2^2) \Phi(\mu_{112}, \Omega_{112}) + \phi(Y - \mu_3, \sigma_3^2) \Phi(- \mu_{113}, \Omega_{113})
\]

where, for \( j \in \{2, 3\}, \)

\[
\mu_{1,j} \equiv E(y^*_j | y^*_1 = Y) = \mu_j + \sigma_j (Y - \mu)/\sigma_j^2,
\]

\[
\Omega_{1,j} \equiv V(y^*_j | y^*_1 = Y) = 1 - \sigma_j^2/\sigma_j^2
\]

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 data sets 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 \( y_{nt}^* = \mu_{nt} + \varepsilon_{nt} \) measures the net benefit to the agent characterizing an action in period \( t \). Typically, \( \mu_{nt} \) is a linear index
function of a \( k \times 1 \) vector of exogenous explanatory variables \( x_{n} \), i.e., \( \mu_{n} = x_{n}'\beta \). The agent chooses one of two actions in each period, denoted by \( y_{n} \in \{0, 1\} \), depending upon the value of \( y_{n}^{*} \):

\[
\tau(y^{*}) = \begin{cases} 
    y_{n} = 1 & \text{if } y_{n}^{*} > 0, \\
    y_{n} = 0 & \text{if } y_{n}^{*} \leq 0,
\end{cases} \quad t = 1, \ldots, T. \tag{12}
\]

Hence, the sample space for \( \tau(y^{*}) \) is \( B = \times_{t=1}^{T} \{0, 1\} \), 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 \( y_{n}^{*} = (y_{n1}^{*}, \ldots, y_{nT}^{*})' \) be the multivariate normal given in equation (3). Then, for individual \( n \) the LDV vector \( \{y_{n}\}, t = 1, \ldots, T, \) has the discrete p.d.f.

\[
f(\beta, \Omega; y) = \Phi(S_{n}, \Sigma S), \quad \text{where } \mu = x'\beta \quad \text{and } S = \text{diag}(2y - 1).
\]

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

By way of illustration, let us consider the specific covariance structure, found very useful in applied work: \(^7\)

\[
\varepsilon_{nt} = \eta_{nt} + \varepsilon_{nt}, \quad \varepsilon_{nt} = \rho \varepsilon_{n,t-1} + \nu_{nt}, \quad |\rho| < 1 \tag{13}
\]

and \( \nu \) and \( \eta \) independent. This implies that

\[
\Omega = \sigma_{\varepsilon}^2 \begin{pmatrix} 
1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\
\rho & 1 & \rho & \ldots & \rho^{T-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1
\end{pmatrix} + \sigma_{\eta}^2 J,T.
\]

The variance parameters \( \sigma_{\varepsilon}^2 \) and \( \sigma_{\eta}^2 \) cannot both be identified, so the normalization \( \sigma_{\varepsilon}^2 + \sigma_{\eta}^2 = 1 \) is used. \(^8\)

The probability of the observed sequence of choices of individual \( n \) is

\[
\Pr \{y_{n}^{*}, \theta, x_{n}\} = \int_{a_{n}y_{n}} \Phi(y_{n}^{*} - \mu_{n}, \Omega_{a}) dy_{a}^{*},
\]

\(^1\) See Hajivassiliou and McFadden (1990); Börsch-Supan et al. (1992) and Hajivassiliou (1993a).

\(^2\) This is the structure assumed in the introductory example – see equation (1) above.
with \( \theta \equiv (\beta, \sigma^2, \rho) \) and

\[
\begin{align*}
\alpha_{nt} &= \begin{cases} 
0 & \text{if } y_{nt} = 1, \\
-\infty & \text{if } y_{nt} = 0,
\end{cases} \\
\beta_{nt} &= \begin{cases} 
+\infty & \text{if } y_{nt} = 1, \\
0 & \text{if } y_{nt} = 0.
\end{cases}
\end{align*}
\]

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 Rruud (1988), Hajivassiliou (1986) and Avery et al. (1983)). 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 \( \theta \) can be written in two ways which we will use to motivate two approaches to approximation of the score by simulation.

\[
\nabla_\theta \ln f(\theta; y) = \frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} = E[\nabla_\theta \ln f(\theta; y^*)|y]
\]

(14)

where \( \nabla_\theta \) is an operator that represents partial differentiation with respect to the elements of \( \theta \). The ratio (14) is simply the derivative of the log-likelihood and

\*Panel data sets in which each agent is observed for the same number of time periods \( T \) are called balanced, while sets with \( T_i \neq T \) for some \( n = 1, \ldots, N \) are known as unbalanced. As long as the determination of \( T_i \) 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 \( T_i \) is determined endogenously through an economic decision, which leads to a multi-period sample selection problem. See Hausman and Wise (1979) for a discussion of this case.
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 $V_\theta \ln f(\theta; y^*)$ 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 $F(\theta; y^* | y)$ denote the conditional c.d.f. of $y^*$ given that $\tau(y^*) = y$. We let

$$E(t(y^*)|y) = \int t(y^*) dF(\theta; y^* | y)$$

(16)

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

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

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

$$= E[\nabla_v \ln f(\theta; y^*) | \tau(y^*) = y]$$

since $(f(\theta; y^*)/f(\theta; y)) \int_{y^* | \tau(y^*) = y} f(\theta; y^*) dy^*$ 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):

$$\frac{\nabla_v \ln f(\theta; y)}{f(\theta; y)} = \Omega^{-1} [E(y^* | y) - \mu],$$

$$\nabla_v \ln f(\theta; y) = \frac{1}{2} \Omega^{-1} \{V(y^* | y) + [E(y^* | \tau(y^*) = y) - \mu]$$

$$\times [E(y^* | \tau(y^*) = y) - \mu]' - \Omega \} \Omega^{-1},$$

(17)

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

$$\nabla_v \ln \phi(y^* - \mu, \Omega) = \Omega^{-1}(y^* - \mu),$$

$$\nabla_\Omega \ln \phi(y^* - \mu, \Omega) = \frac{1}{2} \Omega^{-1} [(y^* - \mu)(y^* - \mu)' - \Omega] \Omega^{-1}.$$
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

$$
\begin{align*}
  z = \begin{cases} 
    y^* & \text{if } \tau(y^*) = y, \\
    \text{unobserved} & \text{otherwise.}
  \end{cases}
\end{align*}
$$

The functional form of these moments depends on the specification of the LDV function $\tau$.

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 $D$. That is, the score depends on only the first two moments of a truncated multivariate normal random variable $z'$ generated by the truncation rule

$$
\begin{align*}
  z' = \begin{cases} 
    y^* & \text{if } \tau(y^*) = y, \quad y^* \in D, \\
    \text{unobserved} & \text{otherwise.}
  \end{cases}
\end{align*}
$$

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:

$$
\nabla_\theta \ln f(\theta; y) = \nabla_\theta \ln f_1(\theta; y) - \nabla_\theta \ln f_2(\theta; y) = E[\nabla_\theta \ln f(\theta; y^*)|\tau(y^*) = y] - E[\nabla_\theta \ln f(\theta; y^*)|y^* \in D]
$$

and (17) becomes

$$
\nabla_\theta \ln F(\theta; y) = \Omega^{-1} \{E[y^*|\tau(y^*) = y] - E[y^*|y^* \in D]\},
$$

$$
\nabla_\theta \ln F(\theta; y) = \frac{1}{2} \Omega^{-1} \{E[(y^* - \mu)(y^* - \mu)^T|\tau(y^*) = y] - E[(y^* - \mu)(y^* - \mu)^T|y^* \in D]\} \Omega^{-1}.
$$

2.7. The computational intractability of LDV models

The likelihood contribution $f(\theta; y, \eta)$ and the score $\nabla_\theta \ln f(\theta; y, \eta)$ are functions of at most $M$-dimensional integrals over the region $D(y) \equiv \{y|\tau(y^*) = y\}$ in the domain of the $M \times 1$ latent vector $y^*_\eta$. 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 $\Omega$ of the unobserved random utilities has $(15 \times 16/2) - 1 = 119$ 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.\textsuperscript{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.\textsuperscript{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

\textsuperscript{11} See Quandt (1986) for a discussion of issues in numerical optimization methods.

