

Simulation-Based Econometric Estimation of Discrete Choice Models

Denis Bolduc

Université Laval

Lecture for
1.205 Advanced Demand Modeling
MIT.

November 2, 2000

Contents

1	Introduction	3
2	Motivation	4
3	Analytical background	21
4	Simulation-Based Estimation	26
4.1	Choice between GSM and MSL . . .	32

1 Introduction

- Simulation-based econometrics specializes on problems for which analytical solutions cannot be obtained. Basically, it is the availability of computer technologies that makes it possible to resolve empirically, problems that are theoretically complex. Two frequently used simulation-based estimation principles are called: MSL, GSM.

A general reference: Gouriéroux, C. and A. Monfort, *Simulation Based Econometric Methods*, Oxford University Press, 1996.

2 Motivation

Review on Classical Methods

Consider estimators $\hat{\beta}_N$ of β (a vector with K components) which are obtained as a solution to the problem:

$$\arg \max_{\beta} \Psi_N (y_1, y_2, \dots, y_N \mid X_1, X_2, \dots, X_N; \beta)$$

- Maximum Likelihood

$$\Psi_N (y \mid X; \beta) = \sum_{n=1}^N \ln f(y_n \mid X_n; \beta). \quad (1)$$

Depending on the problem, $f(y_n \mid X_n; \beta)$ denotes a density, a probability or a combination of the two.

- Example : *Binary Probit*

$$f(y_n \mid X_n; \beta) = \Phi(X_n \beta)^{y_n} [1 - \Phi(X_n \beta)]^{(1-y_n)}$$

- Generalized Method of Moments It may be viewed as an extension of the least-squares criterion to non linear functions of the data. The minimization criterion is:

$$\Psi_N (y | X; \beta) = h' \Omega h \quad (2)$$

where h is a $(N \times 1)$ vector with components h_n defined as:

$$h_n = y_n - E(y_n | X_n; \beta).$$

Ω is a given matrix. With Ω being an identity matrix, we get:

$$\Psi_N (y | X; \beta) = h'h = \sum_{n=1}^N [y_n - E(y_n | X_n; \beta)]^2.$$

– Example : *Binary Probit*

In this case, we obtain:

$$\begin{aligned} y_n &= 1 \text{ or } 0 \text{ depending on choice} \\ E(y_n | X_n; \beta) &= \Phi(X_n \beta). \end{aligned}$$

Statistical Inference

For those two methods, one can show that under certain regularity conditions, the estimator is consistent and asymptotically normally distributed:

$$N^{1/2}(\beta_N - \beta_0) \sim N(0, J_0^{-1} \mathbf{I}_0 J_0^{-1}), \quad (3)$$

where $J_0 = \lim_{N \rightarrow \infty} \left[-\frac{1}{N} \frac{\partial^2 \Psi_N(y|X; \beta_0)}{\partial \beta \partial \beta'} \right]$,

$$\mathbf{I}_0 = \lim_{N \rightarrow \infty} \left[\frac{1}{N} \frac{\partial \Psi_N(y|X; \beta_0)}{\partial \beta} \frac{\partial \Psi_N(y|X; \beta_0)}{\partial \beta'} \right].$$

NOTE: In the ML case, $J_0 = -\mathbf{I}_0$.

Need for simulation

For cases where $f(y_n|X_n; \beta)$ or $E(y_n|X_n; \beta)$ are functions that are hard (or even impossible) to compute analytically.

As noted by Stern (1999), to implement GMM of ML in discrete choice modeling involves evaluating expectations of the form :

$$Eh(U) = \int h(u)f(u)du,$$

where U is a random variable with density $f(u)$. Simulators of the form

$$\hat{h}(U) = \frac{1}{S} \sum_{s=1}^S h(u^s),$$

will turn out to be good objects to replace $Eh(U)$. (Mostly due to the unbiasedness property.) The u^s 's represent draws of U from its distribution.

We now present examples where simulation is needed.

Problematic Examples

- Binary Probit with Random Heterogeneity

Consider the model where we allow for “unobserved” person specific variation in the β :

$$\begin{aligned}U_n &= X_n\beta_n + \varepsilon_n, & \varepsilon_n &\sim \mathbf{N}(0, 1), \\ \beta_n &= Z_n\beta + w_n, & w_n &\sim MVN(0, \Omega),\end{aligned}$$

which implies that $\beta_n \sim MVN(Z_n\beta, \Omega)$.

