COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale University
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 960

SMOOTH UNBIASED MULTIVARIATE PROBABILITY SIMULATORS FOR MAXIMUM LIKELIHOOD ESTIMATION OF LIMITED DEPENDENT VARIABLE MODELS

by Axel Börsch–Supan and Vassilis Hajivassiliou

September 1990

Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models

by

Axel Börsch-Supan, University of Mannheim and NBER

and

Vassilis Hajivassiliou, Cowles Foundation, Yale University

July 1990

Revised September 1990

ABSTRACT

We apply a new simulation method that solves the multidimensional probability integrals that arise in maximum likelihood estimation of a broad class of limited dependent variable models. The simulation method has four key features: the simulated choice probabilities are unbiased; they are a continuous and differentiable function of the parameters of the model; they are bounded between 0 and 1; and their computation takes an effort that is nearly linear in the dimension of the probability integral, independent of the magnitudes of the true probabilities. We also show that the new simulation method produces probability estimates with substantially smaller variance than those generated by acceptance-rejection methods or by Stern's (1987) method. The simulated probabilities can therefore be used to revive the Lerman and Manski (1981) procedure of approximating the likelihood function using simulated choice probabilities by overcoming its computational disadvantages.

* We thank Don Andrews, Al Klevorick, Francois Laisney, Michael Lechner, Daniel McFadden, and Ariel Pakes for helpful comments.

1. INTRODUCTION

It has long been known that classical estimation of limited dependent variable (LDV) models is computationally intractable if one does not impose restrictive correlation structures on the unobservables, because of a need to evaluate high dimensional integrals in likelihood and conditional moment conditions. Recently, McFadden (1989) and Pakes and Pollard (1989) have shown how the multivariate integration problems in a class of LDV models, namely discrete choice models, can be avoided in such models, through the method of simulated moments (MSM), which substitutes unbiased simulators for the choice probabilities in the conditional moment conditions. In a related line of work, Lerman and Manski (1981) proposed the method of simulated maximum likelihood (SML), which used (unbiased) frequency simulators of choice probabilities directly in likelihood function evaluation.

In this paper we apply a new simulation method for choice probabilities that builds on recent progress in Monte Carlo integration techniques by Geweke (1989) and Hajivassiliou and McFadden (1990) and can be employed in MSM and SML estimators. This method has four key features: the simulated choice probabilities are unbiased; they are a continuous and differentiable function of the parameters of the model; they are bounded away from 0 and 1; and their computation takes an effort that is nearly linear in the dimension of the probability integral. Most notably, and in contrast to other simulation methods proposed in the literature, the effort is independent of the true probabilities.

In Section 2, we present the classical LDV Model and explain the multivariate integration problems arising in maximum likelihood estimation. In Section 3, we describe the new unbiased choice probability simulator, which we term the "smooth recursive conditioning" simulator (SRC), and contrast its properties to the ones of the classic frequency simulator used in Lerman and Manski (1981) and to those of the Stern (1987) simulator. The properties of the SRC simulator are derived analytically.

In Section 4, we present Monte Carlo results on the distribution of the three probability simulators. We find that the SRC method completely dominates the other two methods. We show that the SRC method produces probability estimates with substantially smaller variance than those generated by acceptance-rejection or by Stern's (1987) method. This is important because, as we show in Section 5, it allows us to modify the Lerman and Manski (1981) simulation estimation method through the use of the SRC simulator to define a smooth simulated maximum likelihood (SSML) estimator. The small variance of the SRC simulator permits the insertion of the simulated probabilities in a likelihood function without the tough trade-off between a large bias and a heavy computational burden that has handicapped the Lerman and Manski (1981) procedure of approximating the likelihood function by frequency simulated choice probabilities. Section 6 follows with some Monte Carlo results on the distribution of parameters estimated by SSML. These results show that due to the small variance of the simulated probabilities, the bias intrinsic to the SSML procedure becomes negligible even for rather small numbers of replications, in sharp contrast to the Lerman and Manski (1981) procedure. Section 7 concludes with a brief summary.

2. MULTIDIMENSIONAL INTEGRATION PROBLEMS IN LDV MODELS

A broad class of limited dependent variable models can be characterized by a combination of a linear latent variable model and a nonlinear relation between the latent variable and an observable variable that can be expressed as a set of inequalities on linear combinations of the latent variable. To be specific, the linear multivariate latent variable model

$$u_i = X_i\beta + \epsilon_i , \qquad i=1,...,I \tag{1}$$

relates the latent variable $u \in \mathbb{R}^I$ to an $I \cdot K$ matrix of observed explanatory variables X, a K dimensional parameter vector β, and a disturbance $\epsilon \in \mathbb{R}^I$ with a covariance matrix Ω. The

nonlinear function τ

$$(2) \qquad y_i = \tau(u_i), \qquad i=1,\ldots,I$$

relates the latent variable $u \in \mathbb{R}^I$ to an observed variable $y \in \mathbb{R}^I$ and can be described by inequalities between linear combinations of u_i:

$$(3) \qquad \begin{aligned} T(y) &= \{ u \in \mathbb{R}^I \mid y_i = \tau(u_i), i=1,\ldots,I \} \\ &= \{ u \in \mathbb{R}^I \mid a \le Au \le b \}. \end{aligned}$$

The limiting vectors a and $b \in \mathbb{R}^I_{\varpi}$ and the $I \cdot I$ matrix A may depend on y and $X\beta$.[1]

