

# Simulation-Based Econometrics

Paul Ruud

Report by Charles Bellemare, Tilburg University

## 1 Introduction

Simulation based methods replaces intractable numerical optimization problems with simulated counterparts which can be successfully optimized. Say you need an object  $\mu(\theta; Z_n)$  where  $\theta \in \mathbf{R}^k$  is a vector of unknown parameters and  $Z_n$  is a vector of observables for unit  $n$ . Typically,  $\mu(\theta; Z_n)$  is a probability or an expectation. We are interested in the case where  $\mu(\theta; Z_n)$  has no analytical solution given that it's underlying computation requires the evaluation of multiple integrals. Limited Dependent Variable models (LDVs) and models with unobserved heterogeneity in the underlying parameters are the most frequent cases in microeconometrics where multiple integrals appear. One way to go about computing  $\mu(\theta; Z_n)$  would be to use a Riemann integral-type of approach known as numerical quadrature. The accuracy of quadratures for high-dimension integrals as been shown to be at best poor, creating instability in the optimization procedure. Face with this, a microeconometrician can either reduce the complexity of the model in such a way that the dimension of the integral lowers. However, when doing so, one is overturning his beliefs of what the true data-generating process of the data is. The alternative way is to use simulation.

We mentioned that  $\mu(\theta; Z_n)$  was a bunch of integrals

$$\mu(\theta; Z_n) = \int \int \dots \int g(\omega_1, \omega_2, \dots, \omega_s; \theta, y_n) d\omega_1 d\omega_2 \dots d\omega_s \quad (1)$$

The idea of simulation is to draw from what we integrate over. In the above case, this means that we draw from the joint distribution of  $\boldsymbol{\omega} = \{\omega_1, \omega_2, \dots, \omega_s\}$  and plug the draws in the function  $g(\cdot)$ . Formally, say we draw  $R$  times from the distribution of  $\boldsymbol{\omega}$  and stock these draws in the following  $R \times S$  matrix

$$\Psi_n = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1s} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{R1} & \omega_{R2} & \cdots & \omega_{RS} \end{bmatrix}$$

Each row of  $\Psi_n$  represents a draw from the joint distribution of  $\omega$ . If we denote  $\Psi_{ni}$  as the  $i$ th row of  $\Psi_n$ , then a simulator  $\tilde{\mu}_i(\theta; y_n, \Psi_{ni})$  of  $\mu(\theta; Z_n)$  is defined as

$$\tilde{\mu}_i(\theta; y_n, \Psi_{ni}) = g(\omega_{i1}, \omega_{i2}, \dots, \omega_{is}; \theta, y_n)$$

where we simply replaced the sequence  $\omega = \{\omega_1, \omega_2, \dots, \omega_s\}$  in  $g(\cdot)$  with one of the  $R$  simulated sequence  $\Psi_{ni} = \{\omega_{i1}, \omega_{i2}, \dots, \omega_{is}\}$ . We are not yet finished. Notice that integrals appear in (1). The process of integrating a function over continuously distributed variables has its counterpart in discrete averaging. That is, a simulation of the integrals in (1) amounts to averaging over all our  $R$  random draws collected in  $\Psi_n$

$$\begin{aligned} & \int \int \dots \int g(\omega_1, \omega_2, \dots, \omega_s; \theta, y_n) d\omega_1 d\omega_2 \dots d\omega_s \\ \approx & \frac{1}{R} \sum_{i=1}^R g(\omega_{i1}, \omega_{i2}, \dots, \omega_{is}; \theta, y_n) = \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) = \tilde{\mu}(\theta; Z_n, \Psi_n) \end{aligned}$$

Naturally, the more draws we make, the better the approximation of the integrals is likely to be. In fact, we will see that integrals can be approximated to any desired level of accuracy given that we are willing to increase the number of draws<sup>1</sup>.

We have given an intuition on how to simulate some arbitrary function  $\mu(\theta; Z_n)$  but we still have to see where these functions are actually used in econometrics. Two of the most popular inference methods are the Maximum Likelihood (ML) and the Method of Moments (MOM) estimators. Both these approaches benefit from simulated parts of functions of the kind in (1). When necessary, we will divided the sequence  $\{Z_n\}_{n=1}^N \equiv \{(y_n, x_n)\}_{n=1}^N$  where  $y_n$  will refer to a vector of endogenous variables and  $x_n$  as a vector of explanatory variables.

**Example 1 (ML)** *In the ML framework,  $\mu(\theta; Z_n)$  are probabilities such that for a random sample  $y = \{y_1, y_2, \dots, y_N\}$ , we seek to maximize the likelihood of the sample*

$$Q_N^{ML}(\theta) = \sum_{n=1}^N \log \mu(\theta; Z_n)$$

*Under regularity conditions on  $Q_N^{ML}(\theta)$ ,*

$$\hat{\theta}_{ML} \equiv \arg \max_{\theta} \sum_{n=1}^N \log \mu(\theta; Z_n) \xrightarrow{p} \theta_0$$

---

<sup>1</sup>This statement is somewhat strong. Recently, attention has been devoted to variance reduction technics that tries to improve the coverage of the draws. This has the advantage that for a small number of draws, more accurate approximations can be obtained. Antithetics and Halton draws are some examples of these technics. We do not go into this field of research but the interested reader can consult Train (2002).

**Example 2 (MOM)** *In the MOM framework,  $\mu(\theta; \mathbf{z}_n)$  is an expectation such that  $E(g(\mu(\theta; \mathbf{x}_n), \mathbf{y})) = 0$  where  $g(\cdot)$  is a vector of moments;  $g(\mu(\theta; \mathbf{x}_n), \mathbf{y}) \in \mathbb{R}^d$  where  $d \geq k$ , which simply says that there may be more moments than there are parameters to estimate. This implies that  $E(g(\mu(\theta; \mathbf{x}_n), \mathbf{y}))$  may not be made exactly equal to 0. In this case, MOM finds a value of  $\theta$  that brings some squared Euclidean norm of the sample moments close to zero*

$$Q_N^{MOM}(\theta) = \sum_{n=1}^N g(\mu(\theta; \mathbf{x}_n), \mathbf{y}) \Omega(\theta) g(\mu(\theta; \mathbf{x}_n), \mathbf{y})'$$

where  $\Omega(\theta)$  is any positive semi-definite matrix that depends on  $\theta$ . Under regularity conditions on  $Q_N^{MOM}(\theta)$ ,

$$\hat{\theta}_{MOM} \equiv \arg \max_{\theta} \sum_{n=1}^N g(\mu(\theta; \mathbf{x}_n), \mathbf{y}) \Omega(\theta) g(\mu(\theta; \mathbf{x}_n), \mathbf{y})' \xrightarrow{p} \theta_0$$

It is important to realize that the two extremum estimators presented above use all  $N$  observations. The simulation technic actually draws for each unit  $n$  a matrix  $\Psi$ . This is done simply because  $Q_N^{ML}(\theta)$  and  $Q_N^{MOM}(\theta)$  are functions of all units. This means we must simulate  $N$  matrix  $\Psi$  from the distribution of  $\omega$ .

## 1.1 Linking Simulation with Extremum Estimation

We have seen that simulation of difficult objects  $\mu(\theta; \mathbf{z}_n)$  could be achieved via simulation. When these objects are probabilities, we could simulate them and plug them in the objective function of  $Q_N^{ML}(\theta)$ . This method is known as Maximum Simulated Likelihood (SML). When  $\mu(\theta; \mathbf{z}_n)$  is an analytically intractable vector of moments, we can simulate them and plug them in  $Q_N^{MOM}(\theta)$ . This is known as the Method of Simulated Moments (MSM). Can this be so easy? Can we simply simulate and plug-in the objects at the right place and be sure that the solution of the optimization problem converges to the true unknown population vector  $\theta_0$ ? What happens to the properties of ML and MOM when  $\mu(\theta; \mathbf{z}_n)$  are replaced with  $\tilde{\mu}(\theta; \mathbf{z}_n, \Psi_n)$ ?

## 1.2 Laws of Large Numbers

In this subsection, we illustrate using familiar tools the potential problems that might occur when doing simulation. Recall that

$$\tilde{\mu}_i(\theta; \mathbf{z}_n, \Psi_{ni}) = g(\omega_{i1}, \omega_{i2}, \dots, \omega_{is}; \theta, y_n)$$

where  $\Psi_{ni} = \{\omega_{i1}, \omega_{i2}, \dots, \omega_{is}\} \in \Psi_n$ , the  $i$ 'th row of  $\Psi_n$ . It is important to realize that randomness in the simulator comes from  $\mathbf{z}_n$  and  $\Psi_{ni}$ . If we assume that  $\mathbf{z} = \{z_1, z_2, \dots, z_N\}$  is an i.i.d. sequence and that each draw (i.e. each  $\Psi_{ni}$ ) is independent and identically distributed, then

the two sources of randomness in the simulator are uncorrelated. More importantly, since they are both i.i.d, this makes the entire sequence  $\bar{\mu} = \{\tilde{\mu}_i(\theta; Z_n, \Psi_{ni}); \forall n, i\}$  itself an i.i.d sequence. This property is important since it will allow us to use familiar laws of large numbers to prove convergence in probability. This makes it in the interest of the econometrician to sample the  $\Psi$  randomly between observations. The following assumption summarizes the discussion

