LECTURES ON SIMULATION-ASSISTED STATISTICAL INFERENCE

Daniel McFadden
Professor of Economics and Director, Econometrics Laboratory
University of California, Berkeley CA 94720-3880
Tel: 510-643-8428, Fax: 510-642-0638,
E-Mail: mcfadden@econ.berkeley.edu
Internet: http://elsa.berkeley.edu/~mcfadden

EC-squared Conference, Florence, Italy

December 12-14, 1996

REMARKS: Lectures 1-4 were prepared for presentation at Econometrics Days 1995, Umea, Sweden, June 1995. Lecture 2 uses material from McFadden and Ruud (1994). Lecture 3 is extracted from Beckert, Eymann, and McFadden (1994). Lecture 4 uses material from Hajivassiliou, McFadden and Ruud (1995). Lecture 5 is extracted from McFadden (1996). Lecture 6 is extracted from McFadden and Train (1996). The EC-squared presentation is based on Lecture 6.

LECTURE 1. INTRODUCTION TO STATISTICAL COMPUTATION

The computations required in statistical analysis often involve numerical approximations such as those in the list below:

- Data Roundoff and Computer Floating Point Representation
- Approximation of Transcendental Functions
- Matrix Inverse Floating-Point Roundoff
- Standard CDF and Inverse CDF Functions
- Numerical Differentiation
- Numerical Integration (e.g., Expectations)

These approximations can cause bias and inconsistency in statistical estimates, and must be controlled to obtain satisfactory statistical properties. Even highly accurate approximations may introduce non-negligible errors in ill-conditioned problems, or in applications where sample sizes are large and statistical noise is small. Thisted (1988) gives a good general discussion of these issues.

An elementary example shows the impact of approximation error on a statistic: Suppose i.i.d. draws are made from the density $f(x^*) = 2x^*$, $0 < x^* < 1$. Suppose each observation x is recorded to r digits; i.e., if the decimal representation of x^* is $0.d_1d_2\cdots d_rd_{r+1}\cdots$, then the recorded observation x is $0.d_1d_2\cdots d_r$. This truncation may occur in data collection, or in computer floating point representation. Suppose $\bar{x}$ is the sample mean from N observations. Is $\bar{x}$ consistent for the mean $\mu = 2/3$?

The digits deep in a decimal representation are nearly uniformly distributed. (This is the basis for generation of pseudo-random numbers.) Then,

$$(\bar{x} - \mu) = \frac{1}{N}\sum_{n=1}^{N}(x_n^* - \mu) + \frac{1}{N}\sum_{n=1}^{N}(x_n - x_n^*) .$$

Apply a SLLN. The first RHS term converges a.s. to zero. The second RHS term converges a.s. to $-4.5\times10^{-r-1}$. Then, the sample mean is inconsistent unless $r \longrightarrow +\infty$ when $N \longrightarrow +\infty$. If $r \longrightarrow +\infty$, multiply the equation by $\sqrt{N}$ and apply a CLT. Then, $\sqrt{N}(\bar{x} - \mu)$ is asymptotically normal with mean zero (and variance 5/9) only if r grows more rapidly than log N. The table below gives N versus the r required to keep the expectation of the sample mean within <u>one</u> standard deviation of μ:

N	r
100	2
1,000	3
100,000	5
10,000,000	7

Thus, in a Monte Carlo sample of size 10,000 recorded to 3 digits, round-off error is a substantial part of imprecision in an estimate of the mean.

Computational feasibility often limits the precision of approximations. Important research questions are the tolerance of statistical procedures to approximation errors, interaction between approximation method and statistical properties, and design of error-tolerant inference tools.

In statistical inference, it may be possible to use relatively crude approximations that reduce computation time, and still get decent statistical properties. This is true because approximation noise has some of the properties of data noise. The similarities between data noise and approximation noise can be enhanced by introducing some ancillary randomization in numerical approximations, particularly if the ancillary random factors make the approximations unbiased and independent across observations. Then, statistical laws such as the law of large numbers may apply to average out approximation noise, just as they do data noise. Estimators incorporating ancillary random factors will often lack invariance with respect to this randomization, and lose statistical efficiency due to approximation noise. This is the price of computational feasibility.

EXAMPLE: A model for expenditure on food purchased away from home is

$$F^* = \beta_1 + Y\cdot\beta_2 - \sigma\varepsilon \text{ and } F = \max(F^*,0),$$

where Y is total expenditure and F^* is latent desired expenditure, introduced because this expenditure category is often censored at zero. If ε has density $g(\varepsilon)$ with mean zero, variance one, and CDF $G(\varepsilon)$, then one obtains the Tobit model

$$\mathbf{E}\{F|F > 0,\ Y\} = \beta_1 + Y\cdot\beta_2 - \sigma\cdot M\left(\frac{\beta_1 + Y\cdot\beta_2}{\sigma}\right),$$

where

$$M(t) = \mathbf{E}(\varepsilon | \varepsilon<t) = \frac{1}{G(t)} \cdot \int_{-\infty}^{t} \varepsilon \cdot g(\varepsilon) \cdot d\varepsilon \ .$$

If ε is normal, then $M(t) = \phi(t)/\Phi(t)$ is the inverse Mills ratio. If $M(\cdot)$ can be computed conveniently, then a nonlinear regression

$$F = \beta_1 + Y \cdot \beta_2 - \sigma \cdot M(\frac{\beta_1 + Y \cdot \beta_2}{\sigma}) + \xi$$

applied to the subsample with $F > 0$ is consistent for the parameters. In the ε normal case, most statistic packages evaluate $\phi(t)$ and $\Phi(t)$ to limited accuracy at extreme arguments, so the ratio $M(t)$ can be quite imprecise. The table below illustrates the accuracy of approximations in a standard econometric package (the "exact" column comes from an approximation that is known to to be accurate to five significant digits:

t	"Exact" M(t)	Em. Software Internal Approximation to M(t)	Ratio of Em. Software Approx. to $\phi(t)$ and $\Phi(t)$
-5	5.18647	5.18650	5.18650
-6	6.15845	6.15848	6.15848
-7	7.13751	7.13755	7.13756
-8	8.12129	8.12137	8.27397
-9	9.10836	9.10852	$+\infty$
-10	10.09779	10.09809	$+\infty$

This can cause problems in iteration to nonlinear estimators, and in large economic data sets containing outliers.

EXAMPLE: Consider $x,z \in [0,1]$, and a function $f(x,z)$ that is easy to evaluate. You are interested in $\overline{G}_N = N^{-1} \sum_{n=1}^{N} G(z_n)$, where $n = 1,...N$ indexes independent observations and $G(z_n) = \int_0^1 f(x,z_n)dx$, with z_n a sample observation from a uniform distribution on [0,1]. Suppose explicit integration is impossible, and a K-point open trapezoidal rule is used to approximate $G(z_n)$; i.e.,

$$\hat{G}(z_n) = \frac{1}{K}\left(\frac{3}{2}f_1 + f_2 + \ldots + f_{K-2} + \frac{3}{2}f_{K-1}\right),$$

where $f_k = f(k/K,z_n)$. The approximation error is $O(M/2K^2)$, where M is a bound on the derivative of f. In general, this error is not reduced by the averaging used to form the estimator of $\overline{G}_N$, and is not attenuated as N increases.

Now add some randomization. Draw v_{nk} at random from [0,1], and use the approximation $\overline{G}(z_n) = \frac{1}{K}\cdot\sum_{k=1}^{K}\tilde{f}_k$, where $\tilde{f}_k = f((k+v_{nk}-1)/K,z_n)$. The magnitude of the approximation error is unchanged, but it is now unbiased. Hoeffding's inequality implies

$$\text{Prob}\left(\left|N^{-1}\sum_{t=1}^{N}(\hat{G}(z_n) - \overline{G}_N)\right| > \varepsilon\right) < 2\cdot\exp\left(-\frac{N\varepsilon^2K^2}{2M^2}\right).$$

Then, for large N, the approximation with randomization will be far more accurate than straight numerical integration. The table below gives the probability that the approximation error in the formula using randomization will exceed the approximation error in the formula without randomization; this table is constructed for K = 10 and a function for which M = 1.

N	Probability that error in randomized formula exceeds error in nonrandomized formula
100	0.573
1000	3.73×10^{-6}
5000	7.10×10^{-28}

To illustrate, suppose $f(x,z) = \log[1 + \exp(x\cdot z)]$. Then $0 \le f(x,z) \le 1.32$ for $x,z \in [0,1]$, and $0 \le \nabla_x f(x,z) \le 1$. Suppose that a function evaluation takes 10^{-5} seconds. Suppose K = 10 and N = 5000. Estimation of $\overline{G}_N$ by direct numerical integration requires 5×10^4 function evaluations, and takes approximately one-half second. The approximation error is on the order of 0.005. Now suppose randomization is introduced. The number of function evaluations is unchanged, and the probability of an approximation error that exceeds 0.002 is less than 0.001.

The preceding example illustrates that introduction of some randomization can substantially improve the precision of statistical averages with embedded numerical approximations. When one moves to higher-dimensional integrals, the number of function evaluations required for ordinary numerical integration goes up with the power of the dimension, and quickly becomes computationally infeasible. The example above extended to an integral of dimension 8 requires 5×10^{11} function evaluations, and requires 58 days. Then, Monte Carlo techniques may be more accurate, but more importantly may be necessary for computational feasibility.

LECTURE 2. SIMULATION-ASSISTED STATISTICAL INFERENCE

Many econometric estimators are based on the *method of moments.* The model setup is that there is a vector of m functions $g(y,z,\theta)$ of variables (y,z) and a $k\times 1$ parameter vector θ that appears in the underlying conditional density $f(y|z,\theta)$ of y given z, with the property that the expectation of $g(y,z,\theta)$ given z is zero if and only if θ equals its true value θ_o. For a random sample n = 1,...,N of observations (y_n,z_n), a moments estimator solves

$$\hat{\theta}_N = \operatorname{argmin}_\theta \left\| \mathbf{E}_N g(y,z,\theta) \right\| ,$$

where $\|\cdot\|$ is a distance metric and $\mathbf{E}_N$ denotes the empirical (sample) mean.[1] Maximum likelihood estimation fits this setup with $g(y,z,\theta)$ the score of the conditional likelihood function of y given z: $g(y,z,\theta) \equiv \nabla_\theta \log f(y|z,\theta)$. Generalized Method of Moments (GMM) estimators typically start from functions g(y,z) with expectations $\gamma(z,\theta_o) \equiv \mathbf{E}_{y|z}g(y,z)$, and work with the moment conditions $g(y,z,\theta) \equiv g(y,z) - \gamma(z,\theta)$. Linear and nonlinear least squares and instrumental variables estimators are of this form. In some problems, $g(y,z,\theta)$ will be difficult to express analytically or compute numerically, but will be relatively easy to approximate by simulation methods. For example, $g(y,z,\theta)$ may involve an expectation $\gamma(z,\theta)$ that is analytically intractable. It can be approximated by averaging the integrand over a Monte Carlo simulation sample of size r_N drawn from the conditional density $f(y|z,\theta)$ for each trial θ, or by averaging the ratio $g(y,z,\theta)/h(y)$ over a Monte Carlo simulation sample drawn from a convenient density h(y). The idea of *simulation estimators* is to replace analytically intractable pieces in the moment conditions by such simulation approximations to obtain a computationally tractable simulated moment

---

[1] This setup includes both classical method of moments, where there are exactly as many moment conditions as there are unknown parameters, and generalized method of moments where the number of moment conditions can exceed the number of parameters. In the former case, the moments estimator will ordinarily be a root of the sample moment conditions, $0 = \mathbf{E}_N g(y,z,\hat{\theta}_N)$. In the latter case, the distance metric will be of the form $\|x\| = x'\Omega^{-1}x$, where Ω is a positive definite matrix, and the moments estimator will be a root satisfying $\Gamma'\Omega^{-1}\mathbf{E}_N g(y,z,\hat{\theta}_N) = 0$, where $\Gamma = \mathbf{E}_N \nabla_\theta g(y,z,\theta)$.

$g^r(y,z,\theta)$ for each θ. Then, the simulation estimator is $\hat{\theta}_N = \text{argmin}_\theta \left\| \mathbf{E}_N g^{r_N}(y,z,\theta) \right\|$.

An example where simulation may be useful is the general exponential family, which has scores of the form

$$\nabla_\theta l(\theta) = w(z,\theta)\cdot\tau(y,z,\theta) + c(z,\theta) = w(z,\theta)\cdot[\tau(y,z,\theta) - \mathbf{E}\{\tau(y,z,\theta) \mid z,\theta\}]$$

with the second equality following by application of the information identity. If τ is easy to compute, but w and/or $\mathbf{E}\{\tau(y,z,\theta) \mid z,\theta\}$ are hard to compute, then this score is a candidate for simulation-based inference, with simulators for w and $\mathbf{E}\{\tau(y,z,\theta) \mid z,\theta\}$. If the simulation error in w is asymptotically negligible, then the arguments given below establish that R unbiased draws per observation to simulate $\mathbf{E}\{\tau(y,z,\theta) \mid z,\theta\}$ yields an asymptotic efficiency loss of $1 + 1/R$.

Simulation methods have long been used by policy analysts and theorists to provide insights into problems that they cannot solve analytically. The common practice of "tuning" model parameters so that simulation outcomes "match" observed outcomes is a crude form of simulation estimation. In most cases, this has not been recognized as a form of statistical inference, and the statistical properties of the resulting estimators have not been studied. However, the approach suggests a simple but fundamental insight: if one finds Nature's data generation process, then data generated (by simulation) from this process should leave a trail that in all aspects resembles the real data; otherwise, evidence of discrepancies should appear. This match can be made in terms of the score of the data generation process, or of moments associated with this process. However, the match can also be made in terms of the score or other moments of a misspecified data generation process chosen for its simplicity or tractability; this is the idea of indirect inference. A key insight for the simulation process itself is that approximation errors have the same qualitative properties as data errors, so that the statistical tools used to handle data errors can also be applied to handle approximation errors.

Four critical conditions are needed to ensure that simulation estimators have decent statistical properties:

° The underlying classical method of moments estimator must be consistent and asymptotically normal; this requires that the population moments be bounded away from zero outside a neighborhood of the true parameter vector, and that these moments satisfy mild regularity conditions.

° There must be sufficient statistical variability across observations to allow a central limit theorem to operate. This is achieved for the simulation draws by independence, strong mixing with suitable conditions, or exchangeability.

° The simulator must not "chatter" as θ changes; i.e., the simulated moment conditions must be a *stochastically equicontinuous* process in θ. Keeping random number generator seeds the same when θ changes usually accomplishes this.

° The simulator for an observation must be *unbiased*, or else its bias must shrink to zero at a sufficient rate as sample size grows; usually a little faster than $\sqrt{N}$.

In its simplest form, the large sample theory of simulation estimators is a straightforward extension of classical large sample theory. Write the simulated moment condition as

$$0 = N^{1/2}\mathbf{E}_N g^r(y,z,\hat{\theta}) \equiv \underbrace{N^{1/2}\mathbf{E}_N g(y,z,\theta_o)}_{\mathbf{A}} + \underbrace{N^{1/2}\mathbf{E}_N[g^r(y,z,\theta_o) - \mathbf{E}_* g^r(y,z,\theta_o)]}_{\mathbf{B}}$$

$$+ \underbrace{N^{1/2}\mathbf{E}_N[\mathbf{E}_* g^r(y,z,\theta_o) - g(y,z,\theta_o)]}_{\mathbf{C}} + \underbrace{N^{1/2}\mathbf{E}_N[g(y,z,\hat{\theta}) - g(y,z,\theta_o)]}_{\mathbf{D}}$$

$$+ \underbrace{N^{1/2}\mathbf{E}_N[g^r(y,z,\hat{\theta})-g^r(y,z,\theta_o)-g(y,z,\hat{\theta})+g(y,z,\theta_o)]}_{\mathbf{E}} ,$$

where $\mathbf{E}_*$ denotes expectation with respect to the distributions used for the simulation draws, conditioned on (y,z). Term **A** is the asymptotic contribution resulting from data noise; it is asymptotically normal by application of a CLT. Term **B** is the asymptotic contribution of simulation noise. It behaves like data noise, and is also asymptotically normal by a CLT. Let Ω denote the asymptotic covariance matrix of **A** + **B**, and note that Ω is the sum of a classical component arising from data noise (the asymptotic covariance of **A**) and a component arising from simulation noise (the asymptotic covariance of **B**).[2] If $r \to +\infty$ at *any* rate, this is sufficient to guarantee that the asymptotic covariance of **B** is zero, so there is no *asymptotic* contribution of simulation to the imprecision of the estimator.

Term **C** is an asymptotic bias. It is zero if the simulator is unbiased; otherwise it must be controlled as N increases by increasing r. The required rate will depend on the structure of the bias, but a common case occurs when g is a nonlinear function of an embedded expectation, $g(y,z,\theta) \equiv \psi(y,z,\gamma(z,\theta))$, where $\gamma(z,\theta) = \mathbf{E}_{y|z,\theta}y$, and g^r is formed by replacing $\gamma(z,\theta)$ by a simulated approximation, $\hat{\gamma}(z,\theta) = \frac{1}{r}\sum_{j=1}^{r} y_j$ with y_j drawn from the conditional density of y given z and θ. Then a Taylor's expansion gives (for each element of the vector of functions g)

$$g^r(y,z,\theta) - g(y,z,\theta) = \nabla_\gamma\psi(y,z,\gamma(z,\theta))\cdot\{\hat{\gamma}(z,\theta) - \gamma(z,\theta)\}$$

$$+ \frac{1}{2}\{\hat{\gamma}(z,\theta) - \gamma(z,\theta)\}'[\nabla_{\gamma\gamma}\psi(y,z,\gamma(z,\theta))]\{\hat{\gamma}(z,\theta) - \gamma(z,\theta)\} + \text{HOT},$$

where HOT denotes higher-order terms. If $\nabla_{\gamma\gamma}g$ is bounded (by an array M), and V/r denotes the covariance matrix of $\hat{\gamma}(z,\theta) - \gamma(z,\theta)$, then

$$|\mathbf{E}_*\{g^r(y,z,\theta) - g(y,z,\theta)\}| \le \frac{1}{2}\cdot\text{tr}\{M\cdot V/r\} .$$

Hence, the condition for $\mathbf{C} \xrightarrow{p} 0$ is $\sqrt{N}/r \longrightarrow 0$. The borderline case where $\sqrt{N}/r$ approaches a positive constant in general leads to estimators that are asymptotically normal, but with a nonzero mean asymptotic bias. This mean can be eliminated by a higher-order bias correction, at some cost in increased asymptotic variance.[3] Lack

---

[2] An estimator of Ω is $\mathbf{E}_N g^r(y,z,\theta_N)\cdot g^r(y,z,\theta_N)'$, where θ_N is any initial CAN estimator. This estimator will be consistent provided the term **C** goes in probability to zero, so that $\mathbf{E}\,\mathbf{E}_* g^r(y,z,\theta) - \mathbf{E}g(y,z,\theta) \to 0$.

[3] Estimators $\hat{\theta}_{N1}$ and $\hat{\theta}_{N2}$ obtained by simulation with $r_1 = c_1\sqrt{N}$ and $r_2 = c_2\sqrt{N}$ draws, respectively, will be asymptotically normal around θ_o with asymptotic means inversely proportional to c_i. Then, $(c_2\hat{\theta}_{N1} - c_1\hat{\theta}_{N2})/(c_2 - c_1)$ is CAN around θ_o with asymptotic mean zero.

of smoothness in ψ or a simulator for γ that is itself biased may lead to more stringent asymptotic rates for r in order to control bias.

Term **D** behaves in large samples like

$$\mathbf{E}_N \nabla_\theta g(y,z,\theta_o)\cdot N^{1/2}[\hat{\theta}_N - \theta_o] = \Gamma\cdot N^{1/2}[\hat{\theta}_N - \theta_o] + \text{HOT} \;,$$

where $\Gamma = \mathbf{E}\nabla_\theta g(y,z,\theta_o)$ is the Jacobian matrix of the vector of moments.[4] Term **E** is an empirical process that has probability limit zero if the simulator is *stochastically equicontinuous* in θ. Stochastic equicontinuity holds if the simulator is smooth in θ with a bounded derivative. It can also hold in the presence of discontinuities if the jumps are not too numerous; see McFadden (1989) and Pakes and Pollard (1989). In practice, one gets a simulator to be smooth in θ by not switching simulation draws as parameters change; this avoids *chatter* in the simulator.

When all of the conditions above hold, plus regularity conditions that ensure that the HOT can be neglected, the simulated method of moments estimator

$$\hat{\theta}_N = \text{argmin}_\theta \left\| \mathbf{E}_N g^{r_N}(y,z,\theta) \right\| = \text{argmin}_\theta \{\mathbf{E}_N g^{r_N}(y,z,\theta)\}'\Omega^{-1}\{\mathbf{E}_N g^{r_N}(y,z,\theta)\}$$

will be CAN with asymptotic covariance $(\Gamma'\Omega^{-1}\Gamma)^{-1}$. Note that this is the same as the asymptotic covariance matrix for a GMM estimator, except that Ω may now contain a contribution from simulation noise. This is no surprise, since moments containing embedded simulators continue under the regularity conditions outlined above to meet all the requirements of the GMM setup. In particular, Ω in the minimum distance formula above can be estimated at any initial CAN estimator. For estimators such as simulated maximum likelihood where $r \to +\infty$ implies $\Gamma = \Omega$ asymptotically, one should nevertheless <u>not</u> use the formula Ω^{-1}, with Ω estimated using the outer product of the simulated moments, to approximate the sample covariance matrix: finite sample simulation noise will make estimates look <u>more</u> precise when Ω^{-1} is used; in fact they are <u>less</u> precise. It is more accurate for simulation estimators to always use the

---

[4] The array Γ can be estimated by $\mathbf{E}_N\nabla_\theta g(y,z,\theta_N)$, where θ_N is any initial CAN estimator of θ_o. When the moments arise from the gradient of a statistical criterion that is being optimized, then Γ is the hessian of the objective function; maximum likelihood estimation is one example.

formula $\Gamma^{-1}\Omega(\Gamma')^{-1}$, with Γ estimated by the sample mean of the Jacobian of the simulated moments, and Ω estimated by the sample mean of the outer product of the simulated moments, as footnoted earlier.

