Classical Simulation-Based Inference

1 The Canonical LDV Model

$$y_i^* = X_i\beta + \epsilon_i, \quad y_i = \tau(y_i^*). \tag{1}$$

$$\epsilon_i = \Gamma_i\eta, \quad E\epsilon_i\epsilon_i \equiv \Omega_i(\sigma) = \Gamma_i(\sigma)\Gamma_i(\sigma)'. \tag{2}$$

$$D(y_i) = \{y_i^* | y = \tau(y_i^*)\}. \tag{3}$$

1.1 Examples of LDV Models – different $\tau(\cdot)$ functions

Model 1: multinomial probit

Alternative j yields the (random) utility

$$y_{ij}^* = x_{ij}\beta + \epsilon_{ij} \qquad j = 1, \cdots, J$$

and individual i chooses alternative k that satisfies

$$-\infty < y_{ik}^* < \infty, \qquad 0 < y_{ik}^* - y_{ij}^* < \infty, \tag{4}$$

The analyst observes the indicator $y_i \equiv \arg\max_j\{y_{i1}^*, \cdots, y_{ij}^*, \cdots, y_{iJ}^*\}$.

Applications – see Model 3 below.

Model 2: multivariate rank ordered probit

As with model 1, individual i chooses alternative k that offers the highest utility y_{ik}^*. The analyst, however, observes the full ranking of the J alternatives in terms of the utility they yield, i.e., the analyst observes the J-dimensional vector of indices

$$y_i \equiv (k_1, \cdots, k_J)'$$

such that

$$y_{ik_1}^* \leq y_{ik_2}^* \leq \cdots \leq y_{ik_J}^*. \tag{5}$$

Model 3: multiperiod (panel) probit

Binary:

(6)

Multinomial:

$$y_{it} = \arg\max_j \{y^*_{it1}, \cdots, y^*_{itj}, \cdots, y^*_{itJ}\} \quad (7)$$

Illustrations:

- **Applications (Fields):**
  - Finance – Portfolio optimization
  - Marketing – purchasing decisions
  - Political Science – voting behavior
  - Psychology – decision trees
  - Experimental Economics – Bayesian vs. Heuristic Behavior
- **Specific Economic Applications:**
  - The *Incidence* of External Debt Crises of Developing Countries, Hajivassiliou (1994).
  - Unemployment and Liquidity Constraints, Hajivassiliou and Ioannides (1994).
  - Health, Children, and Elderly Living Arrangements, Börsch-Supan, Hajivassiliou, Kotlikoff, and Morris (1992).

Model 4: multiperiod (panel) Tobit

(8)

Illustration: The *Extent* of External Debt Crises of Developing Countries, Hajivassiliou (1994).

1.2 Classical Estimation Methods

$$\ell_i(\theta; y_i) = \int_{D(y_i)} n(y^*_i - X_i\beta, \Omega_i)\ dy^*_i, \quad (9)$$

$$n(\epsilon, \Omega) = (2\pi)^{-T/2}|\Omega|^{-1/2}\exp[-\frac{1}{2}\epsilon'\Omega^{-1}\epsilon] \quad (10)$$

$$\ell_{i\theta}(\theta; y_i) \equiv \frac{\partial \ell_i(\theta; y_i)}{\partial\theta} = \ell_i(\theta; y_i) E\{h(y^*_i - X_i\beta) | y^*_i \in D(y_i)\}, \quad (11)$$

$$s_i(\theta; y_i) \equiv \frac{\partial \ell n \ell_i(\theta; y_i)}{\partial \theta} = E\{h(y_i^* - X_i\beta)|y_i^* \in D(y_i)\}$$

$$\ell_{i\theta}/\ell_i = \frac{\int_{D(y_i)} h(z, X_i, \beta, \Omega_i) n(z - X_i\beta, \Omega_i)\ dz}{\ell_i}. \qquad (12)$$

***i.i.d.* observations across i.**

$$\ell_{i\theta} \equiv \ell_\theta(\theta; y_i) \equiv \frac{\partial \ell(\theta; y_i)}{\partial \theta} = \ell(\theta; y_i) E\{h(y_i^* - X_i\beta)|y_i^* \in D(y_i)\},$$

$$s_i \equiv s(\theta; y_i) \equiv \frac{\partial \ell n \ell(\theta; y_i)}{\partial \theta} = \ell_{i\theta}/\ell_i = E\{h(y_i^* - X_i\beta)|y_i^* \in D(y_i)\}.$$

$$\hat{\theta}_{MLE.1} \equiv \arg\max_\theta \frac{1}{N} \sum_i \ell n \ell_i(\theta),$$

or equivalently,

$$\tilde{\theta}_{MLE.2} \quad \text{solves} \quad \left\{ \frac{1}{N} \sum_i s_i(\theta) = \frac{1}{N} \sum_i [\ell_{i\theta}(\theta)/\ell_i(\theta)] = 0 \right\}.$$

NOTE: at θ^* true, $E\{\frac{\partial \ell n \ell_i(\theta^*)}{\partial \theta}\} = E\{h(y_i^* - X_i\beta^*)|D(y_i)\} = 0.$

1.3 The Intractability of the Classical Estimators

- Cross-sectional MNP with 10 choices.
- 2000 individuals.
- 20 explanatory variables.
- Unconstrained variance/covariance matrix for unobservable utilities.
- **Implication:** each loglikelihood function requires evaluating 2000 integrals of dimension 9.