The most prominent examples in this class are cross-sectional discrete choice models. In this case I denotes the number of alternatives, τ is defined by

$$(4) \qquad \begin{aligned} \tau(u_i) = 1 \qquad &\Leftrightarrow \qquad i = \operatorname*{argmax}_j (u_j,\ j=1,\ldots,I) \\ = 0 \qquad &\text{otherwise,} \end{aligned}$$

and the inequalities in (3) are defined by

$$(5) \qquad A = \begin{bmatrix} -1 & & 1 \\ & -1 & 1 \\ & & \cdots \\ & -1 & 1 \\ & & 1 \\ & & \cdots \\ & & 1\ -1 \end{bmatrix}, \quad a = \begin{bmatrix} X_i\beta - X_1\beta \\ X_i\beta - X_2\beta \\ \cdots \\ X_i\beta - X_{i-1}\beta \\ -\varpi \\ \cdots \\ X_i\beta - X_I\beta \end{bmatrix}, \quad b = \begin{bmatrix} \varpi \\ \\ \\ \\ \\ \\ \varpi \end{bmatrix}.$$

The column of ones in the matrix A is in column $i=\operatorname*{argmax}_j(u_j,\ i=1,..,I)$ and the element $-\varpi$ in the limiting vector a in the corresponding row.[2]

Hajivassiliou and McFadden (1990) elaborate on other examples in this class, such

---

[1] The inequality in (3) is strict if a or b is infinite.

[2] Hence, the dimension I can be reduced by one.

as multiperiod probit and tobit models and multimarket disequilibrium models. They develop the method of simulated scores (MSS) and report on an application to debt crises in less developed countries. Börsch-Supan et al. (1990) investigate a multiperiod-multinomial probit model in which I represents the number of periods times the number of alternatives, and they apply the model to living arrangement choices of the elderly.

The likelihood of an observation (y,X) associated with model (1)–(3) is

$$\ell(y,X;\, g,\beta) = \int_{u \in T(y)} g(u - X\beta)\, du \tag{6}$$

where $g: \mathbb{R}^I \longrightarrow \mathbb{R}$ denotes the joint multivariate density function of the disturbances ϵ_i, i=1,..,I, in the latent variable model (1), and T(y) is the set of latent variables u obeying the nonlinear relation $\tau(\cdot)$ for a given observed dependent variable y, given by (3). Unless the joint density function g and the area of integration T(y) are particularly benign, the integral in (6) will not have a closed form, necessitating approximation solutions.

The following Section builds on recent progress in Monte Carlo integration techniques by Geweke (1989) and Hajivassiliou and McFadden (1990) and discusses the new simulation method to compute the multivariate integral $\ell(y,X;\, g,\beta)$ for models in the class of limited dependent variable models defined by (1)–(3). The method has four key features: it produces unbiased probability estimates; these are a continuous and differentiable function of the parameters of the model; they are bounded between 0 and 1; and the computational effort of the simulation method is nearly linear in the dimension of the probability integral, independent of the magnitudes of the true probabilities.

3. COMPUTING UNBIASED CHOICE PROBABILITIES

Estimating the parameters in (6) is a formidable task because it requires, in the most general case, an evaluation of an I dimensional integral for each observation and each iteration in the maximization process. The dimension I can become very large, particularly

so in the case of mutliperiod-multinomial discrete choice models.

Closed-form solutions for (6) only exist in some instances, whenever the specification of the joint density founction g and the area of integration T(y) are particularly benign, thereby contributing to the popularity of such specifications. The most prominent examples are those in which g is derived from the family of generalized extreme-value (GEV) distributions, e.g., the cross-sectional multinomial logit (MNL) or nested multinomial logit (NMNL) models. Closed-form solutions also exist if multiperiod variants of these models are combined with a one-factor random effect structure in which the random effects are also extreme-value distributed (e.g., McFadden (1984)). But these GEV-type models have the disadvantage of requiring relatively rigid correlation structures, which must be assumed a priori. This is restrictive because the covariance matrix Ω of the ϵ in (1) can have up to $(I+1) \times I/2 - 1$ deep parameters (the upper triangle of Ω including the diagonal, minus one element to scale the parameter vector β). These are many more parameters than GEV-type models can carry. Moreover, the specification of Ω is not constrained by hierarchical structures as it is in NMNL models.[3]

Numerical integration of (6) is computationally infeasible since the number of operations increases exponentially with the dimension I. Approximation methods for the multivariate normal distribution, such as the Clark approximation (Clark, 1961; Daganzo, 1981) or its variant proposed by Langdon (1986), are tractable — their number of operations increases quadratically with I. They remain unsatisfactory, however, since their relatively large bias depends on the (unknown) covariance structure of the latent variable and cannot be controlled by increasing the number of observations.[4]

Simulation methods have been introduced to discrete choice models by Lerman and Manski (1981). The probability (6) is simulated by drawing pseudo-random realizations from the underlying error process (1). The main appeal of simulation methods is that, for

---

[3] Boersch—Supan (1989) investigates the estimates and error structure implied by NMNL models if the true model is multinomial probit.

[4] Horowitz, Sparmonn, and Daganzo (1984) report on the accuracy of the Clark approximation.

any given number of draws, the number of operations increases essentially only linearly with the dimension I, thereby permitting the specification of quite general error structures. We proceed to discuss alternative simulators for the choice probabilities.

3.1 THE FREQUENCY SIMULATOR

The most straightforward simulation method is to simulate the probability (6) by an observed frequency:

$$(7) \qquad \frac{1}{R}\sum_{r=1}^{R} 1(a \le A \cdot u_r \le b) \qquad \text{with } u_r \text{ drawn from } N(X\beta,\Omega),$$

for R draws or replications. The indicator function 1(A) takes the value 1 when event A is true and 0 otherwise.

