1 Introduction

- In recent years, simulation methods have played a very important role in econometrics.
- Principle of integration via simulation ("Monte Carlo integration"): approximate an expectation as a sample average.

  Validity is ensured by law of large numbers. Let $x^1, x^2, \ldots, x^S$ be S i.i.d. draws from some distribution with density $f(X)$. Then

$$EX = \int x f(x) dx \approx \frac{1}{S} \sum_{s=1}^{S} x^s.$$

- Simulation can be a very useful tool for computing integrals, because most integrals can be written as an expectation. Also called "Monte Carlo Integration".
- To simulate from a distribution with CDF F, exploit quantile transform:
  - Draw $Z \sim U[0, 1]$.
  - Transform $X = F^{-1}(Z)$. Then $X \sim F$.

Here we consider several interesting applications of the simulation approach in econometrics.

2 Importance sampling

Importance sampling is a more efficient approach to simulation. In essence, you take draws from an alternative distribution whose support is concentrated in the truncation region. Principle of importance sampling:

$$\int_{\mathcal{F}} s f(s) ds = \int_{\mathcal{G}} s \frac{f(s)}{g(s)} g(s) ds.$$

That is, sampling s from $f(s)$ distribution equivalent to sampling $s * w(s)$ from $g(s)$ distribution, with importance sampling weight $w(s) \equiv \frac{f(s)}{g(s)}$. ($f$ and g should have the same support.)

Simple example You want to simulate the mean of a standard normal distribution, truncated to the unit interval [0,1]. The desired sampling density is:

$$f(x) = \frac{\phi(x)}{\int_0^1 \phi(x)dx}$$

where $\phi()$ denotes the standard normal density.

Brute force simulation: take draws x^s from $N(0,1)$, and only keep draws in [0,1]. Simulated mean is calculated as: $\frac{\sum_{s=1}^S x^s \cdot \mathbf{1}(x^s \in [0,1])}{\sum_{s=1}^S \mathbf{1}(x^s \in [0,1])}$. Inefficient if $\sum_{s=1}^S \mathbf{1}(x^s \in [0,1]) << S$.

Importance sampling: draw from U[0,1], so that $g(x) = 1$ for $x \in [0,1]$. For each draw, importance weight is $w^s = f(x^s) = \frac{\phi(x^s)}{\int_0^1 \phi(z)dz}$ Simulated mean is $\frac{1}{S}\sum_{s=1}^S x^s w^s$. Don't need to reject any draws.

2.1 GHK simulator: get draws from truncated multivariate normal distribution

You want draws from

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \sim TN\left(\vec{\mu}, \Sigma; \vec{a}, \vec{b}\right) \equiv N\left(\vec{\mu}, \Sigma\right) \text{ s.t. } \vec{a} < \vec{x} < \vec{b} \tag{1}$$

where the difficulty is that Σ is not necessarily diagonal (i.e., elements of $\vec{x}$ are correlated).

The most obvious "brute-force" approach to simulation is an acceptance-rejection procedure, where you take draws from $N(\vec{\mu}, \Sigma)$ (the untruncated distribution), but reject all the draws which lie outside the desired truncation region. If the region is small, this procedure can be very inefficient, in the sense that you might end up rejecting very many draws.

Importance sampling from truncated MVN Let $(u_1, \ldots, u_n)'$ denote an n-vector of independent multivariate standard normal random variables. Let $\Sigma^{1/2}$ denote the (lower-triangular) Cholesky factorization of Σ, with elements

$$\begin{bmatrix} s_{11} & 0 & \cdots & 0 & 0 \\ s_{21} & s_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & s_{ii} & 0 & 0 \\ s_{n1} & s_{n2} & \cdots & s_{nn-1} & s_{nn} \end{bmatrix}. \tag{2}$$