**Assumption 1**  $E(\tilde{\mu}_i(\theta; Z_n, \Psi_{ni})) = \mu(\theta; z)$  for all  $i, n$ .

**Linearity of the Simulator in the Objective Function** We have the following double sum

$$\frac{1}{N} \sum_{n=1}^N \left[ \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \right] \quad (2)$$

and we want this sum to converge in probability to  $E(\mu(\theta; Z_n))$ . We are used to see these sums and one of our reflexes could be to say that we must let both  $N$  and  $R$  tend to infinity for convergence in probability to occur. However, the linearity of the double sum allows us to make a stronger statement. We can fix  $R$  to some level, say  $\bar{R}$ . Then,

$$\frac{1}{N} \sum_{n=1}^N \left[ \frac{1}{\bar{R}} \sum_{i=1}^{\bar{R}} \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \right] = \frac{1}{N\bar{R}} \sum_{n=1}^N \sum_{i=1}^{\bar{R}} \tilde{\mu}_i(\theta; Z_n, \Psi_{ni})$$

due to the linearity in which the simulator enters. From our assumptions that all  $\tilde{\mu}_i(\theta; Z_n, \Psi_{ni})$  have the same mean and are independent of each other, it follows that the limit as  $N$  increases with fix  $R$

$$\lim_{N \rightarrow \infty} \frac{1}{N\bar{R}} \sum_{n=1}^N \sum_{i=1}^{\bar{R}} \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \xrightarrow{p} E(\tilde{\mu}(\theta; Z_n, \Psi_n))$$

To convince yourself of this fact, fix  $\bar{R} = 1$ , in which case

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \tilde{\mu}_1(\theta; Z_n, \Psi_{n1}) \xrightarrow{p} E(\tilde{\mu}(\theta; Z_n, \Psi_n)) = \mu(\theta; Z_n)$$

The main message here is that a linear double sum accelerates the convergence in probability for this i.i.d sequence.

**Non-Linearity of the Simulator in the Objective Function** Say we now have the following double sum

$$\frac{1}{N} \sum_{n=1}^N \log \left[ \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \right] \quad (3)$$

We would like it to converge to  $E \log E(\tilde{\mu}(\theta; Z_n, \Psi_n))$  since  $E(\tilde{\mu}(\theta; Z_n, \Psi_n)) = \mu(\theta; Z_n)$ . The log operator is not a linear operator. This implies that for fix  $R$ , the limiting behavior

of the average does not converge in probability to  $E \log E(\mu(\theta; Z_n))$  since the  $E$  and  $\log$  operators do not commute. Take the case where  $\bar{R} = 1$ . It is easy to see from (3) that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \log \tilde{\mu}_1(\theta; Z_n, \Psi_{n1}) \xrightarrow{p} E \log \tilde{\mu}_1(\theta; Z_n, \Psi_{n1}) \neq E \log E(\tilde{\mu}(\theta; Z_n, \Psi_n))$$

unless we also let  $R \rightarrow \infty$  since then,  $\tilde{\mu}(\theta; y_n, \Psi_n) = 1/R \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \xrightarrow{p} E(\tilde{\mu}(\theta; Z_n, \Psi_n))$ ..

## 2 Maximum Simulated Likelihood (MSL)

### 2.1 Consistency of MSL

One of the fundamental lemmas regarding consistency of extremum estimators is the following

Lemma 1 *Let*

1.  $\theta \in \Theta$ , a compact subset of  $\mathbb{R}^k$
  2.  $Q_0(\theta), Q_N(\theta)$  be continuous in  $\theta$
  3.  $\theta_0 \equiv \arg \max_{\theta} Q_0(\theta)$  be unique
  4.  $\hat{\theta}_N \equiv \arg \max_{\theta} Q_N(\theta)$  and
  5.  $Q_N(\theta) \rightarrow Q_0(\theta)$  in probability uniformly in  $\theta \in \Theta$  as  $N \rightarrow \infty$
- Then  $\hat{\theta}_N \xrightarrow{p} \theta_0$

Applying this lemma is not difficult once we have proven the uniform convergence of  $Q_N(\theta)$ . Proving uniform convergence is something more difficult but we assume here that this condition is satisfied for both the  $Q_N^{ML}(\theta)$  and  $Q_N^{MOM}(\theta)$  objective functions<sup>2</sup>.

Definition 1 *Given the log-likelihood function*

$$Q_N^{ML}(\theta) = \sum_{n=1}^N \log \mu(\theta; Z_n)$$

and the unbiased simulator  $\tilde{\mu}_i(\theta; Z_n, \Psi_{ni})$  so that

$$\mu(\theta; Z_n) = E(\tilde{\mu}_i(\theta; Z_n, \Psi_{ni}))$$

the maximum simulated likelihood (MSL) estimator is

$$\hat{\theta}_{MSL} \equiv \arg \max_{\theta} Q_N^{MSL}(\theta)$$

---

<sup>2</sup>Conditions for Uniform Convergence can be found in Ferguson (1996).

where

$$Q_N^{MSL}(\theta) = \sum_{n=1}^N \log \left[ \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}) \right]$$

for some given sequence  $\{\Psi_n\}_{n=1}^N$  where  $\Psi_n = \{\Psi_{n1}, \Psi_{n2}, \dots, \Psi_{nR}\}$

Two important points are worth mentioning. First, the MSL is computed conditional on the sequence of draws  $\{\Psi_n\}_{n=1}^N$ . This is imperative as we are trying to pin down a value of  $\theta$  satisfying the normal equations. If at each iteration one would perform new drawings, our objective function  $Q_N^{MSL}(\theta)$  would be unstable, shifting up and down each time we draw from  $\omega$ . We are then right to say that once  $\{\Psi_n\}_{n=1}^N$  has been drawn, it is added to the vector of observed characteristics  $\{Z_n\}_{n=1}^N$  and this constitutes the data set that we work with. Secondly, for both computational stability and asymptotic distribution theory, it is important that the simulations do not change with  $\theta$ , so that  $Q_N^{MSL}(\theta)$  is well-defined for maximization. To illustrate, suppose we have the following objective function

$$\frac{1}{N} \sum_{n=1}^N \log \left[ \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; Z_n, \Psi_{ni}(\theta)) \right] \quad (4)$$

where  $\Psi_{ni}(\theta)$ , our drawings, are functions of the unknown parameter vector  $\theta$ . If this were the case, then each time  $\theta$  is updated during the iterative procedure, we would have to change our drawings  $\Psi_{ni}$ . This goes back to our first argument, we must treat  $\{\Psi_n\}_{n=1}^N$  as fixed for stability and convergence of the estimator to occur.

In the previous section, we addressed the issue of the linearity of the expectation operator with respect to the simulator. It was shown that objective functions like (4) do not converge in probability to the population expectation when the number of draws  $R$  is fixed. This result was due to the fact that the estimator being a nonlinear function of the simulator is a biased simulation of the MLE.

**Proposition 1** *Let  $\mu(\theta; Z_n)$  be uniformly bounded away from zero for all  $\theta \in \Theta$ , a compact set. Assume typical regularity conditions sufficient for Lemma 1 to hold. Let  $\{\Psi_{ni}\}_{i=1}^R$  be i.i.d. sequence over the index  $i$ . The MSL estimator  $\hat{\theta}_{MSL}$  is consistent if  $R \rightarrow \infty$  as  $N \rightarrow \infty$ .*

**Proof.** See Hajivassiliou and Ruud (1994) p.2417. ■

**Explanation of Proposition 1** The previous result holds for sequences  $\{\Psi_{ni}\}_{i=1}^R$  that are not independent of  $n$ . This implies that it is possible to use a single sequence  $\{\Psi_{ni}\}_{i=1}^R$  and apply it to each unit  $n$  in our sample. The intuition behind the proof is the following. Point 5 of Lemma 1 stipulated that  $Q_N(\theta)$  must converge uniformly to  $Q_0(\theta)$  for  $\hat{\theta}_N \xrightarrow{p} \theta_0$ .

We know that under mild regularity conditions,  $Q_N^{ML}(\theta) \xrightarrow{\text{uniformly}} Q_0(\theta)$ . However, when  $R$  is fixed,  $Q_N^{MSL}(\theta)$  does not converge uniformly to  $Q_N^{ML}(\theta)$ , hence,

$$\begin{aligned} Q_N^{MSL}(\theta) &\overset{\text{uniformly}}{\mathcal{G}} Q_N^{ML}(\theta) \overset{\text{uniformly}}{\rightarrow} Q_0(\theta) \\ &\Rightarrow Q_N^{MSL}(\theta) \overset{\text{uniformly}}{\mathcal{G}} Q_0(\theta) \end{aligned}$$

The implication that consistency fails to occur in the MSL can be seen by replacing  $Q_N(\theta)$  with  $Q_N^{MSL}(\theta)$  in Lemma 1. This leads condition 5 of the lemma to break down and, with it, consistency of the MSL estimator for fixed  $R$ .