Lerman and Manski used this simulator in their simulated maximum likelihood approach. It has two drawbacks. If the true probability is smaller than one, then for any given finite number of draws, R, the probability of a zero frequency count in (7) is strictly positive. Hence, the frequency simulator yields consistent likelihood estimates only if both sample size and number of draws per observation goes to infinity. This is reflected in the computational effort: a very large number of draws is required to obtain reasonably accurate estimates of small probabilities, which results in unacceptably long computer runs (Lerman and Manski (1981)). The second drawback is the discontinuity of the frequency simulator. A small change in the parameter values will change the integer-valued frequency count only in discrete steps. This discontinuity impedes numerical optimization and requires a considerably more complex asymptotic theory to prove consistency and asymptotic normality of simulation estimators.

McFadden (1989) and Pakes and Pollard (1989) simultaneously developed the method of simulated moments (MSM) that circumvents the consistency problem. Rather than simulating the likelihood (6) directly and controlling the simulation error in each

observation, their main idea is to simulate generalized moments and to rely on the law of large numbers to control the simulation error across a large number of observations. They also provide the appropriate asymptotic theory for simulation estimators. McFadden (1989) also suggests several alternative simulators with better numerical properties than the discontinuous frequency simulator.

3.2 THE STERN SIMULATOR

One of the more promising multivariate probability simulators has been proposed by Stern (1987) for the multinomial probit model. This simulator is unbiased and smooth in the model parameters β and Ω.

Smoothness is a key feature of simulators because it keeps the number of random draws small. The intuition is as follows. With only one draw, a discontinuous probability simulator can map the parameters in the model (1)–(3) to only two outcomes, zero or one, while only a large number of draws addresses all values in the [0,1] interval. Loosely speaking, a large number of draws serves to smooth the discontinuity of each single draw. In contrast, each draw of a continuous probability simulator can map the parameters of the model to any value in the interior of the [0,1] interval, and no additional smoothing is necessary. Smoothness also permits the application of standard optimization routines and standard asymptotic theory. Since the smooth probability estimates are bounded away from 0 and 1, no numerical problems emerge when the probabilities are inserted in expressions such as log(p), log(1–p), 1/p, or 1/(1–p). This is in marked contrast to the numerical problems afflicting frequency simulators like (7), since frequency simulators take the values 0 and 1 with positive probability.

As long as the underlying error distribution is multivariate normal, Stern's simulator can be generalized to the entire class of limited dependent variable models defined by (1)–(3) as follows. Let the latent variable u in (1) be normally distributed $N(X\beta,\Omega)$. Define $M = A\Omega A'$ where A is defined by the inequality (3). Stern's basic idea

is to divide Au into two components, one of them composed of independent random variables and the other with a covariance matrix that is as small as possible (in the positive definite sense):

$$(8) \qquad Au = w_1 + w_2 , \qquad w_1 \sim N(0,D) , \qquad D = \mathrm{diag}(d_1,\ldots,d_I) ,$$
$$w_2 \sim N(X\beta, M-D) .$$

Stern proposes an eigenvalue-based algorithm to compute an optimal choice of D.

With this decomposition, the likelihood contribution (6) can be written as

$$(9) \qquad \begin{aligned} \ell(y,X;\beta,\Omega) &= \mathrm{Prob}(a \le A\cdot u \le b) \\ &= \mathrm{Prob}(a-w_1 \le w_2 \le b-w_1) \\ &= \int \mathrm{Prob}(a-w_2 \le w_1 \le b-w_2 \mid w_2)\cdot f(w_2)\, dw_2 \\ &= \int H(w_2)\cdot f(w_2)\, dw_2 , \end{aligned}$$

where $f(w_2)$ is the multivariate density of w_2,

$$(10) \qquad H(w_2) = \prod_{i=1}^{I} \{ \Phi((b_i - X_i\beta)/d_i) - \Phi((a_i - X_i\beta)/d_i) \} ,$$

and $\Phi(\cdot)$ denotes the standard Gaussian c.d.f.[5] Hence, $\ell(y,X;\beta,\Omega)$ can be simulated by

$$(11) \qquad \frac{1}{R}\sum_{r=1}^{R} H(w_{2r}) \qquad \text{with } w_{2r} \text{ drawn from } N(X\beta, M-D),$$

for R replications.

Stern's method employs the additivity property of the normal distribution. Its accuracy, as measured by the variance of the simulated probabilities, depends crucially on the relative magnitudes of w_1 and w_2 in (8). If M is diagonal, w_2=0 and the Stern method is exact. When the correlation among the u_i's becomes stronger, however, the elements of

[5] Due to normality, the density function g is completely characterized by the covariance matrix Ω.

D become smaller and less can be pulled into the independent part w_1. Indeed, the variance of the simulated ℓ is quite large if Ω has large off-diagonal elements.[6]

3.3 THE SMOOTH RECURSIVE CONDITIONING (SRC) SIMULATOR

We now define a multivariate probability simulator that shares the unbiasedness and the smoothness of the Stern simulator, but that produces a substantially smaller variance of the simulated probabilities particularly for the interesting case of highly interdependent u_i. In contrast to the Stern method, it can also be generalized to some non-normal distributions.

The method builds on recent work by Geweke and by Hajivassiliou and McFadden. Geweke (1989) has developed an algorithm to compute random variates from a multivariate, truncated normal distribution by reducing the multidimensional frequency simulator to a recursive sequence of univariate acceptance-rejection draws.[7] His recursive conditioning method has been modified by Hajivassiliou and McFadden (1990) to construct a smooth simulator of the score of the likelihood contribution (6), which yields a method of moments estimator in which the optimal instrument is X, the matrix of explanatory variables.[8] Unfortunately, Geweke's recursive conditioning method does not produce *draws* of the correct multivariate truncated normal distribution, thereby hampering a direct simulation of the scores.

We will show, however, that the *probabilities* generated by the Geweke algorithm are unbiased. In the sequel, we will examine the smooth version of the Geweke algorithm, dubbed the smooth recursive conditioning (SRC) algorithm, when employed in the simulated likelihood approach of Lerman and Manski.