\textsuperscript{12} Clark (1981) 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 $\Omega$. 
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

\[ y^* = \mu + \Gamma \eta, \]  

(20)

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

\[ \{y^* \mid \pi(y^*) = y\} = \{y^* \mid b_0 \leq Ay^* \leq b_1\}, \]  

(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 \( D = \{z \in \mathbb{R}^M \mid b_0 \leq z \leq b_1\} \).

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

\[ \Pr \{D, \mu, \Omega\} = \int 1\{y^* \in D\} \phi(y^* - \mu, \Omega) dy^*. \]  

(22)

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

\[ \nabla_\mu \Pr \{D, \mu, \Omega\} = \Omega^{-1} \int 1\{y^* \in D\} (y^* - \mu) \phi(y^* - \mu, \Omega) dy^*, \]

\[ \nabla_\Omega \Pr \{D, \mu, \Omega\} = \frac{1}{2} \Omega^{-1} \left\{ \int 1\{y^* \in D\} [ (y^* - \mu)(y^* - \mu)' - \Omega ] \phi(y^* - \mu, \Omega) dy^* \right\} \Omega^{-1}. \]  

(23)
Normalized by \( \Pr \{ \mathbf{D}; \mu, \Omega \} \), these equations transform to

\[
\mathbf{V}_n \ln \Pr \{ \mathbf{D}; \mu, \Omega \} = \Omega^{-1} [E(y^*|y^* \in \mathbf{D})] - \mu,
\]

\[
\mathbf{V}_{ij} \ln \Pr \{ \mathbf{D}; \mu, \Omega \} = \frac{1}{2} \Omega^{-1} \{E[(y^* - \mu)(y^* - \mu)^T] | y^* \in \mathbf{D} \} - \Omega \Omega^{-1}, \tag{24}
\]

which are terms in (17).

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

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

\[
\mu_{-i} \equiv E(y^*_{-i}),
\]

\[
\Omega_{-i, -i} \equiv V(y^*_{-i}),
\]

\[
\Omega_{-i, -i} \equiv \text{Cov}(y^*_{-i}, y^*_{-i})
\]

and the conditional moments

\[
\mu_{-ij}(y^*_{-i}) \equiv E(y^*_{-i} | y^*_j) \quad \text{and} \quad \Omega_{-i, -ij} \equiv V(y^*_{-i} | y^*_j).
\]

These conditional moments have the well-known formulas

\[
\mu_{-ij}(y^*_{-i}) = \mu_{-i} + \Omega_{-i, -i} \Omega_{-i, -j}^{-1} (y^*_{-i} - \mu_{-i}),
\]

\[
\Omega_{-i, -ij} = \Omega_{-i, -i} - \Omega_{-i, -j} \Omega_{-j, -j}^{-1} \Omega_{-j, -i}.
\]

The conditional mean and variance of \( y^*_{-i} \) given \( y^*_{-i} \), denoted \( \mu_{-i} \) and \( \Omega_{-i, -i} \), 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

\[
\mu_{i|<i}(y^*_i) = \mu_i + \Omega_{i, <i} \Omega_{<i, <i}^{-1} (y^*_i - \mu_{<i}).
\]

3.2. Censored simulation

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

\[
E[h(y^*, \mathbf{D})] \equiv E \left[ \mathbf{1} | y^* \in \mathbf{D} \right] \left[ \frac{1}{\text{vec}(y^*)} \right] = \left[ \frac{1}{\text{vec}(y^*)} \right]. \tag{25}
\]
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 \( \text{Pr}(\{D; \mu, \Omega\}) \) and its derivatives is to make repeated Monte Carlo draws for \( \eta \), use (20) to calculate \( y^* \) for each \( \eta \), and then form an empirical analogue of (25), by working only with the realizations that fall in set \( D \). Let \( \{\eta_1, \ldots, \eta_R\} \) be \( R \) simulated draws from the \( N(0, I_d) \) distribution and \( \tilde{y}_r = \mu + \Gamma \eta_r \), \( r = 1, \ldots, R \) so that

\[
\tilde{h} = \frac{1}{R} \sum_{r=1}^R h(\tilde{y}_r, D)
\]

is an unbiased simulation of (25). As \( R \) gets larger, the sampling variance of \( \tilde{h} \), \( P(1 - P)/R \), approaches zero and \( \tilde{h} \) converges strongly to \( E[h(y^*, D)] \). The simulation of \( \text{Pr}(\{D; \mu, \Omega\}) \) is simply the observed frequency with which the simulations of \( y^* \) fall into \( D \). Its derivatives with respect to \( \mu \) and \( \Omega \) are functions of the average simulation of \( 1 \{y^* \in D\} y^* \) and \( 1 \{y^* \in D\} y^* y^* \). We will call this the \textit{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”. 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 \( D \). For example, consider parameter values \((\mu_{0r}, \Gamma_{0r})\) chosen so that the \( m \)th element of the \( r \)th simulation equals its lower bound in \( D \):

\[
\tilde{y}_{rm} = \mu_{0m} + \Gamma_{0r} \eta_r = b_{0m}
\]

where \( \Gamma_{0m} \) is the \( m \)th row of \( \Gamma_0 \). Decreasing the parameter \( \mu_m \) from \( \mu_{0m} \) will cause the indicator \( 1 \{\tilde{y}_r \in D\} \) to jump from 1 to 0, and this will result in discrete jumps in the elements of \( h(\tilde{y}_r, D) \) and \( \tilde{h} \). Such discontinuities make computation of estimators and asymptotic distribution theory awkward.\(^{14}\) Second, the number of computations required by the CMC rises inversely with \( \text{Pr}(\{D; \mu, \Omega\}) \), 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.

\(^{13}\) Such a mechanism allows simultaneous operation on adjacent elements of a vector with multiple processors. See Hajivassiliou (1993b) which 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.
techniques, as, for example, the use of control and antithetic variates. See Hendry (1984) for definitions.

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 \( h(y^*, D) \). A simple generalization of this procedure rewrites \( E[h] \) in terms of another sampling distribution \( \gamma \):

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

\( \delta \) is a vector of parameters characterizing the design of the importance sampler \( \gamma(\cdot) \). Note that for \( h(\cdot) = 1_{\{y \in D\}} \), this expression corresponds to \( \Pr\{D; \mu, \Omega\} \). By drawing a random variable \( \tilde{y} \) from the importance p.d.f. \( \gamma \) 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 \( \Pr\{D; \mu, \Omega\} \). The first advantage offered by importance sampling is the ability to substitute sampling from \( \gamma \) for sampling from \( \phi \). In some cases, \( \gamma \) may be sampled more quickly, or, in a more general setting, sampling from \( \phi \) may be impractical.

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

\[
\Pr\{D; \mu, \Omega\} = \Pr\{\tilde{y} \in C\} \int_C 1_{\{\tilde{y} \in D\}} w(\tilde{y}) \gamma(\tilde{y}; \mu, \Omega, \delta) d\tilde{y}.
\]

By drawing from the truncated p.d.f. \( \gamma(\tilde{y}; \mu, \Omega, \delta) \Pr\{\tilde{y} \in C\} \), fewer simulations are "wasted" on outcomes of zero and, in effect, \( \Pr\{\tilde{y} \in C\} w(\tilde{y}) \) approximates \( \Pr\{D; \mu, \Omega\} \).