1. Classical MLE based on Numerical Quadrature will be **inconsistent** and will require **3 months of CRAY-1 CPU**!
2. SSML/GHK and MSS/Gibbs will be **consistent** and **asymptotically normal** and can be calculated with **24 hours of Sparc-10 CPU**.

Comparison: 4^9 vs. 100, i.e., about 2600 longer on the same machine. (or 1 day vs. 10 years!)

1.4 Simulation Estimation Methods

Consider a simulator $\tilde{\ell}_i \equiv \tilde{\ell}_i(\theta, R)$ for the likelihood contribution $\ell_i(\theta)$, based on R independent GHK/SRC simulations. Then, the SSML/GHK estimator is defined by:

$$\hat{\theta}_{SSML} \equiv \arg\max_{\theta} \left\{ \frac{1}{N} \sum_i \ln \tilde{\ell}_i(\theta, R) \right\}.$$

Consider a simulator, $\tilde{s}_i \equiv \tilde{s}_i(\theta, R)$, for the score function $s_i(\cdot)$, satisfying $y_i^* \in D(y_i)$, based on R independent draws according to r_G Gibbs resamplings. Then, the MSS1/GSS estimator:

$$\hat{\theta}_{MSS.1} \quad \text{solves} \quad \left\{ \frac{1}{N} \sum_i \tilde{s}_i(\theta, R, r_G) = 0 \right\}.$$

Consider a simulator $\tilde{\ell}_{i\theta}(\theta, R)$ for the derivative of a likelihood contribution $\ell_{i\theta}(\theta)$, based on R independent GHK/SRC draws, and a simulator $\tilde{\ell}_i(\theta, R)$ for the denominator probability based on the same R GHK/SRC draws. Then, the MSS2/GHK estimator is:

$$\tilde{\theta}_{MSS.2} \quad \text{solves} \quad \left\{ \frac{1}{N} \sum_i [\tilde{\ell}_{i\theta}(\theta, R)/\tilde{\ell}_i(\theta, R)] = 0 \right\}.$$

Not good idea (dominated):

$$\tilde{\theta}_{MSS.3} \quad \text{solves} \quad \left\{ \frac{1}{N} \sum_i [\tilde{\ell}_{i\theta}(\theta, R)/\tilde{\ell}_i(\theta, R_d)] = 0 \right\}.$$

A Simple Explanation of How Simulation Works

NB: Given the non-linearities involved, all estimation methods will involve **iterative search** over the unknown θ.

Consider a trial parameter vector $\theta^{(n)}$ at iteration n.

Maximum likelihood estimation requires the evaluation of

$$L(\theta^{(n)}) = \frac{1}{N} \sum_{i=1}^{N} ln\ell_i(\theta^{(n)}; y_i).$$

Method of Scoring seeks to evaluate

$$S(\theta^{(n)}) = \frac{1}{N}\sum_{i=1}^{N} s_i(\theta^{(n)}; y_i) = \frac{1}{N}\sum_{i=1}^{N} \frac{\ell_{i\theta}(\theta^{(n)}; y_i)}{\ell_i(\theta^{(n)}; y_i)}.$$

Method of moments calculates

$$M(\theta^{(n)}) = \frac{1}{N}\sum_{i=1}^{N} m_i(\theta^{(n)}; y_i) = \frac{1}{N}\sum_{i=1}^{N} w(\theta^{(n)}; X_i)'(y_i - g_{1i}(\theta^{(n)}; X_i),$$

where $g_{1i}(\cdot) \equiv E(y_i; \theta^{(n)})$ and $w(\cdot)$ is an instrument function. Finally, a Pseudo-ML method evaluates the quadratic form

$$Q(\theta^{(n)}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - g_{1i}(\theta^{(n)}; X_i))' \cdot g_{2i}(\theta^{(n)}; X_i)^{-1} \cdot (y_i - g_{1i}(\theta^{(n)}; X_i)),$$

where $g_{1i}(\cdot) \equiv E(y_i; \theta^{(n)})$ and $g_{2i}(\cdot) \equiv V(y_i; \theta^{(n)})$.