Criteria for Estimation and Simulation

The general method of moments setup discussed above, with embedded simulation, can accommodate a number of different estimation criteria. I give a summary that includes the criteria already discussed:

Method of Simulated Moments (MSM)

In a classical GMM criterion $\left(\mathbf{E}_N g(y,z,\theta)\right)' \Omega^{-1} \left(\mathbf{E}_N g(y,z,\theta)\right)$, one can replace g by a simulator g^r and minimize in θ. The resulting MSM estimator will be CAN under regularity conditions like those outlined earlier. In particular, if the simulator is unbiased, then sample averaging does the work of controlling approximation error, so that MSM is CAN without requiring that the approximation be refined with N.

Method of Simulated Scores (MSS)

If $l(y,z,\theta) = \log f(y|z,\theta)$ is the log likelihood of an observation, then the *score* $g(y,z,\theta) = \dfrac{\nabla_\theta f(y|z,\theta)}{f(y|z,\theta)}$ is the efficient vector of moment functions. Computational approximations may be needed in both the numerator and denominator of g. Typically one cannot construct smooth unbiased simulators of the score, due to the denominator. Then, for consistency, the number of denominator draws must grow more rapidly than $\sqrt{N}$. (There is no rate requirement on the numerator, but presumably one will want to refine this simulator at the same time one is refining the simulator for the denominator.)

Maximum Simulated Likelihood Estimation (MSLE)

If $f(y|z,\theta)$ in the log likelihood function is replaced by an approximation, one obtains a pseudo-likelihood function that can be maximized to obtain a MSLE estimator. Consistency requires that the approximation be refined with N at a rate faster than $\sqrt{N}$. MSLE and MSS coincide when a common approximation is used in both the numerator and denominator of the score; otherwise, MSS gives somewhat broader scope for choosing the approximations. On the other hand, MSLE is somewhat easier to work with computationally, since iterative search methods can be used without worrying that the simulated gradient might conflict with the direction of increase of the simulated objective function.

Simulated EM Algorithm (SEM)

Suppose a latent vector $y^* \in \mathbb{R}^m$ has a conditional density $f(y^*|z,\theta_o)$, where z are exogenous variables and θ is a finite parameter vector. Suppose one observes an <u>event</u> *A* from a family $\mathfrak{A}$; the event *A* can be interpreted as a subset of $\mathbb{R}^m$ that contains the latent vector y^*. Then, the probability of *A* can be written

$$g(A|z,\theta_o) = \int_A f(y^*|z,\theta_o)\cdot dy^* .$$

The conditional distribution of y^* given *A* is

$$h(y^*|z,A,\theta_o) = f(y^*|z,\theta_o)/g(A|z,\theta_o)\cdot\mathbf{1}(y^* \in A) .$$

The log likelihood of a random sample is then

$$L(\theta) = \mathbf{E}_N \log g(A|z,\theta) \equiv \mathbf{E}_N \log \int_A f(y^*|z,\theta)\cdot dy^* ,$$

where $\mathbf{E}_N$ denotes empirical expectation. Define a "pseudo-likelihood" function

$$Q(\theta|\theta') = \mathbf{E}_N \mathbf{E}\{\log f(y^*|z,\theta) \ |z,A,\theta'\} \equiv \mathbf{E}_N \int_A [\log f(y^*|z,\theta)]\cdot h(y^*|z,A,\theta')\cdot dy^* .$$

The <u>EM</u> <u>Algorithm</u> starts from a trial value θ', calculates $Q(\theta|\theta')$ as a function of θ (the "E" step), maximizes $Q(\theta|\theta')$ in θ to obtain a new value θ'' (the "M" step), and iterates to convergence.

If the "E" step is analytically intractable, an alternative is to simulate the "E" step:

$$Q(\theta \mid \theta') = \mathbf{E}_N \int_A [\log f(y^* \mid z,\theta)] \cdot h(y^* \mid z,A,\theta') \cdot dy^*$$

$$= \mathbf{E}_N \frac{\displaystyle\int_A \left[\frac{f(y^* \mid z,\theta') \cdot \log f(y^* \mid z,\theta)}{k(y^* \mid A)}\right] k(y^* \mid A) \cdot dy^*}{\displaystyle\int_A \left[\frac{f(y^* \mid z,\theta')}{k(y^* \mid A)}\right] k(y^* \mid A) \cdot dy^*} \approx \mathbf{E}_N \frac{\mathbf{E}_K \left[\dfrac{f(y^* \mid z,\theta') \cdot \log f(y^* \mid z,\theta)}{k(y^* \mid A)}\right]}{\mathbf{E}_K \left[\dfrac{f(y^* \mid z,\theta')}{k(y^* \mid A)}\right]} ,$$

where $\mathbf{E}_K$ denotes an empirical expectation with respect to a simulation sample drawn from $k(y^* \mid A)$. This is an importance sampling approximation to the "E" step. This simulated EM algorithm (SEM) will have the same asymptotic properties as the classical EM algorithm if K for each observation rises more rapidly than $N^{1/2}$. To achieve this, it is essential that the draws from $k(\cdot \mid A)$ not "chatter" as one iterates. SEM is an alternative path to MSLE.

Indirect Estimation and Encompassing

Given a sample (y_n,z_n), n = 1,...,N, fit the parameters ψ of a "stylized" model for y given z. Alternately, start from the empirical distribution of z and a postulated likelihood $f(y \mid z,\theta)$, generate a simulated sample, and fit the stylized model to the simulated sample. Now iterate to $\hat{\theta}$ that "matches" the parameters ψ from the real and the simulated data, using a criterion such as a likelihood ratio criterion for the stylized model. The model $f(y \mid z,\hat{\theta})$ is said to *encompass* the stylized model since it is able to generate data that the stylized model "explains" in the same way that it explains the real data.[5] This is called *indirect estimation* by Gourieroux, Monfort, and Renault (1993), since the fit is carried out in terms of the stylized parameters ψ rather than the deep parameters θ in the postulated data generation process $f(y \mid z,\theta)$.

---

[5] The term is due to Hendry, Mizon, and Richard (see Mizon and Richard, 1986, Hendry, 1988) who extend the idea to comparison of alternative non-nested models by generating simulated samples from *each* "stylized" model, estimating each model using data generated by the other, and asking whether the resulting fits match fits to the real data.

Indirect estimation is an extension of MSM in which the moments are distances between stylized model parameters estimated from real and simulated data. It is an intuitive and powerful way to organize estimation of deep parameters in complex models. For example, suppose the true data generation process contains embedded optimization or fixed point problems, such as solutions to dynamic stochastic programs, engineering assignment algorithms, or equilibria in economic games. Suppose these solutions can be approximated well enough to draw simulated samples for various values of the deep parameters, but analytic characterization of the data generation process is intractable. Then, indirect inference provides a systematic way to obtain CAN estimates of the deep parameters of the data generation process. The indirect inference idea of fitting a stylized model can in turn be extended to very general classes of "sufficient" statistics for $f(y|z,\theta)$.

LECTURE 3. MULTINOMIAL APPROXIMATION AND SEQUENTIAL SIMULATION[6]

One traditional approach to statistical inference, exemplified by minimum chi-square estimators and goodness-of-fit tests, is to characterize observations in terms of a multinomial model over a finite partition of the sample space. This lecture shows first that under mild regularity conditions, maximum likelihood estimators for multinomial models, defined on nested partitions which are refined as sample size increases, are asymptotically efficient. Second, the lecture reviews methods of inference in multinomial models using simulation, and gives conditions under which these methods are asymptotically equivalent to maximum likelihood estimation for a nested sequence of multinomial models. Third, the lecture introduces a *Multinomial Approximation and Sequential Simulation* (MASS) method that converts the problem of simulating high-dimensional multinomial probabilities to one of simulating a nested sequence of low-dimensional probabilities.

The Asymptotic Efficiency of Nested Multinomial Approximations

Consider inference for the following probability model; the notation generally follows that of Neveu (1965) and Ibragimov-Has'minskii (1981). Let **Z**, **Y**, and Θ be subsets of finite-dimensional Euclidean spaces. Let $\mathfrak{Z}$ and $\mathfrak{Y}$ be σ-fields of subsets of **Z** and **Y**, respectively, so that ($\mathbf{Z}$,$\mathfrak{Z}$) and ($\mathbf{Y}$,$\mathfrak{Y}$) are measurable spaces. Let $\mathfrak{Z}\otimes\mathfrak{Y}$ denote the product σ-field of subsets of $\mathbf{Z}\times\mathbf{Y}$. Let ζ denote a probability measure on $\mathfrak{Z}$. Let $\upsilon_{\mathfrak{Z},\theta}$ denote a family, for $\theta \in \Theta$, of conditional probability measures on $\mathfrak{Y}$ given $\mathfrak{Z}$; i.e., $\upsilon_{\mathfrak{Z},\theta}(\cdot|B)$ is a probability on $\mathfrak{Y}$ for each $B \in \mathfrak{Z}$ and $\theta \in \Theta$. Assume that

---

[6] This lecture is extracted from the paper "Efficient Estimation by Multinomial Approximation and Sequential Simulation" by Walter Beckert, Angelika Eymann, and Daniel McFadden. This research was supported by the E. Morris Cox Research Fund, and by the Mathematical Sciences Research Institute, Berkeley. We have benefited from discussions with Peter Bickel, Axel Boersch-Supan, Vassilis Hajivassiliou, Whitney Newey, David Pollard, Paul Ruud, and Keunkwan Ryu. The innovation in this paper of writing moment conditions as a sequence of conditional moments was motivated by the work of Michael Keane on discrete panel data. The step of utilizing increasing sequences of conditional moments was suggested by work of Paul Ruud on adaptive search algorithms. Key elements in the proof of the asymptotic efficiency of the partitioning algorithm were provided by Peter Bickel, Whitney Newey, and Keunkwan Ryu. A number of the results collected here were first presented at the Rotterdam Conference on Simulation Estimators, June 1991.

$\upsilon_{\mathfrak{Z},\theta}$ is absolutely continuous with respect to a σ-finite measure μ on ($\mathbf{Y}$,$\mathfrak{Y}$) for $\theta \in \Theta$, for z in a subset $\mathbf{Z}_o$ of $\mathbf{Z}$ that occurs with probability one. By the Radon-Nikodym theorem, there then exist probability densities $f(y|z,\theta)$ such that

$$\upsilon_{\mathfrak{Z},\theta}(Y|z,\theta) = \int_Y f(y|z,\theta)d\mu \quad \text{for } Y \in \mathfrak{Y},\ \theta \in \Theta, \text{ and } z \in \mathbf{Z}_o.$$

The interpretation of this setup is that z is a vector of explanatory variables, y is a vector of response variables, θ is a parameter vector, and $f(y|z,\theta)$ is the family of conditional densities of y given z, for $\theta \in \Theta$. Assume random sampling, with observations (z_t,y_t) for t = 1,...,N, from a population that has a true parameter vector θ_o. A sequence of realized sample observations is then an element of the product space $\Omega = \prod_{t=1}^{\infty} (\mathbf{Z}\times\mathbf{Y})$, endowed with the product σ-field $\mathfrak{S}$ and product probability measure λ_θ; see Neveu (1965, Corollary, p. 83).

For a function g(y,z), the notation $\mathbf{E}_N g(y,z) \equiv \frac{1}{N}\sum_{t=1}^{N} g(y_t,z_t)$ will be used for the empirical expectation of g given a random sample of size N, and $\mathbf{E}_{\theta_o} g(y,z)$ will denote the expectation of g at the true parameter vector θ_o. The (mean) *full information log likelihood* of the sample is

$$(1) \qquad L_N(\theta) = \mathbf{E}_N \log f(y|z,\theta),$$

and a Full Information Maximum Likelihood (FIML) estimator $\hat{\theta}_N$ maximizes this criterion.

Suppose C_n is a *finite* partition of $\mathbf{Y}$, with $\mu(A) > 0$ for $A \in C_n$. Suppose $C_1,C_2,C_3,...$ is a *nested* sequence of finite partitions that generate $\mathfrak{Y}$; i.e., $C_n \subseteq C_{n+1}$ and $\mathfrak{Y}$ is the smallest σ-field of subsets of $\mathbf{Y}$ that contains $\bigcup_{n=1}^{\infty} C_n$. It is usually rather simple in practical problems to choose partitions C_n that generate a specified $\mathfrak{Y}$. For example, if $\mathfrak{Y}$ is the Borel σ-field of subsets of $\mathbb{R}$, then the real numbers $\pm k/2^n$ for integers $k \le 4^n$ define the end points of half-open intervals in a partition C_n of $\mathbb{R}$, and every closed interval in $\mathfrak{Y}$ is a monotone limit of unions of sets in $\mathcal{C}_n$ as $n \to +\infty$.[7]

Let $\mathfrak{Y}_n$ denote the finite field of subsets of **Y** generated by $\mathcal{C}_n$; then every set $A \in \mathfrak{Y}$ can be written as a monotone limit $A = \bigcap_{n=1}^{\infty} A_n$ of sets $A_n \in \mathfrak{Y}_n$ and $\mu(A_n - A) \to 0$. Define *multinomial response probabilities* on $\mathfrak{Y}_n$:

(2) $$P(A|z,\theta) = \int_A f(y|z,\theta)d\mu \quad \text{for } A \in \mathfrak{Y}_n,\ z \in \mathbf{Z},\ \theta \in \Theta.$$

The $\mathcal{C}_n$ are interpreted as partitions imposed by the analyst to obtain multinomial classifications of response, and $P(A|z,\theta)$ for $A \in \mathcal{C}_n$ are the corresponding multinomial probabilities.[8] Let $C_n(y)$ denote the element of $\mathcal{C}_n$ containing observation y_t. Then, the (mean) *limited information multinomial log likelihood* of the sample is

(3) $$L_N(\theta|\mathfrak{Y}_N) = \mathbf{E}_N \log P(C_N(y)|z,\theta);$$

Limited Information Maximum Likelihood (LIML) estimators $\tilde{\theta}_N$ maximize this criterion. Limited information estimation will be of interest when it is computationally advantageous because evaluation of $f(y|z,\theta)$ at atoms of $\mathfrak{Y}$ is burdensome, and inference using approximations to $P(C_N(y)|z,\theta)$ is easier than inference using approximations to $f(y|z,\theta)$. LIML may also be of interest when questions of robustness or regularity make FIML estimation problematic.

Let $l(y|z,\theta) = \log f(y|z,\theta)$, and $l_N(y|z,\theta) = \log P(C_N(y)|z,\theta)$. Denote the derivatives of l with respect to θ by $\nabla_\theta l$ and $\nabla_{\theta\theta} l$, and the corresponding derivatives of l_N by $\nabla_\theta l_N$ and $\nabla_{\theta\theta} l_N$. When conditions for differentiation under the integral sign

---

[7] One important application is panel Tobit data, in which a real number $y_t > 0$ or the event $\{y_t \le 0\}$ is observed in each of T periods. Then $\mathfrak{Y}$ is the σ-field generated by the Cartesian product of half-closed nondegenerate intervals that either contain or are disjoint from the nonpositive half-line. Finite partitions, consisting of the Cartesian product of partitions of the line into the nonpositive half-line and an increasingly fine division of the positive half-line, converge to $\mathfrak{Y}$ so long as the width of the interval containing each point $y > 0$ eventually converges to zero.

[8] This notation for multinomial response is due to B. Van Praag; see Van Praag and Hopff (1987).

are met, one has

$$(4) \qquad \nabla_\theta l_N(y|z,\theta) = \frac{\int_{C_N(y)} \{\nabla_\theta l(y'|z,\theta)\} f(y'|z,\theta)d\mu}{\int_{C_N(y)} f(y'|z,\theta)d\mu} = \mathbf{E}_\theta\{\nabla_\theta l(y'|z,\theta)|C_N(y)\}$$

and

$$(5) \quad \nabla_{\theta\theta} l_N(y|z,\theta) = \frac{\int_{C_N(y)} \{\nabla_{\theta\theta} l(y'|z,\theta) + [\nabla_\theta l(y'|z,\theta)]^2\} f(y'|z,\theta)d\mu}{\int_{C_N(y)} f(y'|z,\theta)d\mu} - [\nabla_\theta l_N(y|z,\theta)]^2$$

$$= \mathbf{E}_\theta\{\nabla_{\theta\theta} l(y'|z,\theta)|C_N(y)\} + \mathbf{E}_\theta\{[\nabla_\theta l(y'|z,\theta)]^2|C_N(y))\} - [\mathbf{E}_\theta\{\nabla_\theta l(y'|z,\theta)|C_N(y)\}]^2$$

$$= \mathbf{E}_\theta\{\nabla_{\theta\theta} l(y'|z,\theta)|C_N(y)\} + \mathrm{Var}_\theta\{\nabla_\theta l(y'|z,\theta)|C_N(y))\},$$

where the notation x^2 for a column vector x means $x \cdot x'$.

The main result of this section shows that under mild regularity conditions, the FIML and LIML estimators are both consistent and asymptotically normal (CAN), and are asymptotically equivalent. Thus, there is no reason to prefer FIML to LIML on grounds of asymptotic efficiency. The result depends essentially on the condition that the nested partitions C_n generate the σ-field of observable events $\mathfrak{Y}$ associated with the full information observations. However, there is no minimum rate requirement for the approach of $\mathfrak{Y}_n$ to $\mathfrak{Y}$, so that asymptotic efficiency is achieved by the LIML estimator even if the partitions are refined very slowly.[9]

We start from assumptions that are more than sufficient to guarantee that the FIML estimator is consistent and asymptotically normal (CAN). These assumptions are selected because they are simple and cover most practical applications. For CAN results under weaker regularity conditions see Huber (1967), Ibragimov-Has'minskii (1981), or Pollard (1991).

---

[9] This point was suggested by Whitney Newey. It has led to a considerable strengthening of our result on asymptotic efficiency.

A.1 Θ is a compact subset of $\mathbb{R}^k$, and the true parameter vector θ_o lies in a compact set Θ_o contained in the interior of Θ.

A.2 The density f is measurable on $\mathbf{Y}\times\mathbf{Z}\times\Theta$, and $l(y|z,\theta)$ is continuous and three times continuously differentiable in θ, almost surely in z.[10]

A.3 There exists a function $m(y,z) \geq 1$ that dominates $l(y|z,\theta)$, $\nabla_\theta l(y|z,\theta)$, $\nabla_{\theta\theta} l(y|z,\theta)$, and $\nabla_{\theta\theta\theta} l(y|z,\theta)$, and has $\mathbf{E}_{\theta_o}\{m(y,z)^2|z\} = \lambda(z) < +\infty$ almost surely, and hence $\mathbf{E}_{\theta_o}\lambda(z) < +\infty$.

A.4 The parameter vector θ_o is a unique maximum of $\mathbf{E}_{\theta_o} l(y|z,\theta)$, and a unique root of $\mathbf{E}_{\theta_o}\nabla_\theta l(y|z,\theta)$, for $\theta \in \Theta$ and $\theta_o \in \Theta_o$.

A.5 The *Fisher information* matrix $J(\theta_o) = \mathbf{E}_{\theta_o}[\nabla_\theta l(y|z,\theta_o)]^2$ is nonsingular for $\theta_o \in \Theta_o$.

The compactness condition A.1 is inconsistent with some standard models (e.g., normal data without *a priori* bounds on means), but is tacit in applied work where parameters are limited by computer representation if not otherwise. The regularity conditions A.2 and A.3 guarantee the existence of expectations and allow differentiation and expectation operations to be interchanged. They require the density to be smooth and *positive*. This is extremely restrictive in the absence of A.1, but will be satisfied by most applications once A.1 is imposed. Condition A.4 guarantees global identification. Condition A.5 is a regularity condition that strengthens the implications of A.4. The following textbook theorem establishes that FIML is CAN under assumptions A.1-A.5.

Theorem 3.1. *If assumptions* A.1-A.5 *hold, then the FIML estimator is CAN for* $\theta_o \in \Theta_o$, *with*

$$N^{1/2}(\hat{\theta}_N - \theta_o) \xrightarrow{d} N(0,J(\theta_o)^{-1}).$$

---

[10] We need only the weaker condition that the function be twice continuously differentiable, with a second derivative that satisfies a stochastic weak Lipschitz property: there exists a function m(y,z) with $\mathbf{E}_{\theta_o} m(y,z) < +\infty$, a constant $\delta > 0$, and a continuous real-valued function $\kappa(v)$ satisfying $\kappa(0) = 0$ such that $|\theta'-\theta| < \delta$ implies $|\nabla_{\theta\theta} l(y|z,\theta) - \nabla_{\theta\theta} l(y|z,\theta')| \leq m(y,z)\kappa(\delta)$, almost surely in z. If $\nabla_{\theta\theta} l(y|z,\theta)$ is continuously differentiable, then $\kappa(v) = v$.

Proof: See for example Newey and McFadden (1994), Theorem 2.5 and Theorem 3.3. Another argument is given in Amemiya (1985). For consistency, apply his Theorem 4.1.1. Assumptions A.1, A.2, and A.4 imply his conditions (A), (B), (C) respectively, except for uniform convergence of $\mathbf{E}_N l(y|z,\theta)$ to $\mathbf{E}_{\theta_o} l(y|z,\theta)$. The last condition is provided by his Theorem 4.2.1, which requires assumptions A.2 and A.3; this theorem also implies uniform convergence of $\mathbf{E}_N\nabla_\theta l(y|z,\theta)$ and $\mathbf{E}_N\nabla_{\theta\theta} l(y|z,\theta)$. For asymptotic normality, apply his Theorem 4.1.3. His condition (A) is implied by A.2; his condition (B) is implied by the uniform convergence of $\mathbf{E}_N\nabla_{\theta\theta} l(y|z,\theta)$ and A.5; and his condition (C) is implied by the Lindeberg-Levy CLT applied to $N^{1/2}\mathbf{E}_N\nabla_\theta l(y|z,\theta_o)$. Assumption A.3 implies the information equality, $-\mathbf{E}_\theta\nabla_{\theta\theta} l(y|z,\theta) = \mathbf{E}_\theta[\nabla_{\theta\theta} l(y|z,\theta)]^2$, which gives the final form of the asymptotic covariance matrix. ∎

Next consider the LIML estimator. The following result establishes that when the nested sequence of partitions $\mathcal{C}_N$ generate $\mathfrak{Y}$, the LIML estimator is CAN and is asymptotically equivalent to the FIML estimator.