When \( \gamma \) is a good approximation to \( \phi \), 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

\[
V(P_C 1_{\{\tilde{y} \in D\}} w(\tilde{y})) = P_C^2 P_D [V(w(\tilde{y}))|\tilde{y} \in D] + (1 - P_D) E[w(\tilde{y})|\tilde{y} \in D]^2,
\]

where \( P_C \equiv \Pr\{\tilde{y} \in C\} \) and \( P_D \equiv \Pr\{\tilde{y} \in D\} \). In the extreme case that \( \gamma = \phi \),
\( V(\phi(\hat{y}|\hat{y} \in D)) = 0 \) and \( E(\phi(\hat{y}|\hat{y} \in D)^2) = P_D \). Therefore, good approximations to \( \phi \) 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.

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

(26)

for \( \hat{y} \in D \) where \( \sigma_{m|<m} = \sqrt{\Omega_{mm|<m}} \) and

\[
c_{im} = b_{im} - \mu_{m|<m}(\hat{y}_{<m}), \quad i = 0, 1.
\]

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

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

(27)

where the \( \omega \) are independently distributed uniform random variables.\(^{16}\) The GHK simulator is the product

\[
h_{GHK}(\hat{y}) = \prod_{m=1}^M \{ \Phi(c_{1m}, \sigma^2_{m|<m}) - \Phi(c_{0m}, \sigma^2_{m|<m}) \} \cdot \begin{pmatrix} 1 \\ \vec{y} \\ \vec{\gamma}(\hat{y}) \end{pmatrix}
\]

(28)

It is an unbiased simulator of \( E(h) \).

The GHK simulator was developed by Geweke (1992), Hajivassiliou and McFadden (1990) and Keane (1990). Experience suggests that the sampling variance of \( h_{GHK}(\hat{y}) \) is very small so that it approximates \( E(h) \) well in practice. This approximation has the properties of being continuous in \( \mu, \Omega, \) and \( \omega \), which, as already mentioned, is a desirable property of simulators for asymptotic theory and for iterative parameter search. The method of constructing \( \gamma \) in this example can also be extended to a bivariate version using a bivariate normal c.d.f. and standardizing adjacent pairs of elements.

\(^{15}\) Other investigations of the use of Monte Carlo integration in Bayesian analysis are, inter alia, Bauwens (1984), Kook and van Dijk (1978), van Dijk (1987) and West (1990).

\(^{16}\) 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 \( \mu, \Omega, \) and \( \omega \), which, as already mentioned, is a desirable property of simulators for asymptotic theory and for iterative parameter search. The method of constructing \( \gamma \) in this example can also be extended to a bivariate version using a bivariate normal c.d.f. and standardizing adjacent pairs of elements.
disjoint rectangular regions surrounding and including $D$, and being a continuous function of $\omega, \mu, \Omega, h_0,$ and $h_1$. 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 $\Pr\{D; \mu, \Omega\}$.

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. 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.

**Proposition 1**

Suppose $\phi(y^*)$ is a $J$-dimensional density, and one wishes to sample from the conditional density $\phi(y^*|D) = \phi(y^*)\int_D \phi(y^*)dy^*$. Suppose $\gamma(y)$ is a density with a support $A$ from which it is practical to sample, with the property that

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

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

\footnote{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.
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 \( 1/\Pr\{D; \mu, \Omega\} \) if \( \int \rho(\bar{y})d\bar{y} = \Gamma(D) \) is practical to compute. The conditional probability of acceptance, given \( \{\bar{y} \in D\} \), is \( \{\rho(\bar{y})d\bar{y}/\pi = \Pr\{D\}/\pi \), so that the marginal probability of acceptance is \( \Gamma(D) \Pr\{D\}/\pi \). The distribution of the number of trials to get an acceptance is the geometric distribution and its expectation is \( \pi[\Gamma(D) \Pr\{D\}] \). Therefore, if \( t \) is the number of draws made until \( \bar{y} \) is accepted, \( t \Gamma(D)/\pi \) is an unbiased simulator of \( 1/\Pr\{D\} \).

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

\[
\pi = \prod_{m=1}^{M} \left\{ \Phi(b_{1m} - \mu_m|_{<m}(b_{1} < m, \sigma_m^2|_{<m}) - \Phi(b_{0m} - \mu_m|_{<m}(b_{0} < m, \sigma_m^2|_{<m}) \right\},
\]

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

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 approximating 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 \( \mu \) and \( \Omega \). 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 \( D \), which is assumed to be finite. Define a recursive procedure with steps \( j = 1, \ldots, J \) in rounds \( g = 1, \ldots, G \). Let \( \{y^{(j,g)}\} \) be a sequence on \( D \) such that on the \( j \)th step of the \( g \)th

\(^{18}\)This simulator can be generalized in principle to non-normal distributions, provided the corresponding univariate distributions are easy to sample.
round, the $j$th element of $y^{*(j)}_j$ is computed from $y^{*(j),(a)}_j$ by

\[
y^{*(j)}_j = \mu_{j} - \frac{1}{2}(y^{*(j),(a-1)}_j + \sigma_{j} - \frac{1}{2}) \cdot \Phi^{-1}\left[\varphi_{j} \cdot \sigma_{j} - \frac{1}{2}\right] + \sigma_{j} - \frac{1}{2}
\]

where

\[
c_{i}^{(j)} = b_{i} - \mu_{j} - \frac{1}{2}(y^{*(j),(a-1)}_j + \sigma_{j} - \frac{1}{2}), \quad i = 0, 1,
\]

and the $\varphi_{j}$ are independent uniform $[0, 1]$ variates and $\sigma_{j} = \sqrt{\sum_{i=1}^{n} \Omega_{j}^{ij}}$. Repeat this process for $G$ "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 $\{y^{*} | y^{*} \in D\}$ 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 Tierney (1992) for other theoretical results on the Gibbs resampling scheme. We present below Monte Carlo experiments with simulation estimators based on this truncated simulation scheme.

4. Simulation and estimation of LDV models

4.1. 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 are 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

---

19 The usefulness of Gibbs resampling for Bayesian estimation has been recognized by Geweke (1992), Chib (1993), and by McCulloch and Rossit (1993).
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 (GMM). This is the simulated counterpart to the generalized method of moments (GMM) (see Newey and McFadden (1994)). Within the GMM 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.

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

\[ y = \tau_{ij}(y^*) = 1\{y^*_{(i)} = y^*_{(j)}\}, \quad i, j = 1, \ldots, J \]

where the \( \{y^*_{(i)}\} \) correspond to the order statistics of \( y^* \),

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

so that the first row of \( y \) contains indicators of the smallest element of \( y^* \) and so on until the last row indicates the largest element of \( y^* \). The sample space of \( y \) consists of the \( J! = J \times (J - 1) \times \cdots \times 2 \) different \( J \times J \) matrices containing zeros and ones such that only a single entry equals one in each row and column:

\[ B = \left\{ [y_{ij}]_{i,j = 1, \ldots, J} | y_{ij} \in \{0, 1\}, \sum_i y_{ij} = \sum_j y_{ij} = 1 \right\}. \]

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:

\[ D(y) = \{ y^* \in \mathbb{R}^J | y_1 y^* \leq y_2 y^* \leq \cdots \leq y_J y^* \}, \]

where \( y_j \) is the row vector \([y_{j1}, \ldots, y_{jJ}]^T\). Such additional inequalities as \( y_1 y^* \leq y_3 y^* \) are redundant. As in the multinomial choice model, it is convenient to transform the latent \( y^* \) into a vector of \( J - 1 \) differences:

\[ z_j = \Delta_j y^* = [y_{ij} y^* - y_{i+1, j} y^*; i = 1, \ldots, J - 1]. \]
where
\[ \Delta_j \equiv [y_{ij} - y_{i+1,j}; i = 1, \ldots, J - 1; j = 1, \ldots, J] \]
is a \((J - 1) \times J\) differencing matrix. According to this definition, \(D(y) = \{y^* | z_j \leq 0\}\). The transformed random vector \(z_j\) is also multivariate normal and for all \(Y \in B\),
\[ f(\theta; Y) = \Pr\{y = Y; \mu, \Omega\} = \Phi(\Delta_j \mu, \Delta_j \Omega \Delta_j') . \tag{29} \]
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
\[
\begin{bmatrix}
-1 \\
-1/3 \\
1/3 \\
1
\end{bmatrix} \quad \text{and} \quad 
\begin{bmatrix}
1 & 1/2 & 0 & 0 \\
1/2 & 1 & 0 & 0 \\
0 & 0 & 1 & 1/2 \\
0 & 0 & 1/2 & 1
\end{bmatrix}.
\]