IF ANALYTICALLY or NUMERICALLY TRACTABLE:
$\ell_i(\theta^{(n)}; y_i)$, $\ell_{i\theta}(\theta^{(n)}; y_i)$, $s_i(\theta^{(n)}; y_i)$, $g_{1i}(\theta^{(n)}; y_i)$, and $g_{2i}(\theta^{(n)}; y_i)$:
Computer routines can be written to evaluate these expressions as functions of any possible trail parameter vector $\theta^{(n)}$.

1.5 Estimation by Simulation

Relies on simulating routines $\tilde{\ell}_i(\theta^{(n)}; y_i, R)$, $\tilde{\ell}_{i\theta}(\theta^{(n)}; y_i, R)$, $\tilde{s}_i(\theta^{(n)}; y_i, R)$, $\tilde{g}_{1i}(\theta^{(n)}; y_i, R)$, and $\tilde{g}_{2i}(\theta^{(n)}; y_i, R)$, defined as follows:

- Draw a set of R uniform $J-$dimensional random vectors $\tilde{u}_i^1, \cdots, \tilde{u}_i^r, \cdots, \tilde{u}_i^R$. By the assumptions of this model, the disturbance vector ϵ_i is i.i.d. with density function $f_\epsilon(\epsilon_i; \theta^*)$, with $E(\epsilon_i|X_i) = 0$ and $E(\epsilon_i\epsilon_i'|X_i) = \Omega_i(\sigma)$.
- At the given trial parameter vector $\theta^{(n)} = (\beta^{(n)}, \sigma^{(n)})'$, using the inverse of the cumulative distribution function of ϵ, $F_\epsilon^{-1}(\cdot)$, obtain a set of R $\tilde{\epsilon}_i$'s,

$$\tilde{\epsilon}_i^r(\sigma^{(n)}) = F_\epsilon^{-1}(\tilde{u}_i^r; \Omega(\sigma^{(n)}),$$

  which will imply a set of R simulated latent vectors $\tilde{y}_i^{*r}(\theta^{(n)})$, using the specification $y_i = \tau(y_i^*)$.
- From the R simulated $\tilde{y}_i^r$ vectors, calculate the empirical counterparts of the $\ell_i(\cdot)$ etc. functions and thus define the simulators $\tilde{\ell}_i(\theta^{(n)}; y_i, R)$, $\tilde{\ell}_{i\theta}(\theta^{(n)}; y_i, R)$,

$\tilde{s}_i(\theta^{(n)}; y_i, R)$, $\tilde{g}_{1i}(\theta^{(n)}; y_i, R)$,
and $\tilde{g}_{2i}(\theta^{(n)}; y_i, R)$.

- **Keeping the same uniform random variates,** $\tilde{u}_i^r$'s, a new trial parameter vector $\theta^{(n)}$ will imply a new set of simulated $\tilde{\epsilon}_i^r(\sigma^{(n)})$, leading to new $\tilde{y}_i^*$ and hence new values for $\tilde{\ell}_i$, etc.

The iterative search algorithms will keep trying different parameter vectors θ to satisfy the relevant criterion.

1.6 A Cautionary Tale:

SML (Lerman and Manski (1981)):
$\hat{\theta}_{LM} = \arg\max_\theta \frac{1}{N} \sum_i \ln \tilde{\ell}_i(\theta, R)$,
such ℓ_i are simulated unbiasedly ($E\tilde{\ell}_{ir} = \ell_i$) and consistently with R ($\tilde{\ell}_i(\theta, R) \to_p \ell_i(\theta)$ as $R \to \infty$). Empirical choice probabilities as the simulating function $\tilde{\ell}_i$.

This simulator is:

- discontinuous function of the parameters and variates
- and not bounded away from 0 and 1.

1.7 Key Advantages of MSS over Leading Competitor (MSM)

- Applicable to any LDV model that can be written as a set of linear inequality constraints on the underlying latent variables, the distribution of which belongs to the linear exponential class.
- Asymptotically efficient.

2 Smooth Simulators for MSS Estimation

2.1 The Smooth Recursive Conditioning Simulator (GHK)

- Consider the $T \times 1$ random variate vector Y^* distributed as $N(\mu^*, \Omega)$ and consider the event $\mathbf{E} \equiv \{a^* \leq MY^* \leq b^*\}$, where $-\infty \leq a^* < +\infty$, $-\infty < b^* \leq +\infty$, $a^* < b^*$, the matrix M is non-singular, and the matrix Ω is positive definite.

- Define $a \equiv a^* - M\mu^*$, $b \equiv b^* - M\mu^*$, $\mu \equiv M\mu^*$, and let L be the (lower-triangular) Cholesky decomposition of $\Sigma \equiv M\Omega M' \equiv LL'$.

- For a vector e, let $e_{<j}$ denote the subvector of the first $j-1$ components, and for a matrix A, let $A_{j,<j}$ denote a vector containing the first $j-1$ elements of row j.