Theorem 3.2. *Suppose assumptions* A.1-A.5 *hold. Suppose that* $\mathcal{C}_N$ *is a nested sequence of finite partitions of* **Y** *that generate* $\mathfrak{Y}$. *Then, any sequence of LIML estimators* $\tilde{\theta}_N$ *is CAN, with*

$$N^{1/2}(\tilde{\theta}_N - \theta_o) \xrightarrow{d} \mathcal{N}(0,\mathcal{J}(\theta_o)^{-1}) \ .$$

Therefore, the LIML estimators are asymptotically efficient, and hence asymptotically equivalent to the FIML estimator.

Proof: Note that $P(C_N(y)|z,\theta)$ inherits the smoothness properties of $f(y|z,\theta)$, and that as a consequence, $l_N(y|z,\theta)$ inherits the smoothness properties of $l(y|z,\theta)$. Define

$$G_N(\theta) = \mathbf{E}_N\nabla_\theta[l_N(y|z,\theta) - l(y|z,\theta)] \ .$$

The following properties of G_N are easily verified:

(i) $\mathbf{E}(G_N(\theta)|\mathfrak{Y}_{N-1}) = 0$ for all $\theta \in \Theta$; i.e., G_N is a martingale difference sequence.

(ii) $G_N(\theta)$ is equicontinuous in θ; i.e., for each $\varepsilon > 0$ and $\theta \in \Theta$, there exists $\delta > 0$ such that for $|\theta'-\theta| < \delta$, $\sup_N|G_N(\theta') - G_N(\theta)| < \varepsilon$, almost surely in z.

(iii) $\mathbf{E}_{\theta_o}G_N(\theta)^2$ is bounded, uniformly in N, and $N\cdot\mathbf{E}_{\theta_o}G_N(\theta_o)^2 = \mathcal{J}(\theta_o) - \mathcal{J}_N(\theta_o)$, where $\mathcal{J}_N(\theta_o) = \mathbf{E}_{\theta_o}[\nabla_\theta l_N(\theta_o)]^2$ is the Fisher information for the LIML estimator that uses the information $\mathfrak{Y}_N$.

The following argument establishes that $G_N(\theta) \xrightarrow{as} 0$ uniformly on Θ: Given $\varepsilon > 0$, the equicontinuity of G_N on Θ and Θ compact implies G_N is uniformly equicontinuous; i.e., there exists $\delta > 0$ such that $\theta,\theta' \in \Theta$ and $|\theta' - \theta| < \delta$ implies $\sup_n|G_N(\theta') - G_N(\theta)| < \varepsilon/2$. Let Θ_1 be the centers of a finite covering of Θ by open spheres of radius δ. Using conditions (i) and (iii), a theorem of Y. Chow (Stout, 1974) implies that for each $\theta \in \Theta$, $G_N(\theta) \xrightarrow{as} 0$, implying there exists N_o such that $P(\sup_{N\geq N_o}\sup_{\Theta_1}|G_N(\theta)| > \varepsilon/2) < \varepsilon$. But $\sup_{N\geq N_o}\sup_{\Theta}|G_N(\theta)| > \varepsilon$ implies $\sup_{N\geq N_o}\sup_{\Theta_1}|G_N(\theta)| > \varepsilon/2$, so that this event occurs with probability at most ε. Then, $\sup_{\theta\in\Theta}|G_N(\theta)| \xrightarrow{as} 0$. (The preceding argument could have also been carried through with convergence in probability, and this would be sufficient for what follows.)

By A.4, $\psi(\theta) \equiv \mathbf{E}_{\theta_o}\nabla_\theta l(y|z,\theta)$ has a unique root at θ_o. Given $\varepsilon > 0$, let Θ_ε denote an ε-neighborhood of θ_o. Define $\delta = \inf_{\theta\notin\Theta_\varepsilon}|\psi(\theta)|$. By the preceding result, there exists N_o such that for $N \geq N_o$, $\text{Prob}(\sup_{\theta\in\Theta}|G_N(\theta)| > \delta/3) \leq \varepsilon$.

Using Amemiya (1985), Theorem 4.2.1, there exists N_1 such that for $N \geq N_1$,

$$\text{Prob}(\sup_\Theta|\mathbf{E}_N\nabla_\theta l(y|z,\theta) - \psi(\theta)| > \delta/3) < \varepsilon.$$

Then, since $\mathbf{E}_N\nabla_\theta l_N(y|z,\theta) = \mathbf{E}_N\nabla_\theta l(y|z,\theta) + G_N(\theta)$,

$$\text{Prob}(\sup_\Theta|\mathbf{E}_N\nabla_\theta l_N(y|z,\theta) - \psi(\theta)| > 2\delta/3) < 2\varepsilon .$$

Therefore, with probability at least $1 - 2\varepsilon$, $|\mathbf{E}_N\nabla_\theta l_N(y|z,\theta)| > \delta/3$ for $\theta \notin \Theta_\varepsilon$ for $N \geq N_o + N_1$, so that $0 = \mathbf{E}_N\nabla_\theta l_N(y|z,\tilde{\theta}_N)$ implies $\tilde{\theta}_N \in \Theta_\varepsilon$. This establishes that $\tilde{\theta}_N$ is consistent for θ_o.

For asymptotic normality, consider

$$(6) \qquad 0 = N^{1/2}\mathbf{E}_N\nabla_\theta l_N(y|z,\tilde{\theta}_N)$$

$$\equiv N^{1/2}\mathbf{E}_N\nabla_\theta l(y|z,\theta_o) + N^{1/2}G_N(\theta_o) + [\mathbf{E}_N\nabla_\theta l_N(y|z,\tilde{\theta}_N) - \mathbf{E}_N\nabla_\theta l_N(y|z,\theta_o)]$$

Given assumptions A.3, A.4, and A.5, the Lindeberg-Levy CLT implies $N^{1/2}\mathbf{E}_N\nabla_\theta l(y|z,\theta_o) \xrightarrow{d} N(0,J(\theta_o))$. From condition (iii) above, $N^{1/2}G_N(\theta_o)$ is a uniformly bounded martingale difference sequence, and the theorem of Y. Chow implies $N^{1/2}G_N(\theta_o) \xrightarrow{as} 0$. Paraphrasing an argument of Ryu (1993), the terms in (5) satisfy

$$\mathbf{E}_\theta\{\nabla_{\theta\theta} l(y'|z,\theta)|C_N(y)\} \xrightarrow{as} \nabla_{\theta\theta} l(y|z,\theta),$$
$$\mathbf{E}_\theta\{[\nabla_\theta l(y'|z,\theta)]^2|C_N(y))\} \xrightarrow{as} [\nabla_\theta l(y|z,\theta)]^2, \text{ and}$$
$$\mathbf{E}_\theta\{\nabla_\theta l(y'|z,\theta)|C_N(y)\} \xrightarrow{as} \nabla_\theta l(y|z,\theta)$$

by the martingale convergence theorem (Billingsley, 1986, Theorem 35.4). This implies $\nabla_{\theta\theta} l_N(y|z,\theta) \xrightarrow{as} \nabla_{\theta\theta} l(y|z,\theta)$. But $-\mathbf{E}_{\theta_o}\nabla_{\theta\theta} l_N(y|z,\theta_o) = J_N(\theta_o)$ is nondecreasing and bounded by $-\mathbf{E}_{\theta_o}\nabla_{\theta\theta} l(y|z,\theta_o) = J(\theta_o)$ in the positive semidefinite sense (Ibragimov-Has'minskii, Theorem 7.2). Then $\nabla_{\theta\theta} l_N(y|z,\theta)$ is also L_1 convergent (Hall and Heyde, 1.3 Theorem), so that $J(\theta) = \lim_N J_N(\theta)$. Then, the last term in

$$\mathbf{E}_N\nabla_{\theta\theta} l_N(y|z,\theta_o) = \mathbf{E}_N\nabla_{\theta\theta} l(y|z,\theta_o) + \mathbf{E}_N[\nabla_{\theta\theta} l_N(y|z,\theta_o) - \nabla_{\theta\theta} l(y|z,\theta_o)]$$

converges to zero with probability one, implying $\mathbf{E}_N\nabla_{\theta\theta} l_N(y|z,\theta_o) \xrightarrow{as} J(\theta_o)$. Using assumption A.2, a Taylor's expansion of (6) yields

$$(7) \qquad 0 = N^{1/2}\mathbf{E}_N\nabla_\theta l(y|z,\theta_o) + N^{1/2}G_N(\theta_o)$$
$$+ \ [\mathbf{E}_N\nabla_{\theta\theta} l_N(y|z,\theta_o) + \Lambda_N\mathbf{E}_N m(y,z)|\tilde\theta_N - \theta_o|]\cdot N^{1/2}(\tilde\theta_N - \theta_o),$$

where Λ_N is an array with components of magnitude at most one and m(y,z) is a bound on $\nabla_{\theta\theta\theta} l$. Since $\tilde\theta_N \xrightarrow{as} \theta_o$, given $\varepsilon > 0$ there exists N_o such that for $N \geq N_o$ with probability at least $1-\varepsilon$,

$$[\mathbf{E}_N\nabla_{\theta\theta} l_N(y|z,\theta_o) + \Lambda_N\mathbf{E}_N m(y,z)|\tilde\theta_N - \theta_o|] \leq -J(\theta_o)/2,$$

with inequality in the negative semidefinite sense. Assumption A.5 then implies that with at least this probability, $N^{1/2}(\tilde\theta_N - \theta_o)$ in (6) is bounded by a constant times a stochastically bounded random variable. Then

$$N^{1/2}(\tilde\theta_N - \theta_o) = J(\theta_o)^{-1}N^{1/2}\mathbf{E}_N\nabla_\theta l(y|z,\theta_o) + o_p \xrightarrow{d} N(0,J(\theta_o)^{-1}) \ .$$

Since $\tilde\theta_N$ is asymptotically efficient, it must be asymptotically equivalent to $\hat\theta_N$. ■

Simulation Estimation

Consider the score of the LIML estimator

$$(8) \qquad s_N(\theta|y,z) = \mathbf{E}_N\nabla_\theta l_N(y|z,\theta) = \mathbf{E}_N \sum_{A\in \mathcal{C}_N} \chi_A(y)\nabla_\theta \log\ P(A|z,\theta)$$

$$= \mathbf{E}_N \sum_{A\in \mathcal{C}_N} [\chi(y) - P(A|z,\theta)]\cdot W(A|z,\theta) \ ,$$

where $\chi_A(y)$ is an indicator function, one for $y \in A$ and zero otherwise, $W(A|z,\theta)$ is shorthand notation for $\nabla_\theta \log\ P(A|z,\theta)$, and the last equality follows by differentiating the identity $1 \equiv \sum_{A\in \mathcal{C}_N} P(A|z,\theta)$. When $P(A|z,\theta)$ and its gradient are difficult to compute but have relatively convenient simulation approximators, then under very general conditions substitution in (8) of an *unbiased* Monte Carlo simulator for $P(A|z,\theta)$ and an independent simulator for $\nabla_\theta \log\ P(A|z,\theta)$ yields empirical moments whose solution is CAN for θ_o; this is a *Method of Simulated Moments*. This result does not require that the simulation approximation for each observation be improved as sample size increases, as averaging of the unbiased simulator over the sample is sufficient to control the impact on the estimator of the noise introduced by the simulation approximation. To achieve asymptotic efficiency (relative to the LIML estimator for a *fixed* partition $\mathcal{C}$), the simulation approximations must become exact as $N \longrightarrow +\infty$, but there is no minimum rate requirement. Precise statements and proofs of these results for very general classes of simulators, including simulators that are discontinuous in θ, are given in McFadden (1989), Theorem 1 and Lemma 8.

Consider *importance sampling* simulators that have the same smoothness properties in θ as the density f. This simplifies verification of critical regularity conditions of (*asymptotic*) *unbiasedness* and *stochastic equicontinuity*, and in addition covers the simulators that have proven most effective in applications. For example, the Geweke-Hajivassiliou-Keane and Parabolic Cylinder Function simulators for the multinomial probit model can be interpreted as importance sampling methods; see Hajivassiliou-McFadden-Ruud (1996). To define importance sampling simulators, consider an observation (y,z), with y possibly latent, and $A \in \mathcal{C}_N(y)$; a computer algorithm for generating pseudo-random numbers ξ in a space Ξ; and a $\mathfrak{Y}$-measurable mapping $y = \gamma(\xi;A,z,\theta)$ from $\Xi\times\mathbf{Z}\times\Theta$ into A that is chosen by the analyst. Let $g(y|A,z,\theta)$ be the density induced on A by this mapping. The following assumption will be made on the simulation process:

A.6. The transformation $\gamma(\xi;A,z,\theta)$ and induced density $g(y|A,z,\theta)$ have the properties:

(i) Both γ and g are three times continuously differentiable in θ, and are easy to compute.

(ii) There are positive constants c_0 and c_1, independent of $A \in \mathcal{C}_N$, z, and θ, such that for $y \in A$,

$$c_0 \cdot g(y|A,z,\theta) \le \frac{f(y|z,\theta)}{P(A|z,\theta)} \le c_1 \cdot g(y|A,z,\theta);$$

and a positive constant c_2 such that

$$\left|\frac{\nabla_\theta f(y|z,\theta)}{P(A|z,\theta)}\right| \le c_2 \cdot g(y|A,z,\theta) \text{ and } \left|\frac{\nabla_{\theta\theta} f(y|z,\theta)}{P(A|z,\theta)}\right| \le c_2 \cdot g(y|A,z,\theta)$$

(iii) Monte Carlo samples of size r_{0N}, r_{1N}, and r_{2N} respectively are drawn from Ξ for the simulation of $P(A|z,\theta)$, the numerator of $W(A|z,\theta) = \frac{\nabla_\theta P(A|z,\theta)}{P(A|z,\theta)}$, and the denominator of $W(A|z,\theta)$. The sample used to simulate $P(A|z,\theta)$ in the residual $[\chi(y) - P(A|z,\theta)]$ is independent of the samples used to simulate $W(A|z,\theta)$. All these samples are independent across observations. There may be dependence between the samples used to simulate the numerator and denominator of $W(A|z,\theta)$, and dependence across different $A \in \mathcal{C}_N$; e.g., the same draws may be used for different A. There may also be dependence in the samples for different N; e.g., draws made at an observation sample size N can be reused if the observation sample size is increased. The Monte Carlo samples are kept fixed and are <u>not</u> redrawn when θ changes. The simulation sample sizes satisfy $r_N = \min\{r_{0N},r_{1N}.r_{2N}\} \longrightarrow +\infty$ as $N \longrightarrow +\infty$.

When $\mathbf{Y}$ and Θ are compact and f is bounded above and below by positive constants, it is normally easy to satisfy (ii); g uniform on A will do. In applications it may be desirable to use variance reduction techniques such as antithetic sampling or autoregressive methods with negative serial correlation, rather than the random sampling of (iii). More general sampling processes such as exchangeable or mixing processes can also be used as long as the Monte Carlo sampling variance declines at a $1/r_N$ rate. The assumption places no rate restriction on r_N.

For $A \in \mathcal{C}_N$, write

$$P(A|z,\theta) = \int_{y\in A} f(y|z,\theta)d\mu \equiv \int_{y\in A} \left[\frac{f(y|z,\theta)}{g(y|z,\theta)}\right] g(y|z,\theta)d\mu,$$

$$\nabla_\theta P(A|z,\theta) = \int_{y\in A} \nabla_\theta f(y|z,\theta)du \equiv \int_{y\in A} \left[\frac{\nabla_\theta f(y|z,\theta)}{g(y|z,\theta)}\right] g(y|z,\theta)d\mu.$$

Then, under conditions (i)-(iii), *unbiased* simulators for $P(A|z,\theta)$, $\nabla_\theta P(A|z,\theta)$, and $\nabla_{\theta\theta} P(A|z,\theta)$ that inherit the smoothness properties in θ of $f(u|z,\theta)$ are obtained from the (simulated) empirical expectations, denoted by $\mathbf{E_r}$, for $A \in \mathcal{C}_N$:

$$\text{(9)} \qquad P^r(A|z,\theta) = \mathbf{E_r}\left[\frac{f(y|z,\theta)}{g(y|z,\theta)}\right] \equiv \frac{1}{r}\sum_{j=1}^{r}\left[\frac{f(y_j|z,\theta)}{g(y_j|z,\theta)}\right],$$

$$\text{(10)} \qquad \nabla_\theta P^r(A|z,\theta) = \mathbf{E_r}\left[\frac{\nabla_\theta f(y|z,\theta)}{g(y|z,\theta)}\right] \equiv \frac{1}{r}\sum_{j=1}^{r}\left[\frac{\nabla_\theta f(y_j|z,\theta)}{g(y_j|z,\theta)}\right],$$

$$\text{(11)} \qquad \nabla_{\theta\theta} P^r(A|z,\theta) = \mathbf{E_r}\left[\frac{\nabla_{\theta\theta} f(y|z,\theta)}{g(y|z,\theta)}\right] \equiv \frac{1}{r}\sum_{j=1}^{r}\left[\frac{\nabla_{\theta\theta} f(y_j|z,\theta)}{g(y_j|z,\theta)}\right],$$

where $y_j = \gamma(\xi_j;A,z,\theta)$ are the images of pseudo-random draws from Ξ that satisfy A.6. The statistical properties of these simulators are summarized in the following result. Let $\mathbf{E}_\xi$ denote the expectation operator for simulation draws.

Lemma 3.1. *Suppose* A.1-A.6 *hold. Then the simulators* (9) - (11) *are unbiased and strongly consistent for* $P(A|z,\theta)$*,* $\nabla_\theta P(A|z,\theta)$*, and* $\nabla_{\theta\theta}P(A|z,\theta)$ *as* $r \longrightarrow \infty$*, and are thrice, twice, and once continuously differentiable, respectively. They satisfy the bounds*

$$c_o P(A|z,\theta) \le P^r(A|z,\theta) \le c_1 P(A|z,\theta) \le c_1,$$

$$|\nabla_\theta P^r(A|z,\theta)| \le c_2 P(A|z,\theta) \le c_2,$$

$$|\nabla_{\theta\theta} P^r(A|z,\theta)| \le c_2 P(A|z,\theta) \le c_2.$$

They satisfy the moment conditions

$$|\mathbf{E}_\xi[P^r(A|z,\theta) - P(A|z,\theta)]\cdot[P^r(A'|z,\theta) - P(A'|z,\theta)]|$$
$$\le (c_1 - c_o)^2 P(A|z,\theta)P(A'|z,\theta)/r \ ,$$

$$|\mathbf{E}_\xi[\nabla_\theta P^r(A|z,\theta) - \nabla_\theta P(A|z,\theta)]\cdot[\nabla_\theta P^r(A'|z,\theta) - \nabla_\theta P(A'|z,\theta)]|$$

$$\le 4c_2^{\,2} P(A|z,\theta)P(A'|z,\theta)/r \ ,$$

$$|\mathbf{E}_\xi[P^r(A|z,\theta) - P(A|z,\theta)]\cdot[\nabla_\theta P^r(A'|z,\theta) - \nabla_\theta P(A'|z,\theta)]|$$
$$\le 2(c_1 - c_o)c_2 P(A|z,\theta)P(A'|z,\theta)/r \ ,$$

for $A,A' \in \mathcal{C}_N$. *They satisfy the exponential inequalities*

$$Prob(|P^r(A|z,\theta) - P(A|z,\theta)| > \varepsilon P(A|z,\theta)) \le 2\cdot\exp[-2\varepsilon^2 r/(c_1 - c_o)^2].$$

$$Prob(|\nabla_\theta P^r(A|z,\theta) - \nabla_\theta P(A|z,\theta)| > \varepsilon P(A|z,\theta)) \le 2\cdot\exp[-\varepsilon^2 r/2c_2^{\,2}];$$

$$Prob(|\nabla_{\theta\theta} P^r(A|z,\theta) - \nabla_{\theta\theta} P(A|z,\theta)| > \varepsilon P(A|z,\theta)) \le 2\cdot\exp[-\varepsilon^2 r/2c_2^{\,2}].$$

Proof: $\mathbf{E}_\xi P^r(A|z,\theta) = P(A|z,\theta)$, so that the simulator is unbiased. When $r \to +\infty$, a strong law of large numbers guarantees that $P^r(A|z,\theta) \xrightarrow{as} P(A|z,\theta)$ for each θ. Assumption A.6, (ii) and (iii), imply that

$$c_0 P(A|z,\theta) \le f(y_j|z,\theta)/g(y_j|z,\theta) \le c_1 P(A|z,\theta) \ ,$$

so that

$$\begin{aligned} &|\mathbf{E}_\xi[P^r(A|z,\theta) - P(A|z,\theta)]\cdot[P^r(A'|z,\theta) - P(A'|z,\theta)]| \\ &\quad \le \max\{c_1-1, 1-c_0\}^2 P(A|z,\theta)P(A'|z,\theta)/r_{0N} \\ &\quad \le (c_1-c_0)^2 P(A|z,\theta)P(A'|z,\theta)/r_{0N} \ . \end{aligned}$$

Next use Hoeffding's inequality (Pollard, 1984, Appendix B) which states that for independent random variables X_i with mean zero and bounds $a_i \le X_i \le b_i$, one has $\mathrm{Prob}(\sum_{i=1}^{m} X_i > \varepsilon m) < \exp(-\varepsilon^2 m^2/\sum_{i=1}^{m}(a_i-b_i)^2)$. Substituting the bounds for $P^r(A|z,\theta) - P(A|z,\theta)$ gives the exponential inequality in the lemma. Similar arguments apply to the score and hessian simulators. ■

A well-behaved, although biased, simulator for $W(A|z,\theta) \equiv \nabla_\theta \log P(A|z,\theta)$ is

$$(12) \qquad W^r(A|z,\theta) = \frac{\nabla_\theta P^r(A|z,\theta)}{P^r(A|z,\theta)} = \frac{\mathbf{E}_r\left[\dfrac{\nabla_\theta f(y|z,\theta)}{g(y|A,z,\theta)}\right]}{\mathbf{E}_r\left[\dfrac{f(y|z,\theta)}{g(y|A,z,\theta)}\right]} \ .$$

If $y \in A \in \mathcal{C}_N$ and $P(A|z,\theta) > 0$, then Lemma 3.1 implies $W^r(A|z,\theta) \xrightarrow{as} s_N(\theta|y,z)$. The following result establishes an exponential rate of convergence for W^r.