Then we can rewrite (1) as:

$$\vec{x} = \vec{\mu} + \Sigma^{1/2}\vec{u} \sim N(\vec{\mu}, \Sigma) \;\; s.t.$$

$$\begin{pmatrix} \frac{a_1-\mu_1}{s_{11}} \\ \frac{a_2-\mu_2-s_{21}u_1}{s_{22}} \\ \vdots \\ \frac{a_n-\mu_n-\sum_{i-1}^{n-1} s_{ni}u_i}{s_{nn}} \end{pmatrix} < \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} < \begin{pmatrix} \frac{b_1-\mu_1}{s_{11}} \\ \frac{b_2-\mu_2-s_{21}u_1}{s_{22}} \\ \vdots \\ \frac{b_n-\mu_n-\sum_{i-1}^{n-1} s_{ni}u_i}{s_{nn}} \end{pmatrix} \tag{3}$$

The above suggests that the answer is to draw $(u_1, \ldots, u_n)$ **recursively**. First draw u_1^s from $N\left(0, 1; \frac{a_1-\mu_1}{s_{11}}, \frac{b_1-\mu_1}{s_{11}}\right)$, then u_2^s from $N\left(0, 1; \frac{a_2-\mu_2-s_{21}u_1^s}{s_{22}}, \frac{b_2-\mu_2-s_{21}u_1^s}{s_{22}}\right)$, and so on.

Finally we can transform $(u_1^s, \ldots, u_n^s)$ to the desired $(x_1^s, \ldots, x_n^s)$ via the transformation

$$\vec{x}^s = \vec{\mu} + \Sigma^{1/2}\vec{u}^s. \tag{4}$$

Remark 1: It is easy to draw an n-dimensional vector $\vec{u}$ of independent truncated standard normal random variables with rectangular truncation conditions: $\vec{c} < \vec{u} < \vec{d}$. You draw a vector of independent uniform variables $\vec{\tilde{u}} \sim \mathcal{U}[\Phi(\vec{c}), \Phi(\vec{d})]$[1] and then transform $u_i = \Phi^{-1}(\tilde{u}_i)$.

Remark 2: The GHK simulator is an importance sampler. The importance sampling density is the multivariate normal density $N(\vec{\mu}, \Sigma)$ truncated to the region characterized in Eqs. (3). This is a recursively characterized truncation region, in that the

[1] Just draw $\hat{u}$ from $\mathcal{U}[0, 1]$ and transform $\tilde{u} = \Phi(c) + (\Phi(d) - \Phi(c))\hat{u}$.

range of, say, x_3 depends on the draw of x_1 and x_2. Note that truncation region is different for each draw. This is different than the multivariate normal density $N(\vec{\mu}, \Sigma)$ truncated to the region $(\vec{a} \leq \vec{x} \leq \vec{b})$.[2]

For the GHK simulator, the truncation probability for each draw $\vec{x}^s$ is given by:

$$\tau^s \equiv \left[\Phi\left(\frac{b_1 - \mu_1}{s_{11}}\right) - \Phi\left(\frac{a_1 - \mu_1}{s_{11}}\right)\right] \prod_{i=2}^{m} \left[\Phi\left(\frac{b_i - \mu_i - \sum_{j=1}^{i-1} s_{ij} u_j^s}{s_{ii}}\right) - \Phi\left(\frac{a_i - \mu_i - \sum_{j=1}^{i-1} s_{ij} u_j^s}{s_{ii}}\right)\right]. \tag{5}$$

Remark 3: τ^s (even just for one draw: cf. Gourieroux and Monfort (1996, pg. 99)) is an unbiased estimator of the truncation probability Prob $\left(\vec{a} < \vec{x} < \vec{b}\right)$. But in general, we can get a more precise estimate by averaging over w^s:

$$T_{\vec{a},\vec{b}} \equiv \text{Prob}\left(\vec{a} < \vec{x} < \vec{b}\right) \approx \frac{1}{S} \sum_s \tau^s \tag{6}$$

for (say) S simulation draws.

Hence, the importance sampling weight for each GHK draw is the ratio of the GHK truncation probability to the original desired truncation probability: $w^s \equiv \tau^s / T_{\vec{a},\vec{b}} \approx \tau^s / \frac{1}{S} \sum_s \tau^s$.