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 \(\Omega\) contains covariances among the latent \(y^*\) that are zero in the latent logit model. The \(\mu\) and \(\Omega\) parameters are not all identifiable and so we normalize by reducing the parameterization to \(\Delta_j \mu\) and \(\Delta_j \Omega \Delta_j'\) for \(Y = I_4\), the \(4 \times 4\) identity matrix. The first variance in \(\Delta_j \Omega \Delta_j'\) is also scaled to 1. In order to restrict \(\Delta_j \Omega \Delta_j'\) 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 nonzero elements of the Cholesky parameters, the identifiable population parameter vector is \(\theta_0 = [-0.6667, -0.6667, -0.6667, 0.5000, 1.3230, 0.0000, -0.3780, 0.9258]\).

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.
Table 1
Sample statistics for rank ordered probit MLE.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.6667</td>
<td>-0.6864</td>
<td>0.1317</td>
<td>-0.7702</td>
<td>-0.6807</td>
<td>-0.5921</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-0.6667</td>
<td>-0.6910</td>
<td>0.2351</td>
<td>-0.8354</td>
<td>-0.6629</td>
<td>-0.5231</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-0.6667</td>
<td>-0.7063</td>
<td>0.2263</td>
<td>-0.8276</td>
<td>-0.6648</td>
<td>-0.5374</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-0.5000</td>
<td>-0.5135</td>
<td>0.2265</td>
<td>-0.6402</td>
<td>-0.5016</td>
<td>-0.3645</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>1.3230</td>
<td>1.3536</td>
<td>0.3002</td>
<td>1.130</td>
<td>1.317</td>
<td>1.519</td>
</tr>
<tr>
<td>$\theta_6$</td>
<td>0.0000</td>
<td>-0.00127</td>
<td>0.1797</td>
<td>-0.1241</td>
<td>-0.000616</td>
<td>0.09545</td>
</tr>
<tr>
<td>$\theta_7$</td>
<td>-0.3780</td>
<td>-0.4081</td>
<td>0.1909</td>
<td>-0.5158</td>
<td>-0.3891</td>
<td>-0.2765</td>
</tr>
<tr>
<td>$\theta_8$</td>
<td>0.9258</td>
<td>0.9385</td>
<td>0.2461</td>
<td>0.7513</td>
<td>0.9140</td>
<td>1.074</td>
</tr>
</tbody>
</table>

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 approximation.

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

$$
\mu = \begin{bmatrix}
-1 \\
-3/5 \\
-1/5 \\
1/5 \\
3/5 \\
1
\end{bmatrix}
$$

and

$$
\Omega = \begin{bmatrix}
1 & 1/2 & 0 & 0 & 0 & 0 \\
1/2 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 5/4 & 3/4 & 1/4 & 1/4 \\
0 & 0 & 3/4 & 5/4 & 1/4 & 1/4 \\
0 & 0 & 1/4 & 1/4 & 5/4 & 3/4 \\
0 & 0 & 1/4 & 1/4 & 3/4 & 5/4 
\end{bmatrix}
$$

which correspond to $\theta_0 = [-0.4000, -0.4000, -0.4000, -0.4000, -0.4000, -0.5000, 1.414, 0.000, -0.3536, 0.9354, 0.000, -0.1768, -0.6013, 1.052, 0.000, 0.000, 0.000, -0.4752, 0.8799]$ 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 \( \theta \) given the sample of observations \( (y_n, n = 1, \ldots, N) \) be

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

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

\[
\hat{\theta}_{MSL} = \arg\max_{\theta} \tilde{l}_N(\theta),
\]

where

\[
\tilde{l}_N(\theta) = \sum_{n=1}^{N} \ln \tilde{f}(\theta; y_n, \omega_n)
\]

for some given simulation sequence \( \{\omega_n\} \).

It is important to note that the 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 probability function in (29), we used the probability simulator in the first element of \( h_{GHK} \) in (28) to compute the simulated log-likelihood function \( \tilde{l}_N(\theta) \).\(^{20}\) For the simulations of the probability of each observation, we drew a vector of \( J - 1 = 3 \) independently distributed uniform random variables for each \( \omega_n \). For each replication of \( \hat{\theta}_{MSL} \), we drew a new data set

\(^{20}\) 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.
Table 2
Sample statistics for rank ordered probit MSLE using GHK ($J = 4$, $R = 1$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.6667</td>
<td>-0.7230</td>
<td>0.1424</td>
<td>-0.8219</td>
<td>-0.7198</td>
<td>-0.6253</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-0.6667</td>
<td>-0.6077</td>
<td>0.2162</td>
<td>-0.7342</td>
<td>-0.5934</td>
<td>-0.4640</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-0.6667</td>
<td>-0.9555</td>
<td>0.2520</td>
<td>-1.087</td>
<td>-0.9256</td>
<td>-0.7860</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-0.5000</td>
<td>-0.6387</td>
<td>0.1430</td>
<td>-0.7305</td>
<td>-0.6415</td>
<td>-0.5379</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>1.3230</td>
<td>1.2595</td>
<td>0.1741</td>
<td>1.134</td>
<td>1.237</td>
<td>1.353</td>
</tr>
<tr>
<td>$\theta_6$</td>
<td>0.0000</td>
<td>0.0131</td>
<td>0.1717</td>
<td>-0.09063</td>
<td>-0.01013</td>
<td>0.1285</td>
</tr>
<tr>
<td>$\theta_7$</td>
<td>-0.3780</td>
<td>-0.6715</td>
<td>0.2088</td>
<td>-0.7883</td>
<td>-0.6639</td>
<td>-0.5292</td>
</tr>
<tr>
<td>$\theta_8$</td>
<td>0.9258</td>
<td>1.3282</td>
<td>0.2211</td>
<td>1.185</td>
<td>1.301</td>
<td>1.448</td>
</tr>
</tbody>
</table>

The results of this Monte Carlo experiment 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.

The results of this Monte Carlo experiment 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

Table 3
Sample statistics for rank ordered probit MSLE using GHK ($J = 6$, $R = 1$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.4000</td>
<td>-0.4585</td>
<td>0.1504</td>
<td>-0.5565</td>
<td>-0.4561</td>
<td>-0.3664</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-0.4000</td>
<td>-0.2489</td>
<td>0.2059</td>
<td>-0.3898</td>
<td>-0.2460</td>
<td>-0.0940</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-0.4000</td>
<td>-0.5054</td>
<td>0.1710</td>
<td>-0.6056</td>
<td>-0.4957</td>
<td>-0.3891</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-0.4000</td>
<td>-0.4589</td>
<td>0.2013</td>
<td>-0.5779</td>
<td>-0.4551</td>
<td>-0.3216</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>-0.4000</td>
<td>-0.6108</td>
<td>0.1882</td>
<td>-0.6934</td>
<td>-0.6016</td>
<td>-0.5042</td>
</tr>
</tbody>
</table>
is simulated without bias because

\[ E[\hat{\theta}(\theta)] = \mu(\theta) \neq E \left[ \arg \max_{\theta} \bar{\mu}(\theta) \right] = \arg \max_{\theta} \mu(\theta). \]

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 (1994), Theorem 2.1):

**Lemma 1**

Let

1. \( \theta \in \Theta \), a compact subset of \( \mathbb{R}^k \),
2. \( Q_0(\theta), Q_\infty(\theta) \) be continuous in \( \theta \),
3. \( \theta_0 = \arg \max_{\theta \in \Theta} Q_0(\theta) \) be unique,
4. \( \hat{\theta}_N = \arg \max_{\theta \in \Theta} Q_N(\theta) \) 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 \( \theta \) is identifiable, that \( f(\theta; y) \) is continuous at each \( \theta \) in a compact parameter space \( \Theta \), and that \( E[\sup_{\theta \in \Theta} |\ln f(\theta; y)|] < \infty \). We refer the reader to Newey and McFadden (1994, 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:

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

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.