Lemma 3.2. *Suppose the numerator and denominator of* (12) *are simulated with* r_{1N} *and* r_{2N} *draws, respectively, using importance sampling simulators that satisfy* A.6. *There are positive constants* c_3, c_4 *such that*

$$|W^r(A|z,\theta)| \le c_2/c_0 \ ;$$

$$|\mathbf{E}_\xi(W^r(A|z,\theta) - W(A|z,\theta))(W^r(A'|z,\theta) - W(A'|z,\theta))| \le c_3/r_N$$

for $A,A' \in \mathcal{C}_N$*; and*

$$Prob(|W^r(A|z,\theta) - W(A|z,\theta)| > \varepsilon) \le 4\cdot\exp(-\varepsilon^2 c_4 r_N),$$

where $r_N \le \min(r_{1N},r_{2N})$.

Proof: The first bound comes from the A.6 conditions that $f(y|z,\theta)/g(y|z,\theta) \ge c_0 P(A|z,\theta)$ and $|\nabla_\theta f(y|z,\theta)/g(y|z,\theta)| \le c_2 P(A|z,\theta)$. Next, write

$$W^r(A|z,\theta) - W(A|z,\theta)$$

$$= W^r(A|z,\theta)\cdot\frac{P(A|z,\theta) - P^r(A|z,\theta)}{P(A|z_t,\theta)} + \frac{\nabla_\theta P^r(A|z,\theta) - \nabla_\theta P(A|z,\theta)}{P(A|z_t,\theta)} .$$

Since $|W^r(A|z,\theta)| \le c_2/c_0$, one has from the bounds in Lemma 3.1 applied to the terms in this expansion that

$$|\mathbf{E}_\xi(W^r(A|z,\theta) - W(A|z,\theta))(W^r(A'|z,\theta) - W(A'|z,\theta))|$$
$$\le (c_2/c_0)^2(c_1 - c_0)^2/r_N + 4c_2^2/r_N + 4(c_2/c_0)(c_1 - c_0)c_2/r_N \equiv c_3/r_N .$$

For the final inequality, the event $|W^r(A|z,\theta) - W(A|z,\theta)| > \varepsilon$ occurs only if one of the events $|\nabla_\theta P^r(A|z,\theta) - \nabla_\theta P(A|z,\theta)| > \varepsilon P(A|z,\theta)/2$ or $|P^r(A|z,\theta) - P(A|z,\theta)| > \varepsilon P(A|z,\theta)c_0/2c_2$ occurs. From Lemma 3.1, the sum of the probabilities of the the last two events is no greater than $2\cdot\exp[-\varepsilon^2 r_{1N}/8c_2^2] + 2\cdot\exp[-\varepsilon^2 r_{2N}c_0^2/2c_2^2(c_1-c_0)^2] \le 4\cdot\exp(-\varepsilon^2 c_4 r_N)$. ■

The computational efficiency of importance sampling methods is greatest when the expression being averaged is nearly constant, and when the sampling frequency is highest in regions where this expression varies. The (infeasible) ideal choice of g is the conditional distribution of y given $y \in A$, or $f(y|z,\theta)/P(A|z,\theta)$;

computationally practical choices will resemble this density.[11] Experience with applications is that careful choice of the importance sampling distribution g is critical to practical success with simulation estimators.

Summarizing, the *Method of Simulated Moments* (MSM) estimator for the multinomial problem approximates the score (8) by

$$(13) \qquad s_N^r(\theta|y,z) = \mathbf{E}_N \sum_{A\in \mathcal{C}_N} [\chi_A(y) - P^r(A|z,\theta)]\cdot W^r(A|z,\theta)$$

where *independent* simulations are used for the function W and for the probability inside the *residual* $[\chi_A(y) - P^r(A|z,\theta)]$, as stated in A.6. The following result establishes the asymptotic equivalence of $s_N^r(\theta|y,z)$ and $s_N(\theta|y,z)$.

Lemma 3.3. *Suppose* A.1-A.6. *Then*

$$\mathbf{E}_N[s_N^{r_N}(\theta|y,z) - s_N(\theta|y,z)] \xrightarrow{p} 0 \textit{ for each } \theta \in \Theta;$$

$$N^{1/2}\mathbf{E}_N[s_N^{r_N}(\theta_o|y,z) - s_N(\theta_o|y,z)] \xrightarrow{p} 0; \textit{ and}$$

$$N^{1/2}\mathbf{E}_N\nabla_\theta[s_N^{r_N}(\theta_o|y,z) - s_N(\theta_o|y,z)] \xrightarrow{p} 0.$$

---

[11] For example, the compact rectangle $A = \prod_{i=1}^{M}[a_i,b_i]$ and transformation

$$y_i = \gamma(\xi;A,z_t,\theta) \equiv \mu_i+\sigma_i\Phi^{-1}((1-\xi_i)\Phi((a_i-\mu_i)/\sigma_i)+\xi_i\Phi((b_i-\mu_i)/\sigma_i))$$

of uniform random numbers ξ_i for i = 1,...,M, where Φ is the standard normal distribution function, induces a density g that is the product of independent truncated normal densities $(1/\sigma_i)\phi((u_i-\mu_i)/\sigma_i)/[\Phi((b_i-\mu_i)/\sigma_i)-\Phi((a_i-\mu_i)/\sigma_i)]$, with the μ_i and σ_i depending on z_t and θ. When f is multivariate normal, this importance sampling distribution with μ_i and σ_i^2 defined recursively as the first and second moments of f conditioned on draws $y_1,...,y_{i-1}$ is the Geweke-Hajivassiliou-Keane simulator, a practical simulator with good properties.

Proof: Write $s_N^{r_N}(\theta|y,z) - s_N(\theta|y,z) = B_N - C_N + D_N$, with

$$B_N = \sum_{A\in \mathcal{C}_N} [\chi_A(y) - P(A|z,\theta_o)]\cdot[W^{r_N}(A|z,\theta) - W(A|z,\theta)] ,$$

$$C_N = \sum_{A\in \mathcal{C}_N} [P^{r_N}(A|z,\theta) - P(A|z,\theta)]\cdot W^{r_N}(A|z,\theta) ,$$

$$D_N = \sum_{A\in \mathcal{C}_N} [P(A|z,\theta_o) - P(A|z,\theta)]\cdot[W^{r_N}(A|z,\theta) - W(A|z,\theta)] .$$

First consider B_N. For each $\theta \in \Theta$, these random variables are independent across t, with $\mathbf{E}B_N = 0$ and

$$\begin{aligned}\mathbf{E}\, B_N^2 \le & \sum_{A,A'\in \mathcal{C}_N} \mathbf{E}\, |\chi_A(y) - P(A|z,\theta_o)|\cdot|\chi_{A'}(y) - P(A'|z,\theta_o)|\cdot \\ & \cdot|\mathbf{E}_\xi[W^{r_N}(A|z,\theta) - W(A|z,\theta)]\cdot[W^{r_N}(A'|z,\theta) - W(A'|z,\theta)]| \\ & \le (c_3/r_N)\left(\sum_{A\in \mathcal{C}_N} \mathbf{E}\, |\chi_A(y) - P(A|z,\theta_o)|\right)^2 \le 4c_3/r_N .\end{aligned}$$

Using Chebyshev's inequality, $\mathrm{Prob}(|N^{1/2}\mathbf{E}_N B_N| > \varepsilon) \le 4c_3/r_N\varepsilon^2 \longrightarrow 0$, implying $N^{1/2}\mathbf{E}_N B_N \xrightarrow{p} 0$.

Second consider

$$C_N = \mathbf{E}_{\mathbf{r}} \sum_{A\in \mathcal{C}_N} [P(A|z,\theta) - \frac{f(y|z,\theta)}{g(y|z,\theta)}]\cdot W^{r_N}(A|z,\theta).$$

For each $\theta \in \Theta$, these random variables are independent across observations t and simulation draws j with

$$\sum_{A\in \mathcal{C}_N} [P(A|z,\theta) - \frac{f(y|z,\theta)}{g(y|z,\theta)}]\cdot W^{r_N}(A|z,\theta)$$

having mean zero and being bounded by $\sum_{A\in\mathcal{C}_N} P(A|z,\theta)\cdot(c_1-c_0)c_2/c_0 = (c_1-c_0)c_2/c_0$.

Then, Hoeffding's inequality implies

$$\mathrm{Prob}(|N^{1/2}\mathbf{E}_N C_N| > \varepsilon) \le 2\cdot\exp(-\varepsilon^2 r_N c_o^2/2c_o^2(c_1-c_0)^2) \ .$$

Therefore $N^{1/2}\mathbf{E}_N C_N \xrightarrow{p} 0$ for each $\theta \in \Theta$.

Next consider D_N. At θ_o, $D_N = 0$. For other θ, Lemma 3.2 implies

$$\mathbf{E}\ D_N^2 \le (c_3/r_N)\left(\sum_{A\in\mathcal{C}_N} |P(A|z,\theta_o) - P(A|z,\theta)|\right)^2 \le 4c_3/r_N \ .$$

Then, Chebyshev's inequality implies $\mathbf{E}_N D_N \xrightarrow{p} 0$. These results together imply the first two results in the lemma, $\mathbf{E}_N[s_N^{r_N}(\theta|y,z) - s_N(\theta|y,z)] \xrightarrow{p} 0$ for each $\theta \in \Theta$, and $N^{1/2}\mathbf{E}_N[s_N^{r_N}(\theta_o|y,z) - s_N(\theta_o|y,z)] \xrightarrow{p} 0$.

For the last result, write $\nabla_\theta[s_N^{r_N}(\theta_o|y,z) - s_N(\theta_o|y,z)] = F_N + G_N + H_N - I_N$, with

$$F_N = \sum_{A\in\mathcal{C}_N} \nabla_\theta[P(A|z,\theta_o) - P^{r_N}(A|z,\theta_o)]\cdot W(A|z,\theta_o) \ ,$$

$$G_N = \sum_{A\in\mathcal{C}_N} [\nabla_\theta P^{r_N}(A|z,\theta_o)]\cdot[W(A|z,\theta_o) - W^{r_N}(A|z,\theta_o)] \ ,$$

$$H_N = \sum_{A\in\mathcal{C}_N} [\chi_A(y) - P^{r_N}(A|z,\theta_o)]\cdot\nabla_\theta W^{r_N}(A|z,\theta_o) \ ,$$

$$I_N = \sum_{A\in\mathcal{C}_N} [\chi_A(y) - P(A|z,\theta_o)]\cdot\nabla_\theta W(A|z,\theta_o) \ .$$

The covariance of $\nabla_\theta[P(A|z,\theta_o) - P^{r_N}(A|z,\theta_o)]$ and $\nabla_\theta[P(A|z,\theta_o) - P^{r_N}(A|z,\theta_o)]$ is bounded by $c_2^2 P(A|z,\theta_o)P(A'|z,\theta_o)/r_N$, from Lemma 3.1. Then,

$$\mathbf{E}\ F_N^2 \le c_2^2[\mathbf{E}\sum_{A\in\mathcal{C}_N} P(A|z,\theta_o)]\cdot|W(A|z,\theta_o)|]^2/r_N \le c_2^2\mathbf{E}\lambda(z)/r_N.$$

Then $\mathrm{Prob}(|\mathbf{E}_N F_N| > \varepsilon) \le c_2{}^2\mathbf{E}\lambda(z)/r_N\varepsilon^2$, and $\mathbf{E}_N F_N \xrightarrow{p} 0$ by Chebyshev's inequality.

From Lemma 3.2, the expectation of the cross-product of $W(A|z,\theta_o) - W^{r_N}(A|z,\theta_o)$ and $W(A'|z,\theta_o) - W^{r_N}(A'|z,\theta_o)$ is bounded by c_3/r_N. Then

$$\mathbf{E}\ G_N{}^2 \le (c_2/c_o)^2 c_1{}^2[\mathbf{E} \sum_{A\in C_N} P(A|z,\theta_o)]\cdot|W(A|z,\theta_o)|]^2/r_N \le (c_2/c_o)^2 c_1{}^2\mathbf{E}\lambda(z)/r_N.$$

This implies $\mathbf{E}_N G_N \xrightarrow{p} 0$.

The terms H_N and I_N have mean zero. The variance of H_N is bounded by

$$4(c_2/c_o)^2[\mathbf{E} \sum_{A\in C_N} \mathbf{E}|\chi_A(y) - P(A|z,\theta_o)|]^2 \le 16(c_2/c_o)^2,$$

so that Chebyshev's inequality implies $\mathbf{E}_N H_N \xrightarrow{p} 0$. The variance of I_N is bounded by

$$\mathbf{E}\ I_N{}^2 \le \left\{\sum_{A\in C_N} \mathbf{E}|\chi_A(y) - P(A|z,\theta_o)|\cdot|\nabla_\theta W(A|z,\theta_o)|\right\}^2 \le [2\cdot\mathbf{E}\lambda(z)]^2,$$

so that a law of large numbers implies $\mathbf{E}_N I_N \xrightarrow{p} 0$. ■

The following theorem extends the results of McFadden (1989) in the case of smooth importance sampling simulators to establish that MSM is CAN for a nested sequence of partitions C_N that generate $\mathfrak{Y}$, and is asymptotically efficient no matter how slowly $r_N \longrightarrow \infty$. Thus, MSM applied to nested multinomial approximations to general maximum likelihood problems can achieve the same asymptotic statistical precision as direct computation.

Theorem 3. *Suppose* A1-A6 *hold, and let* $\bar\theta_N$ *denote a* MSM *estimator that solves* $\mathbf{E}_N s_n^r(A|z,\bar\theta_n) = 0$ *for a nested sequence of partitions* C_N *that generate* $\mathfrak{Y}$. *Then* $\bar\theta_N$ *is* CAN *with* $N^{1/2}(\bar\theta_N - \theta_o) \xrightarrow{d} N(0,J(\theta_o)^{-1})$, *so that the estimator is asymptotically equivalent to* FIML.

Proof: From Lemma 3.3, for each $\theta \in \Theta$, $\mathbf{E}_N[s_N^{r_N}(\theta|y,z) - s_N(\theta|y,z)] \xrightarrow{p} 0$. Since $s_N^{r_N}(\theta|y,z)$ inherits the smoothness properties of $s_N(\theta|y,z)$, it is equicontinuous. Then, the same argument as in Theorem 2 establishes that

$$\sup_{\theta\in\Theta}|\mathbf{E}_N[s_N^{r_N}(\theta|y,z) - s_N(\theta|y,z)]| \xrightarrow{p} 0,$$

and hence that

$$\sup_{\theta\in\Theta}\left|\mathbf{E}_N[s_N^{r_N}(\theta|y,z) - \psi(\theta)\right| \xrightarrow{p} 0,$$

where $\psi(\theta) = \mathbf{E}_{\theta_o}\nabla_\theta l(y|z,\theta)$. The argument for consistency of $\hat{\theta}_N$ in Theorem 3.2 then establishes the consistency of $\bar{\theta}_N$.

For asymptotic normality, Taylor's expansions of $s_N^{r_N}(\bar{\theta}_n|y,z)$ and $s_N(\tilde{\theta}_n|y,z)$ around θ_o yield

$$0 = N^{1/2}\mathbf{E}_N[s_N^{r_N}(\theta_o|y,z) - s_N(\theta_o|y,z)]$$

$$+ [\mathbf{E}_N\nabla_\theta s_N^{r_N}(\theta_o|y,z) + \Lambda_N\mathbf{E}_N m(y,z)|\bar{\theta}_N - \theta_o|]\cdot N^{1/2}(\bar{\theta}_N - \theta_o)$$

$$- [\mathbf{E}_N\nabla_\theta s_N(\theta_o|y,z) + \Lambda_N'\mathbf{E}_N m(y,z)|\tilde{\theta}_N - \theta_o|]\cdot N^{1/2}(\tilde{\theta}_N - \theta_o),$$

where Λ_N and Λ_N' are arrays with elements at most one. From Lemma 3.3, the first term converges in probability to zero, and $\mathbf{E}_N\nabla_\theta s_N^{r_N}(\theta_o|y,z)$ - $\mathbf{E}_N\nabla_\theta s_N(\theta_o|y,z)$ converges in probability to zero. From Theorem 3.2, $\mathbf{E}_N\nabla_\theta s_N(\theta_o|y,z)$ converges in probability to $-J(\theta_o)$. Therefore, $J(\theta_o)N^{1/2}(\tilde{\theta}_N - \bar{\theta}_N) \xrightarrow{p} 0$, and $\tilde{\theta}_N$ and $\bar{\theta}_N$ are asymptotically equivalent. Since Theorem 3.2 established the asymptotic equivalence of $\tilde{\theta}_N$ and $\hat{\theta}_N$, the asymptotic equivalence of $\bar{\theta}_N$ and $\hat{\theta}_N$ is proven. ■

Sequential Simulation

A multinomial problem with a large numbers of alternatives can be written in terms of a sequence of transitions through a "decision tree". A version of the method of simulated moments estimator can then be formulated that permits consistent estimation of model parameters using unbiased simulators of the unconditional probabilities of the observed nodes in the tree, and practical approximations to terms in the conditional scores at each node. This method can reduce substantially the number of function evaluations required in estimation, while retaining the statistical properties of the simulation estimator described in A.6 and Theorem 3.3. An example illustrates the computational savings: Consider a multinomial probit choice problem for a partition C_N containing $K_N \equiv \mathrm{card}(C_N) = 2^{12}$, or 4096, alternatives. Consider conventional maximum likelihood estimation using 10-point Gaussian quadrature for numerical integration to obtain the choice probabilities and their derivatives. This requires $3\cdot 10^{4096}$ evaluations per observation and iteration, clearly impossible. A conventional MSM estimator in the form described in assumption A.6 requires unbiased simulation of each of the 4096 probabilities, plus approximation to the score of each, on the order of $3 \cdot K_N \cdot r_N$ = 140,088 evaluations per observation and iteration when $r_N = 10$. This is computationally feasible but very burdensome for modest r_N, and increasingly difficult as K_N and r_N increase. The sequential simulation method below requires on the order of $3 \cdot r_N \cdot \log(K_N) = 360$ evaluations per observation and iteration when $r_N = 10$. The relative computational efficiency of the sequential method becomes more dramatic as K_N increases with improved multinomial approximations to a continuous problem.

Let $\mathcal{C}_N$, n = 1,2,3,... denote a nested sequence of finite partitions of **Y** that generate observable events. These partitions define a "decision tree" in which a subject can be pictured as reaching a response y by descending through successive nodes $C_n(y) \in \mathcal{C}_N$. For a node $B \in \mathcal{C}_m$ and $k \geq m$, define $\mathcal{C}_k(B) = \{A \in \mathcal{C}_k \mid A \subseteq B\}$ to be set of level k nodes below B. For $k \leq m$, define $C_k(B)$ to be the element of $\mathcal{C}_k$ that contains B; this generalizes the previous notation of $C_k(y)$ for the element of $\mathcal{C}_k$ that contains $y \in$ **Y**. For $B \in \mathcal{C}_m$ and $k \leq m$, define $\mathcal{C}^k(B) = \mathcal{C}_k(C_{k-1}(B))$ to be the set of nodes immediately below a node $C_{k-1}(B)$. For $D \in \mathcal{C}^k(B)$, define an indicator for the node leading to B, $d(D \mid \mathcal{C}^k(B)) = 1$ if $D = C_k(B)$ and $d(D \mid \mathcal{C}^k(B)) = 0$ otherwise. Note that it is not necessary that the refinement be strict. In most applications, the depth of the tree will increase very slowly with N, so that many transitions from $\mathcal{C}_k$ to $\mathcal{C}_{k+1}$ for k < N will not be strict refinements. For convenience, we will not introduce explicit notation for the depth of nodes. However for computation it is convenient to keep indices of node depth.

For computation, it is useful to consider binomial refinements where each $A \in \mathcal{C}_k$ is in $\mathcal{C}_{k+1}$ or is the union of two sets in $\mathcal{C}_{k+1}$. An alternative to binomial partitioning is to consider "natural" decision trees for the choice process. An example of natural partitioning is the case of discrete panel data, with choice among m alternatives in each of T periods. Then there are $K = m^T$ compound alternatives in **Y**, each described by a profile of T discrete choices. Then, the tree is naturally defined so that the nodes correspond to dynamic transition probabilities: the elements of $\mathcal{C}_k$ are partial profiles of choices through time k, and the transition probabilities from period k to period k+1 equal the conditional probabilities of the branches from a node in $\mathcal{C}_k$.