## 2.2 Asymptotic Distribution of MSL

Once we have identified the conditions under which the estimator of MSL is consistent, we can derive it's limiting distribution. As usual for extremum estimators, the starting point is the set of first order conditions, also known as the normal equations

$$\frac{\partial}{\partial \theta} Q_N^{MSL}(\theta) = 0$$

Applying the mean-value theorem element by element of  $\theta$ , we can make a linear approximation of the normal equations around the true population parameter value  $\theta_0$

$$0 = \frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} Q_N^{MSL}(\theta_0) + \left[ \frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} Q_N^{MSL}(\bar{\theta}) \right] \sqrt{N} (\hat{\theta}_{MSL} - \theta_0)$$

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

$$\frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} Q_N^{MSL}(\bar{\theta}) \xrightarrow{p} \mathcal{J}(\theta_0) \equiv E \left( \frac{\partial^2}{\partial \theta \partial \theta'} Q_0^{ML}(\theta_0) \right)$$

The first term of the expansion is a sum of  $N$  *i.i.d.* terms

$$\frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} Q_N^{MSL}(\theta_0) = \frac{1}{\sqrt{N}} \sum_{n=1}^N \frac{\frac{\partial}{\partial \theta} \tilde{\mu}(\theta_0; y_n, \Psi_n)}{\tilde{\mu}(\theta_0; y_n, \Psi_n)}$$

recalling that  $\tilde{\mu}(\theta_0; y_n, \Psi_n) = \frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta_0; y_n, \Psi_{ni})$ . But this term does not have expectation of 0<sup>3</sup>. To see this, we can add and subtract terms

$$\frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} Q_N^{MSL}(\theta_0) = \frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} Q_N^{ML}(\theta_0) + A_N + B_N$$

where

$$\begin{aligned} A_N &= \frac{1}{\sqrt{N}} \sum_{n=1}^N \left[ \frac{\partial}{\partial \theta} \log \tilde{\mu}(\theta_0; y_n, \Psi_n) - E \left( \frac{\partial}{\partial \theta} \log \tilde{\mu}(\theta_0; y_n, \Psi_n) \right) \right] \\ B_N &= \frac{1}{\sqrt{N}} \sum_{n=1}^N \left[ E \left( \frac{\partial}{\partial \theta} \log \tilde{\mu}(\theta_0; y_n, \Psi_n) \right) - \frac{\partial}{\partial \theta} \log \mu(\theta; Z_n) \right] \end{aligned}$$

---

<sup>3</sup>We could still apply the Lindberg-Levy central limit theorem to show that this term converges to a normal distribution with a non-zero mean.

The term  $A_N$  converges in probability to 0 by a standard law of large numbers, so it is not causing the bias, it simply represents simulation noise added in the problem. The added simulation bias is captured by the  $B_N$  term and due to the fact that the expectation of  $\frac{\partial}{\partial \theta} \log \mu(\theta; Z_n)$  is not equal to  $E\left(\frac{\partial}{\partial \theta} \log \tilde{\mu}(\theta_0; y_n, \Psi_n)\right)$ ,  $B_N$  does not vanish asymptotically<sup>4</sup>. The question now is under what conditions does the bias vanish? The following propositions highlight key results

**Proposition 2** *Let*

1.  $\tilde{\mu}(\theta_0; y_n, \Psi_n)$  be an unbiased simulator for  $\mu(\theta_0; Z_n)$  such that  $V(\tilde{\mu} - \mu|y) = O(R^{-1})$
2.  $s(\theta; y, \mu)$  be a moment function such that  $E(s(\theta_0; y, \mu)) = 0$
3.  $\tilde{s}(\theta; y) \equiv s(\theta; y, \tilde{\mu})$  be a simulator for  $s(\cdot)$  and let  $R/\sqrt{N} \rightarrow \infty$   
If  $\tilde{s}$  is Lipschitz<sup>5</sup> in  $\tilde{\mu}$  uniformly in  $\theta$ , then

$$B_N = \frac{1}{\sqrt{N}} \sum_{n=1}^N (E(\tilde{s}(\theta_0; y)) - s(\theta_0; y, \mu)) \xrightarrow{p} 0$$

**Proof.** See Hajivassiliou and Ruud (1994) p.2419. ■

**Proposition 3** *Let*

1.  $\mu(\theta_0; Z_n)$  be bounded uniformly away from zero and Lipschitz in  $\theta$  on a compact space  $\Theta$
2.  $\tilde{\mu}(\theta_0; y_n, \Psi_n)$  be an unbiased differentiable simulator for  $\mu(\theta_0; Z_n)$  also bounded away from zero and Lipschitz in  $\theta$  on  $\Theta$  such that  $V(\tilde{\mu} - \mu|y) = O(R^{-1})$   
If  $R/\sqrt{N} \rightarrow \infty$  then

$$A_N + B_N \xrightarrow{p} 0$$

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

**Proof.** See Hajivassiliou and Ruud (1994) p.2420. ■

**Explanation of Proposition 2 and 3** These propositions are somewhat very technical especially with regards to the Lipschitz condition. In Proposition 2, the Lipschitz condition allows to set up an upper bound to  $(E(\tilde{s}(\theta_0; y)) - s(\theta_0; y, \mu))$  and this bound is  $O(R^{-1})$ . Since  $\frac{1}{\sqrt{N}}$  multiplies the bound, the product of the order of magnitudes yields  $O_p\left(\frac{R}{\sqrt{N}}\right)$ , the rate of convergence in probability of the sequence  $B_N$ . In Proposition 3, it is mentioned

<sup>4</sup>To have an equality, we would have to fit the expectation operator  $E$  in between the operator  $\log$  and  $\tilde{\mu}$ . To do so, one must let  $R \rightarrow \infty$ .

<sup>5</sup>A function  $f : \mathbb{R}^a \times \mathbb{R}^b \rightarrow \mathbb{R}^c$  is a Lipschitz function in  $\mu$  if and only if there exists a constant  $\phi$  such that  $\|f(\mu^p, p) - f(\mu^q, p)\| \leq \phi \|\mu^p - \mu^q\|$  for all  $\mu^p$  and  $\mu^q \in \mathbb{R}^a$  and all  $p \in \mathbb{R}^b$ .

that under the conditions of the proposition,  $\hat{\theta}_{MSL}$  is asymptotically efficient. Define the two following arrays

$$\begin{aligned}\mathcal{J}^R(\theta) &= -E_N \left( \frac{\partial^2}{\partial\theta\partial\theta'} Q_N^{MSL}(\theta) \right) \\ \mathcal{I}^R(\theta) &= E_N \left( \left\{ \frac{\partial}{\partial\theta} Q_N^{MSL}(\theta) \right\} \left\{ \frac{\partial}{\partial\theta} Q_N^{MSL}(\theta) \right\}' \right)\end{aligned}$$

where  $E_N$  denotes the empirical expectation for our sample of size  $N$ . As  $R \rightarrow \infty$ , both  $\mathcal{J}^R(\theta)$  and  $\mathcal{I}^R(\theta)$  converge to  $\Omega_N(\theta_0) = E_N \left( \left\{ \frac{\partial}{\partial\theta} Q_N^{ML}(\theta_0) \right\} \left\{ \frac{\partial}{\partial\theta} Q_N^{ML}(\theta_0) \right\}' \right)$  so that  $\mathcal{J}^R(\theta)^{-1}$  and  $\mathcal{I}^R(\theta)^{-1}$  are both consistent estimators of the asymptotic covariance matrix. Under this restriction on the rate of increase of the number of replications,  $\sqrt{N}(\hat{\theta}_{MSL} - \theta_0) \xrightarrow{d} N(0, \Omega(\theta_0))$  where  $\Omega(\theta_0) = \lim_{N \rightarrow \infty} \Omega_N(\theta_0)$ , the Cramer-Rao lower bound. It is in this sense that the MSL is asymptotically efficient. However, for finite  $R$ ,  $\mathcal{I}^R(\theta)$  is larger in some matrix sense than  $\mathcal{J}^R(\theta)$  due to simulation noise and  $\mathcal{I}^R(\theta)$  decreases as  $R$  increases. Consequently, using  $\mathcal{I}^R(\theta)^{-1}$  as the estimator for  $\Omega(\theta_0)$  will underestimate the covariance matrix and may suggest erroneously that increasing  $R$  will decrease the precision of the estimator. For this reason, it is recommended (see e.g. McFadden and Train 2000) to use the following robust asymptotic covariance matrix

$$\mathcal{J}^R(\hat{\theta}_{MSL})^{-1} \mathcal{I}^R(\hat{\theta}_{MSL}) \mathcal{J}^R(\hat{\theta}_{MSL})^{-1}$$

see Newey and McFadden 1994. McFadden and Train 2000 give a real life empirical example where the standard errors that come about from using  $\mathcal{I}^R(\theta)^{-1}$  as the covariance matrix estimator may be underestimated by as much as 20%.

### 3 Method of Simulated Moments (MSM)

#### 3.1 Consistency of MSM

The major drawback of MSL is the requirement that the number of draws must tend to infinity for the estimator to be consistent and the number of draws must increase at a faster rate than  $\sqrt{N}$  for the estimator to be asymptotically efficient. This requirement increases the computational requirement of MSL and alternative estimators who reduce the computational requirement would be welcomed. In this vein, the Method of Moments (MOM) discussed in the introduction offers this handy alternative. The MOM estimator is defined by

$$\frac{1}{N} \sum_{n=1}^N \mathbf{w}_n(\mathbf{X}, \hat{\theta}_{MOM}) \left[ y_n - \mu(\hat{\theta}_{MOM}; \mathbf{x}_n) \right] = 0$$

where  $\mu(\theta; \mathbf{x}) = E(y|\mathbf{x}, \theta)$ . Like the case of Maximum Likelihood, consistency of this estimator rest on the uniform convergence of the sample moment conditions to their population counter-

parts. Under the identifying assumption that there exist a single root to the moment conditions, then  $\hat{\theta}_{MOM} \rightarrow \theta_0$ .