It is usually simpler to replace w_n with an equivalent formulation

$$w_n = \Omega^{1/2}u_n, \quad u_n \sim MVN(0, I_K),$$

where $\Omega^{1/2}$ is the Cholesky factorization matrix such that $\Omega^{1/2}\Omega^{1/2'} = \Omega$.*

*_____

Cholesky Decomposition

Let A be a positive definite matrix. Then one can always find a lower triangular matrix P such that $PP' = A$.

Example : let $A = \begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}$. Then $P = \begin{pmatrix} 1 & 0 \\ 2 & 1.4142 \end{pmatrix}$, and

$$PP' = \begin{pmatrix} 1 & 2 \\ 2 & 5.99996 \end{pmatrix}.$$

β_n can then be replaced with the rhs of:

$$\beta_n = Z_n\beta + \Omega^{1/2}u_n, \quad u_n \sim MVN(0, I_K),$$

where β_n is $(K \times 1)$ and is independent of ε_n .
 In this case, the utility formulation becomes:

$$U_n = X_n Z_n \beta + X_n \Omega^{1/2} u_n + \varepsilon_n, \quad \varepsilon_n \sim \mathbf{N}(0, 1),$$

so that given u_n , we have a binary Probit model as follows:

$$f(y_n = 1 | X_n; \beta, u) = \Phi(X_n [Z_n \beta + \Omega^{1/2} u]).$$

The unconditional probability is

$$\begin{aligned} f(y_n = 1 | X_n; \beta) &= \int \dots \int \Phi(X_n [Z_n \beta + \Omega^{1/2} u]) \\ &\quad * \prod_{k=1}^K \varphi(u_k) du_k, \\ &= E(\Phi(X_n [Z_n \beta + \Omega^{1/2} u])), \end{aligned}$$

which involves a K -dimensional integral over u ,
 a $(K \times 1)$ vector of standard normal variates.

- Multinomial Probit

With J alternatives, the MNP model may be written as:

$$\begin{aligned}
 U_n &= X_n\beta + \varepsilon_n, & \varepsilon_n &\sim MVN(0, \Sigma), \\
 y_n &= (y_{1n}, y_{2n}, \dots, y_{Jn})', & \text{where:} \\
 y_{in} &= 1 \text{ if } i \text{ is chosen, } 0 \text{ otherwise.}
 \end{aligned}$$

In a trinomial situation, we would have:

$$\begin{aligned}
 U_{1n} &= X_{1n}\beta + \varepsilon_{1n} \\
 U_{2n} &= X_{2n}\beta + \varepsilon_{2n} \\
 U_{3n} &= X_{3n}\beta + \varepsilon_{3n},
 \end{aligned}$$

where $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}$. In deviation wrt first alternative, we get:

$$\begin{aligned}
 U_{2n} - U_{1n} &= (X_{2n} - X_{1n})\beta - \varepsilon_{1n} + \varepsilon_{2n} \\
 U_{3n} - U_{1n} &= (X_{3n} - X_{1n})\beta - \varepsilon_{1n} + \varepsilon_{3n}.
 \end{aligned}$$

The probability of selecting alternative 1 corresponds to

$$\begin{aligned}
 P_n(1) &= P(U_{1n} > U_{2n}, U_{1n} > U_{3n}) \\
 &= P[\varepsilon_{2n} < (X_{1n} - X_{2n})\beta + \varepsilon_{1n}, \\
 &\quad \varepsilon_{3n} < (X_{1n} - X_{3n})\beta + \varepsilon_{1n}], \\
 P_n(1) &= \int_{\varepsilon_1=-\infty}^{\infty} \int_{\varepsilon_2=-\infty}^{(X_{1n}-X_{2n})\beta+\varepsilon_1} \\
 &\quad \int_{\varepsilon_3=-\infty}^{(X_{1n}-X_{3n})\beta+\varepsilon_1} n(\varepsilon; 0, \Sigma) d\varepsilon,
 \end{aligned}$$