Let $P(D|\mathcal{C}_k(B),z,\theta)$, for a node $B \in \mathcal{C}_{k-1}$ and a node $D \in \mathcal{C}_k(B)$ that is immediately below it in the tree, denote the transition probability between these nodes. For the nodes leading to a node $A \in \mathcal{C}_N$, one has $B = C_{k-1}(A)$ and $D = C_k(A)$, and $\mathcal{C}^k(A) = \mathcal{C}_k(C_{k-1}(A))$, so that the transition probability is $P(C_k(A)|\mathcal{C}^k(A),z,\theta)$. The probability of an element $A \in \mathcal{C}_N$ can be written as a product of the probabilities of transitions through the tree,

$$(14) \qquad P(A|z,\theta) = \prod_{k=1}^{N} P(C_k(A)|\mathcal{C}^k(A),z,\theta) = \prod_{k=1}^{N} \prod_{D\in\mathcal{C}^k(A)} P(D|\mathcal{C}^k(A),z,\theta)^{d(D|\mathcal{C}^k(A))}.$$

Then, the LIML score of an observation $A \in \mathcal{C}_N$ is

$$(15) \qquad \nabla_\theta \log P(A|z,\theta) = \sum_{k=1}^{N} \sum_{D\in\mathcal{C}^k(A)} d(D|\mathcal{C}^k(A))\cdot\nabla_\theta \log P(D|\mathcal{C}^k(A),z,\theta)$$

$$= \sum_{k=1}^{N} \sum_{D\in\mathcal{C}^k(A)} u(D,\mathcal{C}^k(A),z,\theta)\cdot W(D|\mathcal{C}^k(A),z,\theta) \ ,$$

where

$$(16) \qquad u(D,\mathcal{C}^k(A),z,\theta) \equiv [d(D|\mathcal{C}^k(A))\cdot P(C_{k-1}(A)|z,\theta) - P(D|z,\theta)]$$

is a *residual* expressed in terms of unconditional probabilities, and

$$(17) \qquad W(D|\mathcal{C}^k(A),z,\theta) \equiv \left[\nabla_\theta \log P(D|\mathcal{C}^k(A),z,\theta)\right]\cdot P(C_{k-1}(A)|z,\theta)^{-1}$$

$$\equiv \left[\nabla_\theta P(D|\mathcal{C}^k(A),z,\theta)\right]\cdot P(D|z,\theta)^{-1}$$

$$\equiv \left(\frac{\nabla_\theta P(D|z,\theta)}{P(D|z,\theta)} - \frac{\nabla_\theta P(C_{k-1}(A)|z,\theta)}{P(C_{k-1}(A)|z,\theta)}\right)\cdot\frac{1}{P(C_{k-1}(A)|z,\theta)} \ .$$

is a vector of *instruments*, which in the last form is expressed in terms of unconditional probabilities.

When the refinements are binomial, $\mathcal{C}^k(A)$ contains two sets, $C_k(A)$ and $D_k(A) \equiv C_{k-1}(A)\backslash C_k(A)$. Note that $D_k(A)$ is empty if the refinement is not strict. The formula for the LIML score simplifies to

$$\nabla_\theta \log P(A|z,\theta) = \sum_{k=1}^{N} P(D_k(A)|z,\theta)\cdot W_k^*, \tag{18}$$

with the instrument vector

$$W_k^* \equiv [W(C_k(A)|\mathcal{C}^k(A),z,\theta) - W(D_k(A)|\mathcal{C}^k(A),z,\theta)] \tag{19}$$

$$\equiv \left(\frac{\nabla_\theta P(C_k(A)|z,\theta)}{P(C_k(A)|z,\theta)} - \frac{\nabla_\theta P(C_{k-1}(A)|z,\theta)}{P(C_{k-1}(A)|z,\theta)}\right)\cdot\frac{1}{P(D_k(A)|z,\theta)} .$$

Equations (15) or (18) are possible starting points for Method of Simulated Moments (MSM) estimation. The importance sampling simulators in assumption A.6 provide unbiased consistent estimates of the residuals $u(D,\mathcal{C}^k(A),z,\theta)$ in (15) or of $P(D_k(A)|z,\theta)$ in (18), and provide consistent estimates of the instruments (17) or (19). It is essential that the instrument simulators be independent of the simulators of the simulators of $u(D,\mathcal{C}^k(A),z,\theta)$ in (15) or $P(D_k(A)|z,\theta)$ in (18); in particular, $P(D_k(A)|z,\theta)$ appears in both the numerator and denominator of (18), but the numerator probability must have an independent unbiased simulator, and cannot be cancelled out. The conclusions of Theorem 3.3 then apply to establish that estimators using one of these starting points are CAN and asymptotically efficient. The following theorem summarizes these results:

Theorem 3.4. *Suppose* A.1-A.6 *hold, the successively refined finite partitions* $\mathcal{C}_N$ *generate the* σ*-field of observable events* $\mathfrak{Y}$*, and simulators are applied to* (15)-(17), *with an unbiased simulator used for* (16) *and independent consistent simulators used for* (17)*. Then, the multinomial approximation and sequential simulation (MASS) estimators that solve* (15) *are CAN and asymptotically efficient.*

This result has several useful implications and extensions. First, under appropriate identification conditions paralleling A.4, MASS estimation with a *fixed* partition $\mathcal{C}$ and a *fixed* number of Monte Carlo draws per observation will be CAN, although not in general asymptotically efficient. Consequently, estimators with satisfactory statistical properties can be obtained without the computational problems of dealing with very large partitions or Monte Carlo samples. These estimators can be used to provide consistent starting values for further estimation, facilitating iterative search. They may be estimated sequentially, using a stopping rule based on the stability of the estimates with increasing refinements and numbers of repetitions. Because of the nested structure of the problems, classical Wald statistics, using quadratic forms in parameter changes with the (generalized) inverse of a difference of covariance matrices as a weighting matrix, can be used for a stopping criteria. These statistics can also be used as diagnostics for model inconsistencies that may be revealed at higher refinements. One could trade off efficiency for practicality by basing estimation only on the upper levels of the tree, seeking a root of the marginal score of nodes down to an intermediate level k. This has the effect of treating the elements of $\mathcal{C}_k$ as the final objects of choice, with information on choice within these objects margined out. This model is of some independent interest as a consistent model for choice among aggregates of elemental alternatives. The probabilities required near the top of the tree are typically not extreme, so that it is not too difficult to get informative unbiased simulators for them.

There may be computational shortcuts that exploit the tree structure to economize on the unbiased simulation of the probabilities required in (15) or (18). For example, unconditional draws from $f(y|z,\theta)$ yield simultaneously unbiased simple frequency simulators for all the unconditional node probabilities. While these simulators have discontinuities that make iteration difficult, it is possible to adapt importance sampling to provide simultaneous estimates of unconditional probabilities at various nodes in a tree. Suppose $y = \gamma(\xi;\mathbf{Y},z,\theta)$ maps Ξ onto $\mathbf{Y}$, and induces a density $g(y|z,\theta)$ with the property that the unconditional probability $g(A|z,\theta)$ and conditional density $\chi_A(y)g(y|z,\theta)/g(A|z,\theta)$ are easily computed for $A \in \mathcal{C}_N$. (For example, if $\mathbf{Y}$ is compact and the elements of $\mathcal{C}_N$ are rectangles, then random sampling from $\mathbf{Y}$, which induces a uniform density, works.) Then $r_N \longrightarrow +\infty$ draws from $g(y|z,\theta)$ will for each node A in the tree produce a random number of draws $r_N(A)$ that are contained in A. These draws, conditioned on $r_N(A)$, can be treated as importance sampling draws from the conditional density $\chi_A(y)g(y|z,\theta)/g(A|z,\theta)$. Down to any fixed depth in the tree, this method is asymptotically equivalent to applying importance sampling separately to each node. (To get asymptotic equivalence uniformly over the tree as it is refined with increasing N, one must either have the r_N unconditional draws growing rapidly enough so that $r_N\mu(A_N) \longrightarrow +\infty$ for each sequence $A_N \in \mathcal{C}_N$, or one must augment the importance samples with conditional draws as one moves down the observed branch so that the expected numbers of draws at each observed node times the measure of the partition set at this node all go to infinity at some minimum rate.[12] Monte Carlo draws can also be reused as the problem is refined, and retained as the number of draws is increased. This saves on computation, and may stabilize estimates. In general, antithetic Monte Carlo methods that preserve unbiasedness of estimators will improve simulator efficiency.

An issue that is beyond the scope of this lecture is the statistical behavior of MASS estimators for problems where the FIML estimator is ill-conditioned. Clearly, the ability of the multinomial approximation to smooth over singularities in the original problem can make multinomial estimators consistent in cases where FIML is inconsistent. An interesting theoretical question is whether this advantage is retained as the partitions are refined.

---

[12] For example, suppose $\mathbb{Y} = \mathbb{R}^m_+$, and $\mathcal{C}_n$ is constructed from the half-open intervals of the form $[2^{-k_n}\cdot i, 2^{-k_n}\cdot(i+1))$ in each dimension, for $i = 0,\ldots,4^{k_n}$, where k_n is the largest integer satisfying $2^{k_n} \le \log n$. Then, $\mu(A_n) \ge 2^{-mk_n} \le (\log n)^m$, and r_n growing at any fractional power of n will satisfy the condition $r_n\cdot\mu(A_n) \longrightarrow +\infty$.

LECTURE 4. MULTINOMIAL PROBIT BY SIMULATION

A variety of economic problems involve multinomial response. Examples are choice among brands of consumer products such as TV sets, automobiles, and computers; choice of sites for locating industrial plants; choice of occupation; and assignment of jobs in multilateral bargaining between employers and job-seekers. To model such behavior, associate with alternative i in a feasible set C a "payoff" $u_i = z_i\beta + \varepsilon_i$, which in the case of consumer choice may be the indirect utility attached to alternative i and in the case of firm choice may be profit from alternative i. The z_i are observed explanatory variables, and the ε_i are unobserved disturbances. Observed choice is assumed to maximize payoff: $y_i = \mathbf{1}(u_i \geq u_j \text{ for } j \in C)$. The specification $u_i = z_i\beta + \varepsilon_i$ is not intrinsically restrictive, as the utility function can be written as the sum of its expectation plus a deviation from this expectation, and the expectation can usually be approximated globally by a polynomial in observed covariates, with β taken to the the coefficients in this approximation. One source of this model is a <u>random</u> <u>coefficients</u> formulation $u_i = z_i\alpha$, $\mathbf{E}\alpha = \beta$, $\varepsilon_i = z_i(\alpha - \beta)$, implying $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = z_i \cdot \mathrm{Cov}(\alpha) \cdot z_j'$.

For $C = \{1,...,J\}$, define u, z, ε, and y to be $J\times 1$ vectors with components u_j, z_j, ε_j, y_j, respectively. Define a $(J\text{-}1)\times J$ matrix Δ_i by starting from the $J\times J$ identity matrix, deleting row i, and then replacing column i with a vector of -1's. For example,

$$\Delta_1 = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Then alternative i is chosen if $\Delta_i u \leq 0$. The probability of this event is

$$P_i(z,\theta) = \Pr(\Delta_i u \leq 0 \mid z,\theta) \equiv \int_{\Delta_i u \leq 0} f(u \mid z,\theta)du \ ,$$

where $f(u \mid z,\theta)$ is the conditional density of u given z. The parameters θ include the slope parameters β and any additional parameters characterizing the distribution of the disturbances ε. The multivariate integral defining $P_i(z,\theta)$ can be calculated analytically in special cases, notably multinomial logit and its generalizations.

However, for most densities the integral is analytically intractable, and for dimensions much larger than J = 5 is also intractable to evaluate with adequate precision using standard numerical integration methods.

A density that is relatively natural for capturing unobserved effects, and the patterns of correlation of these effects across alternatives, is the multivariate normal distribution with a flexible covariance matrix. This is termed the *multinomial probit* model. If $\varepsilon = z\xi$, where ξ is interpreted as a random variation in "taste" weights across observations with $\xi \sim N(0,\Omega)$, then the transformed variable $w = \Delta_i u$ is multivariate normal of dimension J-1 with mean $\Delta_i z\beta$ and covariance $\Delta_i z\Omega z'\Delta_i'$. Unless $J \leq 5$ or dimensionality can be reduced because ξ has a factorial covariance structure, the resulting MNP response probabilities are impractical to calculate by numerical integration. The method of simulated moments was initially developed to handle this model; see McFadden (1989).

The log likelihood of an observation is

$$l(\theta) = \sum_{i \in C} d_i \cdot \log P_i(z,\theta) ,$$

where d_i is an indicator for the event that i is chosen. The *score* of an observation is then

$$s(\theta) = \sum_{i \in C} d_i \cdot \nabla_\theta \log P_i(z,\theta) \equiv \sum_{i \in C} [d_i - P_i(z,\theta)] \cdot \nabla_\theta \log P_i(z,\theta) ,$$

with the second form holding because $0 \equiv \sum_{i \in C} \nabla_\theta P_i(z,\theta)$. This score can be adapted to MSM or MSS estimation when $P_i(z,\theta)$ is intractable by conventional analysis. Simulators are required for $P_i(z,\theta)$ and $\nabla_\theta \log P_i(z,\theta)$.
Consider the problem of approximating

$$(1) \qquad P \equiv P(\mathbf{B};\mu,\Omega) = \int_{-\infty}^{0} n(v-\mu,\Omega)dv \equiv \mathbf{E}_V \mathbf{1}(V \in \mathbf{B}) ,$$

where V is a m-dimensional normal random vector with mean μ, covariance matrix Ω, and density $n(v - \mu,\Omega)$, and $\mathbf{1}(V \in \mathbf{B})$ is an indicator for $\mathbf{B} = \{V | V < 0\}$. The derivatives of

(1) with respect to μ and Ω are

$$\text{(2)} \qquad \nabla_{\mu}P(\mathbf{B};\mu,\Omega) = \Omega \int_{-\infty}^{+\infty} \mathbf{1}(v\in\mathbf{B})(v-\mu)n(v-\mu,\Omega)dv \equiv \Omega^{-1}\mathbf{E}_{V}\ \mathbf{1}(V\in\mathbf{B})(V-\mu) \ ,$$

$$\nabla_{\Omega}P(\mathbf{B};\mu,\Omega) = \frac{\Omega^{-1}}{2}\cdot\int_{-\infty}^{+\infty} \mathbf{1}(v\in\mathbf{B})[(v-\mu)(v-\mu)'-\Omega]n(v-\mu,\Omega)dv\cdot\Omega^{-1}$$

$$\equiv (1/2)\Omega^{-1}\mathbf{E}_{V}\ \mathbf{1}(V\in\mathbf{B})[(V-\mu)(V-\mu)'-\Omega]\cdot\Omega^{-1} \ .$$

For statistical inference, it is often unnecessary to achieve high numerical accuracy in evaluation of (1) and (2). For example, simulating P by the frequency of the event $\mathbf{1}(v\in\mathbf{B})$ in a number of Monte Carlo draws comparable to sample size will tend to produce statistics in which the variance introduced by simulation is at worst of the same magnitude as the variance due to the observed data. Further, when probabilities appear linearly across observations in an estimation criterion, independent unbiased simulation errors are averaged out. Then, a small, fixed number of draws per probability to be evaluated will be sufficient with increasing sample size to reduce simulation noise at the same rate as noise from the observed data.

<u>Monte Carlo Methods</u>

Crude Frequency Sampling. The random vector V can be written $V = \mu + \Gamma\eta$, where η is an independent standard normal vector of dimension m and Γ is an lower triangular Cholesky factor of Ω, so $\Omega = \Gamma\Gamma'$. Make repeated Monte Carlo draws of η, and <u>fix these throughout the iteration</u>. Calculate $V = \mu + \Gamma\eta$ for trial parameters (μ,Γ) and form empirical analogs of the expectations (1) and (2). This simulator has the advantage of being very fast, but the disadvantages of being discontinuous in θ, with a large relative error for small probabilities.

Importance Sampling. Consider the generic integral

$$H = \int_{-\infty}^{+\infty} \mathbf{1}(v\in\mathbf{B})\cdot h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)dv \ ,$$

where h is an array of polynomials in v; integrals (1)-(2) have this form. Let g(v) be a density with support **B** chosen by the analyst. Then,

$$H = \int_{\mathbf{B}} \{h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)/g(v)\}\cdot g(v)dv$$

and a smooth unbiased simulator of H is obtained by drawing from g, fixing these draws, and then for (μ,Ω) averaging $\{h(v;\mu,\Omega)\cdot n(v-\mu,\Omega)/g(v)\}$ over these draws.

This simulator has the advantage of being smooth in θ and positive for simulated P, which aids iteration to estimates. The method is fast if g(v) is an easy density to draw from. A disadvantage is that it can be inaccurate unless the mass of g is concentrated near the mass of the normal density. Another is that the probability simulator is not necessarily less than one, and censoring to the unit interval disrupts the unbiasedness of the simulator. Also, importance sampling simulators for probabilities that sum to one will often fail to preserve this summing up property. This is unimportant for some simulation methods, but can cause computational difficulties in methods such as MSM that rely on the summing up property.

Geweke-Hajivassiliou-Keane (GHK) *Simulator*. This is an importance sampling simulator that has performed well in comparison with many other simulators; see Hajivassiliou, McFadden, and Ruud (1996). It is based on sampling from recursive truncated normals after a Cholesky transformation. The approach was suggested by Geweke (1989), and has been developed by Hajivassiliou, who proposed the weighting used here. Keene (1988) independently developed a weighting scheme of essentially the same form for a problem of estimating transition probabilities.

Let $v = \mu + \Gamma\eta$, where Γ is the Cholesky factor of Ω. For a vector x, let x_{-i} denote the subvector with component i deleted, $x_{<i}$ denote the subvector with components i and higher deleted, and $x_{>i}$ the subvector with components i and lower deleted. The indicator $\mathbf{1}(v\in\mathbf{B})$ is transformed to $\mathbf{1}(\mu+\Gamma\eta\in\mathbf{B})$, which can be written recursively as the product of indicators of the events $\mathbf{B}_j(\eta_{<j})$ defined by

$$\eta_j < (b_j-\mu_j-\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj}$$

for $j = 1,...,m$. Define $\phi(\eta_j|\mathbf{B}_j(\eta_{<j}) = \phi(\eta_j)\mathbf{1}(\eta_j\in\mathbf{B}_j(\eta_{<j}))/\Phi(\mathbf{B}_j(\eta_{<j}))$, the conditional distribution of η_j given the event $\mathbf{B}_j(\eta_{<j})$. Define a weight

$$\omega(\eta) = \prod_{j=1}^{m} \Phi(\mathbf{B}_j(\eta_{<j})).$$

Then

$$H = \int h(\mu+\Gamma\eta)\omega(\eta)\prod_{j=1}^{m} \phi(\eta_j|\mathbf{B}_j(\eta_{<j}))d\eta \ .$$

The GHK simulator is obtained by drawing and <u>fixing</u> uniform [0,1] variates ζ, then for (μ,Ω) calculating variates

$$\eta_j = \Phi^{-1}(\zeta_j\Phi(-(\mu_j+\Gamma_{j,<j}\eta_{<j})/\Gamma_{jj})) \ ,$$

and then averaging $h(\mu+\Gamma\eta)\omega(\eta)$ over these variates. The advantages of this simulator are that for a broad spectrum of applications, the importance sampling density is concentrated near the ideal truncated multivariate normal density, giving low variance simulators for both probabilities and derivatives that have small relative error even when P is small. The primary disadvantage is that the recursive loops with multiple evaluations of standard normal CDFs and inverse CDFs are computationally costly and may introduce additional approximation errors. Also, GHK simulators will not satisfy the summing up property.

Acceptance/Rejection Methods. These methods provide a mechanism for drawing from a conditional density when practical exact transformations from uniform or standard normal variates are not available. They have the advantage of being unbiased, but this is often offset by the disadvantages of low yield, requiring excessive computation, and discontinuities as parameters change. The following result is standard; see Devroye (1986) or Rubinstein (1981).

Lemma 4.1. *Suppose* $f(x)$ *is an m-dimensional density, and one wishes to sample from the conditional density* $f(\cdot|\mathbf{A})$ *given the event* $x \in \mathbf{A}$. *Suppose* $g(x)$ *is a density from which it is practical to sample, with the property that* $\sup_{\mathbf{A}} f(x)/g(x) \le \alpha < +\infty$. *Assume that either the support of* g *is* $\mathbf{A}$, *or that it is practical to test if* $x \in \mathbf{A}$; *that it is practical to calculate* $f(x)$ *and* $g(x)$; *and that it is practical to calculate a bound* α. *Draw* x *from* g *and* ζ *from a uniform density on* [0,1], *repeat this process until a pair satisfying* $x \in \mathbf{A}$ *and* $f(x) \ge \zeta\alpha g(x)$ *is observed, and accept* *the associated* x. *Then, the accepted points have density* $f(x|\mathbf{A}) \equiv f(x)/f(\mathbf{A})$.

Kernel-Smoothed Simulators. McFadden (1989) suggested replacing the indicator function $\mathbf{1}(v\in\mathbf{B})$ in the crude frequency simulator with a function $\mathbb{K}(v/\omega)$, where $\mathbb{K}(w)$ is a smooth kernel function from $\mathbb{R}^m$ onto [0,1] with $\mathbb{K}(-\infty) = 1$ and $\mathbb{K}(w) \to 0$ if any component of $w \to +\infty$; and ω is a window width parameter. The function $\mathbb{K}(v/\omega)$ approaches $\mathbf{1}(v\in\mathbf{B})$ as $\omega \to 0$. Then, the simulator is an average of $\mathbb{K}(v/\omega)h(v)$, with $v = \mu + \Gamma\eta$, over r Monte Carlo draws of an independent standard normal vector η. This simulator is smooth in parameters, a useful feature when the simulator is used within an iterative optimization.

One interpretation of the kernel-smoothed simulator is that the latent variable model is perturbed to $\Delta_i u = \Delta_i x\beta + \Delta_i\varepsilon + \omega v$, and the disturbance component ωv is then integrated out, conditioned on $\Delta_i\varepsilon$. Possible kernels include a product of independent probits, $\mathbb{K}(w) = \Pi_i\Phi(-w_i)$, and a multinomial logit, $\mathbb{K}(w) = 1/(1 + \Sigma_i e^{w_i})$. Stern () shows that the independent probit kernel can be made exact by "borrowing" variance from $\Delta_i\varepsilon$; i.e., reducing the diagonal of the covariance matrix of $\Delta_i\varepsilon$ by ω^2, where ω is small enough so that that matrix remains positive definite. The multinomial logit kernel is computationally convenient, and is of interest as a utility model in its own right.

The bias in a kernel-smoothed estimator will decline with ω at a rate that depends on the tail behavior of the kernel. The requirement that the bias be asymptotically negligible in the simulation estimator will impose a modest rate condition on ω. For example, with the multinomial logit kernel, it will always be sufficient to have ω decrease with N in inverse proportion to $\sqrt{N}/(\log N)^2$.

Parabolic Cylinder Function Simulator. Consider a spherical transformation $V = \rho\cdot\upsilon$, where υ is in the unit sphere and ρ is a non-negative scalar. The multivariate normal density expressed in terms of the transformed variables is the product of a density on the unit sphere and a conditional density of ρ given υ. The simulator is obtained by drawing υ from a uniform density on the unit sphere, with an importance sampling weight, and by using recursive analytic formulas to obtain moments of the conditional density of ρ given υ; these are termed parabolic cylinder functions. The advantages of the simulator are that it is smooth and fast to compute. In Monte Carlo trials, however, it performs somewhat less well than the GHK simulator for derivatives.