**Distinguishing MOM from ML** It is frequently argued that ML is a special case of MOM since solving the ML estimator requires that one solves the vector or normal equations  $\frac{\partial}{\partial \theta} Q_N^{ML}(\hat{\theta}_{ML}) = 0$  derived from the likelihood function. A subtle difference arises in the presence of multiple roots. In the ML framework, when two optimums  $\hat{\theta}_{ML}^1$  and  $\hat{\theta}_{ML}^2$  satisfy  $\frac{\partial}{\partial \theta} Q_N^{ML}(\hat{\theta}_{ML}) = 0$ , we have a criteria to select between both, namely we will keep the solution that yields the highest likelihood function value. Say,  $Q_N^{ML}(\hat{\theta}_{ML}^1) > Q_N^{ML}(\hat{\theta}_{ML}^2)$ , we would naturally stick with the spirit of maximizing the likelihood and select  $\hat{\theta}_{ML}^1$  as the solution to our problem. In the MOM case, we do not have such a criterion to distinguish between both roots, unless the moments are associated with a likelihood function as we have just been discussing.

## 4 Limited Dependent Variable Models (LDV)

Integrals are almost synonymous to LDV models. The discrete nature of the choice set that these models try to explain can quickly generate multiple integrals that can be estimated without much precision but with great difficulty using numerical quadrature. The classical and most appealing approach consists of simplifying the model in such a way that it is computationally feasible. However, putting such constraints on the model goes against the researchers beliefs of what the true data-generating process is. The most striking example concerns the multinomial choice model.

### 4.1 The Multinomial Choice Model

The case of the multinomial choice model, where we generally model a decision maker faced with a choice amongst a finite set of options, has been at the heart of the simulation literature<sup>6</sup>. In the multinomial setting, a unit  $n$  must choose amongst a set of choices  $C_n = \{c_{n1}, c_{n2}, \dots, c_{nS}\}$  where elements  $c_{nj}$  are binary indicators taking the value of one if unit  $n$  chooses option  $j$ , and 0 otherwise. To ground the decision making into economic theory, we associate with each

---

<sup>6</sup>For a discussion of the simulation of the Multiperiod Multinomial Probit model, see Geweke, Keane, and Runkle (1997).

alternative a random utility function

$$\begin{aligned} u_{n1} &= x'_{n1}\beta + \varepsilon_{n1} \\ u_{n2} &= x'_{n2}\beta + \varepsilon_{n2} \\ &\vdots \\ u_{nS} &= x'_{nS}\beta + \varepsilon_{nS} \end{aligned}$$

where  $\varepsilon_{nj}$  are error terms capturing unobserved characteristics entering the utility unit  $n$  derives from choosing alternative  $j$ . Conditions under which preferences of a unit can be represented by a utility function in this stochastic framework can be found in . Given the vector of unobserved utilities, alternative  $j$  is selected by the optimizing unit if  $u_{nj} > u_{nk}$  for all  $\{k \in S | k \neq j\}$ . This is a statistical event since randomness is introduced thru the unobserved characteristics error terms whom we assume follows some joint distribution. We want to find values of  $\beta$  which maximize the probability that we observe the choice of our unit, his choice of alternative  $j$ . The choice of this alternative is captured by following equivalent events:

$$\Pr(\text{Choose } j) = \Pr(\{c_{n1} = 0, c_{n2} = 0, \dots, c_{nj} = 1, \dots, c_{nS} = 0\} | x_n) \quad (5)$$

$$\equiv \Pr(\{u_{nj} > u_{n1}, \dots, u_{nj} > u_{nj-1}, u_{nj} > u_{nj+1}, \dots, u_{nj} > u_{nS}\} | x_n) \quad (6)$$

$$\equiv \Pr\left(\left\{u_{n1} - u_{nj} < 0, \dots, u_{nj-1} - u_{nj} < 0, \right. \right. \\ \left. \left. u_{nj+1} - u_{nj} < 0, \dots, u_{nS} - u_{nj} < 0\right\} | x_n\right) \quad (7)$$

$$\equiv \Pr\left(\left\{\varepsilon_{n1} - \varepsilon_{nj} < x'_{nj}\beta - x'_{n1}\beta, \dots, \varepsilon_{nj-1} - \varepsilon_{nj} < x'_{nj}\beta - x'_{nj-1}\beta, \right. \right. \\ \left. \left. \dots, \varepsilon_{nS} - \varepsilon_{nj} < x'_{nj}\beta - x'_{nS}\beta\right\} | x_n\right) \quad (8)$$

All these events are equal and some of the description of the same statistical event can be found in all econometrics text books. The first event simply describes our observation rule, that alternative  $j$  is the alternative selected. The second event maps the first event but this time in terms of the underlying random utility framework. Both events have the same probability of accruing, given the observed characteristics of the unit  $x_n$ . It is more convenient to work with the latter formulation of the event. To make this operational, we will need to make assumptions on the random part of the problem, namely the error terms. The assumptions that we impose guide us to the model we will use.

#### 4.1.1 Assuming a Multivariate Normal Distribution

For example, if we assume that the vector of errors  $\epsilon_n = [\varepsilon_{n1}, \varepsilon_{n2}, \dots, \varepsilon_{nS}]'$  follows a multivariate density  $f(\cdot) \equiv N(0, \Omega)$ , where  $\Omega$  is a  $S \times S$  positive definite matrix, then the fact that the joint

density is multivariate normal implies some assumptions on the error terms since we allow for the possibility of covariances in the unobservables, thru the non-diagonal elements of  $\Omega$ . This may make sense if for example we believe that the unobserved factors explaining the utility  $u_{n1}$  are correlated with the unobserved factors affecting the utility level  $u_{nj}$ . If this is the case, then the multivariate normal distribution represents a reasonable approximation to the data-generating process. The  $(S - 1) \times 1$  vector  $\epsilon_{nj}^* = [\epsilon_{n1j}^*, \epsilon_{n2j}^*, \dots, \epsilon_{nSj}^*]'$  follows a density  $f^*(\cdot) \equiv N(0, \Omega^*)$ . In this notation

$\Pr(\text{Choose } j) = \Pr(\{\epsilon_{n1j}^* < x'_{nj}\beta - x'_{n1}\beta, \dots, \epsilon_{nj-1j}^* < x'_{nj}\beta - x'_{nj-1}\beta, \dots, \epsilon_{nSj}^* < x'_{nj}\beta - x'_{nS}\beta\} | x_n)$  is an equivalent expression. If we define the following rectangle in  $\mathbb{R}^{S-1}$

$$\mathbf{B} = \{\epsilon_{n1j}^*, \epsilon_{n2j}^*, \dots, \epsilon_{nSj}^* | \epsilon_{n1j}^* < x'_{nj}\beta - x'_{n1}\beta, \dots, \epsilon_{nj-1j}^* < x'_{nj}\beta - x'_{nj-1}\beta, \dots, \epsilon_{nSj}^* < x'_{nj}\beta - x'_{nS}\beta\}$$

We then have

$$\begin{aligned} \Pr(\text{Choose } j) &= \int_{\mathbf{B}} f^*(\epsilon_{nj}^*) d\epsilon_{nj}^* \\ &= \int_{\mathbb{R}^{S-1}} 1[\epsilon_{nj}^* \in \mathbf{B}] f^*(\epsilon_{nj}^*) d\epsilon_{nj}^* \end{aligned}$$

where  $1[\cdot]$  is the indicator function taking a value of 1 when the expression inside the brackets is true. This implies that we chop off the part of the multivariate density that does not fall inside the rectangle  $\mathbf{B}$ . In most papers where the theory of simulation is laid down, the integration with the indicator function prevails so we keep it here. Since we integrate over the dimensions of  $\mathbb{R}^{S-1}$ , this is a  $S - 1$  integral. When  $S$  tends to a large number, a numerical quadrature approximation will typically break down. Two alternatives there exist, keep the postulated data-generating process the way it is and use simulation to approximate the integrals, or make some assumptions that will simplify the model. One frequently assumes that the unobserved part of utility is independent across alternatives. This gives rise to the multinomial logit model.

#### 4.1.2 Assuming Independence of Irrelevant Alternatives

On the other hand, if we assume that all two binary choice couples are independent, then we can opt for a distribution of  $\epsilon_n = [\epsilon_{n1}, \epsilon_{n2}, \dots, \epsilon_{nS}]'$  that will impose statistical independence between unobservables. One classical way to do so is to make the assumption that all elements of  $\epsilon_n = [\epsilon_{n1}, \epsilon_{n2}, \dots, \epsilon_{nS}]'$  have the following density  $\exp(-\epsilon_{ni})$ . This density is known as an Extreme-Value Type 1 Gumbel distribution and has been introduced by McFadden (1974) to model consumer choice problems. In the appendix we show that this statistical assumption leads to the following closed form expression

$$\begin{aligned} &\Pr(\{u_{nj}^* > u_{n1}^*, \dots, u_{nj}^* > u_{nj-1}^*, u_{nj}^* > u_{nj+1}^*, \dots, u_{nj}^* > u_{nS}^*\} | x_n) \\ &= \frac{\exp(x'_{nj}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)} \end{aligned}$$

which gives rise to the Multinomial Logit model (MNL). The great advantage of the density assumption is to allow for an integral-free choice probability at the expense of assuming that the choice of alternative  $j$  over 1 is not affected by the choice of alternative  $j$  over alternative 4. A heuristic way to see this is to work with odds (probability) ratios. Divide the probability of choosing alternative  $j$  by the probability of choosing alternative 1

$$\frac{\Pr(\text{Choose } j)}{\Pr(\text{Choose } 1)} = \frac{\frac{\exp(x'_{nj}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)}}{\frac{\exp(x'_{n1}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)}} = \frac{\exp(x'_{nj}\beta)}{\exp(x'_{n1}\beta)}$$

and we immediately see that the odds ratio is independent of the characteristics of the other alternative in the choice set, but depends only  $x_{n1}$  and  $x_{nj}$ .

