**Bar-Ilan University**
**Estimation of DP Models**
**March, 2017**

**Moshe Buchinsky**
**Department of Economics**
**UCLA**

# Lecture Note 3

# Simulation Methods

Simulation estimators are estimators that utilize simulated random variables to ease computational difficulties. Many estimators require the evaluation of integrals in order to compute the estimators. If the dimension of the integrals is low (one, two, three, or even four), standard quadrature (i.e., numerical integration) methods can be used. If the dimension is high, then these methods are ineffective and other methods must be employed. Monte Carlo integration and quasi-Monte Carlo integration are two such methods.

As an example, consider maximum likelihood estimation of a panel binary probit model with serial correlation in the latent variables.

## Monte Carlo Integration:

See Ch. 5 of Hammersley and Handscomb (1964), Monte Carlo Methods.

Consider the evaluation of the simple integral

$$I_1 = \int h(x)f(x)dx \ , \quad \text{where} \quad \int h^2(x)f(x)dx < \infty$$

and $f(x)$ is the density of some random variable or random vector $X$. Standard numerical methods approximate $I_1$ by breaking up the domain of $hf$ into a grid and approximating $hf$ by a constant on each interval (or hyper-cube). This approach runs into difficulty in higher dimensional problems. For example, consider the integral

$$I_2 = \int_0^1 \cdots \int_0^1 h(\underset{\sim}{x})f(\underset{\sim}{x})dx_1, \ ..., \ dx_J \ .$$

For a grid of mesh-size $\varepsilon$, one needs $1/\varepsilon^J$ points when the domain is $[0,1]^J$. For example, if $\varepsilon = .01$ and $J = 10$, then $1/\varepsilon^J = 10^{20}$. Computing $I_2$ by approximating $h(\underset{\sim}{x})f(\underset{\sim}{x})$ by a constant on each of $10^{20}$ hyper-cubes is very costly, if not impossible.

Monte Carlo integration is an alternative method of approximating integrals such as $I_2$.

Let $X_1$, ..., $X_R$ be $R$ iid random variables each with density $f$. Then,

$$Eh(X_r) = \int h(x)f(x)dx = I_1 \ , \quad \forall r = 1, \ ..., \ R \ .$$

Furthermore,

$$E\frac{1}{R}\Sigma_{r=1}^{R}h(X_r) = I_1 \ .$$

Thus,

$$\widehat{I}_{CF} = \frac{1}{R}\Sigma_{r=1}^{R}h(X_r)$$

is an unbiased estimator of $I_1$. Here, $CF$ stands for crude frequency. $\widehat{I}_{CF}$ is called the crude frequency simulator of $I_1$. The variance of $\widehat{I}_{CF}$ is

$$\text{Var}(\widehat{I}_{CF}) = \text{Var}(h(X_r))/R = \int (h(x) - I_1)^2 f(x)dx/R \ .$$

We can estimate $Var(\widehat{I}_{CF})$ by

$$\widehat{\text{Var}}(\widehat{I}_{CF}) = \frac{1}{R}\Sigma_{r=1}^{R}(h(X_r) - \widehat{I}_{CF})^2/R \ .$$

Example: Consider the following case:

$$h(x) = \frac{e^x - 1}{e - 1} \quad \text{and} \quad f(x) = \begin{cases} 1 & \text{for } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \ .$$

In this case, $I_1 = .418$. For illustrative purposes, suppose we cannot calculate $I_1$ analytically. Instead, we use the $CF$ simulator to approximate $I_1$. We have $Var^{1/2}(h(X_r)) = .286$. Take $R = 16$. Sixteen $U[0,1]$ random numbers (from a table of random numbers) are .96, .28, .21, .94, .35, .40, .10, .52, .18, .08, .50, .83, .73, .25, .33, .34. These random numbers yield

$$\begin{aligned} &\widehat{I}_{CF} = .357 \ , \\ &\widehat{\text{Var}}(\widehat{I}_{CF})^{1/2} = .07 \ , \quad \text{and} \\ &|\widehat{I}_{CF} - I_1| = .061 \ . \end{aligned}$$

Next, we consider the case where a crude frequency simulator is used in an estimation problem. Suppose an integral of the following type arises in a likelihood function:

$$I_2(\theta) = \int 1(\underset{\sim}{x} \in D_\theta) q(\underset{\sim}{x}, \theta) f(\underset{\sim}{x}) d\underset{\sim}{x} \ ,$$

where $\underset{\sim}{x} \in R^J$, $f(\underset{\sim}{x})$ is a density function, $q(\underset{\sim}{x}, \theta)$ is a function that depends on $\underset{\sim}{x}$ and the unknown parameter $\theta$, and $D_\theta$ is a subset of $R_J$ (possibly unbounded) that may depend on