---

[6] See Section 4.

[7] Keane (1990) independently proposed a variant of this method for the special case of a multiperiod (panel-data) binary probit model.

[8] This was first pointed out by Paul Ruud.

Geweke's recursive conditioning method works as follows for normally distributed u. Let L be the lower diagonal Choleski factor of $M=A\Omega A'$,

$$LL'=A\Omega A' \quad (12)$$

Instead of drawing from the original distribution of the latent variable,

$$u \sim N(X\beta,\Omega) \quad \text{s.t.} \quad a \leq A\cdot u \leq b\,, \quad (13)$$

we draw a random vector

$$e \sim N(0,I) \quad \text{s.t.} \quad a^* \equiv a - AX\beta \leq Le \leq b^* \equiv b - AX\beta\,. \quad (14)$$

The restrictions are recursive due to the triangular structure of L:

$$e_1 \sim N(0,1) \quad \text{s.t.} \quad a_1^* \leq l_{11}\cdot e_1 \leq b_1^* \quad (15a)$$

$$\Leftrightarrow a_1^*/l_{11} \leq e_1 \leq b_1^*/l_{11}$$

$$e_2 \sim N(0,1) \quad \text{s.t.} \quad a_2^* \leq l_{21}\cdot e_1 + l_{22}\cdot e_2 \leq b_2^* \quad (15b)$$

$$\Leftrightarrow (a_2^* - l_{21}\cdot e_1)/l_{22} \leq e_2 \leq (b_2^* - l_{21}\cdot e_1)/l_{22}$$

etc.

The e_i's can therefore be drawn sequentially by a univariate truncated simulator. Finally, the desired random vector u is defined by

$$u = X\beta + A^{-1}Le\,. \quad (16)$$

The random vector u has covariance $A^{-1}LL'A^{-1\prime} = A^{-1}A\Omega A'A^{-1\prime} = \Omega$ and is subject to $A^{-1}a \leq A^{-1}Le \leq A^{-1}b \Leftrightarrow a \leq Au \leq b$.

The draws of u are generally biased. This is evident from a simple two-dimensional example. Suppose $b_1=b_2=\infty$ as is the case in the probit model, and $l_{21}>0$, corresponding to a positive correlation between u_1 and u_2. Draws of e_1 according to the inequality in (15a) will ignore the constraint in (15b), hence will be too large on average. Given an e_1 too

large, e_2, obeying the second constraint (15b), will be too small on average.

In contrast to this, the likelihood contribution (6), i.e., the probability of $a \le A \cdot u \le b$, is correctly simulated by the probability of $a \le L \cdot e \le b$, which in turn is the product of the probabilities that each e_i falls in the respective intervals given by (15):

$$
\begin{aligned}
(17)\quad & \ell(y,X;\beta,\Omega) \\
& = \text{Prob}(a_1^*/l_{11} \le e_1 \le b_1^*/l_{11}) \\
& \cdot \text{Prob}((a_2^* - l_{21}\cdot e_1)/l_{22} \le e_2 \le (b_2^* - l_{21}\cdot e_1)/l_{22} \mid e_1) \\
& \cdots \\
& \cdot \text{Prob}((a_I^* - l_{I1}\cdot e_1 - \ldots - l_{Ii-1}\cdot e_{i-1})/l_{II} \le e_I \le (b_I^* - l_{I1}\cdot e_1 - \ldots - l_{Ii-1}\cdot e_{i-1})/l_{II} \mid e_1,\ldots,e_{i-1}) \\
& = Q_1 \cdot Q_2(e_1) \cdot Q_3(e_1,e_2) \cdot \ldots \cdot Q_I(e_1,\ldots,e_{I-1}) \,.
\end{aligned}
$$

This can be approximated by the simulator $\tilde{\ell}(y,X;\beta,\Omega;R)$

$$
(18)\qquad \frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{I} Q_i(e_{1r},\ldots,e_{i-1,r}) \qquad \text{with } e_{ir} \text{ drawn from truncated } N(0,1),
$$

where R denotes the number of replications.

Lemma:

The simulator $\tilde{\ell}(y,X;\beta,\Omega;R)$ defined by (18) is an unbiased estimator of $\ell(y,X;g,\beta)$.

Proof:

It is sufficient to show the Lemma for R=1. The expected value of $\tilde{\ell}$ is

$$E\tilde{\ell} = \int \tilde{\ell}(e)\, f(e)\, de$$

where f(e) denotes the density that generates the (biased) sequential truncated draws in (15):

$$f(e) = \prod_{i=1}^{I} \phi(e_i) \,/\, Q_i(e_1,\ldots,e_{i-1}) \qquad \text{on the set } T(y) = \{a^* \le L\cdot e \le b^*\}\,,$$

$$= 0 \qquad \text{elsewhere.}$$

By the definition of $\tilde{\ell}$, (18),

$$E\tilde{\ell} = \int_{-\infty}^{\infty} (\prod_{i=1}^{I} Q_i)\cdot(\prod_{i=1}^{I} \phi(e_i)/Q_i)\, de_1\ldots de_I$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{I} \phi(e_i)\, de = \int_{T(y)} \prod_{i=1}^{I} \phi(e_i)\, de = \text{Prob}(a^* \le Le \le b^*) = \ell(y,X;\, \beta,\Omega)\,. \quad \square$$

The univariate truncated normal variates e_i in (15) can be drawn smoothly according to a straightforward application of the integral transform theorem as proposed by Hajivassiliou and McFadden (1990). Let X be distributed according to the univariate uniform distribution on [0,1]. Then

$$(19) \qquad Z \equiv G^{-1}(X) = \Phi^{-1}[\,(\Phi(b)-\Phi(a))\cdot X + \Phi(a)\,]$$

is distributed N(0,1) s.t. $a \le Z \le b$, since the corresponding c.d.f. is

$$(20) \qquad G(z) = \frac{\Phi(z) - \Phi(a)}{\Phi(b) - \Phi(a)}\,,$$

where Φ denotes the univariate normal cumulative distribution function. Note that Z is a continuous and differentiable function of the parameters a and b.

