Econometrics Lunch—Simulation-based Estimation

Bernard Salanié

January 25, 2006
General references:

- **C. Gouriéroux and A. Monfort (1997),** *Simulation-Based Econometric Methods,* Oxford University Press.

  *Econometric Inference Using Simulation Techniques,* John Wiley (collected papers).

- **R. Mariano, T. Schuerman and M. Weeks (2000),**
  *Simulation-based Inference in Econometrics,* Cambridge University Press (some review papers).
Three generations in econometrics

- 60s: closed-form estimators, mostly OLS on linear models.
- 70s and 80s: the advent of numerical optimization, FIML-GMM-PML.
- 90s and beyond: models too complicated to yield a closed-form criterion $\rightarrow$ simulate.
First example that motivated generation 3.

discrete choice model.

individuals \( i = 1, \ldots, n \), alternatives \( j = 1, \ldots, m \)

utility of \( j \) for \( i \): \( U_{ij} = z_{ij} b_j + \varepsilon_{ij} \) where \( \varepsilon_{ij} \) is \( N(0, 1) \), possibly correlations across alternatives \( j \).

\[
\Pr(i \text{ chooses } j) = \Pr(\forall k, U_{ij} \geq U_{ik})
\]

\[
= \Pr(\forall k, \varepsilon_{ij} - \varepsilon_{ik} \geq z_{ik} b_k - z_{ij} b_j)
\]

is an \( (m - 1) \)-dimensional integral of the normal distribution, untractable for \( m > 5 \), which is not unusual for some applications (e.g. transportation).
Second example: model \( y_i = g(x_i, \theta, \varepsilon_i, u_i) \)
where \( u_i \), the unobserved heterogeneity, has some density \( h \) (e.g. panel data with random effects)
Often we can compute in closed form the likelihood conditional on \( u_i \), \( f(y_i, x_i, u_i, \theta) \), but not the unconditional likelihood
\[
\int f(y_i, x_i, u_i, \theta) h(u_i) du_i
\]
Third example: say Tobit with serially correlated latent variable

\[ y_{it}^* = x_{it}\beta_0 + \alpha_0 y_{i,t-1}^* + u_{it} \]

Since \( y_{i,t-1}^* \) is unobserved, we can’t condition on it for estimation \( \rightarrow \) the likelihood is a \( T \)-dimensional integral. In these three examples, only simulation will allow us to estimate the model.
The method of simulated moments


Here take cross-section data for simplicity. Often we have a moment condition like

\[ E(K(y, x)|x) = k(x, \theta_0) \]

Then normally in GMM we do

\[
\min_{\theta} \left\| \frac{1}{n} \sum_{i=1}^{n} (K(y_i, x_i) - k(x_i, \theta)) \right\|_\Omega^2
\]

where \( \|X\|_\Omega^2 = X'\Omega X \), \( \Omega \) well chosen.
Often $k$ is very hard to compute, as $y$ is driven by a complicated model

$$y = g(x, \theta_0, \varepsilon).$$

(Take the multinomial probit with $K(y_i, x_i) \simeq (i \text{ chooses } j)$, so that $k \simeq P_{ij}$)

then we draw $\varepsilon^s$ in the distribution of $\varepsilon$ and compute

$$y^s(\theta) = g(x, \theta, \varepsilon^s)$$

and get the MSM estimator by approximating

$$k(x_i, \theta) \text{ with } \frac{1}{S} \sum_{s=1}^{S} K(y_i^s(\theta), x_i).$$
Is it a good idea?

We just want to approximate an integral, and there are other techniques (see end of slides)
e.g. Gaussian quadrature for some kernel $g$:

$$
\int_0^1 f(x)g(x)dx \equiv \sum_{i=1}^{m} \omega_i f(x_i)
$$

where the $(x_i, \omega_i)$ only depend on $g$.

- much more accurate than Monte Carlo integration for small dimensional integrals
- but in dimensions larger than 3 they become pretty useless.