Hence, the GHK simulator for $\int_{\vec{a} \leq \vec{x} \leq \vec{b}} \vec{x} f(\vec{x}) d\vec{x}$, where $f(\vec{x})$ denote the $N(\vec{\mu}, \Sigma)$ density, is $\frac{1}{S} \sum_{s=1}^{S} \vec{x}^s w(x^s)$, or $\sum_s \vec{x}^s \tau^s / \sum_s \tau^s$.

2.2 Monte Carlo Integration using the GHK Simulator

Clearly, if we can get draws from truncated multivariate distributions using the GHK simulator, we can use these draws to calculate integrals of functions of $\vec{x}$. There are two important cases here, which it is crucial not to confuse.

[2] See Hajivassiliou and Ruud (1994), pg. 2005.

2.2.1 Integrating over untruncated distribution $\mathbf{F}(\vec{\mathbf{x}})$, but $\vec{\mathbf{a}} < \vec{\mathbf{x}} < \vec{\mathbf{b}}$ defines region of integration

If we want to calculate

$$\int_{\vec{a}<\vec{x}<\vec{b}} g(\vec{x}) f(\vec{x}) d\vec{x} \tag{7}$$

where f denotes the $N(\vec{\mu}, \Sigma)$ density, we can use the GHK draws to derive a Monte-Carlo estimate:

$$E_{\vec{a}<\vec{x}<\vec{b}} g(\vec{x}) \approx \frac{1}{S} \sum_s g(\vec{x}^s) * \tau^s. \tag{8}$$

Here the weight is just τ^s (not w^s), because the desired sampling distribution is the *untruncated* MVN density. The most widely-cited example of this is the likelihood function for the multinomial probit model (cf. McFadden (1989)):

Multinomial probit with K choices, and utility from choice k $U_k = X\beta_k + \epsilon_k$. Probability that choice k is chosen is probability that $\nu_i \equiv \epsilon_i - \epsilon_k < X\beta_i - X\beta_k$, for all $i \neq k$. For each parameter vector β, use GHK to draw S (K-1)-dimensional vectors $\vec{\nu}^s$ subject to $\vec{\nu} < (\vec{x\beta})$. Likelihood function is

$$\begin{aligned} \text{Prob}(k) &= \int_{\vec{\nu}} \mathbf{1}\left(\vec{\nu} < (\vec{x\beta})\right) f(\vec{\nu}) d\vec{\nu} \\ &= \int_{\vec{\nu}<(\vec{x\beta})} f(\vec{\nu}) d\vec{\nu} \\ &\approx \frac{1}{S} \sum_s \tau^s. \end{aligned} \tag{9}$$

2.3 Integrating over truncated (conditional) distribution $\mathbf{F}(\vec{\mathbf{x}}|\vec{\mathbf{a}} < \vec{\mathbf{x}} < \vec{\mathbf{b}})$.

The most common case of this is calculating conditional expectations (note that the multinomial probit choice probability is *not* a conditional probability!)[3].

[3]This is a crucial point. The conditional probability of choice k conditional on choice k is trivially 1!

If we want to calculate

$$E_{\vec{a}<\vec{x}<\vec{b}}g(\vec{x}) = \int_{\vec{a}<\vec{x}<\vec{b}} g(\vec{x})f(\vec{x}|\vec{a}<\vec{x}<\vec{b})d\vec{x} = \frac{\int_{\vec{a}<\vec{x}<\vec{b}} g(\vec{x})f(\vec{x})d\vec{x}}{Prob(\vec{a}<\vec{x}<\vec{b})}. \tag{10}$$

As before, we can use the GHK draws to derive a Monte-Carlo estimate:

$$E_{\vec{a}<\vec{x}<\vec{b}}g(\vec{x}) \approx \frac{1}{T_{\vec{a},\vec{b}}}\frac{1}{S}\sum_s g(\vec{x}^s) * \tau^s. \tag{11}$$

The crucial difference between this case and the previous one is that we integrate over a conditional distribution by essentially integrating over the unconditional distribution over the restricted support, but then we need to divide through by the probability of the conditioning event (i.e., the truncation probability).