The combination of the Geweke recursive conditioning method, the above Lemma, and the smooth univariate truncated variate generation algorithm produces an unbiased multivariate probability simulator for (6) that is smooth, i.e., a continuous and differentiable function of the model parameters β and Ω. This smoothness is essential for computational efficiency as was pointed out at the beginning of the previous subsection.

Apart from an initial Choleski decomposition and several matrix multiplications, most computational effort is in drawing the univariate truncated normal variates according

to the steps in (15).[9] This effort is linear in I, the dimension of the probability integral (6). Moreover, if $M=A\Omega A'$ is a diagonal matrix, L is also diagonal. In this case, $Q_i(e_1,...,e_{i-1})$ is independent of $e_1,...,e_{i-1}$, and the simulator (18) is exact. This is also true for the Stern simulator (11). The simulator (18), however, avoids the main disadvantage of (11), namely its poor performance for highly correlated u_i.

We finally note, that the simulator (18) is applicable also for some non-normal distributions, as long as the univariate draws from the conditional one-dimensional distributions correspond to the multivariate distribution of ϵ in (1). This useful property is not shared by the Stern simulator which relies on the very special additivity property of the normal distribution.

4. MONTE CARLO RESULTS: DISTRIBUTION OF PROBABILITIES

To evaluate the relative accuracy of the three simulators, we take four examples from Stern (1987). They exemplify choice probabilities generated from a multinomial probit model with five choices. In this case, the latent variable u_i in (1) represents the utility of choice i. Because only relative utilities matter in a choice among discrete alternatives, the actual dimension of the problem is I=4, and instead of specifying $X\beta$ and Ω we assume different values for $\Delta X\beta$ and $\Delta\Omega$, defined as first differences with respect to, say, the first alternative.

Table 1 presents the distribution of simulated choice probabilities for the three simulators defined by (7), the frequency simulator; (11), the simulator proposed by Stern (1987) with his optimal choice of D; and (18), the SRC simulator proposed in Subsection 3.3. Each distribution is based on 10000 draws: 100 realizations of each probability estimate, where in turn each estimate is based on R=100, the number of replications.

Example 1 features mild correlation and a very small choice probability. The

[9] For reliable results, it is crucial to compute the cumulative normal distribution function and its inverse with high accuracy.

frequency simulator has a very large standard deviation, about 75 percent of the mean. The Stern simulator (11) performs better, with a standard deviation about 2.2 times smaller than the frequency simulator (7). However, the SRC simulator (18) is substantially more accurate. Its standard deviation is more than eleven times smaller than in the Stern case, and 25 times smaller than the frequency simulator.

Example 2 has a slightly higher correlation and is less diagonally dominant than Example 1. The standard deviation of the Stern simulator is now only 1.8 times that of the frequency simulator, while the SRC simulator has an eight times higher accuracy.

Example 3 features some very large correlation coefficients. In this case, the relative accuracy of the Stern method is only slightly better than the frequency simulator: its standard deviation is only 1.2 times smaller. However, the SRC simulator out-performs both the frequency and Stern's simulator with a six times smaller standard deviation than the Stern simulator.

Example 4 features a probability close to 0.5 and only mild correlation. In this benign case, the performance of the three simulators is very similar. Still, the SRC simulator has a standard deviation that is 2.8 times smaller than the Stern method and 3.4 times smaller than the frequency simulator.

Table 2 investigates the performance of the simulators (11) and (18) with different numbers of replications. We use example 4, the most unfavorable for the SRC simulator relative to the Stern simulator. The distributions are based on 100 realizations of the simulators, each realization based on R=1, 3, 10 and 100 replications.

The standard deviation of the simulators decreases sharply with the number of replications. But even with one single draw, the SRC simulator has a coefficient of variation of less than .27, while that of the Stern simulator is more than .71. The relative accuracy stays approximately constant: across all replication numbers, the standard deviation of the SRC simulator is at least 2.6 times smaller than that of the Stern simulator.

Finally, Table 3 investigates the performance of the SRC simulator at very small

probabilities. Even for probabilities very close to zero, few replications are sufficient to generate accurate probability estimates.

5. ESTIMATION PROCEDURE:

SMOOTH SIMULATED MAXIMUM LIKELIHOOD ESTIMATION (SSML)

Of central interest is the estimation of the parameters β and Ω from a sample of observations (y_n, X_n), n=1,...,N. The loglikelihood function corresponding to the multivariate choice probabilities (6) is

$$(21) \qquad \mathcal{L}(\beta,\Omega) = \sum_{n=1}^{N} \log \ell(y,X;\beta,\Omega) .$$

The simulated maximum likelihood (SML) approach replaces $\ell(y,X;\beta,\Omega)$ by a probability simulator $\tilde{\ell}$, e.g., defined by (7),(11), or (18). However, because of the nonlinearity of the logarithm

$$(22) \qquad E\tilde{\ell} = \ell \implies E(\log\tilde{\ell}) - \log(E\ell) \approx -\mathrm{Var}(\tilde{\ell})/2\ell^2 < 0 .$$

Hence, the SML approach is biased as long as $\tilde{\ell}$ has a positive variance, i.e., for any finite number of replications in case of the three simulators defined above. This is — in a less extreme form — the consistency problem that was mentioned in Subsection 3.1 when the frequency simulator was discussed. The bias goes to infinity for very small probabilities ℓ. In the discrete choice context, the likelihood contribution ℓ is the predicted choice probability of an observation. In a well specified model, this probability should be large, the bias therefore relatively small.