where $n(\varepsilon; 0, \Sigma)$ denotes the density of a multivariate normal distribution with mean vector 0 and variance-covariance matrix Σ evaluated at ε . In the general case with J alternatives, the choice probability associated with alternative 1 is written as:

$$\begin{aligned}
 P_n(1) &= \int_{\varepsilon_1=-\infty}^{\infty} \int_{\varepsilon_2=-\infty}^{(X_{1n}-X_{2n})\beta+\varepsilon_1} \\
 &\quad \dots \int_{\varepsilon_J=-\infty}^{(X_{1n}-X_{Jn})\beta+\varepsilon_1} n(\varepsilon; 0, \Sigma) d\varepsilon, \quad (4)
 \end{aligned}$$

a J -dimensional integral. (Can be decreased by 1).

Maximum likelihood :

The log lik. function is :

$$\Psi_N (y | X; \beta) = \sum_{n=1}^N \ln P_n(i_n),$$

where i_n designates the alternative chosen by n .

Generalized Method of Moments:

The criterion function is

$$\Psi_N (y | X; \beta) = h' \Omega h,$$

where h is a $(NJ \times 1)$ vector that vertically stacks the $(J \times 1)$ vectors $h_n = y_n - E(y_n)$. The matrix Ω is a $(NJ \times NJ)$ given matrix which, because the obs. are independent, has a block diagonal structure. The $E(y_n)$ is a $(J \times 1)$ of probabilities with elements $E(y_{in}) = P_n(i)$. Both estimation approaches require the evaluation of multidimensional integrals.

Numerical Integration

- Unidimensional Integrals

When an integral does not exist in closed form, it may be numerically approximated. Approximations to an integral of the form:

$$F(x) = \int_L^U f(x)dx,$$

are variants of the well-known Simpson's rule. One popular technique is the Gaussian quadrature, the Gauss-Legendre being the most popular. That integral gets approximated using :

$$F(x) \approx \sum_{l=1}^L f(a_l)w_l,$$

where w_l and a_l are quadrature weights and abscissa, respectively. More details may be found in references listed in GREENE's book. The more points are used (L) the more accurate the approximation is.

- Bidimensional Integrals

Let's say that we wish to evaluate $E(XY)$, where X, Y are jointly distributed with density $f(x, y)$. Then,

$$E(XY) = \int_x \int_y xy f(x, y) dy dx$$

would be evaluated as

$$E(XY) \approx \sum_i \sum_j a_i a_j f(a_i, a_j) w_i w_j.$$

Note : Given a sample of N obs. of (X, Y) 's, $1/N \sum_{n=1}^N X_n Y_n$ would provide an unbiased estimator for $E(XY)$.

- Computing Multidimensional Normal Integrals

- To use numerical techniques (Gaussian Quadrature) in situations involving more than three dimensions is too demanding in computing time if numerical integration is used and not accurate enough is approximations are used.

Table 1: Computing times (points against dimensions)

# points	Dimensions		
	3	4	5
10	1 sec.	6 sec.	1 m. 07 sec.
20	3 sec.	1 m. 20 sec.	32 m.
40	15 sec.	20 m. 50 sec.	16 hrs. 40 min.

(Computed on a Sun SparcStation 2.) In the last case, $40^5 = 1.6$ billion points are visited.

Monte Carlo Integration

- – To understand this integration technique is critical because it serves as the backbone for most simulators used for estimation. Consider the computation of the integral:

$$F(x) = \int_L^U \pi(x) g(x) dx \quad (5)$$

$$= K \int_L^U \pi(x) \frac{g(x)}{K} dx$$

$$= K \int_L^U \pi(x) h(x) dx$$

$$= K E_{h(x)}[\pi(x)], \quad (6)$$

where $K = \int_L^U g(x) dx$ is a normalizing constant which ensures that $\int_L^U h(x) dx = 1$.

If it is simple to sample x^s values from the $h(x)$ distribution, then a good approximation for the integral in (5) is :

$$\hat{F}(x) = K \frac{1}{S} \sum_{s=1}^S \pi(x^s).$$

As we know, the above will be an unbiased and consistent estimator of the integral.