- Draw sequentially $e_1 \sim N(0,1)$ s.t. $a_1 \leq l_{11} \cdot e_1 \leq b_1$, $e_2 \sim N(0,1)$ s.t. $a_2 \leq l_{21} \cdot e_1 + l_{22} \cdot e_2 \leq b_2$, $\cdots$, and $e_T \sim N(0,1)$ s.t. $a_T \leq l_{T1} \cdot e_1 + \cdots + l_{TT} \cdot e_T \leq b_T$. These univariate truncated normal variates are drawn according to the following smooth scheme: Let U be a uniform (0,1) random variable and let $\Phi(\cdot)$ denote the standard normal $N(0,1)$ cumulative distribution function. Define the random variable $e \equiv \Phi^{-1}((\Phi(b) - \Phi(a)) \cdot U + \Phi(a))$, where $-\infty \leq a < b \leq \infty$. As Proposition 1 proves, e is distributed $N(0,1)$ conditional on $a \leq e \leq b$.

- Now let $e \equiv (e_1, \cdots, e_T)'$ and define

  $Q_1 \equiv \text{Prob}(a_1/l_{11} \leq e_1 \leq b_1/l_{11})$,

  $Q_t(e_1, \cdots, e_{t-1}) \equiv \text{Prob}((a_t - L_{t,<t} \cdot e_{<t})/l_{tt} \leq e_t \leq (b_t - L_{t,<t} \cdot e_{<t})/l_{tt} e_1, \cdots, e_{t-1})$.

$$
\begin{aligned}
\ell(y, X; \beta, \Omega) &= \int_{a^*(y) \leq M(y) \cdot z \leq b^*(y)} n(z - X\beta, \Omega)\, dz \\
&= \text{Prob}[a^*(y) \leq M(y) \cdot Y \leq b^*(y);\ Y \sim N(X\beta, \Omega)] \\
&= \text{Prob}[a(y, X, \beta, \Omega) \leq L(y, \Omega) \cdot \nu \leq b(y, X, \beta, \Omega); \nu \sim N(0, I)].
\end{aligned}
$$

$$
\tilde{\ell}(e; y, X; \beta, \Omega; R) = \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T} Q_t(e_{1r}, \cdots, e_{t-1,r}).
$$

As Lemma 1 establishes, the simulator $\tilde{\ell}(e; y, X; \beta, \Omega; R)$ is:

1. an unbiased estimator of $\ell(y, X; \beta, \Omega)$;
2. a smooth, i.e., a continuous and differentiable function of the model parameters β and Ω and the underlying uniform random deviates.

2.2 An Outline of the GHK Method

$$P(y^* \in A) = \int_{a^*<Qy^*<b^*} n(y^*)dy^* = \int_{a<Ls^*<b} n(s^*)ds^* = \int_{a<Le^*<b} \frac{n(e^*)}{g(e^*)} g(e^*)de^*$$

But:

$$g(e^*) \equiv \frac{n(e^*)}{p(a < Le^* < b)} = \frac{\Pi_j^M n(e_j^*)}{\Pi_j^M Prob(\underline{\lambda}_j < \tilde{e}_j < \bar{\lambda}_j)}$$

Therefore:

$$\frac{n(e^*)}{g(e^*)} = \Pi_{j=1}^M Prob(\underline{\lambda}_j < \tilde{e}_j < \bar{\lambda}_j)$$

since $n(e^*) = n(s^*)$.

Definitions:

$$y^* \sim N(\mu, \Sigma), \quad s^* \sim N(0, I), \quad LL' = Q\Sigma Q'^* - Q\mu, \quad b = b^* - Q\mu$$

and e^* is drawn according to the *sequential scheme*:

$$\tilde{e}_j^* \sim N(0,1) \text{ truncated on } \underline{\lambda}_j < \tilde{e}_j < \bar{\lambda}_j$$

where $u_j \sim U[0,1]$ and

$$e^* \equiv \Phi^{-1}\left([\Phi(b) - \Phi(a)] \cdot u_j + \Phi(a)\right).$$

2.3 The Gibbs Resampling Simulator (GSS)

Developed for and has been applied to the problems of image reconstruction, neural networks and expert systems.