Because the bias depends crucially on the variance of the probability simulator, it is important to keep this variance small. As the results of the preceding section have shown, the SRC simulator (18) is for this reason a much more suitable simulator than the Stern simulator (11) or the frequency simulator (7).

The smoothness of the SRC simulator is also a key feature for a convenient

computation of the optimal parameter values of model (1)–(3). Because $\tilde{\ell}(y,X; \beta,\Omega; R)$ is a continuous and differentiable function of these parameters, the first order conditions of maximizing (21) with respect to β and Ω are well defined, and conventional numerical methods with a high speed of convergence such as one of the conjugate gradient methods or quadratic hillclimbing can be employed. This is in sharp contrast to the frequency simulator that generates a discontinuous objective function with the associated numerical problems.

Finally, we stress that the computational effort of the SSML approach increases almost linearly with the dimension of the latent variable vector in (1). This near linearity permits applications to discrete choice problems with large choice sets, to panel data models with a large number of panel waves, or a combination of the two.

6. MONTE CARLO RESULTS: DISTRIBUTION OF SSML ESTIMATES

In order to assess the bias that is generated by taking the logarithm in (21), we investigate the distribution of the parameter estimates in a simple three-alternative probit model. We created a 500 observation data set (y,X) specified by model (1) and (4) with a dependent choice variable y and a single explanatory variable X. The explanatory variable was realized as independently and normally distributed draws with mean [1,0,0] and standard deviation [2,2,2]. This large variation in the explanatory variable was chosen to generate a wide range of predicted choice probabilities.

Given the explanatory variables X, the dependent choice variable y was defined by

$$(23) \qquad y = \underset{j}{\operatorname{argmax}}(u_j \mid j=1,\ldots,3) ,$$

where the latent variable vector u was drawn from the multivariate normal distribution $N(X\beta,\Omega)$ with true parameters

$$(24) \qquad \beta = 1.0, \qquad \Omega = \begin{bmatrix} 2.0 & 0.7 & 0.0 \\ 0.7 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

Parameter estimation of the particular 500 observation sample using the Hausman-Wise (1978) method with high accuracy numerical integration yields

$$(25) \qquad \hat{\beta} = 0.925, \qquad \hat{\Omega} = \begin{bmatrix} 2.062 & 0.695 & 0.0 \\ 0.695 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

At these parameters, the predicted choice probabilities range from less than 0.01 to above 0.999.

Table 4 presents the distribution of β, estimated by inserting the SRC simulator into the loglikelihood function (22), for alternative numbers of replications. The bias in the parameter β is substantial for very small numbers of replications. However, the bias is below its standard deviation for R=10 and becomes negligible for R=20. This is in sharp contrast to the Lerman and Manski method. They report a minimum number of draws R_o=1000, and they used much larger numbers for the accurate simulation of probabilities close to zero and one.[10]

7. CONCLUSIONS

We revived the Lerman and Manski (1981) procedure of approximating the likelihood function by simulated choice probabilities by employing a new probability simulator that is bounded away from 0 and 1, is smooth in the parameters to be estimated, and has a small variance relative to the frequency simulator employed by Lerman and Manski and the simulator by Stern (1987).

Due to the small variance of the SRC simulator, the bias intrinsic to the simulated

---

[10] Lerman and Manski (1981), Footnote 10.

maximum likelihood approach stays small. In our Monte Carlo experiment, 20 replications were sufficient to produce a negligible bias. The experiment was designed to represent the case of a large variation in the predicted choice probabilities, particularly small predicted choice probabilities. Because the bias is largest for small predicted choice probabilities, a well specified model of real data should feature even less bias than our Monte Carlo example.

Due to the smoothness of the simulator, conventional and fast optimizers can be employed for the computation of maximum likelihood estimates. The method is feasible for discrete choice models with large choice sets and/or panel data applications with a large number of panel waves. Börsch-Supan et al. (1990) report on the application of the SML approach on a multinomial-multiperiod probit model with unobserved heterogeneity and autocorrelated errors. The class of models (1)–(3) includes many other limited dependent variable models such as tobit and switching regression models with their multiperiod generalizations.

REFERENCES

Börsch-Supan, A., V. Hajivassiliou, L.J. Kotlikoff, and J.N. Morris, 1990, Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors, NBER Working Paper.

Börsch-Supan, A., 1989, Recent Developments in Flexible Discrete Choice Models: Nested Logit Analysis versus Simulated Moments Probit Analysis. in: M. M. Fischer et. al. (eds.), Behavioral Modelling of Spatial Decisions and Processes, Amsterdam: North Holland.

Chamberlain, G., 1984, Panel Data, in: Z. Griliches and M. Intriligator (eds.), Handbook of Econometrics, vol. 2, Amsterdam: North Holland.

Clark, C., 1961, The Greatest of a Finite Set of Random Variables, Operations Research 9, pp. 145–162.

Daganzo, C., 1981, Multinomial Probit, New York: Academic Press.

Finney, D., 1971, Probit Analysis. Cambridge University Press, New York .

Geweke, J., 1989, Efficient Simulation from the Multivariate Normal Distribution Subject to Linear Inequality Constraints and the Evaluation of Constraint Probabilities, mimeo, Duke University.

Hajivassiliou, V., 1990, Smooth Simulation Estimation of Panel Data LDV Models, Discussion Paper, Yale University.

Hajivassiliou, V. and D. McFadden, 1990, The Method of Simulated Scores for the Estimation of LDV Models: With An Application to the Problem of External Debt Crises, Discussion Paper, Yale University.

Hausman, J.A. and D.A. Wise, 1978, Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences, Econometrica 47:403–426.