If sampling from $h(x)$ is not simple and if there exists another density $I(x)$ with the following properties: it is defined over the same support, it is somehow not too different from $h(x)$, it is easy to sample from, and $\pi(x) h(x) / I(x)$ is bounded and smooth over the support of x , then:

$$\begin{aligned} F(x) &= K \int_L^U \pi(x) h(x) dx \\ &= K \int_L^U \frac{\pi(x) h(x)}{I(x)} I(x) dx, \end{aligned}$$

can be well approximated using:

$$\tilde{F}(x) = K \frac{1}{S} \sum_{s=1}^S \frac{\pi(x^s) h(x^s)}{I(x^s)}.$$

The $I(x)$ density is known as the importance function and the technique is called importance sampling.

- In the case of the MNP model, good solutions exist. Recall the MNP choice probability associated with alternative 1 in equation (4) and note that it can be written as:

$$P_n(1) = \int \cdots \int \mathbf{1}_{[U_{1n} > U_{jn}, \forall j \neq 1]} n(\varepsilon; 0, \Sigma) d\varepsilon, \quad (7)$$

making then the integral unbounded. The choice probability has the form of an expectation.

Because of that,

$$\frac{1}{S} \sum_{s=1}^S \mathbf{1}_{[U_{1n}^s > U_{jn}^s, \forall j \neq 1]} = \frac{S_1}{S} \quad (8)$$

would be an unbiased simulator for the probability where one would take S random draws from the appropriate distribution. This is the Lerman and Manski (1976) frequency simulator. In order to be able to compute it, one needs to know how to draw numbers from a given distribution.

Implementation : For a given subject n , draw S values of the ε_n vector from $MVN(0, \Sigma)$. Count the number of times $U_{in}^s > U_{jn}^s, \forall j \neq i$, to get S_i . A simulator for the probability $P_n(i)$ is the frequency S_i/S computed for a given subject n . In the trinomial case, it would be computed using the equations :

$$\begin{aligned} U_{1n}^s &= X_{1n}\beta + \varepsilon_{1n}^s \\ U_{2n}^s &= X_{2n}\beta + \varepsilon_{2n}^s \\ U_{3n}^s &= X_{3n}\beta + \varepsilon_{3n}^s. \end{aligned}$$

The most severe weakness of this simulator is its non-differentiability wrt the model parameters. The normal probability simulators that have been suggested in the late 80's replace the indicator function in (7) with smooth (differentiable) simulators with good statistical properties. The best of all so far is the GHK simulator.

It is described in details in Bolduc(1999, TR-B). It has the form of an expectation and it can be calculated as:

$$g_n(\mathbf{1}|C) = \frac{1}{S} \sum_{s=1}^S \prod_{j=1}^J \Phi_{jn}(a_{jn}^s)$$

where a_{jn}^s are recursively computed. Choice probability simulators in the MNP context are thoroughly presented in Hajivassiliou, McFadden and Ruud , *Journal of Econometrics*, 1996. (See also Stern, 1999 sect. 2.2 to 2.4).

3 Analytical background

In order to implement simulation-based econometric methods, one needs to learn few basic techniques. It is essential to become familiar with : 1) random number generation and 2) simulation of variates from given distributions.

- Random Number Generation

To make sampling from any distribution, one usually only needs a $U(0, 1)$ random generator. Actually, random number generators are small computer programs. They are pseudo-random number generators, because they are just deterministic recursions. Starting with an initial condition (called a seed), it will give exactly the same sequence of numbers.

$$\text{newseed} = (a * \text{oldseed}) \% m \quad ; \quad x = \text{newseed} / m;$$

where : $a=397204094$ and $m=2^{31}-1$. Starting seed is any positive integer $< m$.

- Sampling from Continuous Distributions

Once a sequence of $U(0, 1)$ is available, there are several ways to transform them to a desired distribution. The most useful: *inversion technique*. Let x be a univariate random variable with known cumulative distribution function $F(\cdot)$. It is well known that $F(x)$ will take a value between 0 and 1 with an equal probability. This suggests:

1. Draw a q value from a $U(0, 1)$ distribution
2. Invert $F(x) = q$ to get $x = F^{-1}(q)$.

The x thus obtained will come from F . (Antithetic acceleration, Stern 1999, p. 12).