$\theta$. For example, suppose we want to calculate

$$\begin{aligned} &P_\theta(\underset{\sim}{Y} \geq \underset{\sim}{0}) \quad \text{where} \\ &\underset{\sim}{Y} \sim N(\mu, \Omega) \ , \quad \theta = (\mu', \, \text{vec}h(\Omega)')' \ . \end{aligned}$$

Let $\underset{\sim}{X} \sim N(0, \, I_J)$, $\Omega = \Gamma\Gamma'$, and $\underset{\sim}{Y} = \mu + \Gamma \underset{\sim}{X}$. Then

$$\begin{aligned} P_\theta(\underset{\sim}{Y} \geq \underset{\sim}{0}) \ &= \ P(\Gamma X \geq -\mu) \\ &= \ \int 1(\underset{\sim}{x} \geq -\Gamma^{-1}\mu) \prod_{j=1}^{J} \phi(x_j) dx_j \ . \end{aligned}$$

This is of the form $I_2(\theta)$ with $D_\theta = \{\underset{\sim}{x} : \underset{\sim}{x} \geq -\Gamma^{-1}\mu\}$, $q(\underset{\sim}{x}, \theta) = 1$, and $f(\underset{\sim}{x}) = \prod_{j=1}^{J} \phi(x_j)$.

If we can draw $R$ iid draws $\underset{\sim}{X}_1$, ..., $\underset{\sim}{X}_R$ from the density $f(\cdot)$, then the crude frequency simulator of $I_2(\theta)$ is

$$\widehat{I}_{CF}(\theta) = \frac{1}{R} \Sigma_{r=1}^{R} 1(\underset{\sim}{X}_r \in D_\theta) q(\underset{\sim}{X}_r, \, \theta) \ .$$

In the case where $I_2(\theta) = P_\theta(\underset{\sim}{Y} \geq 0)$, we need to be able to draw iid $N(0, 1)$ random variables. Standard packages, such as GAUSS and Matlab, have built in random number generators for $N(0, 1)$ random variables, as well as $U[0, 1]$ random variables.

If one needs to draw a real-valued random variable with continuous distribution function $F$, one can use the probability integral transform:

$$F^{-1}(U_r) \sim F \ \text{ if } \ U_r \sim U[0, 1] \ .$$

(To prove this, write $P(F^{-1}(U_r) \leq x) = P(U_r \leq F(x)) = F(x)$.) If the inverse df $F^{-1}(\cdot)$ has closed form, this method works well.

Note that $\widehat{I}_{CF}(\theta)$ is often a discontinuous function of $\theta$ even if $I_2(\theta)$ is continuous. For example, this is true when $I_2(\theta) = P_\theta(\underset{\sim}{Y} \geq \underset{\sim}{0})$. This occurs because

$$\widehat{I}_{CF}(\theta) = \frac{1}{R} \sum_{r=1}^{R} 1(\underset{\sim}{X}_r \geq -\Gamma^{-1}\mu)$$

is a step function with steps that depend on $\mu$ and $\Gamma$. Discontinuity of $\widehat{I}_{CF}(\theta)$ is a nuisance because it prevents one from using standard optimization methods (which rely on derivatives) to maximize the likelihood function when the likelihood function is evaluated using the $CF$ simulator. (A common optimization algorithm for discontinuous criterion functions is the Nelder–Meade simplex method.) It also prevents one from using standard asymptotic distribution theory, which relies on derivatives (but more sophisticated asymptotic distribution

can be applied).

## Importance Sampling:

The crude frequency simulator generally is not very efficient. That is, its variance is relatively large compared to that of more sophisticated methods. Three more sophisticated methods are (i) importance sampling, (ii) control variate methods, and (iii) antithetic variate methods.

To explain importance sampling, we return to approximation of $I_1$. We can rewrite $I_1$ as

$$I_1 = \int \frac{h(x)f(x)}{g(x)} g(x)dx \ .$$

Suppose $g(\cdot)$ is a density and we can draw iid random variables $X_1$, ..., $X_R$ with density $g$. Then,

$$\widehat{I}_{IS} = \frac{1}{R}\Sigma_{r=1}^{R} \frac{h(X_r)f(X_r)}{g(X_r)}$$

is an unbiased estimator of $I_1$. Its variance is

$$\text{Var}(h(X_r)f(X_r)/g(X_r))/R = \int (h(x)f(x)/g(x) - I_1)^2 g(x)dx/R \ .$$