Horowitz, J., J. Sparmonn, and C. Daganzo, 1984, An Investigation of the Accuracy of the Clark Approximation for the Multinomial Probit Model, Transportation Science 16:382–401.

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

Langdon, M.G., 1984, Methods of Determining Choice Probability in Utility Maximizing Multiple Alternative Models, Transportation Research B 18.

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

McFadden, D., 1984, Qualitative Response Models, in: Z. Griliches and M. Intriligator (eds.), Handbook of Econometrics, Amsterdam: North Holland.

McFadden, D., 1989, A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration, Econometrica 57.

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

Stern, S., 1987, A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models, mimeo, University of Virginia.

TABLE 1

Distribution of Probability Simulators: (Alternative Specifications of $\Delta X\beta$ and $\Delta\Omega$)

Example 1:

$\Delta X\beta =$ [-1.0,-.75,-.5,-.2]

$$\Delta\Omega = \begin{bmatrix} 1.0 & & & \\ .2 & 1.0 & & \\ .3 & .4 & 1.0 & \\ .1 & .3 & .5 & 1.0 \end{bmatrix}$$

Percentile	Simulator		
	FREQ	STERN	SRC
1%	0.00	0.00894	0.02290
5%	0.00	0.01226	0.02297
10%	0.00	0.01417	0.02327
25%	0.01	0.01837	0.02364
33%	0.01	0.01957	0.02377
Median	0.02	0.02351	0.02398
66%	0.03	0.02744	0.02432
75%	0.03	0.02882	0.02453
90%	0.05	0.03268	0.02499
95%	0.06	0.03598	0.02540
99%	0.09	0.04923	0.02599
Mean	0.0228	0.02378	0.02409
Std.Dev.	0.0171	0.00773	0.00068

Example 2:

$\Delta X\beta =$ [0., 0., 0., 0.]

$$\Delta\Omega = \begin{bmatrix} 1.0 & & & \\ .2 & 1.0 & & \\ .2 & .4 & 1.0 & \\ .2 & .4 & .6 & 1.0 \end{bmatrix}$$

Percentile	Simulator		
	FREQ	STERN	SRC
1%	0.05	0.10447	0.14243
5%	0.10	0.12312	0.14331
10%	0.12	0.12625	0.14555
25%	0.13	0.13959	0.14722
33%	0.14	0.14488	0.14827
Median	0.14	0.15286	0.14956
66%	0.17	0.15804	0.15187
75%	0.18	0.16433	0.15347
90%	0.20	0.17661	0.15646
95%	0.21	0.18357	0.15877
99%	0.24	0.21171	0.16221
Mean	0.1506	0.15515	0.15037
Std.Dev.	0.0355	0.01978	0.00444

NOTE: FREQ is the frequency simulator defined by equation (7), STERN is the simulator (11) proposed by Stern (1987), and SRC is the smooth simulator defined by equations (17)-(19).

TABLE 1 (continued)

Distribution of Probability Simulators:
(Alternative Specifications of $\Delta X\beta$ and $\Delta\Omega$)

Example 3:

$$\Delta X\beta = [1.0,\ 1.0,\ 1.0,\ 1.0]$$

$$\Delta\Omega = \begin{bmatrix} 1.0 & & & \\ .9 & 1.0 & & \\ .0 & .0 & 1.0 & \\ .0 & .0 & .95 & 1.0 \end{bmatrix}$$

Percentile	Simulator		
	FREQ	STERN	SRC
1%	0.52	0.54609	0.62888
5%	0.55	0.56634	0.63375
10%	0.57	0.57878	0.63640
25%	0.62	0.61931	0.64394
33%	0.63	0.63253	0.64513
Median	0.65	0.64899	0.64747
66%	0.67	0.66387	0.65216
75%	0.68	0.68304	0.65275
90%	0.72	0.70928	0.65725
95%	0.73	0.72285	0.65974
99%	0.75	0.73334	0.66536
Mean	0.6463	0.64706	0.64773
Std.Dev.	0.0536	0.04573	0.00773

Example 4:

$$\Delta X\beta = [1.5,\ .75,\ .5,\ .75]$$

$$\Delta\Omega = \begin{bmatrix} 1.0 & & & \\ .5 & 1.0 & & \\ .2 & .5 & 1.0 & \\ .1 & .2 & .5 & 1.0 \end{bmatrix}$$

Percentile	Simulator		
	FREQ	STERN	SRC
1%	0.41	0.41884	0.47243
5%	0.42	0.44022	0.47612
10%	0.43	0.44936	0.48010
25%	0.47	0.47882	0.48758
33%	0.49	0.48395	0.49021
Median	0.50	0.49919	0.49611
66%	0.52	0.51789	0.50192
75%	0.53	0.52536	0.50694
90%	0.56	0.55565	0.51593
95%	0.59	0.56592	0.52341
99%	0.63	0.59952	0.53192
Mean	0.5023	0.50147	0.49716
Std.Dev.	0.0481	0.03940	0.01394

NOTE: FREQ is the frequency simulator defined by equation (7), STERN is the simulator (11) proposed by Stern (1987), and SRC is the smooth simulator defined by equations (17)-(19).