Gibbs Sampling. This simulator is based on a Markov chain that utilizes computable univariate truncated normal densities to construct transitions, and has the desired truncated multivariate normal as its limiting distribution. Implemented in a form that approximates a sample from the truncated multivariate normal density, the simulator is constructed in the following steps: Start from any $v^{(0)} \in \mathbf{B}$. Define a recursive procedure with steps i = 1,...,m in rounds j = 1,...,r. Suppose at step i in round j, $v^{(j-1)}$ and $v_{<i}^{(j)}$ have been determined. Define

$$v_i^{(j)} = \kappa_{ij} + \sigma_i\Phi^{-1}(\zeta_{ij}\Phi(-\kappa_{ij}/\sigma_i)) \ ,$$

where the ζ_{ij} are independent uniform [0,1] variates,

$$\kappa_{ij} = \mu_i + \Omega_{i,-i}\Omega_{-i,-i}^{-1}\begin{bmatrix} v_{<i}^{(j)} - \mu_{<i}^{(j)} \\ v_{>i}^{(j-1)} - \mu_{>i}^{(j-1)} \end{bmatrix},$$

and

$$\sigma_i = \left[\Omega_{ii} - \Omega_{i,-i}\Omega_{-i,-i}^{-1}\Omega_{-i,i}\right]^{1/2} \ .$$

Then, simulation sample expectations formed with respect to the draws $v^{(j)}$ for $j \le r$ converge almost surely to their population counterparts as $r \to +\infty$. The Gibbs sampler has the advantage of being smooth in parameters. However, its convergence can be very slow for problems with highly correlated variates.

Factor-Analytic MNP. For dynamic applications such as multiperiod binomial probit with autocorrelation and other applications with large dimension, alternatives to the MNP setup with a unrestricted covariance matrix may be more practical. McFadden (1984, 1989) suggests a "factor analytic" MNP with a components of variance structure, starting from

$$u_i = z_i\beta + \sum_{k=1}^{K} \lambda_{ik}\xi_k + \sigma_i\nu_i \ ,$$

where $\xi_1,...,\xi_K,\nu_1,...,\nu_J$ are independent standard normal, with the ξ_k interpreted as levels of unobserved factors and the λ_{ik} as the loading of factor k on alternative i. The λ's are identified by normalizations and exclusion restrictions. For multinomial response with a large number of alternatives, the response probabilities for this specification are

$$P_i(z,\theta) = \int_{\nu_i=-\infty}^{+\infty} \int_{\xi=-\infty}^{+\infty} \phi(\nu_i)\cdot \prod_{k=1}^{K} \phi(\xi_k)$$

$$\times \prod_{j\neq i} \Phi\left(\frac{(z_j-z_i)\beta + \Sigma_k[\Lambda_{jk}-\Lambda_{ik}]\cdot\xi_k + \sigma_i\nu_i}{\sigma_j}\right)\cdot d\nu_i d\xi_1 \cdots d\xi_K \ .$$

Estimation of this model requires numerical integration or simulation in K+1 dimensions, and provides J·(K+1) parameters, less normalization and exclusion restrictions, to describe the covariance structure. A simulator formed using Monte Carlo draws of $\nu_i,\xi_1,\cdots,\xi_K$ is continuous, positive, and yields simulated probabilities that sum to one. Judicious choice of factor structures will often be able to capture the primary correlation patterns, even in problems with large J, with a small number of factors. Furthermore, correlations are difficult to estimate precisely, so that it may be infeasible in typical data sets to estimate the parameters of a correlation structure containing more than a few factors.

For discrete panel data, with binomial observations over J periods coded as $(s_1,\ldots,s_J)$, with $s_j = \text{sign}(u_j)$ and $u_j = z_j\beta + \Sigma_k\Lambda_{jk}\cdot\xi_k + \sigma_j\nu_j$, the response probabilities are

$$P(s_1,\ldots,s_J | z_1,\ldots,z_J,\beta) = \int_{\xi=-\infty}^{+\infty} \prod_{k=1}^{K} \phi(\xi_k)\cdot \prod_{j\neq i} \Phi\left(\frac{s_i(z_i\beta + \Sigma_k\Lambda_{ik}\cdot\xi_k)}{\sigma_j}\right)\cdot d\xi_1 \cdots d\xi_K \ ,$$

and an integral of dimension K is required. Simulation based on Monte Carlo draws of $\xi_1,\cdots,\xi_K$ provides well-behaved unbiased simulators for MSM or MSLE.

LECTURE 5. MONTE CARLO MARKOV CHAIN METHODS

The Gibbs sampler is a special case of a *Markov Chain Monte Carlo* (MCMC) method for constructing draws from a distribution. Since these methods are potentially quite useful for simulation inference, I will summarize their structure and properties. I will concentrate on MCMC methods that use what is termed a Metropolis-Hastings (MH) kernel; see Roberts and Smith (1994), Tierney (1994), Gilks *et al* (1996), Robert (1996). Presentation of these methods requires some general terminology from the theory of Markov chains. Consider $(\mathbf{X},\mathfrak{B},\mu)$, where $\mathbf{X}$ is a convex subset of $\mathbb{R}^n$ that has a nonempty interior, $\mathfrak{B}$ is the Borel σ-algebra of subsets of $\mathbf{X}$, and μ is Lebesgue measure. A (*stochastic*) *Markov transition kernel* is a mapping $P{:}\mathbf{X}\times\mathfrak{B} \to [0,1]$ with the properties that $P(\cdot,B)$ is measurable for each $B \in \mathfrak{B}$, and $P(\varepsilon,\cdot)$ is a probability measure on $(\mathbf{X},\mathfrak{B})$ for each $\varepsilon \in \mathbf{X}$. A *Markov chain* is a sequence of random variables ε^0, ε^1,... with the ε^t drawn recursively from $P(\varepsilon^{t-1},\cdot)$. The probability measure of ε^t for a Markov chain starting from ε^0 is denoted $P^t(\varepsilon^0,\cdot)$. Note that $P^t(\varepsilon,B) = \int_{\mathbf{X}} P(y,B)\cdot P^{t-1}(\varepsilon,dy)$. A probability measure π on $(\mathbf{X},\mathfrak{B})$ is an *invariant distribution* for P if $\pi(\cdot) = \int P(\varepsilon,\cdot)d\pi(\varepsilon)$. The kernel P is π-*irreducible* if starting from any $\varepsilon^0 \in \mathbf{X}$, there is a positive probability of eventually reaching any set $B \in \mathfrak{B}$ that has $\pi(B) > 0$. The kernel is *aperiodic* if there is no partition of $\mathbf{X}$ such that the Markov chain cycles through the partition elements with probability one. The kernel is *Harris recurrent* if for every $\varepsilon^0 \in \mathbf{X}$ and $B \in \mathfrak{B}$ with $\pi(B) > 0$, a Markov chain starting from ε^0 will visit B infinitely often with probability one. The kernel is *ergodic* if for any probability measure ν on $(\mathbf{X},\mathfrak{B})$, $\left\|\int_{\mathbf{X}} P^t(\varepsilon,\cdot)d\nu(\varepsilon) - \pi(\cdot)\right\| \to 0$, where $\|\cdot\|$ denotes the total variation norm.[13]

The kernel is *uniformly recurrent* if $\operatorname{Sup}_{\varepsilon\in\mathbf{X}} \mathbf{E}S(\varepsilon,B) < +\infty$ for each $B \in \mathfrak{B}$ with $\pi(B) > 0$, where $S(\varepsilon,B)$ is the (random) minimum time $t \ge 1$ required to visit the set B starting from ε. A (signed) measure ω is *continuous* with respect to a measure η if

[13] The *total variation norm* is $\|\nu - \mu\| = \operatorname{Sup}_{B\in\mathfrak{B}} [\nu(B) - \mu(B)] - \operatorname{Inf}_{B\in\mathfrak{B}} [\nu(B) - \mu(B)]$ for measures ν and μ on $(\mathbf{X},\mathfrak{B})$. If ν and μ are probability measures, $\|\nu - \mu\| = 2\cdot\operatorname{Sup}_{B\in\mathfrak{B}} [\nu(B) - \mu(B)]$.

$B \in \mathfrak{B}$ and $\eta(B) = 0$ imply $\omega(B) = 0$. The Lebesgue decomposition theorem states that any bounded measure λ on $(\mathbf{X},\mathfrak{B},\mu)$ has a unique decomposition into a μ-*continuous* measure ω and a μ-*singular* measure ρ whose mass is concentrated on sets of μ-measure zero. The Radon-Nikodym theorem states that if ω is a bounded μ-continuous measure, then there exists a measurable function $w:\mathbf{X} \to \mathbb{R}$ such that $\omega(B) = \int_B w(\varepsilon)d\mu(\varepsilon)$. Let $\omega(\varepsilon,B) = \int_B w(\varepsilon,y)d\mu(y)$ denote the μ-continuous part of a Markov transition kernel $P(\varepsilon,\cdot)$, and term $\omega(\varepsilon,\cdot)$ *positive* if $w(\varepsilon,\cdot) > 0$ almost everywhere.

Term a stochastic Markov transition kernel $P(\varepsilon,B)$ a *sampler* if it can be written

$$P(\varepsilon,B) = \int_B w(\varepsilon,y)\cdot\mu(dy) + \delta_\varepsilon(B)\cdot[1 - R(\varepsilon)],$$

where w is a non-negative measurable function on $\mathbf{X}\times\mathbf{X}$, δ_ε is a Dirac measure with unit mass at ε, and $R(\varepsilon) = \int_{\mathbf{X}} w(\varepsilon,y)\cdot\mu(dy) \le 1$. This kernel corresponds to a process which, starting from ε, will with probability $1 - R(\varepsilon)$ return the same vector, and otherwise will return a vector $y \neq \varepsilon$ with a probability density $w(\varepsilon,y)/R(\varepsilon)$. A sampler is *reversible* for a density p if $p(\varepsilon)\cdot w(\varepsilon,y) = p(y)\cdot w(y,\varepsilon)$ on $\mathbf{X}\times\mathbf{X}$.

A *Metropolis-Hastings* (MH) *sampler* for a probability density $f(\cdot)$ is defined by a conditional density $q(y|\varepsilon)$ on $\mathbf{X}\times\mathbf{X}$ and $w(\varepsilon,y) = \text{Min}\{q(y|\varepsilon),f(y)\cdot q(\varepsilon|y)/f(\varepsilon)\}$. This kernel is associated with a transition process in which y is sampled from $q(y|\varepsilon)$, then the process moves to y with probability $\alpha(\varepsilon,y) = \text{Min}\{1,q(\varepsilon|y)\cdot f(y)/q(y|\varepsilon)\cdot f(\varepsilon)\}$, and otherwise stays at ε. This sampler is reversible for the density f, since $f(\varepsilon)\cdot w(\varepsilon,y) = \text{Min}\{f(\varepsilon)\cdot q(y|\varepsilon),f(y)\cdot q(\varepsilon|y)\}$ is symmetric in ε and y. An *independence* Metropolis-Hastings sampler has $q(y|\varepsilon) = g(y)$, independent of ε.

The Metropolis-Hastings sampler starts from an arbitrary point in $\mathbf{X}$, and proceeds recursively. Suppose at step t-1, the draw is ε^{t-1} and $f_{t-1} = f(\varepsilon^{t-1})$. Draw $\tilde{\varepsilon}^t$ from the conditional density $q(\cdot|\varepsilon^{t-1})$, and define $q_{t+} = q(\tilde{\varepsilon}^t|\varepsilon^{t-1})$ and $q_{+t} = q(\varepsilon^{t-1}|\tilde{\varepsilon}^t)$, Calculate $\alpha(\varepsilon^{t-1},\tilde{\varepsilon}^t) = \text{Min}\{1,q_{+t}f_t/q_{t+}f_{t-1}\}$. Draw a uniform [0,1] random number ζ. If $\zeta \le \alpha(\varepsilon^{t-1},\tilde{\varepsilon}^t)$, set $\varepsilon^t = \tilde{\varepsilon}^t$; otherwise, set $\varepsilon^t = \varepsilon^{t-1}$. Expectations with respect to $f(\cdot)$ are approximated by means over the ε^t for $t \le r$.

The following lemmas together establish for a probability density $f(\cdot)$ that is positive on $\mathbf{X}$, such as a truncated multivariate normal, and a conditional density $q(\cdot|\varepsilon)$ that is positive and bounded on $\mathbf{X}$, the Metropolis-Hastings sampler produces a sequence of draws whose empirical density converges weakly to $f(\cdot)$.

Lemma 5.1. (Tierney, 1994, Theorem 1, Corollary 2 to Theorem 2) *Suppose P is π-irreducible and aperiodic, and π is a μ-continuous invariant distribution of P. Then, π is the unique invariant distribution of P. If the μ-continuous part of* $P(\varepsilon,\cdot)$ *is positive for all* $\varepsilon \in \mathbf{X}$, *then P is Harris recurrent and ergodic.*

Lemma 5.2. (Tierney, 1994, Theorem 3; Roberts and Smith, 1994, Theorem 1) *Suppose the assumptions of Lemma 5.1, with the μ-continuous part of P positive for all* $\varepsilon \in \mathbf{X}$. *Suppose π is continuous with respect to μ, so that it has a density p satisfying* $\pi(B) = \int_B p(\varepsilon)d\mu(\varepsilon)$. *Suppose* h *is a π-integrable function. Then, for any* $\varepsilon^0 \in \mathbb{R}^n$, $\mathbf{E}_r h \equiv \frac{1}{r}\sum_{t=1}^{r} h(\varepsilon^t) \xrightarrow{p} \mathbf{E}h \equiv \int h(\varepsilon)\cdot p(\varepsilon)d\mu(\varepsilon)$ *almost surely as* $r \to +\infty$.

Lemma 5.3. (Orey, 1971, Theorem 7.2; Tierney, 1994, Proposition 2) *Suppose the assumptions of Lemma 5.2, and suppose P satisfies the* <u>minorization</u> <u>condition</u> *that for some* $\rho \in (0,1)$, $P(\varepsilon,B) \geq (1-\rho)\cdot\pi(B)$ *for all* $\varepsilon \in \mathbf{X}$ *and* $B \in \mathfrak{B}$. *Then P is uniformly recurrent, and there exists* $K > 0$ *such that* $\left\|\int_{\mathbf{X}} P^t(\varepsilon,\cdot)d\nu(\varepsilon) - \pi(\cdot)\right\| \leq K\cdot\rho^t$.

Lemma 5.4. (Tierney, 1994, Theorem 5; Chan, 1993, Theorem 1) *Suppose the assumptions of Lemma 5.3. Suppose* h *is a real π-square integrable function. Then,* $\sqrt{r}(\mathbf{E}_r h - \mathbf{E}h)$ *is asymptotically normal with a finite variance* σ^2.

Lemma 5.5. (Tierney, 1994) *If a sampler* $P(\varepsilon,B)$ *is reversible with respect to a probability density p, then p is an invariant of the process; i.e.,*

$$\int_{\mathbf{X}} P(\varepsilon,B)\cdot p(\varepsilon)d\mu(\varepsilon) = \int_B p(y)d\mu(y) \ .$$

Proofs: The last result in Lemma 5.1 follows from Corollary 2 of Tierney (1994), as inspection of his proof shows that it holds for any sampler, not just the Metropolis-Hastings sampler. This establishes the property of Harris recurrence, which in turn yields the result in Lemma 5.2. Lemma 5.5 is established by

$$\int_{\mathbf{X}} P(\varepsilon,B)\cdot p(\varepsilon)d\mu(\varepsilon) = \int_{\varepsilon\in\mathbf{X}}\int_{y\in B} w(\varepsilon,y)\cdot p(\varepsilon)d\mu(\varepsilon)\cdot\mu(dy)$$

$$+ \int_{\varepsilon\in\mathbf{X}}\delta_{\varepsilon}(B)\cdot[1 - R(\varepsilon)]\cdot p(\varepsilon)d\mu(\varepsilon) = \int_B p(\varepsilon)d\mu(\varepsilon) + \int_{\varepsilon\in\mathbf{X}}\int_{y\in B} w(\varepsilon,y)\cdot p(\varepsilon)d\mu(\varepsilon)\cdot\mu(dy)$$

$$- \int_{\varepsilon\in B}\int_{y\in\mathbf{X}} w(\varepsilon,y)\cdot p(\varepsilon)d\mu(\varepsilon)\cdot\mu(dy) = \int_B p(\varepsilon)d\mu(\varepsilon). \quad \blacksquare$$

The results above do not establish a rate of convergence. However, a modification of the MH sampler that corresponds to computational practice yields stronger results. Let M denote a large number, say the largest number representable in standard floating point computer calculations, and define $\mathbf{X} = \{\varepsilon \mid -M \le \varepsilon \le M\}$. Construct the quantities $\tilde{\varepsilon}_j^t$, g_t, and f_t as above, but now define the Markov chain as

$$\varepsilon^t = \begin{cases} \tilde{\varepsilon}^t & \text{if } \tilde{\varepsilon}^t \in \mathbf{X} \text{ and } \zeta \le f_t\cdot q_{+t}/f_{t-1}\cdot q_{t+} \\ \varepsilon^{t-1} & \text{otherwise} \end{cases}.$$

This is now a Metropolis-Hastings sampler on the compact set $\mathbf{X}$; its invariant is the truncated density $\mathbf{1}(\varepsilon\in\mathbf{X})\cdot f(\varepsilon)/\int_{\mathbf{X}} f(y)d\mu(y)$. Note that the sampler requires only the ratio of this density at two arguments, so that the normalizing denominator does not appear. To the level of practical computer accuracy, this density will be indistinguishable from the untruncated density $f(\cdot)$. On the compact set X, there is a lower bound $1 - \rho > 0$ on $w(\varepsilon,y)/f(y) = \mathrm{Min}\{q(y|\varepsilon)/f(y), q(\varepsilon|y)/f(\varepsilon)\}$, implying the minorization condition $P(\varepsilon,B) \ge (1-\rho)\cdot\pi(B)$. Consider a continuous bounded function $h(\varepsilon)$ for $\varepsilon \in \mathbf{X}$. Then, Lemma 5.3 establishes there exists $K > 0$ such that the bias in the modified MH sampler goes away at a $1/r$ rate; specifically,

$$|\mathbf{E}_r h(\varepsilon^t) - \mathbf{E}h(\varepsilon)| \le 2K/(1-\rho)r,$$

where K is a product of the bound in Lemma 5.3 and the bound on $h(\cdot)$. Lemma 5.4 establishes that the scaled difference $\sqrt{r}(\mathbf{E}_r h(\varepsilon^t) - \mathbf{E}h(\varepsilon))$ is asymptotically normal with a variance σ^2. To estimate σ^2, consider R independent repetitions of the Markov chain of length r, with realizations ε^{tk}, and define

$$\mathbf{E}_{rR}h \equiv \frac{1}{rR}\sum_{k=1}^{R}\sum_{t=1}^{r} h(\varepsilon^{tk})$$

and

$$\hat{\sigma}^2 = \frac{1}{R}\sum_{k=1}^{R}\left(\frac{1}{r}\sum_{t=1}^{r} h(\varepsilon^{tk}) - \mathbf{E}_{rR}h\right)^2 \quad .$$

These estimators are consistent when $R,r \to +\infty$.

While the last results guarantee rapid asymptotic convergence, they do not establish the rate ρ or the constant K. Roberts (1992) and Zellner and Min (1995) propose convergence criteria that may be useful. The statistical properties of the sampler can be estimated by constructing $k = 1,...,R$ independent chains with stopping times r_k. Let $\mathbf{E}_r h^{(k)}$ denote the estimate of $\mathbf{E}h$ obtained from chain k. Then, a regression

$$\mathbf{E}_r h^{(k)}\cdot\sqrt{r_k} = \beta_0\cdot\sqrt{r_k} + \beta_1/\sqrt{r_k} + \xi_r$$

will provide an "accelerated" estimate $\hat{\beta}_0$ of the almost sure limit $\mathbf{E}h$ of the sampler. The coefficient β_1 provides information on bias, but Monte Carlo experiments suggests that it is not accurate enough to be used for first-order bias correction. The estimate of β_0 will be consistent for $\mathbf{E}h$ as R and $\min_k r_k$ approach infinity.

LECTURE 6. MIXED MNL MODELS FOR DISCRETE RESPONSE[14]

Define a *mixed multinomial logit (MMNL)* model as a MNL model with random coefficients drawn from a density $k(\cdot)$:

$$P_C(i|x.s) = \int L_C(i|x,\theta)\cdot k(\theta)d\theta \qquad \text{with} \qquad L_C(i|x,\theta) = e^{x_i\theta}/\sum_{j\in C} e^{x_j\theta} \ .$$

The x_i are functions of observed attributes of alternative i, possibly interacted with observed characteristics of the decision-maker. The MMNL model was introduced by Talvitie (1974), Westin (1974), and Westin and Gillen (1978). There is a lengthy literature investigating various aspects of this model; see Beggs (1988), Dubin and Zeng (1991), Enberg, Gottschalk, and Wolf (1990), Follmann and Lambert (1989), Formann (1992), Gonul and Srinivasan (1993), Jain, Vilcassim, and Chintagunta (1994), Montgomery, Richards, and Braun (1986), Steckel and Vanhonacker (1988), Warren and Strauss (1979), Train and Revelt (1995), Train and Brownstone (1996), and Train (1996). Chesher and Santos (1995) have developed specification tests for MMNL that are relatives of the ones proposed here. Estimation of MMNL by MSLE or MSM is particularly tractable when it is easy to draw from the density k. A kernel-smoothed simulator for multinomial probit that uses a MNL kernel can be reinterpreted as a simulator of a MMNL model.

McFadden and Train (1996) establish the following results, loosely stated:

° Under mild regularity conditions, any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a mixed MNL model.

° Practical estimation of a parametric mixing family $k(\cdot|\beta)$ can be carried out by MSLE when Monte Carlo draws can be made from k, or from an importance sampling density h, so that the simulation approximations to $P_C(i)$ and its derivatives are stochastically equicontinuous.