An example of this comes from structural common-value auction models, where:

$$v(x,x) \equiv \mathcal{E}\left(v|x_1 = x, \min_{j\neq 1} x_j = x\right) =$$

$$\underbrace{\int\cdots\int}_{x_k\geq x,\ \forall k=3,\ldots,n} \mathcal{E}\left(v|x_1,\ldots,x_n\right)\ dF\left(x_3,\ldots,x_n|x_1=x,x_2=x,x_k\geq x,k=3,\ldots,n;\theta\right) =$$

$$\frac{1}{T_x}\ \underbrace{\int\cdots\int}_{x_k\geq x,\ \forall k=3,\ldots,n} \mathcal{E}\left(v|x_1,\ldots,x_n\right)\ dF\left(x_3,\ldots,x_n|x_1=x,x_2=x;\theta\right) \tag{12}$$

where F here denotes the conditional distribution of the signals $x_3,\ldots,x_n$, conditional on $x_1 = x_2 = x$, and T_x denotes the probability that $(x_k \geq x, k = 3,\ldots,n|x_1 = x, x_2 = x;\theta)$.

If we assume that $\vec{x} \equiv (x_1,\ldots,x_n)'$ are jointly log-normal, it turns out we can use the GHK simulator to get draws of $\tilde{x} \equiv \log\vec{x}$ from a multivariate normal distribution subject to the truncation conditions $\tilde{x}_1 = \tilde{x}, \tilde{x}_2 = \tilde{x}, \tilde{x}_j \geq \tilde{x}, \forall j = 3,\ldots,n$. Let $\mathcal{A}(x)$ denote the truncation region, for each given x.

Then we approximate:

$$v(x,x) \approx \frac{1}{T_{\mathcal{A}(x)}}\frac{1}{S}\sum_s \mathcal{E}\left(v|\tilde{x}^s\right) * \tau^s \tag{13}$$

where $T_{\mathcal{A}(x)}$ is approximated by $\frac{1}{S}\sum_s \tau^s$.

3 MCMC simulation

Source: Chib and Greenberg (1995)

3.1 Background: First-order Markov chains

- Random sequence $X_1, X_2, \ldots, X_n, X_{n+1}, \ldots$
- First-order Markov: $P(X_{n+1}|X_n, X_{n-1}, X_{n-2}, \ldots) = P(X_{n+1}|X_t) = P(X'|X)$. History-less. Denote this transition distribution as $P(x, dy) = Pr(X' \in dy|X = x)$.
- Invariant distribution: Π is distribution, π is density

$$\Pi(dy) = \int P(x, dy)\pi(x)dx.$$

- Markov chain converges to invariant distribution: for starting value x, we have

$$\begin{aligned}
p^{(1)}(x, A) &= P(x, A) \\
p^{(2)}(x, A) &= \int_y P^{(1)}(x, dy)P(y, A) \\
p^{(3)}(x, A) &= \int_y P^{(2)}(x, dy)P(y, A) \\
&\ldots\ldots \\
p^{(n)}(x, A) &= \int P^{(n-1)}(x, dy)P(y, A) \approx \Pi(A)
\end{aligned}$$

  That is, for n large enough, each realization of X_n drawn according to $P(x, dy)$ is drawn from the marginal distribution $\Pi(dy)$. (Initial value x does not matter.)

- Markov chain theory mainly concerned about: for a given kernal $P(x, dy)$, what is invariant distribution Π?
- MCMC simulation goes backwards: given a marginal distribution Π, can we create a Markov process with some kernel function, such that Π is the invariant distribution?