## 4.2 A general presentation of LDV models

We have seen the derivation of choice probabilities associated with a special type of model, that of the multinomial choice problem. This model has the special feature that we observe  $S$  binary indicators for each individual and only a single one of these binary variables can take a value of 1. A special case of the multinomial choice problem is the binomial choice problem commonly framed as the logit and probit models. The value of  $u_{ni}$  tells us if the binary variable associated with alternative  $i$  is 0 or 1. We can define a function  $\tau(\cdot) : \mathbb{R} \rightarrow \{0, 1\}$  that maps the value of the utility level  $u_{n1}$  associated with alternative  $i$  to the value  $y_{ni}$ , its associated binary random variable. This was the case of the multinomial choice model. Remember, we only observe the set  $\mathbf{y}_n = \{y_{n1}, y_{n2}, \dots, y_{nS}\}$  for each unit  $n$ . We can plug in the observation rule  $\tau(\cdot)$  the value of  $u_{n1}$  and get a unique outcome, 0 or 1. This makes the observation rule a function. However, from the observation of a 0 or a 1 for alternative  $i$ , an infinity of  $u_{n1}$  may have generated that observed  $y_{n1}$ . This makes the  $\tau(\cdot)$  mapping not one-to-one. In the rest of this section, we will denote the utility levels of each alternative as  $y_{ni}^*$  and the vector of all utilities by  $\mathbf{y}_n^* = \{y_{n1}^*, y_{n2}^*, \dots, y_{nS}^*\}$

**Definition 2** *A LDV model occurs when the observation rule  $\tau(y_{ni}^*) = y_{n1}$  is not a one to one mapping. Alternatively, one could say that LDV models occur when the observation rule  $\tau(y_{ni}^*)$  is not invertible.*

All LDV models can be nested in this definition, given some minor adjustments to the observation rule.

**Example 3 (Multinomial Choice)**  $y_{ni} = \tau(\mathbf{y}_n^*) = 1 \{ \text{Max}(\mathbf{y}_n^*) = y_{ni}^* \}$ . *If we assume that the unobservables follows a multivariate normal distribution, then we have previously derived*

that the probability of choosing alternative  $j$  is

$$\begin{aligned}\Pr(\text{Choose } j) &= \int_{\mathbf{B}} f^*(\epsilon_{nj}^*) d\epsilon_{nj}^* \\ &= \int_{\mathbb{R}^{S-1}} \mathbf{1}[\epsilon_{nj}^* \in \mathbf{B}] f^*(\epsilon_{nj}^*) d\epsilon_{nj}^*\end{aligned}$$

We just need to estimate for all observations who chose  $j$  the probability of their choice given their observed characteristics and do that for all alternatives. Finally, take the log of these probabilities and form the log-likelihood function.

**Example 4 (Censored Models)**  $y_n = \tau(y_n^*) = \mathbf{1}\{y_n^* \geq 0\} \cdot y_n^*$ . From this observation rule, we see that the researcher observes the outcome of  $y_n^*$  only for those values above zero. All the other cases are coded 0 which implies that the distribution (i.e. density) of the observed data will be a mixture, with a discrete and a continuous part. What we mean by that is there will be a point, here  $y_n = 0$  which will have a positive probability of appearing:  $\Pr(y_n = 0) = a > 0$ . In the continuous part of the distribution, we know that there are an infinite number of possible values that  $y_n$  can take so that  $\Pr(y_n = y_n^* | y_n^* > 0) = 0$ . For example, what is the probability that  $y_n < Y$ ? This is simply  $\int_{-\infty}^Y f(y_n^*) dy_n^*$ . We can break this integral into two parts

$$\begin{aligned}\int_{-\infty}^Y f(y_n^*) dy_n^* &= \int_0^Y f(y_n^*) dy_n^* + \int_{-\infty}^0 f(y_n^*) dy_n^* \\ &\equiv \Pr(0 < y_n < Y) + \underbrace{\Pr(y_n = 0)}_a\end{aligned}$$

Again, remember that  $\Pr(y_n = b \neq 0) = 0$ . If we want to construct the log-likelihood for a random sample drawn from a population, then we must know the associated density of the data. Above, we have derived the CDF of the data. The associated pdf is the derivative of the CDF with respect to the variable of integration  $y_n^*$ .

$$\begin{aligned}&\mathbf{1}\{y_n^* \geq 0\} \cdot f(y_n^*) + \mathbf{1}\{y_n^* < 0\} \cdot \Pr(y_n = 0) \\ &= \mathbf{1}\{y_n^* \geq 0\} \cdot \underbrace{f(y_n^*)}_{\text{Part A}} + \mathbf{1}\{y_n^* < 0\} \cdot \underbrace{\int_{-\infty}^0 f(y_n^*) dy_n^*}_{\text{Part B}}\end{aligned}$$

for observations with  $\mathbf{1}\{y_n^* \geq 0\}$ , the log of Part A is relevant in the log-likelihood function while for those observations satisfying  $\mathbf{1}\{y_n^* < 0\}$ , the log of Part B enters the log-likelihood function. Taking the derivative of the CDF with respect to  $y_n^*$  may not seem obvious but it is reasonable if we interpret the integral as a Stieltjes-Lebesgue integral which contains a mixtures of discrete and continuous measures.

A closely related model is the sample selection model.

**Example 5 (Sample-Selection)**  $y_n = \tau(y_n^*) = \begin{pmatrix} \tau_1(y_{1n}^*) \\ \tau_2(y_{2n}^*) \end{pmatrix} = \begin{pmatrix} 1 \{y_{1n}^* \geq 0\} \\ 1 \{y_{1n}^* \geq 0\} \cdot y_{2n}^* \end{pmatrix}$ . Although it looks very different than the Censored regression case, it is just a generalization of the latter. The main difference is that in the Censored regression case, it is assumed that the same latent process  $y_n^*$  guided both the censoring and the observed values of  $y_n$  for positive  $y_n^*$  while in the Sample-Selection model, we allow for a different behavior to guide the Censoring and the observed outcome. This different behavior appears by separating the latents in to parts,  $y_{1n}^*$  and  $y_{2n}^*$  where we now observe  $y_{2n} = y_{2n}^*$  depending on the values of  $y_{1n}^*$  and no longer depending on the values of  $y_{2n}^*$  itself. Since now  $y_n = [y_{1n}, y_{2n}]$ , the sample space of observed values take the following form

$$\mathbf{B} = \{(y_{1n} = 0, y_{2n} = 0) \cup (y_{1n} = 1, y_{2n}), y_{2n} \in \mathbf{R}\}$$

We need to evaluate the density of both sub-groups of events  $(y_{1n} = 0, y_{2n} = 0)$  and  $(y_{1n} = 1, y_{2n})$ , take the logarithm of the respective expressions and form the log-likelihood function. Our observation rule tells us that the event  $(y_{1n} = 0, y_{2n} = 0)$  occurs depending on the value of  $y_{1n}^*$  and not both  $y_{1n}^*$  and  $y_{2n}^*$ . The density of this event is

$$\int_{-\infty}^0 f\left(\frac{y_{1n}^* - \mu_1}{\sigma_1}\right) dy_{1n}^* = \Phi\left(\frac{-\mu_1}{\sigma_1}\right) \quad (9)$$

The other possible observed outcome in our sample is  $(y_{1n} = 1, y_{2n}) = 1 \{y_{1n}^* \geq 0\} \cdot y_{2n}^*$ . Our observation rule says that this occurs when  $y_{1n}^* \geq 0$ . If we denote the joint normal distribution of  $y_{1n}^*$  and  $y_{2n}^*$  by  $g(y_{1n}^* - \mu_1, y_{2n}^* - \mu_2, \Omega)$ , we have that the event is described by the following integral

$$\int_{-\infty}^0 g(y_{1n}^* - \mu_1, y_{2n}^* - \mu_2, \Omega) dy_{1n}^*$$

To come to an analytic form, we rewrite the joint distribution  $g(\cdot) = \phi(y_{1n}^* - \mu_1^{cl2}, \Omega^{cl2}) \cdot \phi(y_{2n}^* - \mu_2, \sigma_2)$  in terms of the product of a marginal and a marginal density. Here,  $\mu_1^{cl2}$  and  $\Omega^{cl2}$  represent the conditional expectation and variance of  $y_{1n}^*$  given  $y_{2n}^*$ . It is usually the case that we condition on the variable  $y_{2n}^*$  since we do not integrate over it and this feature allows us to take some of the stuff out from the integral. By doing so, the preceding integral becomes

$$\begin{aligned} & \phi(y_{2n}^* - \mu_2, \sigma_2) \cdot \int_{-\infty}^0 \phi(y_{1n}^* - \mu_1^{cl2}, \Omega^{cl2}) dy_{1n}^* \\ \equiv & \phi(y_{2n}^* - \mu_2, \sigma_2) \cdot \Phi(y_{1n}^* - \mu_1^{cl2}, \Omega^{cl2}) \end{aligned} \quad (10)$$

**Example 6** In summary, for observations with  $(y_{1n} = 0, y_{2n} = 0)$ , compute the log of (9) and for observations with  $(y_{1n} = 1, y_{2n})$ , compute the log of (10). Then, for both subsets, form the log-likelihood function.