TABLE 2

Distribution of Probability Simulators: (Example 4, Alternative Numbers of Replications)

Number of Replications: R=1

Percentile	STERN	SRC
1%	0.00005	0.25053
5%	0.00268	0.30877
10%	0.01432	0.33497
25%	0.13740	0.39466
33%	0.25300	0.42191
Median	0.48756	0.48385
66%	0.71905	0.54611
75%	0.83464	0.59162
90%	0.97089	0.68864
95%	0.98954	0.73999
99%	0.99935	0.82959
Mean	0.49074	0.49833
Std.Dev.	0.35042	0.13387

Number of Replications: R=3

Percentile	STERN	SRC
1%	0.06723	0.33782
5%	0.16197	0.37485
10%	0.23206	0.40125
25%	0.35429	0.44558
33%	0.39944	0.46212
Median	0.50775	0.49755
66%	0.60033	0.52762
75%	0.63660	0.54490
90%	0.77384	0.59608
95%	0.83626	0.62534
99%	0.93800	0.67204
Mean	0.50051	0.49733
Std.Dev.	0.20397	0.07477

Number of Replications: R=10

Percentile	STERN	SRC
1%	0.24452	0.40795
5%	0.32358	0.43241
10%	0.35651	0.44397
25%	0.42664	0.46831
33%	0.44939	0.47662
Median	0.49934	0.49706
66%	0.54535	0.51461
75%	0.57498	0.52474
90%	0.63815	0.54879
95%	0.68004	0.56590
99%	0.75349	0.61200
Mean	0.49901	0.49716
Std.Dev.	0.10945	0.04152

Number of Replications: R=100

Percentile	STERN	SRC
1%	0.41710	0.46144
5%	0.43796	0.47241
10%	0.45230	0.47816
25%	0.47190	0.48690
33%	0.48068	0.48993
Median	0.49631	0.49530
66%	0.50949	0.50132
75%	0.51849	0.50530
90%	0.54024	0.51299
95%	0.55198	0.51712
99%	0.58402	0.52711
Mean	0.49571	0.49560
Std.Dev.	0.03513	0.01362

NOTE: STERN is the simulator (11) proposed by Stern (1987), SRC is the smooth simulator defined by equations (17)-(19).

TABLE 3

Accuracy of SRC Simulator at Small Probabilities

$\Delta X\beta = [-x,-x,0]$, $\Delta\Omega = \begin{bmatrix} 3.0 & & \\ 0.7 & 2.0 & \\ 0.5 & 0.3 & 1.0 \end{bmatrix}$

x	Numerical Integration	Continuous Recursive Probability Simulator					
		R=2	R=5	R=10	R=20	R=50	R=100
0.0	0.18838e+00	0.1893e+00 (0.0329e+00)	0.1898e+00 (0.0215e+00)	0.1902e+00 (0.0145e+00)	0.1901e+00 (0.0103e+00)	0.1891e+00 (0.0057e+00)	0.1888e+00 (0.0038e+00)
1.0	0.69869e-01	0.7042e-01 (0.1340e-01)	0.7056e-01 (0.0864e-01)	0.7062e-01 (0.0572e-01)	0.7055e-01 (0.0402e-01)	0.7018e-01 (0.0226e-01)	0.7003e-01 (0.0148e-01)
2.0	0.15701e-01	0.1587e-01 (0.0333e-01)	0.1589e-01 (0.0212e-01)	0.1588e-01 (0.0138e-01)	0.1586e-01 (0.0096e-01)	0.1578e-01 (0.0054e-01)	0.1574e-01 (0.0036e-01)
3.0	0.20252e-02	0.2052e-02 (0.0469e-02)	0.2053e-02 (0.0296e-02)	0.2051e-02 (0.0190e-02)	0.2046e-02 (0.0131e-02)	0.2035e-02 (0.0075e-02)	0.2030e-02 (0.0050e-02)
4.0	0.14505e-03	0.1473e-03 (0.0362e-03)	0.1473e-03 (0.0227e-03)	0.1470e-03 (0.0144e-03)	0.1465e-03 (0.0099e-03)	0.1458e-03 (0.0057e-03)	0.1454e-03 (0.0038e-03)
5.0	0.56534e-05	0.5749e-05 (0.1503e-05)	0.5748e-05 (0.0936e-05)	0.5730e-05 (0.0593e-05)	0.5711e-05 (0.0402e-05)	0.5682e-05 (0.0236e-05)	0.5664e-05 (0.0157e-05)
6.0	0.11847e-06	0.1206e-06 (0.0331e-06)	0.1206e-06 (0.0205e-06)	0.1201e-06 (0.0129e-06)	0.1197e-06 (0.0087e-06)	0.1191e-06 (0.0051e-06)	0.1187e-06 (0.0034e-06)
7.0	0.13251e-08	0.1351e-08 (0.0387e-08)	0.1349e-08 (0.0239e-08)	0.1344e-08 (0.0150e-08)	0.1338e-08 (0.0100e-08)	0.1332e-08 (0.0059e-08)	0.1327e-08 (0.0040e-08)

TABLE 4

Distribution of SSML Estimators: (Alternative Numbers of Replications)

Monte Carlo Data Set [500 Observations]: $\beta = 0.925$, $\Omega = \begin{bmatrix} 2.0 & 0.7 & 0.0 \\ 0.7 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$

Distribution of β [Based on 100 Realizations]:

Percentiles	R=1	R=3	R=6	R=10	R=20
1%	0.66167	0.76337	0.83462	0.86093	0.89159
5%	0.68356	0.78429	0.85367	0.86630	0.89466
10%	0.69626	0.80365	0.85692	0.87005	0.89841
25%	0.72614	0.81759	0.86567	0.88251	0.91280
33%	0.73202	0.82828	0.86841	0.89277	0.91642
Median	0.74736	0.83822	0.87698	0.90196	0.92434
66%	0.77041	0.84946	0.89082	0.91172	0.93363
75%	0.77597	0.86050	0.90186	0.91707	0.93605
90%	0.79580	0.87786	0.91930	0.93883	0.94622
95%	0.80411	0.90133	0.92769	0.94137	0.96520
99%	0.83851	0.92168	0.95149	0.96841	0.97846
Mean	0.74835	0.83921	0.88299	0.90235	0.92497
Std.Dev.	0.03752	0.03210	0.02494	0.02414	0.01914