Depending on the choice of $g(\cdot)$, this variance may be smaller than that of $\widehat{I}_{CF}$. In particular, if the shape of $g(\cdot)$ mimics that of $h(\cdot)f(\cdot)$ well, then $h(\cdot)f(\cdot)/g(\cdot)$ will be close to a constant and $Var(h(X_r)f(X_r)/g(X_r))$ will be close to zero. If $g(x) = ch(x)f(x)$ for a constant $c$, then $Var(h(X_r)f(X_r)/g(X_r))$ is identically zero and $\widehat{I}_{IS} = I_1$. But, for $g$ to be a density, one needs $\int g(x)dx = \int ch(x)f(x)dx = 1$. That is, one needs to take $c = 1/\int_0^1 h(x)f(x)dx = 1/I_1$. Hence, in order to draw from the density $ch(x)f(x)$, one needs to know $c$, and equivalently, $I_1$. In consequence, drawing from $g(\cdot) = ch(\cdot)f(\cdot)$ is not feasible. Nevertheless, we would like to choose $g(\cdot)$ such that $g(\cdot)$ is close to being proportional to $h(\cdot)f(\cdot)$ and such that it is easy to simulate rv's with density $g(\cdot)$.

**Example:** Suppose $h(x) = \frac{e^x - 1}{e-1}$, $f(x) = 1(x \in [0,1])$, and we take $g(x) = 2x$. Then, the df $G$ and the inverse df $G^{-1}$ corresponding to the density $g$ are

$$G(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ x^2 & \text{for} \quad 0 < x < 1 \\ 1 & \text{for} \quad x \geq 1 \end{cases} \quad \text{and} \quad G^{-1}(x) = \begin{cases} 0 & \text{for} \quad x \leq 0 \\ x^{1/2} & \text{for} \quad 0 < x < 1 \\ 1 & \text{for} \quad x \geq 1 \ . \end{cases}$$

We can draw $R$ iid random variables $X_1$, ..., $X_R$ by drawing $R$ iid $U[0,1]$ random variables

$\xi_1$, ..., $\xi_R$ and letting $X_r = G^{-1}(\xi_r) = \xi_r^{1/2}$ for $r = 1$, ..., $R$. Then,

$$\widehat{I}_{IS} = \frac{1}{R}\Sigma_{r=1}^{R}\frac{e^{X_r}-1}{(e-1)2X_r} = \frac{1}{R}\Sigma_{r=1}^{R}\frac{e^{\xi_r^{1/2}}-1}{(e-1)2\xi_r^{1/2}} \ .$$

In this example,

$$\mathrm{Var}(\widehat{I}_{CF})/\mathrm{Var}(\widehat{I}_{IS}) = 29.9 \ ,$$

so the importance sampling simulator is much more efficient than the crude frequency simulator.

## Control Variate Method:

Given a function $m(\cdot)$ on $[0,1]$, we can rewrite $I_1$ as

$$I_1 = \int h(x)f(x)dx = \int m(x)f(x)dx + \int (h(x)-m(x))f(x)dx \ .$$

Suppose $m(\cdot)$ is such that (i) $\int m(x)f(x)dx$ can be calculated analytically, i.e., $\int m(x)f(x)dx$ is known, and (ii) $m(\cdot)$ mimics $h(\cdot)$ sufficiently well that $h(x)-m(x)$ is close to a constant and, hence, $Var(h(X_r)-m(X_r))$ is close to zero, where $X_r$ has density $f$. Then, the control variate simulator

$$\widehat{I}_{CV} = \int m(x)f(x)dx + \frac{1}{R}\Sigma_{r=1}^{R}(h(X_r)-m(X_r))$$

will be relatively efficient. Its variance is

$$\begin{aligned}\mathrm{Var}(\widehat{I}_{CV}) &= \mathrm{Var}(h(X_r)-m(X_r))/R \\ &= [\mathrm{Var}(h(X_r)) + \mathrm{Var}(m(X_r)) - 2\mathrm{Cov}(h(X_r)),\, m(X_r))]/R \ .\end{aligned}$$

In this case, $\frac{1}{R}\Sigma_{r=1}^{R}m(X_r)$ is called a control variate for $\frac{1}{R}\Sigma_{r=1}^{R}h(X_r)$, the $CF$ simulator. A good control variate is one that is positively correlated with $\widehat{I}_{CF}$. Also, one needs to know $\int m(x)f(x)dx$ in order for $\frac{1}{R}\Sigma_{r=1}^{R}m(X_r)$ to serve as a control variate.