**Example 7 (Mixtures)**  $y_n = \tau(y_n^*) = 1 \{y_{1n}^* \geq 0\} \cdot y_{2n}^* + 1 \{y_{1n}^* < 0\} \cdot y_{3n}^*$ . This is an interesting class of models which, again, generalizes the sample-selection model. From the observation

rule, we see that  $y_n = y_{2n}^*$  if  $y_{1n}^* \geq 0$ , otherwise we observe  $y_{3n}^*$ . A very important fact is that we cannot observe the indicator  $1\{y_{1n}^* \geq 0\}$  and thus,  $y_n$  is a mixture of  $y_{2n}^*$  and  $y_{3n}^*$ . The density of  $y_n$  will then be the sum of the densities of two events,  $1\{y_{1n}^* \geq 0\} \cdot y_{2n}^*$  and  $1\{y_{1n}^* < 0\} \cdot y_{3n}^*$ . We have seen in the standard selection model how to derive the density for  $1\{y_{1n}^* \geq 0\} \cdot y_{2n}^*$ . By symmetry, the same procedure can be applied for  $1\{y_{1n}^* < 0\} \cdot y_{3n}^*$  and we find that

$$f(y_n) = \phi(y_{2n}^* - \mu_2, \sigma_2) \cdot \Phi(y_{1n}^* - \mu_1^{c|2}, \Omega^{c|2}) + \phi(y_{3n}^* - \mu_3, \sigma_3) \cdot \Phi(y_{1n}^* - \mu_1^{c|3}, \Omega^{c|3})$$

which shows that the density of  $y_n$  is a mixture of the densities of  $y_{2n}^*$  and  $y_{3n}^*$  weighted by probabilities.

All these are examples in which the LDV models arise. Notice that we took care in specifying bold face characters in the  $\tau(\cdot)$  function so that  $\tau(\mathbf{y}_n^*)$  implies that there exist a vector of unobservables random variables  $\mathbf{Y}_n^* = \{y_{n1}^*, y_{n2}^*, \dots, y_{nS}^*\}$  entering the observation rule. In the multinomial case, the value taken by any of the  $S$  binary indicators is a function of all  $S$  unobserved utility levels  $\{y_{ni}^* | i = 1, \dots, S\}$ . Second important point is to note that all examples have in common that you cannot invert the observation rule function to identify a unique  $\mathbf{y}_n^*$  that generated the observed  $\mathbf{y}_n$ . A last remark, some of the above models do not seem to involve high-order integrals. This is because we have presented the simplest type of models. Both the truncation, censoring, sample-selection can be extended. Form example, one could estimate a multinomial tobit model, in which case multiple integrals appear in the density of the model. Similarly, one can estimate a sample-selection model where the observed  $y_{2n}$  depends on more than one selection (latent) variable. Again, multiple integrals appear in this formulation.

## 5 Unobserved heterogeneity-Taste Variations

### 5.1 Non-Linear Model Version A

Assume you have the following non-linear regression function

$$y_n = m(x_n' \boldsymbol{\theta}) + \sigma e_n$$

where  $e_n | x_n \sim \mathbf{N}(0, 1)$  for a random sample  $n = 1, 2, \dots, N$  and  $\boldsymbol{\theta}$  is a  $K \times 1$  vector of parameters. Suppose you do not want to assume that all individuals in your sample have the same parameter vector  $\boldsymbol{\theta}$ . In fact, we can assume that for each individual  $n$ , there is a unique vector  $\boldsymbol{\theta}$ . We then have a sequence  $\{\theta_1, \theta_2, \dots, \theta_N\}$  where all parameters are assumed to be drawn from a unique distribution  $\mathbf{N}(\boldsymbol{\theta}, \Omega)$  is assumed to follow a spherical normal distribution. This implies that all elements in  $\boldsymbol{\theta}$  are uncorrelated, or, that all the off-diagonal elements of  $\Omega$  are zero. We can rewrite the above equation as

$$y_n = m(x_n' \boldsymbol{\theta}_n) + \sigma e_n$$

Given we assumed that  $e_n|x_n$  is normally distributed, we can subtract and divide appropriately

$$e_n = \frac{y_n - m(x'_n \boldsymbol{\theta}_n)}{\sigma} \sim \mathbf{N}(0, 1)$$

given  $\boldsymbol{\theta}_n$ . The likelihood of this model is

$$\mathbf{Q}(\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_N\}, \sigma) = \sum_{n=1}^N \phi\left(\frac{y_n - m(x'_n \boldsymbol{\theta}_n)}{\sigma}\right)$$

Now, to include the random coefficients,

$$\sum_{n=1}^N \left[ \int \phi\left(\frac{y_n - m(x'_n (\boldsymbol{\theta} + \Omega^{\frac{1}{2}} \mathbf{u}_n))}{\sigma}\right) \prod_{k=1}^K \phi(u_k) du_k \right]$$

This is the linear model setup. Notice that we integrate over a  $K$  dimensional space since we assume that all parameters of the vector  $\boldsymbol{\theta}$  are random. We get that the joint distribution of all parameters is simply the  $K$ -product of univariate normal densities due to our assumption that the random coefficients were independent of each other. This example is discussed at more length in Gouriéroux and Montfort (1996).

## 5.2 Non-Linear Model Version B

The version that we discussed is a Non-Linear Least-Squares (NLS) setup. In LDV models, taste variations can also be incorporated in a similar manner as in the NLS. We start from the multinomial choice problem of Section 4 where we assume the IIA hypothesis on the error terms. This gives rise to the following probability of choosing alternative  $j$

$$\Pr(\text{Choose } j) = \frac{\exp(x'_{nj}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)}$$

If we assume that the vector of parameters  $\beta$  follows a distribution  $f(\beta)$ , then, we can interpret the above probability as being conditional on a particular realization of  $\beta$

$$\Pr(\text{Choose } j|\beta) = \frac{\exp(x'_{nj}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)}$$

We want to estimate the unconditional probability of choosing any alternative  $j$ . To make this probability unconditional, we multiply the conditional probability by the marginal distribution of  $\beta$

$$\Pr(\text{Choose } j) = \int_{\mathbf{B}} \frac{\exp(x'_{nj}\beta)}{\sum_{i=1}^S \exp(x'_{ni}\beta)} f(\beta) d\beta$$

where  $\mathbf{B}$  is the range over which  $\beta$  is allowed to vary. We integrate since this creates a weighted average of the probability of choosing alternative  $j$  for all possible values of  $\beta$ . We can choose any distribution we want for the density of  $f(\beta)$  and it does not necessarily have to be continuous. A radical example of a discrete case would be that  $\beta$  is allowed to take on only two values,  $\beta_1$  and  $\beta_2$  each with probability  $\frac{1}{2}$ . Then the probability of choosing alternative  $j$  is

$$\Pr(\text{Choose } j) = \sum_{i=1}^2 \frac{\exp(x'_{nj}\beta_i)}{\sum_{i=1}^S \exp(x'_{ni}\beta_i)} \cdot \frac{1}{2}$$

This model is called the Mixed Logit Model (Train, 2002) and can approximate well any random utility model. This makes it a powerful alternative to the multinomial probit model. Like in Version Am the dimension of  $\beta$  implies an equivalent number of integrals. For example, if there are 3 parameters, then  $\mathbf{B} \in \mathbb{R}^3$  could be a reasonable space from which the parameters can take likely values. We could restrict the space of  $\mathbf{B}$  is, for example, we do not have reasons to believe that the first parameter may be, say negative. Then,  $\mathbf{B} \in \mathbb{R}_+ \times \mathbb{R}^2$ . In any case, we need to be able to draw from the region  $\mathbf{B}$  in order to use simulation technics.

## 6 Simulators for Probabilities and Expectations

In section 1, we mentioned that one of the frequent objects of applied econometrics was to compute objects of the following kind

$$\mu(\theta; \mathbf{Z}_n) = \int \int \dots \int g(\omega_1, \omega_2, \dots, \omega_s; \theta, y_n) d\omega_1 d\omega_2 \dots d\omega_s$$

Depending on the problem,  $\mu(\theta; \mathbf{Z}_n)$  can be thought of as a probability, which leads to likelihood methods, or as an expectation, which lead to method of moments estimation technics. We argued that the main difficulty in performing ML of MOM was that it required the numerical evaluation of multiple integrals. We heuristically argued that one needs simply to collect random draws from the joint distribution of  $\boldsymbol{\omega} = \{\omega_1, \omega_2, \dots, \omega_s\}$ , that is we draw from what we integrate over. We then simply replaced the draws in the simulator function  $\tilde{\mu}_i(\theta; y_n, \Psi_{ni}) = g(\omega_{i1}, \omega_{i2}, \dots, \omega_{is}; \theta, y_n)$  and simply averaged over all draws that we made

$$\frac{1}{R} \sum_{i=1}^R \tilde{\mu}_i(\theta; \mathbf{Z}_n, \Psi_{ni}) \equiv \tilde{\mu}(\theta; \mathbf{Z}_n, \Psi_n) \xrightarrow{p} \mu(\theta; \mathbf{Z}_n)$$

This description is obviously very crude since it takes for granted that it is easy to draw from the joint distribution of  $\boldsymbol{\omega} = \{\omega_1, \omega_2, \dots, \omega_s\}$ . In reality and especially in LDV models, drawing from the appropriate region of  $\boldsymbol{\omega} = \{\omega_1, \omega_2, \dots, \omega_s\}$  over which the integration takes place is rather a difficult task. In this section, we outline some of the most frequently used simulators for probabilities and expectations and briefly outline how the drawings are actually performed

and what are the possible loopholes that one must guard against when applying these methods. The following subsections will increase in levels of complexity according to the increasing task. We start with drawing from univariate and univariate-truncated densities. Then, we move on to drawings from multivariate normal densities. The highlight will be the discussion of the GHK simulator.