- Let the $T \times 1$ variate random vector Z describe the distribution of $Y^* \sim N(X\beta, \Omega)$ truncated on the event $\mathbf{E} \equiv a^* \leq M \cdot Y^* \leq b^*$.
- Assume, **without loss of empirical generality**, that the truncation region (a^*, b^*) is compact, equivalent to $-\infty < a < b < +\infty$.
- Gibbs sampling: Markovian updating scheme: Given an arbitrary starting set of values $Z_1^{(0)}, Z_2^{(0)}, \cdots, Z_T^{(0)}$, we draw $Z_1^{(1)} \sim [Z_1|Z_2^{(0)}, \cdots, Z_T^{(0)}]$, then $Z_2^{(1)} \sim [Z_2|Z_1^{(1)}, Z_2^{(0)} \cdots, Z_T^{(0)}]$, $Z_3^{(1)} \sim [Z_3|Z_1^{(1)}, Z_2^{(1)}, Z_3^{(0)} \cdots, Z_T^{(0)}]$, $\cdots$, and

so on, up to
$Z_T^{(1)} \sim [Z_T | Z_1^{(1)}, \cdots, Z_{T-1}^{(1)}]$. Thus each variable is "visited" in the "natural" order and a cycle in this scheme requires T random variate generations. After r_G such iterations we would arrive at $Z^{(r_G)} \equiv (Z_1^{(r_G)}, \cdots, Z_T^{(r_G)})$.

Proposition 3: $Z^{(r_G)}$ asymptotically has the true joint distribution of Z as r_G grows without bound.

- Let $Z_r^{(r_G)}$ be a vector drawn according to the Gibbs scheme after r_G resamplings. Define a simulator for the logarithmic score, s_i, by $\tilde{s}_i(Z^{(r_G)}, y, X, \beta, \Omega, n, R) \equiv \frac{1}{R}\sum_r h(Z_r^{(r_G)}, y, X, \beta, \Omega)$, where R is the (finite) number of terminal simulations drawn, and r_G the number of Gibbs resamplings used for each simulation.

2.4 Results:

1. $\tilde{s}_i$ is a continuous function of parameters and random draws;
2. $\tilde{s}_i$ is unbiased for the true s_i asymptotically with r_G;
3. the MSS/GSS estimator is CUAN provided r_G rises at a rate at least as fast as $\log N$.

Theory of GSS

Geman and Geman (1984): *finite* sites and states problem. Given our interest in the continuous normality case, we need new results. We exploit results in Orey (1971) about the behavior of **Markov chains**:

Consider a set A with positive Lebesgue measure. We give five definitions:

Definition 1 *A Markov process is* irreducible *if the probability that the process ever visits the set A, starting from any x, is positive.*

Definition 2 *A Markov process is* recurrent *if the probability that it ever visits the set A, starting from any point x, is* 1.

Definition 3 *A Markov process is* aperiodic *if with positive probability it goes from x to set A in one move.*

Definition 4 *A Markov process is* uniformly recurrent *if the probability of reaching state A within n transitions is bounded below by a positive number, uniformly in the starting point x.*

Definition 5 *A density $f(x)$ is an* invariant *of the Markov process if it describes the distribution of the outcomes of the process irrespective of the number of transitions.*

In the Gibbs sampler application, one transition corresponds to one updating cycle: start from $(Y_1^{(0)}, \cdots, Y_J^{(0)})$, draw $\tilde{Y}_1$ from $[\tilde{Y}_1|Y_2^{(0)}, \cdots, Y_J^{(0)}]$, draw $\tilde{Y}_2$ from $[\tilde{Y}_2|\tilde{Y}_1, Y_3^{(0)}, \cdots, Y_J^{(0)}], \cdots$, draw $\tilde{Y}_j$ from $[\tilde{Y}_j|\tilde{Y}_1, \cdots, \tilde{Y}_{j-1}, Y_{j+1}^{(0)}, \cdots, Y_J^{(0)}], \cdots$, draw $\tilde{Y}_J$ from $[\tilde{Y}_J|\tilde{Y}_1, \cdots, \tilde{Y}_{J-1}]$, where the $\tilde{Y}'s$ are drawn from the correct univariate conditional normal truncated density, as described in Proposition 2(a). These drawings are done according to the scheme of Proposition 1. Specifically, let $[\tilde{Y}_j|\tilde{Y}_{-j}]$ denote the conditional distribution of $\tilde{Y}_j$ conditional on the $(J-1)\times 1$ vector excluding the $j-th$ random variable. From Proposition 2(a), $\tilde{Y}_j|\tilde{Y}_{-j} \sim N(\mu_{j|-j}, \Sigma_{j|-j})$ conditional on $a^* \leq M \cdot \tilde{Y} \leq b^*$, where $\mu_{j|-j} = \mu_j + \Omega_{j,-j} \cdot \Omega_{-j,-j}^{-1} \cdot (\tilde{Y}_{-j} - \mu_{-j})$, $\mu_k \equiv (X\beta)_k$, and $\Sigma_{j|-j} = \Omega_{jj} - \Omega_{j,-j} \cdot \Omega_{-j,-j}^{-1} \cdot \Omega_{-j,j}$. Then it follows that the truncated multivariate normal distribution Y conditional on the compact region $a^* \leq MY \leq b^*$ will be an invariant of this process.