...and there are ways to improve the precision of Monte Carlo integration in several important cases.
Asymptotic properties of MSM

it can be shown that when $S$ is fixed and $T \to \infty$, 

- the MSM estimator is consistent and asymptotically normal
- its asymptotic variance is of the form $(a + \frac{b}{S})$
- so the efficiency loss due to the simulations can be made small by increasing $S$. 
Simulated Pseudo-Maximum Likelihood

References:


Take an NLLS example: \( E(y|x) = m(x, \theta_0) \).

Often \( m(x, \theta) \) is very difficult to compute. So instead we simulate

\[
y^s(\theta) = g(x, \varepsilon^s, \theta)
\]
Asymptotic properties: although $m^S(x, \theta)$ is unbiased, it enters nonlinearily in the criterion we minimize. This implies that with fixed $S$ and $n \to \infty$, the SPML estimators have a bias of order $1/S$.

We can correct the objective function so as to get a consistent estimator for fixed $S$. To do this, compute the Monte-Carlo variance

$$v^S(x, \theta) = \frac{1}{S-1} \sum_{s=1}^{S} \left( y^s(\theta) - m^S(x, \theta) \right)^2$$

and minimize over $\theta$

$$\sum_{i=1}^{n} \left( (y_i - m^S(x_i, \theta))^2 - \frac{1}{S} v^S(x_i, \theta) \right)$$
Example: model with unobserved heterogeneity. Then, often

\[ f(y, x, \theta, m, \sigma) = \int f(y, x, m + \sigma u, \theta) h(u) du \]

which cannot be integrated (except in some very special cases). The solution is to draw \( u^s \) from \( h \), to let

\[ f^S(y, x, \theta, m, \gamma) = \frac{1}{S} \sum_{s=1}^{S} f(y, x, u^s, \theta, m, \gamma) \]

and to solve

\[ \max_{\theta, m, \gamma} \sum_{i=1}^{n} \log f^S(y, x, \theta, m, \gamma) \]
Other example: models with latent variables (logit, probit, tobit...).

Often very hard with lagged $y_{i,t-1} \longrightarrow \text{use}$

$$\frac{\partial \log f}{\partial \theta}(y, x, \theta) = E \left( \frac{\partial \log f^*}{\partial \theta}(y^*, x, \theta) | y = h(y^*) \right)$$

Where $f$ is the (hard to evaluate) likelihood of the model and $f^*$ is the likelihood of the (unusable) latent-variable model.

So that if we can simulate $y^s*$ in $h^{-1}(y)$, we can obtain a SML estimator. This is called the method of simulated scores (MSS).

Asymptotics: $f^S$ is an unbiased simulator of $f$, but log is nonlinear $\longrightarrow$ again, we have a (hopefully small) bias in $1/S$. 
Fermanian-Salanié, *Econometric Theory* 2004:
Given a (parametric!) model \( y = g(x, \theta, \varepsilon) \),
draw \( y^s(x, \theta) \), \( s = 1, \ldots, S \)
and nonparametrically estimate the density:

\[
    f^S(y, x, \theta) = \frac{1}{Sh} \sum_{s=1}^{S} K \left( \frac{y - y^s(x, \theta)}{h} \right).
\]

Consistent and asymptotically efficient if \( S \) goes to infinity and \( h \)
goes to zero (at the right rates . . . ) as sample size goes to infinity.
Indirect Inference


Old idea: minimum-distance estimation, resurfaced as asymptotic LS, then (with simulations) as indirect inference.

Assume our model \((M)\) has a complicated likelihood function \(f(y, x, \theta)\) that we cannot maximize. On the other hand, model \((M')\) is reasonably close to model \((M)\) and has a simple likelihood function \(f'(y, x, \beta)\) → it is easy to estimate.

We could do

\[
\max_{\beta} \sum_{i=1}^{n} \log f'(y, x, \beta)
\]

to estimate \((M')\). But of course, we get an estimator \(\hat{\beta}\) that gives us little information on \(\theta_0\). *Or does it?*
Indirect inference proceeds in three steps:

1. For given $\theta$, simulate model ($M$) to get $y^s(\theta)$
2. Estimate model ($M'$) on simulated data:

$$
\max_{\beta} \sum_{i=1}^{n} \sum_{s=1}^{S} \log f'(y^s(\theta), x, \beta)
$$

to get an estimator $\hat{\beta}^S(\theta)$

3. "Calibrate" to make $\hat{\beta}^S(\theta)$ as close to $\hat{\beta}$ as possible, i.e. with some positive definite matrix $\Omega$, do

$$
\min_{\theta} \left\| \hat{\beta}^S(\theta) - \hat{\beta} \right\|^2_{\Omega}
$$

which gives the indirect inference estimator $\hat{\theta}^S$. 
For this to work, we must have $\dim \beta \geq \dim \theta$. Moreover, let $b$ be the “binding function”:

$$b(\theta) = \arg \max_{\beta} E \log f'(y(\theta), x, \beta)$$

$b(\theta)$ is the pseudo-value.

Then we need the equation $b(\theta) = b(\theta_0)$ to have the only solution $\theta = \theta_0$:

model $(M)$ is identified from model $(M')$

The indirect inference estimator $\hat{\theta}^S$ is consistent and asymptotically normal for fixed $S$, with efficiency loss in $1/S$.

The choice of the auxiliary model matters a great deal... just like the choice of instruments in GMM.
Econometrics of finance
The data (say the price of a stock) is generated by a continuous-time process

\[ dy_t = g(\theta, y_t)dt + h(\theta, y_t)dW_t \]

where \( W_t \) is a Brownian motion. We only observe \( y_1, \ldots, y_T \). The problem is how to discretize the model for estimation purposes.
In very rare cases, we can compute the exact discretization. For instance, take the geometric diffusion used by Black-Scholes

\[ \frac{dy_t}{y_t} = \mu dt + \sigma dW_t \]

Then Itô's formula gives

\[ d(\log y)_t = \left( \mu - \frac{\sigma^2}{2} \right) dt + \sigma dW_t \]

which integrates to

\[ \log y_t = \log y_{t-1} + \mu - \frac{\sigma^2}{2} + \sigma \varepsilon_t \]
Usually we can’t integrate, so we can use the naïve discretization

\[ y_t = y_{t-1} + g(\beta, y_{t-1}) + h(\beta, y_{t-1}) \varepsilon_t \]

as model \((M')\), which can be estimated through weighted NLLS. This is not the true model, so the ‘pseudo-true-value is not \(\theta_0\), but things may work well if it is a (locally) one-to-one function of \(\theta_0\).
Another example: dynamic programming models

Stationary example:

\[ V(s_t) = \max_{c_t} (u(c_t, \theta, x_t, \varepsilon_t) + \beta(x_t, \theta)E(V(s_{t+1}|s_t, c_t, \theta))) \].

You observe the three time-series \((c_{it}, s_{it}, x_{it})\) for several agents, so you “can” estimate nonparametrically (inter alia)

- the transition process \((s_t, c_t, x_t) \longrightarrow s_{t+1}\)
- the choice process \((s_t, x_t) \longrightarrow c_t\).

you can then fit a flexible parametric form for these time series processes: \(l(s_{t+1}, c_t|s_t, x_t; \beta)\)

...and you “can” simulate the dynamic model for any \(\theta\)
...and use indirect inference to get your estimator for \(\theta\)
...+ overidentification tests since in general \(\dim \beta \gg \dim \theta\).
the building block usually is a congruential random number generator (RNG).
This starts from an integer “seed” $n_0$ chosen by the user or the computer and computes recursively

$$n_{m+1} \equiv an_m + b \pmod{c}$$

where $a$, $b$ and $c$ are well-chosen large numbers.
After a while, this settles into an (almost) uniformly distributed sequence in $[0, c-1]$. After dividing by $c$, we get random draws from the uniform $U[0,1]$.
A RNG for a distribution $F$ can be obtained by doing $v = F^{-1}(u)$, where $u$ follows $U[0,1]$.
Careful: choose a RNG with a large periodicity $c$. There are also more clever “direct” tricks to draw from the normal, the Student, the Beta...
E.g. for Student $t_n$: draw $x_i$ for $i = 1, \ldots, n + 1$ in $N(0, 1)$ i.i.d, and generate

$$y = \frac{x_1}{\sqrt{\frac{\sum_{i=2}^{n+1} x_i^2}{n}}}.$$
Tougher cases

You want to draw from $f(x)$ and

- you can compute $f(x)$ in every $x$
- but you cannot draw directly (e.g. $F^{-1}$ is very nasty)
- and you can draw from $g(x)$, another density such that

$$\forall x, \frac{f(x)}{g(x)} \leq \text{some } M \quad (B).$$

Acceptance-rejection algorithm:

1. draw $x$ from $g$
2. accept it with proba $f(x)/Mg(x)$
3. or go back to step 1.
Acceptance-rejection works in principle, but condition (B) requires $g$ to have thicker tails than $f$, not so great; and you need to get $M$ about right. There is a very efficient variant that works by adapting $g$ in several steps:

*Adaptive Rejection Sampling* works whenever $f$ is log-concave—which is “very reasonable” unless it has clusters of probability.
Monte Carlo integration

= what we have been doing from the start: approximate

$$E h(X) = \int h(x) f(x) dx$$

by drawing $$x_1, \ldots, x_n$$ from $$f$$ and computing

$$\frac{1}{n} \sum_{i=1}^{n} h(x_i).$$

It has variance $$V h/n$$; can we do better?

1. Antithetic variables often work well: use near symmetries in $$h$$, e.g:

$$V \left( \frac{h(x) + h(-x)}{2} \right) = \frac{V h}{2} + \text{cov}(h(x), h(-x)).$$

useful (in this example) when $$h$$ is “almost odd.

2. Importance sampling: draw from $$g$$ and use

$$\frac{1}{n} \sum_{i=1}^{n} h(x_i) f(x_i)/g(x_i),$$

clever in theory but often unreliable in practice.
Weighted Monte Carlo Integration

Idea: Riemann sum integration,

\[
\int_{a}^{b} w(x)dx \simeq \frac{b-a}{n} \sum_{i=1}^{n} w(a + i(b-a)/n).
\]

Bring together with Monte Carlo integration:

1. draw the points \(x_1, \ldots, x_n\) randomly in \(f\)
2. order them: \(x_{(1)} < \ldots < x_{(n)}\)
3. compute

\[
\sum_{i=1}^{n-1} h(x_{(i)}) f(x_{(i)}) (x_{(i+1)} - x_{(i)}).
\]
(originally) Bayesian motivation: posterior ← data + prior

\[ f(\theta|y) \sim f(y|\theta)f(\theta) \]

We are usually concerned with its moments

\[ \int \theta^k f(\theta|y) d\theta; \]

but often dim \( \theta \) is large, so we want to do Monte Carlo integration

... provided we can draw from \( f(\theta|y) \).
Drawing from the posterior

Often we can evaluate $f(\theta|y)$ at any point, but not draw from it directly.
But there are two MCMC (Markov Chain Monte Carlo) tricks:

- Metropolis-Hastings;
- Gibbs sampling.
Say you want to draw from some distribution with pdf $f(x)$; choose some density $q(x, y)$ (here, symmetric, for simplicity); given a draw $x^s$, draw $y$ from $q(y|x^s)$; accept it with probability

$$\min \left( \frac{f(y)}{f(x)}, 1 \right)$$

otherwise redraw from $q(y|x^s)$.
After a burn-in phase, draws are as if from $f$. 
Gibbs sampling

If \( x = (y, z) \) and you can draw from \( f(y|z) \) and \( f(z|y) \) but not \( f(y, z) \):
given \( x^s = (y^s, z^s) \), draw 
y\(^{s+1}\) from \( f(y|z^s) \), 
z\(^{s+1}\) from \( f(z|y^s) \) \( \longrightarrow x^{s+1} \)
again, draws converge to as if from \( f \).
Why is Gibbs sampling so useful?

E.g. in all models with latent Markov chains (Hamilton in macro, volatility models in finance. . . )

\[ y_t = G(s_t, u_t, \theta) \]

We need to draw from \( l(s_1, \ldots, s_T | y_1, \ldots, y_T) \), often very hard, but it is much easier to draw from distributions like

\[ l(s_t | s_{-t}, y_1, \ldots, y_T). \]

(here if \( u_t \) is i.i.d and independent of \( (s_t) \), you just need to condition on \((s_{t-1}, y_T, s_{t+1}).\)