## 6.1 Drawing from univariate densities

Let  $F(e) = \Pr_F(E < e)$  be a univariate CDF with associated pdf  $f(e)$  and let  $e$  be realizations of the random variable  $E$ . A fundamental theorem of statistics states that the distribution of any CDF  $F(e)$  follows a uniform distribution  $U(0, 1)$ . We are interested in drawing realizations of  $e$  from its underlying distribution. The trick is to invert the CDF in order to get the draws. The intuition is the following. Draw  $\mu$  from a  $U(0, 1)$  distribution. By the fundamental theorem of statistics that we discussed above, let this  $\mu$  be defined as follows

$$\mu = F(e^*)$$

for some  $e^*$ . Since the function  $F(\cdot)$  is a one-to-one mapping from the space of  $e$  to the  $[0, 1]$  interval, for each value of  $\mu$ , there exists a unique value  $e^*$  such that the mapping above holds. In this sense, a draw from  $f(e)$  is determined indirectly by  $\mu$ . If we know the inverse function  $F^{-1}$ , then it is a simple matter to compute

$$e^* = F^{-1}(\mu)$$

and we will have succeeded in recovering a draw  $e^*$  from a  $U(0, 1)$  draw  $\mu$ . Examples of CDFs which have an analytical inverse form are

**Exponential Distribution**  $F_\theta(e) = 1 - \exp(-\theta e)$  for non-negative values of  $e$  and  $\theta > 0$ .

Inverting this function gives

$$e^* = F^{-1}(\mu) = -\frac{1}{\theta} \log(1 - \mu)$$

where  $\mu \sim U(0, 1)$ .

**Weibull Distribution**  $F_\theta(e) = 1 - \exp(-e^\theta)$  for non-negative values of  $e$  and  $\theta > 0$ . Inverting this function gives

$$e^* = F^{-1}(\mu) = [-\log(1 - \mu)]^{\frac{1}{\theta}}$$

where  $\mu \sim U(0, 1)$ .

**Cauchy Distribution**  $F_\theta(e) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{e}{\theta}\right)$  for any values of  $e$  and  $\theta$ . Inverting this function gives

$$e^* = F^{-1}(\mu) = \theta \tan\left(\pi\left(\mu - \frac{1}{2}\right)\right)$$

where  $\mu \sim U(0, 1)$ .

**Extreme Value 1 Distribution**  $F(e) = \exp(-\exp(-e))$  for any value of  $e$ . Note that this distribution does not depend on any unknown parameters. Inverting the function gives

$$e^* = F^{-1}(\mu) = -\ln(-\ln(\mu))$$

where  $\mu \sim U(0, 1)$ .

We do not list the normal distribution since it does not have a closed-form expression. However, most computer packages provide simulators from the univariate standard normal and this does not require immediate use of drawings from uniform distributions. Furthermore, since the univariate normal density simulator comes with mostly all commercial software packages, you can use it to simulate other univariate densities in a similar way as we used the uniform random draws to simulate the exponential, weibull, cauchy and extreme value 1 distributions. We briefly mention two useful univariate densities that you can simulate with the standard normal draws

**Normal distribution**  $N(\mu, \sigma)$  You want to simulate the realization of a random variable  $e \sim N(\eta, \sigma^2)$ . Use a computer package built-in routine and compute

$$e^* = \eta + \sigma\mu$$

where  $\mu \sim N(0, 1)$ .

**Log-Normal Distribution** You want to simulate the realization of a random variable  $e \sim LN(\eta, \sigma^2)$  where  $\log e \sim N(\eta, \sigma^2)$ . Use a computer package built-in routine and compute

$$e^* = \exp(\eta + \sigma\mu)$$

where  $\mu \sim N(0, 1)$ .

### 6.1.1 Drawing from Truncated Univariate Densities

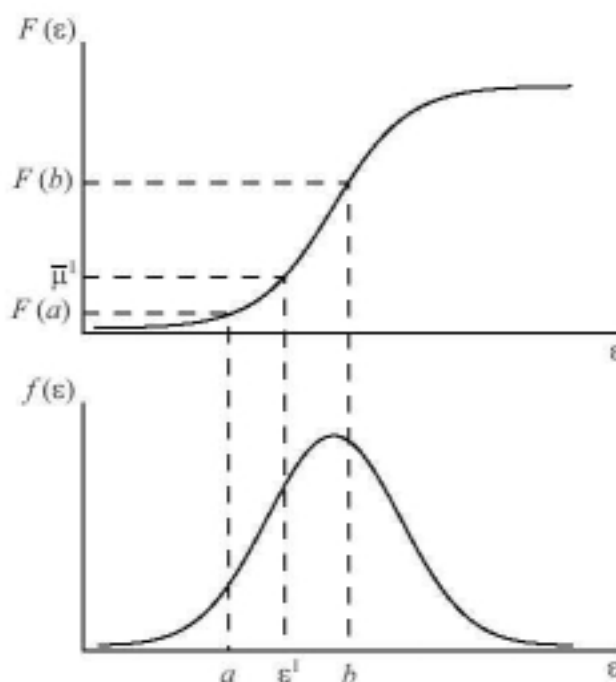
Say you want to draw from the distribution  $F(e)$  but for values of  $e$  restricted between boundaries  $a$  and  $b$ . That is, we want to draw  $e$  from  $f(e)$  conditional on  $a < e < b$ . We can evaluate the CDF at these boundary values by computing  $F(a)$  and  $F(b)$ . The procedure is as follows. Draw  $\mu \sim U(0, 1)$  and compute the following average

$$\bar{\mu} = \mu F(a) + (1 - \mu) F(b)$$

By construction, the  $e$  quantile associated with  $\bar{\mu}$  will be in between  $a$  and  $b$ . Then, we simply proceed as we in the univariate case and compute the inverse

$$e^* = F^{-1}(\bar{\mu})$$

The following graph taken from Train (2002) illustrates the procedure



## 6.2 Drawing from Multivariate Normal Densities

We have that a draw from  $N(\eta, \sigma^2)$  can be easily achieved as  $e^* = \eta + \sigma\mu$  where  $\mu \sim N(0, 1)$ . We can generalize this procedure to the case where we are interested in drawing from the multivariate normal  $N(\eta, \Sigma)$ , where  $\eta$  is the  $m \times 1$  vector of means and  $\Sigma$  the corresponding  $m \times m$  covariance matrix of the random variables. We would like to keep the thing as simple as in the univariate case, so we would look like something of the following form

$$e^* = \eta + \tau(\Sigma)\mu$$

where  $\mu$  is a  $m \times 1$  vector of i.i.d. standard normal density draws and  $\tau(\cdot)$  denotes a transformation matrix applied to  $\Sigma$ . In the univariate case,  $\eta$  and  $\mu$  are both scalars and the transformation of the variance,  $\tau(\sigma^2) = \sigma$ , was simply taking the square root of the variance. The major difference that hampers the direct application of this procedure to the multivariate case is that one cannot simply take the square root of a matrix like  $\Sigma$ . However, the matrix-equivalent transformation of taking the square root is to decompose  $\Sigma$  in the product of two parts,  $LL'$  where, for the case of  $m = 3$

$$L = \begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix}$$

This in fact implies that

$$\begin{aligned} \underbrace{\begin{pmatrix} e_1^* \\ e_2^* \\ e_3^* \end{pmatrix}}_{\mathbf{e}^*} &= \underbrace{\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}}_{\boldsymbol{\eta}} + \underbrace{\begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix}}_{\tau(\boldsymbol{\Sigma})} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}_{\boldsymbol{\mu}} \\ &= \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} + \begin{pmatrix} s_{11}\mu_1 \\ s_{21}\mu_1 + s_{22}\mu_2 \\ s_{31}\mu_1 + s_{32}\mu_2 + s_{33}\mu_3 \end{pmatrix} \end{aligned}$$

We can see that the correlation between say  $e_1^*$  and  $e_2^*$  is driven by the fact that both simulated terms have  $\mu_1$  in common. Further, it is easy to check that  $V(\mathbf{e}^*) = (\mathbf{L}\boldsymbol{\mu})(\mathbf{L}\boldsymbol{\mu})' = \mathbf{L}\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{L}' = \mathbf{L}\mathbf{L}' = \boldsymbol{\Sigma}$ .

### 6.2.1 Drawing from Truncated Multivariate Normal Densities

This is one of the most difficult densities to draw from, yet, it is fundamental to the computation of the multinomial probit model. Recall that the probability that a unit chooses alternative  $j$  is

$$\begin{aligned} \Pr(\text{Choose } j) &= \int_{\mathbf{B}} f^*(\boldsymbol{\epsilon}_{nj}^*) d\boldsymbol{\epsilon}_{nj}^* \\ &= \int_{\mathbb{R}^{S-1}} \mathbf{1}[\boldsymbol{\epsilon}_{nj}^* \in \mathbf{B}] f^*(\boldsymbol{\epsilon}_{nj}^*) d\boldsymbol{\epsilon}_{nj}^* \end{aligned}$$

where  $\mathbf{B}$  is a rectangle in the  $\mathbb{R}^{S-1}$  Euclidean space. Each dimension of the multivariate density has its own truncation point. For example, in the bivariate case, one may wish to draw from the zone  $\mathbf{B} = [a, b] \times [c, d] \subset \mathbb{R}^2$  associated with the integral

$$\int_a^b \int_c^d \text{biv}(\epsilon_{nj1}^*, \epsilon_{nj2}^*, \rho) d\epsilon_{nj1}^* d\epsilon_{nj2}^*$$

with  $\text{biv}(\cdot)$  symbolizing the bivariate standard normal density. There are several ways to do this.