In our example, we might take $m(x) = x$ to be a control variate, since we can calculate $\int_0^1 xdx = \frac{1}{2}$ easily and since $m(x)$ and $h(x)$ resemble each other. With this choice,

$$\mathrm{Var}(\widehat{I}_{CF})/\mathrm{Var}(\widehat{I}_{CV}) = 60.4 \ ,$$

so the $CV$ simulator works well.

Note that the CV method can be used to improve the efficiency of any simulator, not just a $CF$ simulator. For example, one could use the CV method to improve the efficiency of an importance sampling simulator.

## Antithetic Variate Method:

This method is based on using the $CF$ simulator plus a second simulator that is unbiased for $I_1$ and is negatively correlated with the $CF$ simulator. The antithetic variate simulator is then the average of the two simulators. Let $\widehat{I}_{AV}$ be an antithetic variate. Then, the antithetic variate simulator is

$$\widehat{I}_{AVS} = (\widehat{I}_{CF} + \widehat{I}_{AV})/2 \ .$$

The $AV$ simulator is unbiased for $I_1$ and has variance

$$\text{Var}(\widehat{I}_{AVS}) = \text{Var}(\widehat{I}_{CF}) + \text{Var}(\widehat{I}_{AV}) + 2\text{Cov}(\widehat{I}_{CF},\, \widehat{I}_{AV}) \ .$$

$Var(\widehat{I}_{AVS})$ is small when $Cov(\widehat{I}_{CF},\, \widehat{I}_{AV})$ is negative.

The $AV$ method does not require that the first simulator is the $CF$ simulator. One could construct an $AV$ simulator when starting with an $IS$ or $CV$ simulator.

In our example with $h(x) = \frac{e^x - 1}{e - 1}$, an antithetic variate for the $CF$ simulator is

$$\widehat{I}_{AV} = \frac{1}{R}\Sigma_{r=1}^{R} h(1 - X_r) \ ,$$

where $X_r \sim U[0,1]$. Note that $1 - X_r \sim U[0,1]$. Also, $h(X_r)$ and $h(1 - X_r)$ are negatively correlated when $h$ is monotone (as it is).

In our example,

$$\text{Var}(\widehat{I}_{CF})/\text{Var}(\widehat{I}_{AVS}) = 62 \ ,$$

where $\widehat{I}_{AVS} = (\widehat{I}_{CF} + \widehat{I}_{AV})/2$.

## An Example of Binary Choice Model:

We will use the binary probit model to illustrate different simulation methods. In this model, we observe $(Y_1,\, X_1)$, ..., $(Y_n,\, X_n)$, where

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } \; Y_i^* > 0 \\ 0 & \text{otherwise,} \end{cases} \\ Y_i^* &= X_i'\theta_0 + U_i \ , \\ U_i &\sim \text{iid}\, N(0,1) \ . \end{aligned}$$

The likelihood function is

$$\begin{aligned} & \prod_{i=1}^n P_\theta(Y_i = 1)^{Y_i}(1 - P_\theta(Y_i = 1))^{1-Y_i} \\ = & \prod_{i=1}^n \Phi(X_i'\theta)^{Y_i}(1 - \Phi(X_i'\theta))^{1-Y_i} \ . \end{aligned}$$

The (normalized) log-likelihood function is

$$\ell_n(\theta) = \frac{1}{n}\Sigma_{i=1}^n (Y_i \log \Phi(X_i'\theta) + (1-Y_i)\log(1 - \Phi(X_i'\theta))) \ .$$

We pretend (for the purposes of the example) that we cannot compute $\Phi(X_i'\theta)$ by standard quadrature methods. Instead, we will use simulation methods.

The method of simulated likelihood (MSL) just replaces the unknown integrals $\Phi(X_i'\theta)$ in the likelihood function by simulated versions. The MSL estimator maximizes the simulated likelihood function. Generally we want to carry out the simulations such that the same sequence of underlying pseudo-random variables is used for each value of $\theta$. We have

$$\Phi(X_i'\theta) = \int 1(u \leq X_i'\theta)\phi(u)du \ .$$

Suppose $\varepsilon_{i1}$, ..., $\varepsilon_{iR}$ are iid $N(0,1)$ pseudo-random variables. Then, the crude frequency simulator of $\Phi(X_i'\theta)$ is

$$\widehat{\Phi}_{Ri}(X_i'\theta) = \frac{1}{R}\Sigma_{r=1}^R 1(\varepsilon_{ir} \leq X_i'\theta) \ .$$

The simulated likelihood estimator $\widehat{\theta}_{Rn}$ maximizes the simulated likelihood function or simulated log likelihood function:

$$\widehat{\ell}_{Rn}(\theta) = \frac{1}{n}\Sigma_{i=1}^n (Y_i \log \widehat{\Phi}_{Ri}(X_i'\theta) + (1-Y_i)\log(1 - \widehat{\Phi}_{Ri}(X_i'\theta))) \ ,$$

where $\widehat{\Phi}_{Ri}(X_i'\theta)$ is some simulator of $\Phi(X_i'\theta)$ such as the $CF$ simulator above.