**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.

In the rank ordered probit model and similar discrete I.DV 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 $\omega$ satisfy any restrictions on dependence across observations. The following proposition, taken from Lee (1992), 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 $y \in \mathcal{B}$, the sample space of $y$. Assume that the set of regularity conditions in the

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$-0.6667$</td>
<td>$-0.6795$</td>
<td>0.1366</td>
<td>$-0.7726$</td>
<td>$-0.6774$</td>
<td>$-0.5840$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>$-0.6667$</td>
<td>$-0.6528$</td>
<td>0.2267</td>
<td>$-0.7913$</td>
<td>$-0.6268$</td>
<td>$-0.5029$</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>$0.5000$</td>
<td>$0.5771$</td>
<td>0.2159</td>
<td>$-0.7076$</td>
<td>$-0.5641$</td>
<td>$-0.4412$</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>1.3230</td>
<td>1.3582</td>
<td>0.2549</td>
<td>1.1863</td>
<td>1.3184</td>
<td>1.5036</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>$0.0000$</td>
<td>$0.0121$</td>
<td>0.2089</td>
<td>$-0.1380$</td>
<td>$-0.0157$</td>
<td>$0.1275$</td>
</tr>
<tr>
<td>$\theta_6$</td>
<td>$-0.3780$</td>
<td>$-0.5034$</td>
<td>0.2016</td>
<td>$-0.6256$</td>
<td>$-0.4975$</td>
<td>$-0.3753$</td>
</tr>
<tr>
<td>$\theta_7$</td>
<td>$0.9258$</td>
<td>1.1334</td>
<td>0.2454</td>
<td>0.9505</td>
<td>1.1412</td>
<td>1.2814</td>
</tr>
</tbody>
</table>

**Table 5**

Sample statistics for rank ordered probit MSLE using GHK ($J = 6$, $R = 5$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$-0.4000$</td>
<td>$-0.4088$</td>
<td>0.1256</td>
<td>$-0.4893$</td>
<td>$-0.4053$</td>
<td>$-0.3227$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>$-0.4000$</td>
<td>$-0.3059$</td>
<td>0.1776</td>
<td>$-0.4200$</td>
<td>$-0.2966$</td>
<td>$-0.1846$</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>$-0.4000$</td>
<td>$-0.4554$</td>
<td>0.1387</td>
<td>$-0.5373$</td>
<td>$-0.4553$</td>
<td>$-0.3615$</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>$-0.4000$</td>
<td>$-0.4288$</td>
<td>0.1661</td>
<td>$-0.5369$</td>
<td>$-0.4219$</td>
<td>$-0.3142$</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>$-0.4000$</td>
<td>$-0.5046$</td>
<td>0.1773</td>
<td>$-0.6211$</td>
<td>$-0.4976$</td>
<td>$-0.3872$</td>
</tr>
</tbody>
</table>
paragraph after Lemma 1 hold. Let \( \{\omega_n\} \) be an i.i.d. sequence over the index \( r \). The MSL estimator \( \hat{\theta}_{\text{MSL}} \equiv \arg \max_{\theta} (1/N) \sum_{n=1}^{N} \ln \tilde{f}(\theta; y_n, \omega_n) \) is consistent if \( R \to \infty \) as \( N \to \infty \).

**Proof**

By a uniform law of large numbers and the lower bound of \( \tilde{f} \),

\[
\sup_{r, \theta} \left| \frac{\tilde{f}(\theta; y_n, \omega_n)}{f(\theta; y_n)} - 1 \right| \overset{p}{\to} 0, \quad \text{as} \quad R \to \infty,
\]

so that

\[
\sup_{\theta} \frac{1}{N} \left| \tilde{I}_n(\theta) - l_n(\theta) \right| \overset{p}{\to} 0, \quad \text{as} \quad R \to \infty, \quad N \to \infty.
\]

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

\[
\sup_{\theta} \left| \frac{1}{N} l_n(\theta) - E[\ln f(\theta; y)] \right| \overset{p}{\to} 0 \quad \text{as} \quad N \to \infty,
\]

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

Q.E.D.

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 \( \{\omega_n\} \), 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 \( \{\omega_n\} \) are independent across the observation index \( n \), as well as the replication index \( r \), then we preserve the independence of the \( \tilde{f}(\theta; y_n, \omega_n) \) and its derivatives across \( n \), permitting the application of familiar laws of large numbers and central limit theorems. When \( \tilde{f} \) is
differentiable in $\theta$, we can make a familiar linear approximation for $\hat{\theta}_{\text{MSL}}$:

$$0 = \frac{1}{\sqrt{N}} \nabla_{\theta} \tilde{f}(\theta_0) + \left[ \frac{1}{N} \nabla_{\theta}^2 \tilde{f}(\theta) \right] \sqrt{N}(\hat{\theta}_{\text{MSL}} - \theta_0),$$

(30)

where the elements of $\theta$ lie on the line segment between $\hat{\theta}_{\text{MSL}}$ and $\theta_0$. The consistency of $\hat{\theta}_{\text{MSL}}$ implies the consistency of $\theta$ which in turn implies that

$$\frac{1}{N} \nabla_{\theta}^2 \tilde{f}(\theta) \xrightarrow{P} E[\nabla_{\theta}^2 \ln f(\theta_0; y)] = \mathcal{J}(\theta_0),$$

(31)

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

terms

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

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

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,

$$\frac{1}{\sqrt{N}} \nabla_{\theta} \tilde{f}(\theta_0) = \frac{1}{\sqrt{N}} \nabla_{\theta} \tilde{f}(\theta_0) + A_N + B_N,$$

(32)

where

$$A_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \{ \nabla_{\theta} \ln \tilde{f} - E_\omega[\nabla_{\theta} \ln \tilde{f}] \},$$

$$B_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \{ E_\omega[\nabla_{\theta} \ln \tilde{f}] - \nabla_{\theta} \ln f \}.$$  

(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 $\hat{\theta}_{\text{MSL}}$. $B_N$ 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 $\bar{\mu}(\theta; y, \omega)$ be an unbiased simulator for $\mu(\theta; y)$ such that $V(\bar{\mu} - \mu|y) = O(R^{-1})$.

Let $s(\theta; y, \mu)$ be a moment function such that $E[s(\theta_0; y, \mu)] = 0$. Consider the simula-
tor $\tilde{s}(\theta; y) = s(\theta; y, \tilde{\mu})$ and let $R/\sqrt{N} \to \infty$. If $\tilde{s}$ is Lipschitz in $\tilde{\mu} \in \mathcal{S}$ uniformly in $\theta$, then the simulation bias

$$B_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left\{ E_n [\tilde{s}(\theta; y)] - s(\theta; y, \mu) \right\} \overset{P}{\to} 0.$$

**Proof**

If $\tilde{s}$ is Lipschitz in $\tilde{\mu}$ uniformly in $\theta$ then

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

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

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

so that

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

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

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

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

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

The difference between simulated and exact scores can be written

\[ A_n + B_n = \frac{1}{\sqrt{N}} \sum_n \frac{1}{f} \left[ \nabla \tilde{f} - \tilde{f} \nabla \ln f \right] \]

\[ = O(1) \left( \frac{1}{\sqrt{N}} \sum_n \left[ \nabla \tilde{f} - \tilde{f} \nabla \ln f \right] \right). \]

By the Chebychev inequality,

\[ \Pr \left( \left| \frac{1}{\sqrt{N}} \sum_n \left[ \nabla \tilde{f} - \tilde{f} \nabla \ln f \right] \right| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \sum_n \left[ \nabla \tilde{f} - \tilde{f} \nabla \ln f \right] = O(\sqrt{N}/R) \]

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

Q.E.D.