Note: The inversion technique can also be used with discrete random variables. Let Y be a discrete random variable taking the value i with probability p_i , and let $P_i = pr(Y \leq i) = \sum_{j=1}^i p_j$. Let $Z \sim \text{Uniform}(0, 1)$ and let $Y = i$ iff $P_{i-1} < Z \leq P_i$, with $P_0 = 0$, then Y is distributed as desired.

Examples :

- – *Exponential distribution*

$$F(x) = 1 - \exp(-\beta x), \quad x \geq 0, \beta > 0, \text{ and}$$

$$\text{therefore, } x = -\frac{1}{\beta} \ln(1-q), \quad q \sim U(0, 1),$$

is exponential.

- *Standard Normal*

$\Phi(x) = q$, implies that $x = \Phi^{-1}(q)$, is a random normal variate generator. The standard normal CDF $\Phi(\cdot)$ is available in most computer packages but not the inverse standard normal CDF.

- *Truncated Standard Normal*

Let u be a standard normal variate with density ϕ and CDF Φ . If u is forced to belong to an interval $D = (L, U)$, then $u | u \in D$ will be the truncated normal with density:

$$g(u | u \in D) = \frac{\phi(u)}{\Phi(U) - \Phi(L)}.$$

Two simple methods exist for drawing from this truncated distribution.

* Inversion Technique

The last equation allows one to write:

$$F(x) = \frac{\Phi(x) - \Phi(L)}{\Phi(U) - \Phi(L)},$$

and therefore, the sampler becomes:

$$x = \Phi^{-1}[q \{ \Phi(U) - \Phi(L) \} + \Phi(L)].$$

* Crude Acceptance-Rejection (CAR) Method

This technique is the simplest to implement. To get draws for $u | u \in D$, draw u variates from the standard normal distribution and keep only those values of u that satisfy the constraint $u \in D$. This is less efficient than the inversion technique because you may need to draw quite a large number of u values in order to be able to get the required number of draws. With the inversion technique, each draw is automatically qualified.

Remark: Once it is possible to draw u values from that truncated distribution, it then becomes simple to compute expected values. For instance:

$$E(u | u \in D) = \int_L^U u \frac{\phi(u)}{\Phi(U) - \Phi(L)} du, \quad (9)$$

could be evaluated using a Gaussian quadrature or, more simply, taking the mean of a set of S values u^s drawn from the truncated distribution. This would produce:

$$E(u | u \in D) \approx \frac{1}{S} \sum_{s=1}^S u^s.$$

This is an acceptable way to proceed, because empirical means of random samples are unbiased and consistent estimators of their population counterparts. As seen above, to replace integrals with empirical means is known as *Monte Carlo Integration*.

– *Multivariate Normal*

To sample x from a $MVN(\mu, \Omega)$, use the Cholesky factorization matrix $\Omega^{1/2}$, and a u vector of standard normal variates as follows: $x = \mu + \Omega^{1/2}u$. (Check that x has the right distribution.)

4 Simulation-Based Estimation

Given the background just covered, we are now ready to approach simulation-based econometric estimation from a general perspective. Basically we will show how the classical ML and GMM settings can be modified to accommodate situations where $f(y_n|x_n; \beta)$ or $E(y_n|x_n; \beta)$ are hard to compute analytically.

- Maximum Simulated Likelihood (MSL)

In the MSL setting, the log. likelihood function

$$\Psi_N(y | x; \beta) = \sum_{n=1}^N \ln f(y_n | x_n; \beta).$$

is replaced with:

$$\Psi_{SN}(y | x; \beta) = \sum_{n=1}^N \ln \left[\tilde{f}(y_n | x_n; \beta) \right]$$

where

$$\tilde{f}(y_n | x_n; \beta) = \frac{1}{S} \sum_{s=1}^S \tilde{f}(y_n | x_n, u_n^s; \beta)$$

is an unbiased simulator for $f(y_n | x_n; \beta)$.