- Let $p(x, y)$ denote density function corresponding to kernel function $P(x, dy)$ (ie. $P(x, A) = \int_A p(x, y)dy$). For a given π, if the following relationship is satisfied, then the kernel $P(x, dy)$ achieves π as the invariant distribution:

$$\pi(x)p(x, y) = \pi(y)p(y, x). \tag{14}$$

  This is a "reversibility" condition: interpreting $X = x$ and $X' = y$, it (roughly) implies that the probability of transitioning from x to y is the same as transitioning from y to x.

  (Note that in the – strange? – case where $f_{X'|X} = f_{X|X'}$, then both sides of Eq. (14) represent two alternative ways of writing the joint density of $f_{X',X}$.)

  Then π is the invariance density of $P(x, dy)$:

$$\begin{aligned}
\int P(x, A)\pi(x)dx &= \int \int_A p(x, y)dy\pi(x)dx \\
&= \int_A \int p(x, y)\pi(x)dxdy \\
&= \int_A \int p(y, x)\pi(y)dxdy \\
&= \int_A \left[\int p(y, x)dx\right] \pi(y)dy \\
&= \int_A \pi(y)dy = \Pi(A).
\end{aligned}$$

- Can we find such a magic function $p(x, y)$ which satisfies (14)?
- Consider any conditional density function $q(x, y)$. Suppose that

$$\pi(x)q(x, y) > \pi(y)q(y, x)$$

  so condition (14) fails. We will "fudge" $q(x, y)$ so that (14) holds:

3.2 Metropolis-Hastings approach

- Introduce the "fudge factor" $\alpha(x, y) \leq 1$, such that

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x).$$

When Eq. (14) is violated such that LHS>RHS, you want to set $\alpha(x,y) < 1$ but $\alpha(y,x) = 1$. Vice versa, if LHS<RHS, then you want $\alpha(x,y) = 1$ but $\alpha(y,x) < 1$. We can summarize this as:

$$\alpha(x,y) = \min\left[\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right]. \tag{15}$$

Correspondingly, define the Metropolis-Hastings kernel as:

$$P_{MH}(x,dy) = q(x,y)\alpha(x,y)dy + \underbrace{\left[1 - \int q(x,y)\alpha(x,y)dy\right]}_{r(x)} \mathbb{1}(x \in dy). \tag{16}$$

For a given value x, the Markov chain moves to $y \neq x$ with probability $q(x,y)\alpha(x,y)$, and stays at $y = x$ with probability $\left[1 - \int q(x,y)\alpha(x,y)dy\right] = r(x)$.

- Note that this kernel achieves π as the invariant distribution. As above:

$$\begin{aligned}
\int P_{MH}(x,A)\pi(x)dx &= \int \int_A q(x,y)\alpha(x,y)dy\pi(x)dx + \int_A r(x)\pi(x)dx \\
&= \int_A \int q(x,y)\alpha(x,y)\pi(x)dxdy + \int_A r(x)\pi(x)dx \\
&= \int_A \int q(y,x)\alpha(y,x)\pi(y)dxdy + \int_A r(x)\pi(x)dx \\
&= \int_A \left[\int q(y,x)\alpha(y,x)dx\right]\pi(y)dy + \int_A r(x)\pi(x)dx \\
&= \int_A \left[1 - r(y)\right]\pi(y)dy + \int_A r(x)\pi(x)dx \\
&= \int_A \pi(y)dy = \Pi(A).
\end{aligned}$$

- This MH kernel can be implemented via simulation: start with x^0, then
  - Draw "candidate" y_1 from the conditional density $q(x^0, y_1)$.
  - Set $x^1 = \begin{cases} y_1 & \text{with probability } \alpha(x^0, y_1) \\ x^0 & \text{with probability } 1 - \alpha(x^0, y_1) \end{cases}$
  - Now draw candidate y_2 from the conditional density $q(x^1, y_2)$.
  - Set $x^2 = \begin{cases} y_2 & \text{with probability } \alpha(x^1, y_2) \\ x^1 & \text{with probability } 1 - \alpha(x^1, y_2) \end{cases}$

– And so on.