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 \( \hat{\theta}_{\text{MSL}} \) negligible.

There is a simulation design that uses the same total number \((N \times R)\) of simulations of \( \omega \) as the independent design, but applies every simulation of \( \omega \) to every observation of \( y \). That is, the simulated log-likelihood function is generated according to the double sum

\[ \tilde{f}(\theta) = \sum_{n=1}^{N} \sum_{m=1}^{R} \ln \tilde{f}(\theta, y_n, \omega_m). \]

The motivation for this approach is to take advantage of all \( N \times R \) simulations that must be drawn when \( R \) independent simulations are made for each observation. Lee (1992) 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 \( N \times R \) 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 \( O(N^{1.5}) \), 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 Larouque and Salanié (1989) is another early simulation estimation approach for LDV models. This
method, originally developed for the mixture models of Section 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 \( r(y^*) = y \) yield a mixture model with the first two moments \( g_1(x_n, \theta) = E(y|x_n, \theta) \) and \( g_2(x_n, \theta) = E((y - E(y))^2|x_n, \theta) \). 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 \( \hat{\theta}_{\text{SPML}} \) is defined by:

\[
\hat{\theta}_{\text{SPML}} = \arg \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \psi(y_n, \tilde{g}_1(\cdot), \tilde{g}_2(\cdot)),
\]

(34)

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

Laroque and Salanić (1989) prove that for \( x_n \in X \in \Theta \) compact, and \( \tilde{g}_j(\cdot) \) sufficiently continuous on \( X \times \Theta \), then \( \hat{\theta}_{\text{SPML}} \xrightarrow{p} \theta_{\text{PML}} \) as \( R \to \infty \).\(^{21}\) It should be noted that for particular 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 \( \psi(\cdot) = (y_n - g_1(\cdot))^2 \), 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).

\(^{21}\) 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 (1994) for a discussion of CMD and the closely related generalized method of moments (GMM).
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 \( \theta \) to be estimated. Let \( E(y|\theta, x) = \mu(\theta; x) \). 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 \( \hat{\theta}_{\text{MOM}} \) is defined by

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

The consistency of such estimators rests on the uniform convergence of the sample averages to their population counterparts for any value of \( \theta \) as the sample size approaches infinity. When the unique root of the population equations is \( \theta_0 \), the population value of \( \theta \), the root of the sample equations, converges to \( \theta_0 \). The limiting distribution of \( \hat{\theta}_{\text{MOM}} \) is derived from the linear expansion

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

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

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

where

\[
\frac{1}{N} \sum_{n=1}^{N} w_n(\theta_0) V[e_n(\theta_0)|x_n] w_n(\theta_0) \xrightarrow{P} \Sigma_{\text{MOM}}.
\]

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 \( \hat{\theta}_{\text{MOM}} \). Thus under regularity
conditions,

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

where

$$\frac{1}{N} \sum_{n=1}^{N} w_n(\bar{y}) \mathbf{V}_n e_n(\bar{y}) \overset{p}{\to} \mathbf{H}.$$

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 $\bar{\mu}(\theta; x, \omega) = 1/R \sum_{r=1}^{R} \mu(\theta; x, \omega_r)$ be an unbiased simulator so that $\mu(\theta; x) = E[\bar{\mu}(\theta; x, \omega)|x]$ where $\omega$ is a simulated random variable. The method of simulated moments estimator is

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

where

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

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 (1994), 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.
Table 6
Sample statistics for rank ordered probit CMD ($J = 4$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.6667</td>
<td>-0.6906</td>
<td>0.1481</td>
<td>-0.7792</td>
<td>-0.6918</td>
<td>-0.5948</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-0.6667</td>
<td>-0.7887</td>
<td>0.3714</td>
<td>-0.9496</td>
<td>-0.7109</td>
<td>-0.5431</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-0.6667</td>
<td>-0.6953</td>
<td>0.2223</td>
<td>-0.8347</td>
<td>-0.6594</td>
<td>-0.5366</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-0.5000</td>
<td>-0.6683</td>
<td>0.4271</td>
<td>-0.8962</td>
<td>-0.5688</td>
<td>-0.3679</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>1.3230</td>
<td>1.4143</td>
<td>0.4384</td>
<td>1.118</td>
<td>1.337</td>
<td>1.633</td>
</tr>
<tr>
<td>$\theta_6$</td>
<td>0.0000</td>
<td>0.1764</td>
<td>0.0505</td>
<td>-0.1957</td>
<td>0.08331</td>
<td>0.5563</td>
</tr>
<tr>
<td>$\theta_7$</td>
<td>-0.3780</td>
<td>-0.3077</td>
<td>0.2765</td>
<td>-0.4703</td>
<td>-0.3207</td>
<td>-0.1747</td>
</tr>
<tr>
<td>$\theta_8$</td>
<td>0.9258</td>
<td>0.7714</td>
<td>0.3356</td>
<td>0.5955</td>
<td>0.7980</td>
<td>0.9834</td>
</tr>
</tbody>
</table>

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 $y$:

$$\sum_{n=1}^{N} \frac{y_{jn}}{N} - \Pr[y_{ij} = 1; \hat{\mu}, \hat{\Omega}] = 0, \quad i,j = 1, \ldots, J - 1.$$

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 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 seem 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
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 $\hat{\mu}(\theta; x, \omega)$ 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 $\omega_{nr} = \omega_{n}, n = 1, \ldots, N.$

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

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

**Proof**

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>-0.6667</td>
<td>-0.6976</td>
<td>0.1905</td>
<td>-0.7915</td>
<td>-0.6809</td>
<td>-0.5798</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-0.6667</td>
<td>-0.9790</td>
<td>0.9576</td>
<td>-1.099</td>
<td>-0.7619</td>
<td>-0.5654</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-0.6667</td>
<td>-0.8561</td>
<td>0.6394</td>
<td>-1.008</td>
<td>-0.6900</td>
<td>-0.4813</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>-0.5000</td>
<td>-0.7083</td>
<td>0.6392</td>
<td>-0.8918</td>
<td>-0.5559</td>
<td>-0.3327</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>1.3230</td>
<td>1.4733</td>
<td>0.8402</td>
<td>1.086</td>
<td>1.323</td>
<td>1.662</td>
</tr>
<tr>
<td>$\theta_6$</td>
<td>0.0000</td>
<td>0.0780</td>
<td>0.6268</td>
<td>-0.3091</td>
<td>0.03749</td>
<td>0.4616</td>
</tr>
<tr>
<td>$\theta_7$</td>
<td>-0.3780</td>
<td>-0.3828</td>
<td>0.5423</td>
<td>-0.5758</td>
<td>-0.3099</td>
<td>-0.1110</td>
</tr>
<tr>
<td>$\theta_8$</td>
<td>0.9258</td>
<td>0.8500</td>
<td>0.6837</td>
<td>0.5023</td>
<td>0.7341</td>
<td>0.9745</td>
</tr>
</tbody>
</table>
\[ s_N(\theta) - \tilde{s}_N(\theta) = \frac{1}{N} \sum_{n=1}^{N} w_n(\theta) [\tilde{\mu}(\theta; x_n, \omega_n) - \mu(\theta; x_n)] \]

\[ = \frac{1}{NR} \sum_{n=1}^{N} \sum_{s=1}^{R} w_n(\theta) [\tilde{\mu}(\theta; x_n, \omega_{ns}) - \mu(\theta; x_n)], \]

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.i.d. sequence so that a uniform law of large numbers applied to (37) implies \( s_N(\theta) - \tilde{s}_N(\theta) \to^P 0 \) 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 (1992)) implies \( s_N(\theta) - \tilde{s}_N(\theta) \to^P 0 \) as \( N \to \infty \). Therefore, in either case, by continuity, \( \| s_N(\theta) - \tilde{s}_N(\theta) \| \to^P 0 \) uniformly in \( \theta \) and Lemma 1 implies the result. Q.E.D.