– *Example: The Logit Kernel Model*

This model is written as:

$$\begin{aligned} U_n &= X_n \beta + \varepsilon_n + \nu_n, \\ \varepsilon_n &\sim MVN(0, \Sigma), \\ \nu_n &\sim \text{Gumbel i.i.d.} \end{aligned} \tag{10}$$

In this case, one can write:

$$P_n(i) = \int \dots \int \Lambda_n(i|\varepsilon) n(\varepsilon; 0, \Sigma),$$

For this last expression, an unbiased simulator is:

$$\tilde{P}_n(i|C) = \frac{1}{S} \sum_{s=1}^S \Lambda_n(i | \varepsilon_n^s),$$

where

$$\Lambda_n(i | \varepsilon_n^s, C) = \frac{\exp(X_{in}\beta + \varepsilon_{in}^s)}{\sum_{j=1}^J \exp(X_{jn}\beta + \varepsilon_{jn}^s)}$$

denotes the MNL choice probability given a value ε_n^s of ε_n , in (10).

– *Statistical Inference*

If one is using a simulator that is unbiased and consistent, then it is clear that as $S \rightarrow \infty$, MSL and ML become identical problems and therefore, the asymptotic results are identical to those obtained in our review of the classical methods. Now a few remarks:

1. If S is fixed and small, and $N \rightarrow \infty$, the MSL estimator for β is inconsistent because the simulator enters non linearly into the likelihood function .
2. Given the results of our previous calculations where 1.6 billion points were visited, the computing cost of doing 1000 simulations draws still look insignificant. Until we get more definitive answers on the ultimate value for S , it is advisable to perform sensitivity analyses with S varying from small to huge values.

- Generalized Method of Simulated Moments (GMSM)

GMM finds the value of β that makes $E(y_n|x_n; \beta)$ as close as possible to y_n .

- The GMSM estimator that we call β_{GMSM} is obtained when we replace in all those equations the hard to compute functions $E(y_n|x_n; \beta)$ with an unbiased simulator.

- *Example*

In the MNP model, to simulate $P_n(1)$, we could use :

$$\frac{1}{S} \sum_{s=1}^S \mathbf{1}_{[U_{1n}^s > U_{jn}^s, \forall j \neq 1]} = \frac{S_1}{S}$$

which is the Lerman and Manski frequency simulator.

– *Statistical Inference*

Let the simulated criterion function be

$$\Psi_{SN} (y | x; \beta) = h' \Omega h$$

with $h = y - \tilde{E}(y|x; \beta)$, where y is a vector that vertically stacks the $(J \times 1)$ vectors of choice y_n . $\tilde{E}(y|x; \beta)$ is arranged accordingly. See Gourieroux and Monfort for a derivation. To use Ω is the same as using GLS over OLS.

Remarks:

- * As mentioned in Sect 3.1 of Stern's paper, GMSM has good properties for a finite S as opposed to MSL which needs $S \rightarrow \infty$ to achieve optimality.
- * Both MSL and GMSM lead to a var. cov. matrix of the estimators that is larger than the usual ML and GMM matrices because of the extra noise introduced by the simulation. This extra contribution disappears as $S \rightarrow \infty$.

4.1 Choice between GSM and MSL

- MSL

- Advantages

- * Easy to implement because it is still just like optimization of a standard ML criterion.
 - * If S is large, MSL is as efficient as ML and ML properties are well known.
 - * Objective function only depends on the probability of the chosen alternative.
 - * Well behaved.
 - * Table 1 of Stern indicates that MSL is significantly faster than GSM.
 - * GHK is the choice prob. simulator to favor.

– Disadvantages

- * Needs large S to get ride of the bias introduced because of the natural log.
- * An MSL formulation with very large draws may take as long to estimate as a GSM one with small number of draws where one has to calculate the probability of each alternative in the choice set.

● GSM

– Advantages

- * Have good properties with finite value of S . Because simulators enter linearly the first-order condition h . Precision is insured because the sum is taken over individual observations as well.
- * The criterion function exploits the minimum distance principle that leads to least squares type estimators and is easy to implement.

– Disadvantages

- * Behavior of criterion function as you iterate may not be as smooth as one would like. Erratic behavior has many times been noticed. This is from empirical experience.
- * In order to compute this function, one has to calculate the probability of each alternative in the choice set. MSL needs only the probability of the chosen alternative.