Proposition 6 *For compact support $B \equiv [a,b]$, $-\infty < a < b < \infty$, the joint density of $(Y_1^{(n)}, \cdots, Y_J^{(n)})$ converges in L_1 norm to the true joint density, $n(z - X\beta, \Omega, \bar{a}, \bar{b})$ at a geometric rate in n.*

Proof (sketch):

- Define $p(n,x,y)$ for $(x,y) \in B$ to be the density of $Y^{(n)}$ starting from $Y^{(0)} = x$; (given constructively above)
- Also by construction, p is continuous on B, $p(1,x,y) > 0$, and $p(n,x,y) = \int p(n-1,x,z) \cdot p(1,z,y)dz > 0$ for $n > 1$. Since, by assumption, B is compact, $p(1,x,y)$ is bounded positive on B.
- It follows the process is *uniformly recurrent* with the truncated multivariate normal $n(z - X\beta, \Omega, \bar{a}, \bar{b})$ as its invariant.
- Then, Theorem 7.2 in Orey (1971) implies that the L_1 distance $\|p(n, Y^{(0)}, y) - n(z - X\beta, \Omega, \bar{a}, \bar{b})\|$ converges to 0 at a geometric rate as $n \to \infty$. □

Need compact support: For example, consider the one-factor model $\epsilon_i = \alpha\zeta + u_i$, where ζ and u_i are independent standard normal variates, and α is a parameter. In this model, as $\alpha \to \infty$, the rate of convergence of the Gibbs sampler from an initial density to the limiting density is slower and slower.

3 Main Asymptotic Properties of MSS Estimation

Summary of Conclusions:

1. The SSML/GHK estimator will be consistent and uniformly asymptotically normal (CUAN) with the number of observations $N \to \infty$, as long as R rises at least as fast as $\sqrt{N}$.

2. The MSS/GSS estimator will be CUAN as $N \to \infty$, for any *finite* number of simulations R provided the number of Gibbs resamplings r_G used to calculate each simulation rises at least as fast as $\log N$.

3. The MSS/GHK-Ratio estimator (using the GHK/SRC simulator to simulate the numerator likelihood derivatives and the denominator likelihood probabilities R times) will be CUAN as long as R rises at least as fast as $\sqrt{N}$.

4. MSS/AR (based on acceptance-rejection arguments – Devroye (1986)): CUAN and asymptotically efficient as $N \to \infty$ for any (finite) R. **Problem:** Discontinuous function.

NOTE: Same underlying random variates, used to simulate the $h(\cdot)$, $\ell(\cdot)$, and $\ell(\cdot)$ functions, be used throughout the iterative searches.

Theorem 7 *Assume that:*

1. *the parameter θ is contained in a compact set Θ, and that the true value θ^* is in the interior of Θ;*

2. *the score $s_i(\theta)$ is continuously differentiable on Θ;*

3. *the score and its derivatives, and the simulated score, are dominated by a function independent of θ with finite first and second order moments;*

4. *$E_i s_i(\theta) = 0$ if and only if $\theta = \theta^*$, and that $J = -E_i s_{i\theta}(\theta^*)$ is positive definite, where E_i denotes expectation with respect to the distribution of the observations;*

5. *observations and simulators are i.i.d. across observations;*

6. (a) *the simulation bias*

$$B_N(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [\mathbf{E}_i \tilde{s}_i(\theta) - s_i(\theta)], \tag{13}$$

converges to zero in probability, uniformly in θ*, and*

(b) *the simulation residual process*

$$\zeta_N(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i(\theta), \text{ with} \tag{14}$$

$$\xi_i(\theta) \equiv [\tilde{s}_i(\theta) - \mathbf{E}_i \tilde{s}_i(\theta) - \tilde{s}_i(\theta^*) + \mathbf{E}_i \tilde{s}_i(\theta^*)].$$

is stochastically equicontinuous

Then, MSS satisfies $\hat{\theta}_N \xrightarrow{p} \theta^*$ *and* $\sqrt{N}(\hat{\theta}_N - \theta^*) \xrightarrow{d} Z \sim \mathcal{N}(0, J^{-1} + J^{-1}QJ^{-1})$*, where* $Q = E[\tilde{s}_i(\theta^*) - \mathbf{E}_i \tilde{s}_i(\theta^*)][\tilde{s}_i(\theta^*) - \mathbf{E}_i \tilde{s}_i(\theta^*)]'$.

Proof (sketch):

Regularity assumptions: functions $p_k(x,\theta)$, $c_k(x,\theta)$, $a(x,\theta)$, and $B(x,\theta)$ are all continuously differentiable in θ, and that these functions and their derivatives are dominated by a square-integrable function $m(x)$.