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 \( \tilde{\theta}_{\text{MSM}} \) exists:

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

where we have denoted the simulated residual by \( \tilde{e}_n(\theta) = y_n - \tilde{\mu}(\theta; x_n) \) and \( \tilde{\theta} \) lies between \( \theta_{\text{MSM}} \) and \( \theta_0 \). 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 \( \tilde{\theta}_{\text{MSM}} \):

\[ \frac{1}{\sqrt{N}} \sum_{n=1}^{N} w_n(\theta_0) \tilde{e}_n(\theta_0) \xrightarrow{d} N(0, \Sigma_{\text{MSM}}), \]

where

\[ \frac{1}{N} \sum_{n=1}^{N} w_n(\theta_0) V[\tilde{e}_n(\theta_0) | x_n] w_n(\theta_0) \to^P \Sigma_{\text{MSM}}. \]

Also, as before,

\[ \frac{1}{N} \sum_{n=1}^{N} \tilde{e}_n(\theta) \nabla_{\theta} w_n(\theta) \to^P 0. \]

so that under regularity conditions,
\[
\sqrt{N}(\hat{\theta}_{\text{MSM}} - \theta_0) \xrightarrow{d} N(0, H^{-1} \Sigma_{\text{MSM}} H^{-1}),
\]
where
\[
\frac{1}{N} \sum_{n=1}^{N} w_n(\bar{\theta}) V_n(\bar{\theta}) \rightarrow^p H.
\]

The equivalence of the \(H\) matrices also rests on the unbiased simulation of \(\mu\). If \(\mu(\theta; x) = E[\bar{\mu}(\theta; x, \omega)|x]\), then \(V_n(\mu(\theta; x)) = V_n E[\bar{\mu}(\theta; x, \omega)|x] = E[V_n(\bar{\mu}(\theta; x, \omega)|x)]\) 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 \(\hat{\theta}_{\text{MOM}}\) and \(\hat{\theta}_{\text{MSM}}\). 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 \(\hat{\theta}_{\text{MSM}}\) is contained in the object (37). This term is generated conditional on the realizations of \(y\) and is, by definition, distributed independently of the classical moment function.

Inflating the simulation noise by \(\sqrt{N}\) and evaluating it at \(\theta_0\), we can apply a central limit theorem to it to obtain the following result.

**Proposition 7**

\[
\Sigma_{\text{MSM}} = \Sigma_{\text{MOM}} + (1/R) \Sigma_S \text{ where }
\]
\[
\frac{1}{N} \sum_{n=1}^{N} w_n(\theta_0) V[\bar{\mu}(\theta_0; x, \omega_m) - \mu(\theta_0; x_m)] w_n(\theta_0)^T \rightarrow^p \Sigma_S.
\]

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 \(\bar{\mu}\) is obtained by averaging simulations of the data generating process itself, \(\Sigma_S = \Sigma_{\text{MOM}}\) and \(\Sigma_{\text{MSM}} = (1 + 1/R) \Sigma_{\text{MOM}}\). 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 \(R = O(N^z), z > 0\), then \(\sqrt{N}(\hat{\theta}_{\text{MOM}} - \hat{\theta}_{\text{MSM}}) \xrightarrow{d} 0\).
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 \( H = \Sigma_{\text{MSM}} \) 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:

\[
    w_n = [V_{\theta}\mu(p, \omega, x_n)]^{-1} - V[y_n - \mu(p, \omega, x_n)]^{-1}.
\]

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 \( \mu \) is differentiable in \( \theta \), then independent simulations of the \( V_{\theta}\mu \) are unbiased simulators of the instruments. Otherwise, discrete numerical derivatives can be employed. The covariance matrix can be estimated using the sample variance of \( \mu \) 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 \( \mu \), 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 nor-

\(^{23}\) A Taylor series expansion suggests that positive correlation between the numerator and denominator of a ratio can yield a smaller variance than independent simulation.
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.\textsuperscript{24}

**Definition 8. Method of simulated scores**

Let the log-likelihood function for the unknown parameter vector $\theta$ given the sample of observations $(y_n, n = 1, \ldots, N)$ be $l_\theta(\theta) = \sum_{n=1}^{N} \ln f(\theta; y_n)$. Let $\mu(\theta; y, \omega) = (1/R) \sum_{r=1}^{R} \hat{\mu}(\theta; y, \omega_r)$ be an asymptotically (in $R$) unbiased simulator of the score function $\mu(\theta; y) = \nabla \ln f(\theta; y)$ where $\omega$ is a simulated random variable. The method of simulated scores estimator is $\hat{\theta}_{MSL} \equiv \arg \min_{\theta \in \Theta} \| \hat{S}(\theta) \|$ where $\hat{S}(\theta) \equiv (1/N) \sum_{n=1}^{N} \hat{\mu}(\theta; y_n, \omega_n)$ 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.

**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 faster 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 Section 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 $\{\eta_t\} \sim N(0, I_M)$ in order to obtain truncated multivariate normal random variables for rank ordered probit estimation. Given the observation $y$, one seeks a simulation from $D(y)$, as defined in Example 8. Let the simulation of $y^*$ be $\tilde{y}_t(\mu, \Gamma) \equiv \mu_t + \Gamma_t \eta_t$ at the parameter values $(\mu, \Gamma)$. 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 $\{\tilde{y}_t(\mu, \Gamma)\}$. See Hajivassiliou and McFadden (1990) and McFadden and Ruud (1992) for treatments of the special

\textsuperscript{24} The term was coined by Hajivassiliou and McFadden (1990).
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 discontinuities 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 Section 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. 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 $r$th iteration, the EM algorithm solves

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

where the function $Q$ is

$$
Q(\theta^1, \theta^0; y) = E_{\theta^1}[\ln f(\theta^1; y^*)|y], \quad (40)
$$

where $E_{\theta^1}[\cdot|y]$ indicates an expectation measured with respect to $f(\theta^0; y^*|y)$.

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

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

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 $\ln f(\theta; y^*)$.

---

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.
(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

$$
\mu_t = \frac{1}{N} \sum_{n=1}^{N} E_{\theta_t}(y_n^* | y_n) \quad \text{and} \quad \Omega_t = \frac{1}{N} \sum_{n=1}^{N} E_{\theta_t}[(y_n^* - \mu_t)(y_n^* - \mu_t)' | y_n],
$$

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 $\theta$ can be imposed at convergence. See Ruud (1991) for a discussion of these points.

**Example 13. SEM estimation**

In this example, we apply the SEM procedure to the rank ordered probit model of our previous examples. We simulated an (approximately) unbiased $\hat{Q}$ 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 $D(y)$ in Example 8. The simulated estimators were computed according to (42), after replacing the expectations with the averages of independent simulations.