What is needed in order for $\widehat{\theta}_{Rn}$ to be consistent? Remember the consistency result of Lecture 1 for extremum estimators. Assumptions EE, ID, and W–CON are sufficient for $\widehat{\theta}_{Rn}$ to converge in probability to the value $\theta^*$ that minimizes $Q(\theta)$, the probability limit of $Q_n(\theta) = \widehat{\ell}_{nR}(\theta)$. What is $Q(\theta)$? If we use different (and independent) simulated rv's for each observation, then $\widehat{\ell}_{Rn}(\theta)$ is an average of iid rv's and a law of large numbers can be applied. If we use the same simulated rv's for each observation, i.e., if $\varepsilon_{ir}$ is the same for all $i$, then a law of large numbers does not apply. We recommend using different simulated rv's for each observation. In this case, if $R$ is fixed as $n \to \infty$, then $Q_n(\theta)$ converges in probability to

$$\begin{aligned} Q_{(R)}(\theta) &= EY_i \log \widehat{\Phi}_{Ri}(X_i'\theta) + E(1-Y_i)\log(1-\widehat{\Phi}_{Ri}(X_i'\theta)) \\ &= E_X[\Phi(X_i'\theta_0)E_\varepsilon \log \Phi_{Ri}(X_i'\theta)] + E_X[(1-\Phi(X_i'\theta_0))E_\varepsilon \log(1-\widehat{\Phi}_{Ri}(X_i'\theta_0))] \ , \end{aligned}$$

where $E_X$ denotes expectation with respect to $X_i$ and $E_\varepsilon$ denotes expectation with respect to $\varepsilon_{i1}$, ..., $\varepsilon_{iR}$. Note that

$$E_\varepsilon \widehat{\Phi}_{Ri}(X_i'\theta) = \Phi(X_i'\theta) \ ,$$

but

$$E_\varepsilon \log \widehat{\Phi}_{Ri}(X_i'\theta) < \log E\Phi_{Ri}(X_i'\theta) = \log \Phi(X_i'\theta)$$

and

$$E_\varepsilon \log(1-\widehat{\Phi}_{Ri}(X_i'\theta)) < \log(1-E_\varepsilon\widehat{\Phi}_{Ri}(X_i'\theta)) = \log(1-\Phi(X_i'\theta))$$

by Jensen's inequality. In consequence, the limit function $Q_{(R)}(\theta)$ does not equal the limit function for the case where $\Phi(X_i'\theta)$ can be calculated exactly. The latter is

$$Q^*(\theta) = E_X\Phi(X_i'\theta_0)\log\Phi(X_i'\theta) + E_X(1-\Phi(X_i'\theta_0))\log(1-\Phi(X_i'\theta)) \ .$$

There is no reason why $Q_{(R)}(\theta)$ is uniquely minimized at $\theta_0$. In general, it will not be and $\widehat{\theta}_{Rn}$ will be inconsistent when $R$ is held fixed as $n \to \infty$.

To obtain a consistent estimator one needs to let $R$ depend on $n$ and have $R = R_n \to \infty$ as $n \to \infty$. In this case, $\widehat{\ell}_{R_n n}(\theta)$ will converge in probability to $Q^*(\theta)$, the same function as when the likelihood function can be computed exactly. Thus, in this case, the MSL estimator $\widehat{\theta}_{Rn}$ will be consistent. In practice, this means that for given $n$, $R$ needs to be relatively large for the MSL estimator to avoid large bias.

To obtain asymptotic normality of $\widehat{\theta}_{Rn}$, one needs $R$ to diverge to infinity at a certain rate as $n \to \infty$. This imposes a more stringent requirement for $R$ to be large than that required for consistency. In practice, a suitable choice for $R$ depends on how accurate the

simulator is for given $R$. The better the simulator, the smaller $R$ can be.

Although different simulated rv's should be used for different observations, the same simulated rv's should be used for every value of $\theta$ for a given observation. The latter choice circumvents the "chattering effect" on the likelihood that occurs if different simulated rv's are used for different values of $\theta$. It also reduces the number of different rv's that need to be simulated.