Corollary 8 *If the simulation process is unbiased, or if the bias in an observation is dominated by a positive function independent of* θ *whose expectation is of order* $(1/\sqrt{N})$*, then the simulation bias converges to zero.*

In the canonical LDV model, the simulation process is:

1. unbiased for MSS/AR for any R;

2. the bias is suitably dominated for MSS/GHK-Ratio as long as $\frac{\sqrt{N}}{R} \to 0$; and

3. the bias is suitably dominated for MSS/GSS as long as $\frac{\log N}{r_G} \to 0$).

Corollary 2. *Assume that the simulator* $\tilde{s}_i(\theta)$ *is probably Lipschitz on* Θ*. Then, the simulation residual process is stochastically equicontinuous.*

In the canonical LDV model:

1. the GHK simulator is continuously differentiable;

2. the Gibbs simulator is continuously differentiable; and

3. the AR simulator is *probably Lipschitz.*

This implies that *Stochastic Equicontinuity* holds for all three MSS estimators.

Proof of Corollary 2 (sketch):

Without loss of generality, assume $\Theta \in [0,1]^k$. For any integer j, partition this cube into 2^{kj} small cubes with sides of length 2^{-j}. Let Θ_j be a set containing one point selected from each cube that intersects Θ. These points can be selected so that $Q_\delta(\theta) \leq K\delta^\gamma$ for $\theta \in \Theta_j$. Define $\theta_j(\theta)$ to be the mapping from θ into the point in Θ_j that is in the same region of the partition; then $|\theta - \theta_j(\theta)| \leq 2^{-j} \equiv \beta_j < 1$.

Define the function

$$B_{ij}(\theta) = CASES$$

and note that this function is region-wise constant on partition j. Using the Lipschitz hypothesis, one has $|\tilde{s}_i(\theta) - \tilde{s}_i(\theta_j(\theta))| \leq B_{ij}(\theta)$. Also, for j large enough so that $\beta_j < \delta_o$,

$$E\ B_{ij}(\theta)^2 \leq E_i\{(1 - Q_{\beta_j}(\theta_j(\theta))m_i^2\beta_j^2 + Q_{\beta_j}(\theta_j(\theta))2m_i^2\}$$

$$\leq E_i\{m_i^2\beta_j^2 + 2m_i^3\beta_j\} \leq 3\beta_j E_i m_i^3 \equiv \delta_j^2.$$

Define $\delta_j^2 = 2E_i m_i^3$ for $\beta_j \geq \delta_o$. Then,

$\sum_{j=2}^{\infty}[\log N_j]^{1/2}(\delta_{j-1} - \delta_j) = \sum_{j=2}^{\infty}[kj\ log2]^{1/2}(2^{-(j-1)} - 2^{-j}) \cdot 3E_i m_i^3 < +\infty$. Then, the condition for the Ossiander result holds, and stochastic equicontinuity follows.

Consider the special case. GHK and GSS are continuously differentiable on Θ, so they are Lipschitz with probability one.

Now consider the AR simulator. Given a fixed sequence of random generators v_r for $r = 1, 2, \cdots$, the acceptance-rejection procedure can be described as one in which trials are rejected until the criterion $v_r \in D(x_i, \theta, d_i)$ is met, then $\tilde{s}_i(\theta) = h(v_r, \theta, x_i, d_i, y_i)$ for the accepted v_r. Given $\theta \in \Theta$ and $\delta > 0$, let $N_\delta(\theta)$ denote a δ-neighborhood of θ. Let $R_\delta(x_i, \theta, d_i)$ denote the probability that a trial will lead to rejection for all $\theta' \in N_\delta(\theta)$, equal to the integral of the truncated standard normal density over the intersection of $D(x_i, \theta', d_i)^c$ for θ' in the neighborhood. Let $A_\delta(x_i, \theta, d_i)$ denote the probability that a trial will lead to acceptance for all $\theta' \in N_\delta(\theta)$, equal to the integral of the truncated standard normal density over the intersection of $D(x_i, \theta', d_i)$ for θ' in the neighborhood. The probability of acceptance on the same trial for all $\theta' \in N_\delta(\theta)$ is then $A_\delta(x_i, \theta, d_i)/(1 - R_\delta(x_i, \theta, d_i))$.