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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population value</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$-0.4000$</td>
<td>$-0.3827$</td>
<td>0.1558</td>
<td>$-0.4907$</td>
<td>$-0.3848$</td>
<td>$-0.2757$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>$-0.4000$</td>
<td>$-0.4570$</td>
<td>0.3271</td>
<td>$-0.5992$</td>
<td>$-0.4089$</td>
<td>$-0.2455$</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>$-0.4000$</td>
<td>$-0.4237$</td>
<td>0.2262</td>
<td>$-0.5351$</td>
<td>$-0.3756$</td>
<td>$-0.2766$</td>
</tr>
<tr>
<td>$\theta_4$</td>
<td>$-0.4000$</td>
<td>$-0.4268$</td>
<td>0.2710</td>
<td>$-0.5319$</td>
<td>$-0.3891$</td>
<td>$-0.2580$</td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>$-0.4000$</td>
<td>$-0.4300$</td>
<td>0.2622</td>
<td>$-0.5535$</td>
<td>$-0.3794$</td>
<td>$-0.2521$</td>
</tr>
</tbody>
</table>
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 \( y = \tau(y^*) \) can generally be written in the ratio form:

\[
\frac{\nabla_\theta f(\theta; y)}{f(\theta; y)} = \frac{\nabla_\theta dF(\theta; y^*)}{dF(\theta; y^*)} = \frac{E(\nabla_\theta dF(\theta; y^*) | \tau(y^*) = y)}{Pr(\tau(y^*) = y)},
\]

where \( F(\theta; y^* | y) \) is the conditional c.d.f. of \( y \) given \( \tau(y^*) = y \). See Section 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. An MSS estimator can be constructed by simulating separately the numerator and denominator of the score expressions:

\[
\hat{s}_n(\theta) = \frac{1}{N} \sum_{n=1}^{N} \frac{\tilde{d}(\theta; y_n, \omega_{1n})}{\tilde{p}(\theta; y_n, \omega_{2n})},
\]

where \( \tilde{d}(\theta; y_n, \omega_{1n}) = (1/R_1) \sum_{r=1}^{R_1} \tilde{d}(\theta; y_n, \omega_{1n}) \) is an unbiased simulator of the derivative function \( \nabla_\theta f(\theta, y) \) and \( \tilde{p}(\theta; y_n, \omega_{2n}) = (1/R_2) \sum_{r=2}^{R_2} \tilde{p}(\theta; y_n, \omega_{2n}) \) is an unbiased function of the probability expression \( f(\theta; y) \). 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 \( \tilde{p}(\theta; y, \omega) \) of \( f(\theta; y) \) does not yield an unbiased simulator of

\[26\, \text{See Hajivassiliou (1993e) for a survey of the development of simulation estimation methods for LDV models.}\]
the reciprocal $1/f(\theta; y)$ in the simulator $1/\hat{\beta}(\theta; y, \omega_2)$, $R_2$ 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 $\tilde{d} = \nabla_y \hat{\beta}$.

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 $B$ that is countable and finite, we can always write $y$ as a vector of dummy variables for each of the possible outcomes so that

$$E_0(y_j) = \Pr\{y_i = 1; \theta\} = f(\theta; Y) \text{ if } Y_i = 1, Y_j = 0, j \neq i.$$ 

Thus,

$$E_0 \left[ \frac{\nabla_0 f(\theta; y)}{f(\theta; y)} \right] = 0 = \sum_{y \in B} f(\theta; Y) \frac{\nabla_0 f(\theta; Y)}{f(\theta; Y)}$$

and the score can be written

$$\frac{\nabla_0 f(\theta; y)}{f(\theta; y)} = \sum_{y \in B} 1\{y = Y\} \frac{\nabla_0 f(\theta; Y)}{f(\theta; Y)}$$

$$= \sum_{y \in B} \left[ 1\{y = Y\} - f(\theta; Y) \right] \frac{\nabla_0 f(\theta; Y)}{f(\theta; Y)}. \tag{44}$$

Provided that the “residual” $1\{y = Y\} - f(\theta; Y)$ and the “instrumental variables” $\nabla_0 f(\theta; Y)/f(\theta; Y)$ 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 values.

For example, we can alter (43) to

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

where $\omega_1$ and $\omega_2$ 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 the other.

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

$$
\mathbf{s}_{\text{MSS}}(\theta) = R^{-1} \sum_{r=1}^{R} \nabla_\theta l^\ast[\theta; \mathbf{Y}^\ast(\theta; Y; \omega)].
$$

Then $\Sigma_s$, the contribution of simulation to the covariance matrix of the estimator, has a useful interpretation:

$$
\Sigma_s = (\Sigma_s^* - \Sigma_0)/R,
$$

where $\Sigma_s^* = E_{\omega} \{ \nabla_\theta l^\ast(\theta; \mathbf{Y}^\ast) [\nabla_\theta l^\ast(\theta; \mathbf{Y}^\ast)]' \}$ 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 $\mathbf{s}_{\text{MSS}}(\theta) = g(\theta; Y, \omega_2) - \tilde{g}(\theta; \omega_1, \omega_2)$ (see (45)). In that case, $\Sigma_s$ simplifies to $V\{\sqrt{N}[\tilde{g}(\theta; \omega_1, \omega_2)]\}$. In general, the simulation process makes $R$ independent replications of the simulations $\{\omega_r; r = 1, \ldots, R\}$, so that

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

and $\Sigma_s = R^{-1} V_{\omega} \{ \tilde{g}(\theta; \omega_1, \omega_2) \}$. In an important special case of censored simulation, the simulation process makes $R$ independent replications of the modeled data generating process, $\{ \mathbf{Y}(\theta; \omega_r); r = 1, \ldots, R \}$, so that

$$
\tilde{g}(\theta; \omega_1, \omega_2) = R_2^{-1} \sum_{r=1}^{R} g[\theta; \mathbf{Y}(\theta; \omega_r), \omega_2]
$$

and $\Sigma_s = R^{-1} V_{\omega} \{ \tilde{g}(\theta, \omega_1, \omega_2) \} = \Sigma_M/R$. 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: $E[\mathbf{s}_{\text{MSS}}(\theta) | Y] = \nabla_\theta l(\theta; Y)$.

Of course in most cases, the MSM estimator will have a simulation bias.
for the score. The asymptotic variance of the MSM estimator is
\[
\Sigma_M = \lim_{N \to \infty} \mathbb{V}(\hat{s}_{\text{MSM}}(\theta_0) - \nabla \ell(\theta_0; Y))
\]
\[
= \lim_{N \to \infty} \mathbb{V}(\hat{s}_{\text{MSM}}(\theta_0)) + \Sigma_0
\]
\[
= \Sigma_W + \Sigma_0,
\]
where \( \Sigma_W = \Sigma_M/R \) and \( \Sigma_W \) 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
\[
R^{-1} \Sigma_0^{-1} [\Sigma_0 + (R + 1) \Sigma_W - (\Sigma_W - \Sigma_0)] \Sigma_0^{-1}.
\]
This expression gives guidance about the conditions under which censored simulation is likely to dominate truncated. It is already obvious that if \( \Sigma_W \) is 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 \( \Sigma_0 \) 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 \( E[\hat{s}(\theta) | Y] = \nabla \ell(\theta; Y), \)
\[
E[g(\theta; Y, \omega_3) - \tilde{g}(\theta, \omega_1, \omega_2)] = \nabla \ell(\theta; Y)
\]
\[
\Rightarrow E[\tilde{g}(\theta, \omega_1, \omega_2)] = E[g(\theta, \tilde{Y}(\theta; \omega), \omega')] = 0 \quad \forall \theta.
\]
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 \( E_{\omega}[e(\theta)] \neq \nabla \ell(\theta; Y) \) in such a way that the loss in efficiency in \( \Sigma_W \) is offset by low \( \Sigma_0 \) and low \( \Sigma_W. \)

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

\[\footnote{The actual difference in asymptotic covariance matrices is more complicated than the formula above however, because \( G \neq \Sigma_W \neq \Sigma_0. \)}\]
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 \( V_{af}(\theta, Y) / f(\theta, Y) \). 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.

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

\[
E[s(\tilde{\theta}; Y, X) | X] = 0.
\]

When the moment function \( s \) is computationally burdensome, an approximation \( g(\tilde{\theta}; Y, X, \omega) \) becomes a feasible alternative. The additional argument \( \omega \) 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 \( \tilde{g}(\tilde{\theta}; X, \omega', \omega) \), where \( \omega' \) represents the simulated component. Thus, we consider estimators \( \tilde{\theta} \) that satisfy

\[
g(\tilde{\theta}; Y, X, \omega) - \tilde{g}(\tilde{\theta}; X, \omega', \omega) = 0. \tag{47}
\]

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 Wald statistic, measuring the statistical significance of the bias term, can be computed (see Engle (1984)). 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 Daniel 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