According to the Markov chain theory, for N large enough we have approximately:

$$\Pi(A) \approx \frac{1}{N} \sum_{i=\tau>1}^{\tau+N} \mathbb{1}(x^i \in A).$$

Here τ refers to the length of an initial "burn-in" period, when the Markov chain is still converging.

- What are choices for the conditional density $q(\theta, \theta')$? Two common options are:
  - Random walk: $q(x, y) = q_1(y - x)$, where q_1 is a density function, eg. $N(0, \sigma^2)$. According to this approach, the candidate $y = x + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, and so it is called a random walk density.
  - Independent draws: $q(x, y) = q_2(y)$, where q_2 is a density function. The candidate y is independent of x. (However, because the candidate is sometimes rejected, with probability $1 - \alpha(x, y)$, the resulting random sequence still exhibits dependence.)
  - Gibb's sampler: Consider the case when x and y are both multidimensional, ie. $x = (x_1, x_2)$; $y = (y_1, y_2)$, and consider some joint density $g(x)$ for x and y, with associated conditionals $g(x_1|x_2)$ and $g(x_2|x_1)$. Then set

    $$q(x, y) = g(y_2|y_1) \cdot g(y_1|x_2).$$

    That is, for a given $x = (x_1, x_2)$, first you draw y_1 according to the conditional density $g(y_1|x_2)$, then you draw y_2 according to the conditional density $g(y_2|y_1)$.

3.3 Application to Bayesian posterior inference

- For Bayesian inference, the desired density is the posterior density $f(\theta|\vec{z})$ (where $\vec{z}$ denotes the data information). We want to construct a Markov-chain, using the MH idea, such that its invariant distribution is the posterior distribution of $\theta|\vec{z}$.

- We use θ and θ' to denote the "current" and "next" draws of θ from the Markov chain.

- Recall that $f(\theta|\vec{z}) \propto f(\vec{z}|\theta)f(\theta)$, the likelihood function times the prior denstiy. Hence, the MH fudge factor is

$$\begin{aligned}\alpha(\theta, \theta') &= \min\left[\frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}, 1\right] \\ &= \min\left[\frac{f(\vec{z}|\theta')f(\theta')q(\theta', \theta)}{f(\vec{z}|\theta)f(\theta)q(\theta, \theta')}, 1\right].\end{aligned}$$

  So you are more likely to accept a draw θ' relative to θ if the likelihood ratio is higher.

- For a given $q(x, y)$, drawing a sequence $\theta^1, \theta^2, \ldots, \theta^n, \ldots$ such that, for n large enough, each θ^n can be treated as a draw with marginal distribution equal to $\pi(\theta|\vec{z})$. Hence, for instance, the posterior mean can be approximated as:

$$\mathbb{E}[\theta|\vec{z}] \approx \frac{1}{N}\sum_{i=\tau}^{\tau+N}\theta^i.$$

- In the multidimensional case when $\theta = (\theta_1, \theta_2)$, when $q(\theta, \theta')$ is the Gibb's sampler, then you could take the joint density of θ to be the desired posterior density $f(\theta|\vec{z})$ itself, and take the Gibb's sampler to be

$$q(\theta, \theta') = f(\theta_2'|\theta_1'; \vec{z}) \cdot f(\theta_1'|\theta_2; \vec{z}).$$

  This is useful when the full posterior $f(\theta|\vec{z})$ is difficult to sample from directly, but the individual conditionals are easy to characterize and sample from.

  Moreover, when data augmentation is performed, you could also use Gibb's sampling, and interpret θ_1 as the parameters of the model, while θ_2 are the latent variables.

References

Chib, S., and E. Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm," *American Statistician*, 49, 327–335.

Gourieroux, C., and A. Monfort (1996): *Simulation-Based Econometric Methods*. Oxford University Press.

Hajivassiliou, V., and P. Ruud (1994): "Classical Estimation Methods for LDV Models Using Simulation," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden. North Holland.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57, 995–1026.