**Crude Frequency Simulator** This would imply drawing from a multivariate normal density  $\mathbf{e}^*$  as done in the previous subsection

$$\mathbf{e}^* = \boldsymbol{\eta} + \tau(\boldsymbol{\Sigma})\boldsymbol{\mu}$$

then check if the draw  $\mathbf{e}^*$  is inside the required zone  $\mathbf{B}$ . If it is inside, record a value of one, otherwise, record a value of zero. Do this  $R$  times and just take the proportion of 1's as the estimate of the probability. Formally, this simulator can be written as

$$\Pr(\mathbf{e}^* \in \mathbf{B}) \simeq \frac{1}{R} \sum_{i=1}^R \mathbf{1}[\mathbf{e}_i^* \in \mathbf{B}]$$

where  $e_i^*$  denotes one of the  $R$  draws that we take from the multivariate normal density. The main problem with this simulator is that it jumps and is discontinuous in the parameters that we seek to estimate<sup>7</sup>. Furthermore, it may take an enormous amount of computer time to get sufficient draws to estimate small probabilities since small probabilities are associated with small rectangles  $\mathbf{B}$ . This implies that it is likely that small probabilities are estimated to be 0 due to the lack of draws inside the rectangle, and this is particularly troublesome in the MSL case where we take the logs of these simulated probabilities. Needless to say that the log of 0 is undefined. A more complete discussion of this point can be found in Train (2002).

**Importance Sampling** This procedure transforms a difficult density to draw from into an auxiliary density from which it is easy to draw from. For example, suppose that we wish to draw from

$$\int t(e) g(e) de$$

where  $g(e)$  is the density to draw from and  $t(e)$  is some statistic but that to draw from  $g(e)$  is very hard *but easy to compute*. However, if you have an auxiliary density  $f(e)$  with the same range as  $g(\cdot)$ , but which it is easy to draw from, then we can do importance sampling. The idea is easy, multiply and divide the integrand by  $f(e)$

$$\int \frac{t(e) g(e) f(e)}{f(e)} de$$

It is easy to see that this multiplication-division does not change the value of the integral, hence it is an acceptable thing to do. Now draw  $e_r^*$  from  $f(e)$  and compute

$$\frac{t(e_i^*) g(e_i^*) f(e_i^*)}{f(e_i^*)}$$

Do this  $R$  times and average the results and this gives the estimated probability. An example where such a sampling scheme may be useful is when one needs to draw from a multivariate truncated normal density. We have seen in the previous sections how to draw from univariate truncated normal densities and we can do this  $m$  times, where  $m$  corresponds to the dimension of  $e^*$ . Then,  $f(e_i^*) = \prod_{i=1}^m t.u.n(e_i^*)$ , where  $t.u.n(e_i^*)$  means truncated univariate normal density and we take the product given the draws have been done independently from each other.

---

<sup>7</sup>McFadden (1989) suggested to use a smooth version of the Crude Frequency Simulator where the indicator function is replaced by a multivariate kernel  $w(e_i^*)$  where observations of  $e_i^*$  close to the region  $\mathbf{B}$  but not in it are considered in the probability, although weighted by how far away from  $\mathbf{B}$  they are. This procedure basically expands the region around  $\mathbf{B}$  in which draws are considered to be informative of the probability of falling in  $\mathbf{B}$ . Like all kernels, a bandwidth  $h$  controls the size of the extra zone and as  $h \rightarrow 0$ ,  $w(e_i^*) \rightarrow 1 [e_i^* \in \mathbf{B}]$ .

**GHK Simulator** This is the most popular simulator of truncated normal densities. Its popularity originates from the paper of Hajivassiliou, McFadden, and Ruud (1996) who showed that this simulator outperforms a wide-class of simulators to compute multivariate normal rectangles and derivatives. It is a little bit more complicated to discuss so we go at a slower pace and start with the case of a trivariate truncated normal distribution. Denote  $\mathbf{e}^* = [e_1^*, e_2^*, e_3^*]'$ . We are interested in the following probability

$$\Pr(\mathbf{e}^* < \mathbf{b}) = \Pr(e_1^* < b_1, e_2^* < b_2, e_3^* < b_3)$$

where, in the previous notation, one could see the rectangle to  $\mathbf{b}$  defined as  $\mathbf{B} = [-\infty, b_1] \times [-\infty, b_2] \times [-\infty, b_3]$ . The GHK simulator decomposes this joint probability in the product of conditionals and marginals

$$\Pr(\mathbf{e}^* < \mathbf{b}) = \Pr(e_1^* < b_1) \cdot \Pr(e_2^* < b_2 | e_1^* < b_1) \cdot \Pr(e_3^* < b_3 | e_1^* < b_1, e_2^* < b_2) \quad (11)$$

Now, let's say that we draw  $\mathbf{e}^*$  using the Choleski decomposition

$$\begin{aligned} \begin{pmatrix} e_1^* \\ e_2^* \\ e_3^* \end{pmatrix} &= \begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \\ &= \begin{pmatrix} s_{11}\mu_1 \\ s_{21}\mu_1 + s_{22}\mu_2 \\ s_{31}\mu_1 + s_{32}\mu_2 + s_{33}\mu_3 \end{pmatrix} \end{aligned}$$

From these three equalities, we can rewrite (11) as

$$\begin{aligned} \Pr(\mathbf{e}^* < \mathbf{b}) &= \underbrace{\Pr(s_{11}\mu_1 < b_1)}_{\text{Term 1}} \cdot \underbrace{\Pr(s_{21}\mu_1 + s_{22}\mu_2 < b_2 | s_{11}\mu_1 < b_1)}_{\text{Term 2}} \cdot \\ &\quad \underbrace{\Pr(s_{31}\mu_1 + s_{32}\mu_2 + s_{33}\mu_3 < b_3 | s_{11}\mu_1 < b_1, s_{21}\mu_1 + s_{22}\mu_2 < b_2)}_{\text{Term 3}} \end{aligned}$$

We have three terms to evaluate. Each of these will be evaluated in turn, starting with the first term.

1.  $\Pr(s_{11}\mu_1 < b_1)$  can simply be computed as  $\Phi(b_1)$  where  $\Phi(\cdot)$  denotes the CDF of the standard normal density. No need for simulations here, simply compute

$$\Phi(b_1) \quad (12)$$

2.  $\Pr(s_{21}\mu_1 + s_{22}\mu_2 < b_2 | s_{11}\mu_1 < b_1) = \int_{-\infty}^{b_1/s_{11}} \Phi(b_2 - s_{21}\mu_1/s_{22}) \cdot \phi(\mu_1) d\mu_1$ . The integral here appears because we condition on all values of  $\mu_1$  lower than  $b_1/s_{11}$ . Since

there is an integral and that computers cannot integrate, we must resort to simulation. To proceed, we need to draw from a truncated normal density in  $[-\infty, b_1/s_{11}]$ . This can be done in a similar way as described in the section of drawing from truncated univariates. Label this draw  $\mu_{r1}^*$ . It is then easy to compute

$$\Phi(b_1 - s_{21}\mu_{i1}^*/s_{22}) \quad (13)$$

Repeat this step  $R$  times and average the results. This gives a simulation of the second term.

3.  $\Pr(s_{31}\mu_1 + s_{32}\mu_2 + s_{33}\mu_3 < b_3 | s_{11}\mu_1 < b_1, s_{21}\mu_1 + s_{22}\mu_2 < b_2)$  is the last part of the problem. This expression is analogous to the preceding one, except that we are conditioning now on two values,  $\mu_1$  lower than  $b_1/s_{11}$  and  $\mu_2$  lower than  $\mu_2 < (b_2 - s_{21}\mu_1)/s_{22}$ . This generates the following integral

$$\int_{-\infty}^{b_1/s_{11}} \int_{-\infty}^{(b_2 - s_{21}\mu_1)/s_{22}} \Phi((b_3 - s_{31}\mu_{r1}^* - s_{32}\mu_{i2}^*)/s_{33}) \phi(\mu_1) d\mu_1 \phi(\mu_2) d\mu_2$$

What do we do in practice? We have already drawn  $\mu_{i1}^*$ . Using them, we can draw  $\mu_{r2}^*$  from the truncated are  $[-\infty, (b_2 - s_{21}\mu_{i1}^*)/s_{22}]$ . Given this, we compute

$$\Phi((b_3 - s_{31}\mu_{i1}^* - s_{32}\mu_{i2}^*)/s_{33}) \quad (14)$$

and repeat the operation  $R$  times, then, average the results.

This is the end of the process. The final step is simply to multiply (12), (13) and (14) together

$$\Pr(e^* < b) = \Phi(b_1) \cdot \left( \frac{1}{R} \sum_{i=1}^R \Phi(b_1 - s_{21}\mu_{i1}^*/s_{22}) \right) \cdot \left( \frac{1}{R} \sum_{i=1}^R \Phi((b_3 - s_{31}\mu_{i1}^* - s_{32}\mu_{i2}^*)/s_{33}) \right)$$

## References

- [1] Thomas S. Ferguson. *A Course in Large Sample Theory*. Texts in Statistical Science. Chapman and Hall, London, UK, 1996.
- [2] John F. Geweke, Michael P. Keane, and David E. Runkle. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics*, 80:125–165, 1997.
- [3] Christian Gouriéroux and Alain Monfort. *Simulation-Based Econometric Methods*. CORE Lectures Series. Oxford University Press, Oxford, England, 1996.
- [4] Vassilis Hajivassiliou, Daniel McFadden, and Paul Ruud. Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. *Journal of Econometrics*, 72:85–134, 1996.

- [5] Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.
- [6] Kenneth Train. Discrete choice methods with simulation. Forthcoming, Cambridge University Press, 2002.