Suppose that $p_k(x, \theta) \cdot v \leq c_k(x, \theta)$ for $k = 1, \cdots, K$ defines the set $D(x, \theta, d)$. The compactness of the support of v, the continuous differentiability of p_k and c

in θ, and the dominance assumption, implies by Taylor's expansions that

$$|p_k(x, \theta') \cdot v - p_k(x, \theta) \cdot v| \leq m(x) \cdot |\theta' - \theta| \leq m(x) \cdot \delta,$$

$$|c(x, \theta') - c(x, \theta)| \leq m(x) \cdot |\theta' - \theta| \leq m(x) \cdot d.$$

It then follows that the probability that the simulator has a discontinuity in $N_\delta(\theta)$ satisfies

$$Q_{i\delta}(\theta) = 1 - A_\delta(x_i, \theta, d_i)/(1 - R_\delta(x_i, \theta, d_i))$$
$$\leq \frac{P(\{v | p_k(x, \theta) \cdot v - c_k(x, \theta)| \leq 2m(x) \cdot \delta,\ k = 1, \cdots, K\})}{P(\{v | p_k(x, \theta) \cdot v - c_k(x, \theta) \leq 0,\ k = 1, \cdots, K\})}.$$

But $p_k(x, \theta) \cdot v$ is standard normal, implying

$$Q_{i\delta}(\theta) \leq \frac{\sum_{k=1}^{K} [\Phi(c_k(x_i, \theta) + 2m(x_i)\delta) - \Phi(c_k(x_i, \theta) - 2m(x_i) \cdot \delta)]}{\prod_{k=1}^{K} \Phi(c_k(x_i, \theta))}.$$

This implies that the AR simulator is probably Lipschitz. □

3.1 Schematic Outline of Proof

$$\hat{\theta}_{MSS}: \quad \frac{1}{\sqrt{N}} \sum_i \tilde{s}_i(\hat{\theta}_{MSS}) = 0$$

$$\frac{1}{\sqrt{N}} \sum_i \tilde{s}_i(\hat{\theta}_{MSS}) =$$

$$\frac{1}{\sqrt{N}} \{$$

$$s_i(\theta^*) \tag{15}$$

$$+\tilde{s}_i(\theta^*) - s_i(\theta^*) \tag{16}$$

$$+s_i(\hat{\theta}) - s_i(\theta^*) \tag{17}$$

$$+\tilde{s}_i(\hat{\theta}) - s_i(\hat{\theta}) - \tilde{s}_i(\theta^*) + s_i(\theta^*) \tag{18}$$

$$\}$$

Classical Terms:

$$(15): \text{ Asymptotically Normal}$$

$$(17): \text{ proportional to } \sqrt(N)(\hat{\theta} - \theta^*)$$

Simulation-Induced Terms:

$$(16)\text{: Simulation Bias} \equiv \frac{1}{\sqrt{N}} \{E\tilde{s}_i(\theta) - s_i(\theta)\}$$

$$(18)\text{: Simulation Residual Process} \equiv$$

$$\frac{1}{\sqrt{N}} \{E\tilde{s}_i(\theta) - s_i(\theta) - E\tilde{s}_i(\theta^*) + s_i(\theta^*)\}$$

4 Simulation-Based Testing

Consider the classic inference problem on an unknown parameter vector θ of dimension $p \times 1$. We wish to devise tests for (a) a set of r linear hypotheses denoted by $R\theta = q$, where the known matrix R is of dimension $r \times p$ and q denotes the $r \times 1$ vector of known real numbers; and for (b) a set of r nonlinear restrictions denoted by $g(\theta) = 0$ where 0 has dimension $r \times 1$ and the (at least twice continuously differentiable) function $g(\cdot)$ from $R^p \to R^r$.

Suppose that, using the simulation methods discussed above, we have obtained a simulation-based estimator for θ which possesses good asymptotic properties, e.g., CUAN, Asymptotically Best CUAN, etc. The fundamental point made by Hajivassiliou (2001) was that to devise test procedures based on the classic "Trinity" of testing approaches, we will need to evaluate *only once* the "trinity" test statistics using the simulation-based estimator $\hat{\theta}$. This is in sharp contrast to carrying out simulation-based *estimation*, where the estimation criterion function (Method-of-Moments distance metric, Log Likelihood Function, etc.) will need to be evaluation thousands or millions of times before we converge to the solution.

Consequently, we can afford to perform a very large number of replications in our evaluation of the simulated test statistic based on $\hat{\theta}$, therefore guaranteeing very high precision in the evaluation-by-simulation of the simulated statistics.

4.1 Simulated Wald Statistics

4.2 Simulated Likelihood Ratio Statistcs

4.3 Simulated Lagrange Multiplier Statistics

5 Research in Progress and Future Directions

- **Distributed Computing – "Supercomputing for Poor Researchers."** Hajivassiliou (1995); Traub (1995); Traub and Pashkov (1995).
- **Deterministic vs. Random Rules:** Wozniakowski (1991), Traub and Wozniakowski (1995), Hajivassiliou (1995), Hajivassiliou and Rust (1995).
- **Finance:** pricing derivatives. Traub and Pashkov (1994). Pricing 10-year Mortgage Cash flows – 360 dimensional integrals.
- **Investment:** Financing decisions of firms. Hajivassiliou and Ioannides (1995).
- **Entry and Exit:** Endogenous Attrition of firms. Corres, Hajivassiliou and Ioannides (1994).