° A mixed MNL model with normally distributed coefficients can approximate a multinomial probit model as closely as one pleases.

[14] This lecture is extracted from the paper "Mixed Multinomial Logit Models for Discrete Response" by Daniel McFadden and Kenneth Train.

° The adequacy of a mixing specification can be tested simply as an omitted variable test with appropriately defined artificial variables.

° Nonparametric estimation of a random utility model for choice can be approached by successive approximations by MMNL models with finite mixing distributions; e.g., *latent class* models.

An application to a problem of demand for alternative vehicles shows that MMNL provides a flexible and computationally practical approach to discrete response analysis.

<u>A General Approximation Property of MMNL</u>

Economic theory often suggests that discrete responses are the result of optimization of payoffs to decision-makers: utility for consumers, profit for firms. The following discussion will be phrased in terms of utility-maximizing consumers. When unobserved heterogeneity in the population of consumers is accounted for, one has a class of response models based on random utility maximization (RUM). To fix ideas, assume a random utility model $U(z,s,\zeta,\varepsilon)$, where z is a vector of observed attributes of the alternative, s is a vector of observed characteristics of the decision-maker, ζ is a vector of unobserved attributes of the alternative, and ε is a vector of unobserved variables characterizing tastes. A primitive postulate of preference theory is that tastes are established prior to a specific choice context; this implies that <u>the distribution of ε cannot depend on</u> (z,ζ). In general, one might expect the distribution of ε to depend on s, say with a CDF $H(\varepsilon|s)$. But one can always represent ε in the form $\varepsilon = h(p,s)$, where p is a vector of uniform [0,1] random variables with the same dimension as ε, and then write the utility function as $U(z,s,\zeta,h(p,s))$.[15] Without loss of generality, absorb this transformation into the definition of U and consider the random utility model $U(z,s,\zeta,\varepsilon)$ with ε uniformly distributed on a unit hypercube. In general, one would expect the distribution of ζ

---

[15] Suppose ε is of dimension m, and define $p_i = \dfrac{H(\varepsilon_1,\varepsilon_{i-1},\varepsilon_i,+\infty,...,+\infty|s)}{H(\varepsilon_1,\varepsilon_{i-1},+\infty,+\infty,...,+\infty|s)}$ for $i \le m$. The random variables $(p_1,..,p_m)$ are independently uniform [0,1]. The system of equations can be solved recursively for $\varepsilon_i = h^i(p_1,...,p_i,s)$.

to depend on z. Once again, ζ can be represented as a transformation of a uniformly distributed vector, and this transformation can be absorbed into the definition of *U*. Then, finally, there is no loss of generality in writing the random utility function as $U(z,s,\varepsilon)$, with ε uniformly distributed on a unit hypercube, independently of z and s. The following result establishes that MNL mixtures can closely approximate very broad classes of random utility models:

Theorem 6.1. *Suppose alternatives* i *come from a finite master choice set* $C_O = \{1,...,J\}$. *Suppose discrete responses maximize a utility function* $U^*(z_i,s,\varepsilon)$ *that is a continuous function of its arguments, where* z_i *are observed attributes of alternative* i, s *are observed characteristics of the decision-maker, and* ε *are unobserved variables that can be assumed without loss of generality to be uniformly distributed on a unit hypercube* Ω. *Assume that* $\mathbf{z} = (z_1,...,z_J)$ *varies in a compact set* $\mathbf{Z}^J$, *that* s *varies in a compact set* $\mathbf{S}$, *and that there is zero probability of ties between alternatives given* $\mathbf{z} \in \mathbf{Z}^J$ *and* $\mathbf{s} \in \mathbf{S}$. *Let* $P^*_C(i)$ *denote the choice probabilities generated by maximization of* U^* *from any choice set* $C \subseteq C_O$. *If* η *is a small positive scalar, then there exists a random utility function whose choice probabilities* $P_C(i)$ *are of MMNL form such that for all* $s \in \mathbf{S}$, $\mathbf{z} \in \mathbf{Z}^J$, and $C \subseteq C_O$, $P^*_C(i)$ *and* $P_C(i)$ *differ by at most* η.

Proof: The continuous function U^* has a Bernstein-Weierstrauss polynomial approximation U^{*k} on $\mathbf{Z}\times\mathbf{S}\times\Omega$ that satisfies $|U^* - U^{*k}| \le 1/k$. Form $U^k(z_i,s,\varepsilon) = U^{*k}(z_i,s,\varepsilon) + \nu_i/k^2$, where the ν_i are i.i.d. Extreme Value Type I. Define

$$A_k(\mathbf{z},s) = \{\varepsilon\in\Omega \mid |U^*(z_i,s,\varepsilon) - U^*(z_j,s,\varepsilon)| \le 5/k \text{ for some } i \ne j\} \ .$$

Then $A_k(z,s)$ is monotone decreasing to the set of ε that result in a tie. By hypothesis, this limit set has measure zero. Then there exists $k(\mathbf{z},s) \ge -\log(\eta/2)$ such that for $k \ge k(\mathbf{z},s)$, the measure of $A_k(\mathbf{z},s)$ is less than $\eta/2$. Consider the event of a preference reversal,

$$B_k(\mathbf{z},s) = \{\varepsilon,\nu \mid \exists\ i,j \ni U^*(z_i,s,\varepsilon) > U^*(z_j,s,\varepsilon) \text{ but } U^k(z_i,s,\varepsilon) < U^k(z_j,s,\varepsilon)\}$$

If $B_k(\mathbf{z},s)$ occurs and $A_k(\mathbf{z},s)$ does not, then it must be true that

$$U^{*k}(z_i,s,\varepsilon) + \nu_i/k^2 < U^{*k}(z_j,s,\varepsilon) + \nu_j/k^2,$$

$$U^*(z_i,s,\varepsilon) - U^{*k}(z_i,s,\varepsilon) < 1/k \qquad \text{and} \qquad U^{*k}(z_j,s,\varepsilon) - U^*(z_j,s,\varepsilon) > 1/k \ ,$$

$$5/k < U^*(z_i,s,\varepsilon) - U^*(z_j,s,\varepsilon) \ .$$

Together, these inequalities imply $\nu_i - \nu_j < -3k$. Then, the probability of $B_k(\mathbf{z},s)$ is no greater than the probability of $A_k(\mathbf{z},s)$ plus the probability of $\nu_i - \nu_j < -3k$; this sum is less than η.

Given $\mathbf{z}$, s, and $k = k(\mathbf{z},s)$, continuity of U^* implies an open neighborhood $N_k(\mathbf{z},s) \subseteq \mathbf{Z}^J\times\mathbf{S}$ such that if not $A_k(\mathbf{z},s)$, then $3/k < U^*(z_i',s'\varepsilon) - U^*(z_j',s',\varepsilon)$ for $(\mathbf{z}',s') \in N_k(\mathbf{z}.s)$. If the event

$$B_k'(\mathbf{z},s) = \{\varepsilon,\nu \,|\, \exists\ i,j \ni U^*(z_i',s',\varepsilon_i) > U^*(z_j',s',\varepsilon_j)\ \&\ U^k(z_i',s',\varepsilon_i) < U^k(z_j',s',\varepsilon_j)\}$$

occurs, but the event $A_k(\mathbf{z},s)$ does not, then $\nu_i - \nu_j < -k$. Then the probability of $B_k'(z,s)$ is less than η.

The neighborhoods $N_k(\mathbf{z},s)$ constitute an open covering of the compact set $\mathbf{Z}^J\times\mathbf{S}$. Then there exists a finite sub-covering with centers $(\mathbf{z}^i,s^i)$, $i \le n$. Choose $k = \max_{i\le n} k(\mathbf{z}^i,s^i)$, and write the polynomial approximation U^k in the form

$$U^k(z,s,\varepsilon) = x(z,s)\cdot\theta(\varepsilon) + \nu_i/k^2 \ ,$$

where x_i is a vector of the z_i and s parts of the terms in the polynomial and θ is a vector of the ε parts. The construction then guarantees that with probability at least $1 - \eta$, U^* and U^k order the alternatives in C_o the same. Therefore, $P_C(i)$ from U^k and $P_C^*(i)$ from U^* differ by at most η. ■

This result establishes that for most economic optimizing models, MMNL is sufficiently flexible to provide a satisfactory econometric approximation. The proof provides no useful indication of how to choose parsimonious mixing families, or how many simulation draws from the mixing distribution are needed to obtain acceptable approximations to $P_C(i)$. However, Monte Carlo studies indicate that fairly simple mixing structures, with random coefficients following a factor analytic structure of relatively low dimension, and relatively simple mixing families, such as latent class models with relatively few classes, are sufficiently flexible to capture quite complex patterns of substitution; see Bolduc, Fortin, and Gordon (1996).

Simulation of the MMNL Model

Assume $C = \{1,...,J\}$. A tractable empirical form for the MMNL model $P_C(i) = \int L_C(i;\theta)\cdot k(\theta)d\theta$ is obtained by taking $\theta = \beta + \Lambda\zeta$, where β is a $k\times 1$ vector of "mean" coefficients, Λ is a $k\times m$ matrix of factor loadings, with normalizations and exclusion restrictions for identification, and ζ is a $m\times 1$ vector of factor levels that are assumed to be independently distributed with zero means and a "standard" density $f(\zeta)$. (This specification includes models with alternative-specific random effects: take x_j to include alternative-specific dummies and introduce factors that load on these dummies.) The associated score of an observation satisfies

$$\nabla_\beta \log P_C(i) = \frac{\int (x_i - x_C(\zeta)) L_C(i;\beta+\Lambda\zeta)\cdot f(\zeta)d\zeta}{\int L_C(i;\beta+\Lambda\zeta)\cdot f(\zeta)d\zeta} = x_i - \sum_{j\in C} x_j \cdot \mathbf{E}_\zeta(L_C(j;\beta+\Lambda\zeta)\,|\,i) ,$$

$$\nabla_\Lambda \log P_C(i) = \frac{\int (x_i - x_C(\zeta))\zeta' L_C(i;\beta+\Lambda\zeta)\cdot f(\zeta)d\zeta}{\int L_C(i;\beta+\Lambda\zeta)\cdot f(\zeta)d\zeta} = \mathbf{E}_\zeta\left\{x_i\zeta' - \Sigma_{j\in C} x_j\cdot\zeta' L_C(j;\beta+\Lambda\zeta)\,|\,i\right\} ,$$

where $\mathbf{E}_\zeta(\cdot\,|\,i)$ is the expectation with respect to the conditional density of ζ given that i is chosen: $L_C(i;\beta+\Lambda\zeta)\cdot f(\zeta)/\int L_C(i;\beta+\Lambda\zeta')\cdot f(\zeta')d\zeta'$.

Estimation by MSLE can be carried out by averaging the MNL formula $L_C(i;\beta+\Lambda\zeta)$ with respect to r Monte Carlo draws from $f(\zeta)$, and keeping these draws fixed throughout the optimization. (If it is difficult to make Monte Carlo draws from f, then an importance sampling procedure can be used.) This yields unbiased simulators of $P_C(i)$, but of course the simulator of the log likelihood is not unbiased. MSLE estimators that are relatively free of simulation bias in finite samples are likely to require r considerably larger than $\sqrt{N}$.

The score for an observation can be rewritten

$$s_N(\beta,\Lambda) = \sum_{i\in C} \{d_i - \mathbf{E}_\zeta L_C(i;\beta+\Lambda\zeta)\}\cdot W(i,\beta,\Lambda) ,$$

the cross-product of the *generalized residual* $d_i - \mathbf{E}_\zeta L_C(i;\beta+\Lambda\zeta)$ and an *instrument vector* $W(i,\beta,\Lambda)$ which is itself the score,

$$W(i,\beta,\Lambda) = \begin{bmatrix} \nabla_\beta \log P_C(i) \\ \mathrm{vec}\{\nabla_\Lambda \log P_C(i)\} \end{bmatrix} .$$

For MSM estimation, any instrument vector that is fully correlated with $W(i,\beta,\Lambda)$ can be used to obtain estimators that are consistent asymptotically normal, but in general less than fully efficient. In particular, MSM estimation can be carried out using an unbiased simulator for $\mathbf{E}_\zeta(L_C(i;\beta+\Lambda\zeta)$ in the generalized residual, and any statistically independent instrument vector that is fully correlated with the score, such as

$$W^r(i,\beta,\Lambda) = \begin{bmatrix} \dfrac{\mathbf{E_r}(x_i - x_C(\zeta))L_C(i;\beta+\Lambda\zeta)}{\mathbf{E_r}L_C(i;\beta+\Lambda\zeta)} \\ \\ \mathrm{vec}\left\{\dfrac{\mathbf{E_r}(x_i - x_C(\zeta))\zeta' L_C(i;\beta+\Lambda\zeta)}{\mathbf{E_r}L_C(i;\beta+\Lambda\zeta)}\right\} \end{bmatrix} ,$$

where $\mathbf{E_r}$ denotes empirical expectation over a simulation sample of r Monte Carlo draws from $f(\zeta)$. Large r will be needed to simulate $W(i,\beta,\Lambda)$ accurately. However, it is possible to obtain a computationally convenient instrument vector that is fairly highly correlated with $W(i,\beta,\Lambda)$, and will as a consequence yield moderately efficient MSM estimates at low computational cost. Using the approach of Talvitie (1974), make a first-order Taylor's expansion of the multinomial logit function $L_C(i|\beta+\Lambda\zeta)$ in ζ around $\zeta = 0$, and ignore higher-order terms. This yields the approximation

$$L_C(i|\beta+\Lambda\zeta) \approx L_C(i|\beta)\left\{1 + (x_i - x_C^o)'\Lambda\zeta\right\} ,$$

where $x_C^o = \sum_{i\in C} x_i L_C(i|x,\beta)$. Because the Taylor's expansion is not uniformly convergent, this is not a good approximation to the MMNL response probability. However, it does provide the basis for an easily computed approximation to $W(i,\beta,\Lambda)$. Substituting the approximation above for $L_C(i|\beta+\Lambda\zeta)$ yields

$$W(i,\beta,\Lambda) \approx \begin{bmatrix} x_i - x_C^o \\ \text{vec}\left\{(x_i - x_C^o)(x_i - x_C^o)'\Lambda - \Sigma_k L_C(k|x,\beta)(x_k - x_C^o)(x_k - x_C^o)'\Lambda\right\} \end{bmatrix}.$$

For preliminary estimation, β can be set to simple MNL coefficient estimates and Λ can be any matrix of full column rank that respects the exclusion restrictions present in the specification of Λ.

Toward Nonparametric Estimation of Random Utility Models

When multinomial choice probabilities are smooth functions of covariates, and cannot be placed a priori in parametric families, it is possible to approach estimation of these probabilities as a conventional non-parametric problem. For example, the system

$$\begin{bmatrix} d_{1n} \\ d_{2n} \\ \vdots \\ d_{Jn} \end{bmatrix} = \begin{bmatrix} P_C(1|x_n) \\ P_C(2|x_n) \\ \vdots \\ P_C(J|x_n) \end{bmatrix} + \begin{bmatrix} \xi_{1n} \\ \xi_{2n} \\ \vdots \\ \xi_{Jn} \end{bmatrix}$$

for observations $n = 1,...,N$, with d_{in} is an indicator for the observed choice, can be treated as a non-parametric regression problem. The only potential complications are heteroscedasticity in the disturbances, which in principle can be handled using a two-stage estimation procedure; and discrete covariates, which require enumeration of all configurations.[16] However, a significant drawback is that the nonparametric estimates are not necessarily consistent with random utility maximization.

---

[16] First-stage regression estimates $\hat{P}_C(i|x)$ can be used to estimate the covariances $\text{cov}(\xi_i,\xi_j) = \delta_{ij}\cdot\hat{P}_C(i|x) - \hat{P}_C(i|x)\cdot\hat{P}_C(j|x)$, and a generalized least squares transformation of the data can be carried out prior to second-stage nonparametric regression.

Economists may want to impose such consistency as a maintained hypothesis; see Matzkin (1994).

An approach to nonparametric estimation that maintains consistency with random utility maximization is suggested by Theorem 6.1. The idea is to estimate a MMNL model with finite mixtures (also called *latent class models*), with the number of mixing points growing slowly with sample size. Assume a random utility function $U(z,s,\varepsilon)$ that maps observable $z \in Z \subseteq \mathbb{R}^m$ and $s \in \mathbf{S} \subseteq \mathbb{R}^k$, and unobservable ε in a unit hypecube, into [0,1]. (Since the utility function is ordinal, confining its range to the unit interval is no restriction. As noted in the introduction to Theorem 6.1, we can assume without loss of generality that ε is uniformly distributed, independently of z and s. Assume a master set $C_o = \{1,...,J\}$ of possible choice alternatives, and let $\mathbf{z} = (z_1,...,z_J) \in \mathbf{Z}$ with $\mathbf{Z} = Z^J$.

Paralleling the construction in the proof of Theorem 6.1, approximate U^* uniformly within υ by a Bernstein-Weierstrauss polynomial plus an additive i.i.d. Extreme Value I disturbance to obtain $U^{\#}(z_i,s,\varepsilon,\nu_i) = x(z_i,s)'\Theta w(\varepsilon) + \nu_i \cdot \upsilon^2$, where x and w are vectors of known functions of their arguments and Θ is a diagonal matrix of unknown coefficients. When U^* is Lipschitz with a known bound M, it is possible to establish *a priori* the order of the polynomial necessary to achieve this degree of accuracy; e.g., McFadden and Mundlak (1978, p.236) establish $(m+k)JM^2/\upsilon^2$ suffices. Now consider estimating a MMNL model using simulation, with r_N draws of ξ from the uniform distribution. The tolerance υ determines the degree of bias in the MMNL approximation to the true choice probabilities, and $\upsilon \to 0$ at any rate is sufficient to guarantee that there is zero asymptotic approximation bias. Increasing r_N at a rate faster than $\sqrt{N}$ is sufficient to eliminate asymptotic contributions from simulation noise. Finally, there will be a sufficiently slow rate of decrease in υ so that the variance of the estimated choice probabilities goes to zero. This argument is summarized in the following result:

Theorem 6.2 *Suppose choice from subsets of a finite master choice set* C_o *are generated by a uniformly Lipschitz random utility function* $U^*(z,s,\varepsilon)$ *with* ε *uniform on* $[0,1]^m$, $z \in Z \subseteq \mathbb{R}^m$, $s \in \mathbf{S} \subseteq \mathbb{R}^k$, *and* $Z\times\mathbf{S}$ *compact. Then consistent MSLE or MSM estimation of the choice probability functions can be carried out by estimating a RUM-compatable latent class MMNL model*

$$P_C^{\#}(i) = \frac{1}{r_N} \sum_{j=1}^{r_N} \frac{\exp(x^k(z_i,s)'\Theta w^k(\varepsilon^j)/k^2)}{\sum_{j\in C} \exp(x^k(z_j,s)'\Theta w^k(\varepsilon^j)/k^2)} ,$$

where $x^k(z,s)$ *and* $w^k(\varepsilon)$ *are vectors of polynomials of order up to* k *and* Θ *is a diagonal matrix of parameters, the* ε^j *are random draws with* r_N *increasing more rapidly than* $\sqrt{N}$, *and* k *is increasing to infinity with sample size at a sufficiently slow rate so that the variance of the estimated probabilities converge to zero.*

Note that this theorem does not claim that Θ is identified; in general, some identifying normalizations will be needed. The nonlinear mapping of the estimated utility function to choice probabilities complicates the usual arguments for asymptotic normality, optimal rates, and cross-validation methods. Further research will be required to establish analogous results for this problem. One promising avenue is to seek a generalization of a result of Cenkov (1982) which shows for a broad class of nonparametric regression problems, orthogonal series approximations with truncation based on a stopping rule using T-statistics is approximately optimal. If this approach can be made to work, then MMNL approximation in the form above combined with LM tests of the form given in Theorem 6.2 would provide a practical basis for nonparametric estimation within the family of continuous random utility models.

Approximating MNP

Consider the MNP model derived from a RUM with $u = Z\beta + \varepsilon$, with u a $J\times1$ vector of utilities for the alternatives in C_o and ε multivariate normal with mean zero and covariance matrix Σ. In many applications, it is useful to start from a random coefficients interpretation of the RUM, with $u = Z\alpha$ and $\alpha = \beta + \Lambda\zeta + D\nu$, where Λ is a $k\times m$ matrix of factor loadings, D is a diagonal matrix, and ζ and ν are vectors of independent standard normal variates. Then, $\Sigma = Z(D^2 + \Lambda\Lambda')Z'$. This model places no restrictions on Σ when $m = J-1$ and the factors are loaded solely on alternative-specific effects. Identification requires restrictions on Λ, typically exclusion restrictions, and on D, since only parameters that appear in normalized utility differences can be identified from discrete choice behavior. The choice probabilities from this model can be written

$$P^{*}_{C}(i) = \int_{\nu_i}\int_{\zeta} \prod_{j\in C_{-i}} \Phi\left(\frac{\Delta_i(Z\beta + \Lambda\zeta) + D_{ii}\nu_i}{D_{jj}}\right)\cdot\left(\prod_{k=1}^{m} \phi(\zeta_k)\cdot d\zeta_k\right)\cdot\phi(\nu_i)\cdot d\nu_i ,$$

where C_{-i} denotes the set C with alternative i deleted. It is possible to estimate this model using direct numerical integration when m is small, or using MSME or MSM with Monte Carlo integration when m is large. Alternately, perturb the RUM to $u = Z(\beta + \Lambda\zeta + D\nu) + \omega\cdot\eta$, where the η are independent Extreme Value Type I distributed and ω is a small window width, to obtain

$$P_C(i) = \int_{\nu}\int_{\zeta} L_C(i;(\beta + \Lambda\zeta + D\nu)/\omega)\cdot\left(\prod_{k=1}^{m} \phi(\zeta_k)\cdot d\zeta_k\right)\cdot\left(\prod_{j\in C}^{m} \phi(\nu_j)\cdot d\nu_j\right) .$$

As in the proof of Theorem 6.1, by taking ω small one can make the probability negligible that the perturbed model will order alternatives differently than the MNP model. In estimation, ω can be absorbed into β, Λ, and D. Once the simulation draws of ζ and ν are made for each observation and then fixed for the remainder of the analysis, the response probability is simply an average of MNL probabilities. Brownstone and Train (1996) find in a application that the MMNP model can approximate MNP probabilities more accurately than a direct GHK simulator, when both are constrained to use the same amount of computer time. This finding is supported by Ben-Akiva and Bolduc (1996), who find in Monte Carlo experiments that MMNP gives approximation to MNP probabilities that are comparable to the GHK algorithm.

It is also possible to approximate nested MNL models closely by considering a MMNL model with a factor that loads on a node-specific dummy for each node in the nested logit tree.

Specification Testing

Because the MMNL model requires use of simulation methods, it is useful to have a specification test based on MNL model estimates that determine if mixing is needed. The next result describes a Lagrange Multiplier test for this purpose. This test has the pivotal property that its asymptotic distribution under the null hypothesis that the correct specification is MNL does not depend on the structure of the mixing distribution under the alternative.

Theorem 6.3. *Consider choice from a set* C = {1,...,J}. *Let* x_i *be a vector of attributes of alternative* i. *Suppose from a random sample* n = 1,...,N *one estimates the parameter* $\hat{\theta}$ *in the simple MNL model*,

$$L_{Cn}(i;\theta) = e^{x_{in}\theta} / \sum_{j\in C} e^{x_{jn}\theta} ,$$

using maximum likelihood; constructs artificial variables for selected components r *of* x_{in},

$$z_{rin} = (x_{rin} - x_{rCn})^2/2 \qquad \textit{with} \qquad x_{rCn} = \sum_{j\in C} x_{rjn} \cdot L_{Cn}(j;\hat{\theta}) \ ;$$

and then uses a Wald or Likelihood Ratio test for the hypothesis that the artificial variables z_{rin} *should be omitted from the MNL model. This test is asymptotically equivalent to a Lagrange multiplier test of the hypothesis of no mixing against the alternative of a mixed MNL model* $P_C(i) = \int L_C(i;\theta)\cdot k(\theta)d\theta$ *with mixing in the selected components* r *of* θ. *The degrees of freedom equals the number of artificial variables* z_{rin} *that are linearly independent of x*.

Proof: The Lagrange Multiplier testing problem is one in which a natural parameterization in terms of the standard deviations of the mixing density leads to a log likelihood whose score is identically zero under the null. Then, it is necessary to reparameterize the model, as in Lee and Chesher (1986) and Newey and McFadden (1995), to circumvent this problem. Write the MMNL model as

$$P_C(i) = \int L_C(i;\beta+\upsilon^{1/2}\odot\zeta)\cdot k(\zeta)d\zeta \ ,$$

where β and υ are vectors of parameters, $\upsilon^{1/2}$ is the vector of square roots of the components of υ, ζ is a vector of random variables that has mean zero, component variances of one, and full rank over the specified components r, and $\upsilon^{1/2}\odot\zeta$ denotes the component-by-component product. Let $x_{Cn} = \Sigma_{j\in C} x_{jn}\cdot L_{Cn}(j;\beta+\upsilon^{1/2}\odot\zeta)$. Then

$$\nabla_\beta P_{Cn}(i) = \int L_{Cn}(i;\beta+\upsilon^{1/2}\odot\zeta)\cdot(x_{in} - x_{Cn})\cdot k(\zeta)d\zeta \ ,$$

$$\nabla_{\upsilon_r} P_{Cn}(i) = 0.5 \cdot \upsilon^{-1/2} \cdot \int L_{Cn}(i;\beta+\upsilon^{1/2} \odot \zeta) \cdot (x_{in} - x_{Cn}) \cdot \zeta_r \cdot k(\zeta) d\zeta \ .$$

Taking the limit as the $\upsilon_r \to 0$, and using L'Hopital's rule on $\nabla_{\upsilon_r} P_{Cn}(i)$, one obtains

$$\nabla_{\beta} P_{Cn}(i) = L_{Cn}(i;\hat{\beta}) \cdot (x_{in} - x_{Cn}) \ ,$$

$$\nabla_{\upsilon_r} \log P_{Cn}(i) = L_{Cn}(i;\hat{\beta}) \cdot (z_{rin} - z_{rCn}) \ .$$

The sample mean of $\nabla_{\beta} \log P_{Cn}(i)$ is zero at the maximum likelihood estimator $\hat{\beta}$ of the simple MNL model, and the Lagrange Multiplier statistic tests whether the vector of sample means of $\nabla_{\upsilon_r} \log P_{Cn}(i)$ for the selected r are zero. As in McFadden (1987), this test is equivalent to a Lagrange Multiplier test for the null hypothesis that the variables z_{rin} have zero coefficients in the MNL model, and thus asymptotically equivalent to a Likelihood Ratio or Wald test for this hypothesis. □

To examine the operating characteristics of the test for mixing in the MNL coefficients, I carried out two simple Monte Carlo experiments for choice among three alternatives, with random utility functions $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. The disturbances ε_i were i.i.d. Extreme Value Type I. In the first experiment, the covariate were distributed as described in the table below:

Variable	Alternative 1	Alternative 2	Alternative 3
x_1	±1/2 w.p. 1/2	0	0
x_2	±1/2 w.p. 1/2	±1/2 w.p. 1/2	0

The parameter $\alpha_2 = 1$ under both the null and the alternative. The parameter $\alpha_1 = 0.5$ under the null hypothesis, and under the alternative $\alpha_1 = 0.5 \pm 1$ w.p. 1/2. I carried out 1000 repetitions of the test procedure for a sample of size N = 1000 and choices generated alternately under the null hypothesis and under the alternative just described, using likelihood ratio tests for the omitted variable z_{1ni}. The results are given in the table below:

Nominal Significance Level	Actual Significance Level	Power Against the Alternative
10%	8.2%	15.6%
5%	5.0%	8.2%

The nominal and actual significance levels of the test agree well.[17] The power of the test is low, and an examination of the estimated coefficients reveals that the degree of heterogeneity in tastes present in this experiment does not cause estimates coefficients to deviate significantly from their expected values. Put another way, this pattern of heterogeneity is difficult to distinguish from added extreme value noise.

In the second experiment, the covariates are distributed as shown in the table below:

Variable	Alternative 1	Alternative 2	Alternative 3
x_1	$\pm 1/2$ w.p. 1/2	$\pm 1/2$ w.p. 1/2	0
x_2	$\pm 1/2$ w.p. 1/2	$\pm 1/2$ w.p. 1/2	0

The utility function is again $u_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \varepsilon_i$. Under the null hypothesis, $\alpha_1 = \alpha_2 = 1$, while under the alternative $(\alpha_1,\alpha_2) = (2,0)$ w.p. 1/2 and (0,2) w.p. 1/2. Again, 1000 repetitions of the tests are made for N = 1000 under the null and the alternative; the results are given in the table below:

Nominal Significance Level	Actual Significance Level	Power Against the Alternative
10%	9.7%	52.4%
5%	3.9%	39.8%

In this case where mixing is across utility functions of different variables, the test is moderately powerful. It remains the case in this example that the estimated

[17] The standard error for the actual significance level is 0.0095 for a test at the nominal 10 percent level, and 0.0069 for a test at the nominal 5 percent level.

coefficients in the MNL model without mixing are close to their expected values.

Testing the Adequacy of a Mixing Distribution. Suppose one has estimated a MMNL model in which the MNL parameters are mixed according to a base density *g*, and the object is to test whether *additional* mixing is needed to describe the sample. The choice probability under the alternative is

$$P_C(i) = \int\{\int L_C(i;\beta + \Gamma\eta + \lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}\cdot h(\nu)d\nu \ ,$$

where β is a k×1 vector, η is q×1 with mean zero and covariance matrix I_q, Γ is a k×q factor loading matrix, ν is k×1 with mean zero, unit variances, and a covariance matrix Ω, λ is a k×1 vector of variances, k-r of which are maintained at zero, $\lambda^{1/2}$ denotes the component-wise square root, and $\odot$ denotes component by component direct product. The null hypothesis is that the data are generated by this model with $\lambda = 0$; i.e., a mixed MNL model with q latent factors determining the choice probabilities, versus the alternative that up to r additional factors, with density $h(\cdot)$, are needed.

Recall that $\log L_C(i|\gamma) = x_i'\gamma - \log \sum_{j\in C} \exp(x_j'\gamma)$, $\nabla_\gamma \log L_C(i|\gamma) = x_i - x_C$, where $x_C = \sum_{j\in C} x_j\cdot L_C(j|\gamma)$, and $\nabla_{\gamma\gamma}\log L_C(i|\gamma) = -\sum_{j\in C}(x_j - x_C)(x_j - x_C)'L_C(i|\gamma)$. Differentiating $P_C(i)$,

$$\nabla_\beta P_C(i) = \int\{\int(x_i\text{-}x_C)L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}h(\nu)d\nu \ ,$$

$$\nabla_\Gamma P_C(i) = \int\{\int(x_i\text{-}x_C)\eta' L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}h(\nu)d\nu \ ,$$

$$\nabla_{\lambda_m} P_C(i) = 0.5\lambda_m^{-1/2}\int\{\int(x_{mi}\text{-}x_{mC})L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}\nu_m h(\nu)d\nu \ .$$

To evaluate the last derivative under the null, use L'Hopital's rule. Note that the derivative of $2\lambda_m^{1/2}\nabla_{\lambda_m}P_C(i)$ with respect to λ_m is

$$0.5\lambda_m^{-1/2}\int\{\int(x_{mi}-x_{mC})^2 L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}\nu_m^2 h(\nu)d\nu$$

$$-\ 0.5\lambda_m^{-1/2}\int\{\int\sum_{j\in C} x_m(x_{mj}-x_{mC})L_C(j;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}\nu_m^2 h(\nu)d\nu$$

$$= 0.5\lambda_m^{-1/2}\int\{\int(z_{mi}-z_{mC})^2 L_C(i;\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)\cdot g(\eta)d\eta\}\nu_m^2 h(\nu)d\nu\ ,$$

where $z_{mi} = (x_{mi}-x_{mC})^2$ and $z_{mC} = \sum_{j\in C} z_{mj}\cdot L_C(j|\beta+\Gamma\eta+\lambda^{1/2}\odot\nu)$.

Hence, at $\delta = 0$,

$$x_C(\eta) = \sum_{j\in C} x_j\cdot L_C(j|\beta+\Gamma\eta)$$

$$z_{mi}(\eta) = (x_{mi}-x_{mC}(\eta))^2/2$$

$$z_{mC}(\eta) = \sum_{j\in C} z_{mj}(\eta)\cdot L_C(j|\beta+\Gamma\eta)$$

$$P_C(i) = \int L_C(i;\beta+\Gamma\eta)\cdot g(\eta)d\eta\ ,$$

$$\nabla_\beta P_C(i) = \int(x_i-x_C(\eta))L_C(i;\beta+\Gamma\eta)\cdot g(\eta)d\eta\ ,$$

$$\nabla_\Gamma P_C(i) = \int(x_i-x_C(\eta))\eta' L_C(i;\beta+\Gamma\eta)\cdot g(\eta)d\eta\ ,$$

$$\nabla_{\lambda_m} P_C(i) = \int(z_{mi}(\eta)-z_{mC}(\eta))L_C(i;\beta+\Gamma\eta)\cdot g(\eta)d\eta\ .$$

For comparison, suppose one had the base model in variables x and wanted to test whether additional variables z_{mi} that interact with individual characteristic η belong in the model. In this case, the model under the alternative is $P_C(i) = \int\tilde{L}_C(i;\beta + \Gamma\eta,\theta)\cdot g(\eta)d\eta$, where

$$\log\ \tilde{L}_C(i;\beta + \Gamma\eta,\theta) = x_i'(\beta + \Gamma\eta) + z_i(\eta)'\theta - \log \sum_{j\in C}\exp\{x_j'(\beta + \Gamma\eta) + z_j(\eta)'\theta\}\ .$$

The derivatives under the null hypothesis $\theta = 0$ are the same as before for $\nabla_\beta P_C(i)$ and for $\nabla_\Gamma P_C(i)$. Finally,

$$\nabla_\theta P_C(i) = \int (z_{mi}(\eta) - z_{mC}(\eta)) L_C(i;\beta+\Gamma\eta)\cdot g(\eta)d\eta \ ,$$

also as before. Therefore, a LM test for the hypothesis $\lambda = 0$ is equivalent to a LM test for the hypothesis $\theta = 0$ for the auxiliary variables $z_i(\eta)$. This test is most readily computed by first estimating the base model, using say a simulation procedure with specified starting seeds, then regressing (over observations) the constant one on the scores $\nabla_\beta \log P_C(i)$, $\nabla_\Gamma \log P_C(i)$, and $\nabla_{\lambda_m} \log P_C(i)$ for m = 1,...r, and testing whether the sum of squared residuals is significant according to a chi-square distribution with r degrees of freedom. This testing procedure is summarized in the following theorem:

Theorem 6.4. *Suppose the base model* $P_C(i) = \int L_C(i;\beta + \Gamma\eta)\cdot g(\eta)d\eta$ *has been estimated by MSLE, using Monte Carlo draws* η^{kn} *from* $g(\cdot)$ *for* $k = 1,...,r_N$ *and* $n = 1,...,N$. *Construct the variables*

$$L_{Cn}(i|\beta+\Gamma\eta^{kn}) = \exp(x_{in}(\beta+\Gamma\eta^{kn}))/\sum_{j\in C}\exp(x_{jn}(\beta+\Gamma\eta^{kn}))$$

$$x^k_{Cn} = \sum_{j\in C} x_{jn}\cdot L_{Cn}(j|\beta+\Gamma\eta^{kn}) \ ,$$

$$z^k_{min} = (x_{min} - x^k_{mCn})^2/2 \ ,$$

$$z^k_{mCn} = \sum_{j\in C} z^k_{mjn}\cdot L_{Cn}(j|\beta+\Gamma\eta^{kn}) \ ,$$

$$v_{in} = r_N^{-1} P_C(i)^{-1}\cdot \sum_{k=1}^{r_N} (x_{in} - x^k_{Cn})\cdot L_{Cn}(i;\beta+\Gamma\eta^{kn}) \ ,$$

$$w_{in} = r_N^{-1} P_C(i)^{-1}\cdot \sum_{k=1}^{r_N} (x_{in} - x^k_{Cn})\cdot(\eta^{kn})' L_{Cn}(i;\beta+\Gamma\eta^{kn}) \ ,$$

$$y_{min} = r_N^{-1} P_C(i)^{-1} \cdot \sum_{k=1}^{r_N} (z_{min} - z_{mCn}^{k}) \cdot L_{Cn}(i;\beta+\Gamma\eta^{kn}) \ ,$$

where all parameters are set to the base model estimates. A regression of one on the variables v_{in}, w_{in}, y_{min} *for* $m = 1,...,r$ *and a sum-of-squared-residuals test for the significance of variables in this regression is asymptotically equivalent to a Lagrange Multiplier test for additional mixing of dimension r in the coefficients* $m = 1,...,r$ *of* x_{in}.

In light of the Monte Carlo results in the base case of no mixing, one can expect this test to have relatively low power. Hence, for use as a diagnostic for model specification, one will want to err on the side of admitting too much potential heterogeneity, and use a rejection region with a large nominal significance level.

<u>An Application: Demand for Alternative Vehicles</u>

The State of California suffers from air pollution generated by conventional gasoline-powered vehicles, and the State is in the process of mandating quotas for alternative-fueled vehicles: methanol, compressed natural gas (CNG), or electric. An important policy question is consumer acceptance of these alternative vehicles, and the extent to which subsidies will be necessary to stimulate consumer demand to the levels required by the quotas. Brownstone *et al* (1996) have carried out a conjoint analysis study of preferences between alternative vehicles. The study has 4654 respondents, each of whom was asked to choose among six alternatives. The alternatives were described in terms of the variables defined in Table 6.1. An experimental design was used to select the offerings of six alternatives from 120 possible profiles, distinguished by four fuels (gasoline, methanol, CNG, electric), five sizes (mini, subcompact, compact, midsize, large), and six body types (regular car, sports car, truck, van, station wagon, sports utility vehicle).

Table 6.2 gives a MMNL model estimated by Brownstone and Train (1996). This model includes four random effects, associated with the following variables: Dummy for non-EV, Dummy for non-CNG, Size, and Luggage Space. The panel of the table headed "Variables" gives estimates of the β parameters, and the panel headed "Random Effects" gives the factor loading Λ on standard normal factors, with an independent factor for each of the random effects above. Then, the coefficients are estimates of the standard deviations of these random effects. The estimation uses 250 replications per observation, and MSLE. The parameter estimates show strong random effects, with magnitudes large enough to suggest that they are capturing correlation structure in unobservables in addition to variation in tastes. The variables and random effects included in this model are the result of an ad hoc model selection procedure. A likelihood ratio test at the five percent level shows that this model fits significantly better than a simple MNL model (given in Table 6.3). The table gives estimates of the standard errors of the coefficients for 250 replications, and also for 50 replications. The columns headed "Asymptotic" use the result that for the number of repetitions increasing more rapidly than $\sqrt{N}$, the asymptotic covariance matrix is the inverse of the information matrix when there is no simulation. This matrix is estimated by the outer product of the score; this overestimates the information in the sample because of the presence of non-negligible simulation noise in a finite sample. The columns headed "Robust" use the formula $\Gamma^{-1}\Omega\Gamma^{-1}$, where Ω is estimated by the outer product of the gradient and Γ is estimated by the hessian, with each of these formed using sample averages at the estimated coefficients. This formula coincides with the usual estimator of covariances in GMM estimation. It is less likely than the asymptotic formula to understate standard errors. The estimates show that the robust standard errors fall with number of repetitions, as expected. By contrast, the asymptotic standard errors rise with number of replications; this indicates the degree to which their failure to handle simulation noise properly biases the results. In general, using the asymptotic covariance formula with 250 replications results in a ten to twenty percent underestimate of standard errors of coefficients.

Table 6.3, Model 1, is a simple MNL model $L_C(i|\beta)$ fitted to the data; these estimates are taken from Brownstone and Train (1996). Model 2 adds the artificial variables defined in Theorem 6.2; i.e., given the base model $L_{Cn}(i;\beta) = e^{x_{in}\beta}/\sum_{j\in C} e^{x_{jn}\beta}$ and $x_{Cn} = \Sigma_{j\in C} x_{jn}\cdot L_{Cn}(j;\beta)$, with β set equal to its MNL estimator, define the artificial variables $z_{min} = (x_{min}-x_{mCn})^2$ for variables m where heterogeneity is suspected, and estimate the MNL model with the original x variables and the additional artificial variables. The list of artificial variables may include variables m which have the coefficient β_m constrained to zero in the base MNL model; these are interpreted as pure random effects. A likelihood ratio test at the five percent significance level rejects the null hypothesis of no mixing. The individual T-statistics for the artificial variables are not necessarily a reliable guide to the location of significant mixing, due to lack of independence, and due to the possibility of correlation across alternatives in unobserved attributes. However, the results (based on T-statistics exceeding one in magnitude) suggest that there may be taste variation in the following variable coefficients: Non-EV, Non-CNG, Size, Luggage space, Operating Cost, and Station Availability. The first four of these were included in the Brownstone-Train model in Table 6.2; the last two are additional factors where mixing may be present.

Table 6.4 gives a MMNL model which includes the six random effects identified as possibly significant by the artificial variable test in Table 6.3, using T-statistics greater than one in magnitude as the selection criterion. The MMNL estimates show that there is significant mixing in each of these factors. Likelihood ratio tests show that this model is a significant improvement on the model in Table 6.2. Further exploration with additional factors in the MMNL model finds that there are several factor combinations that will fit as well or marginally better than the model in Table 6.4, and that some of these combinations will place weight on factors that were excluded by the artificial variable selection procedure, and will lower the significance of some of the factors previously included. These results reflect in part the relatively poor identification of factor structure from observed data on covariances, but may also indicate that other specification issues, such as true omitted variables, need to be addressed.

Table 6.1. Variable Definitions

Variable	Definition
Price/ln(income)	Purchase price (in thousands of dollars) divided by log household income (in thousands)
Range	Hundreds of miles that the vehicle can travel between refuelings/rechargings
Acceleration	Tens of seconds required to reach 30 mph from stop
Top speed	Highest attainable speed in hundreds of MPH
Pollution	Tailpipe emissions as fraction of those for new gas vehicle
Size	0 = mini, 0.1 = subcompact, 0.2 = compact, 0.3 = mid-size or large
"Big enough"	1 if household size $\geq$ 2 and vehicle is mid or large
Luggage space	Fraction of luggage space in comparable new gas vehicle
Operating cost	Cost per mile of travel (tens of cents): home recharging for electric vehicle, station refueling otherwise
Station availability	Fraction of stations that can refuel/recharge vehicle
Sports utility vehicle	1 for sports utility vehicle, 0 otherwise
Sports car	1 for sports car, 0 otherwise
Station wagon	1 for station wagon, 0 otherwise
Truck	1 for truck, 0 otherwise
Van	1 for van, 0 otherwise
Dummy for EV	1 if electric vehicle (EV)
Commute < 5 & EV	1 if electric vehicle and commute < 5 miles/day
College & EV	1 if electric vehicle and some college education
Dummy for CNG	1 if compressed natural gas (CNG) vehicle
Dummy for methanol	1 if methanol vehicle
College & methanol	1 if methanol vehicle and some college education
Non-EV dummy	1 if not electric vehicle
Non-CNG dummy	1 if not compressed natural gas vehicle

Table 6.2: Mixed Logit for Alternative-Fueled Vehicle Choice
(Brownstone & Train, 1996)

	Parameter Estimates	Standard Error 250 replic. Asymp	Robust	Standard Error 50 replic. Asymp	Robust
Variables					
Price/ln(income)	-0.264	0.0435	0.0452	0.0412	0.0525
Range	0.517	0.0581	0.0685	0.0511	0.1022
Acceleration	-1.062	0.1859	0.1990	0.1738	0.2519
Top speed	0.307	0.1150	0.1184	0.1131	0.1188
Pollution	-0.608	0.1392	0.1420	0.1357	0.1546
Size	1.435	0.5082	0.4991	0.4945	0.5156
"Big Enough"	0.224	0.1126	0.1166	0.1113	0.1220
Luggage Space	1.702	0.4822	0.5854	0.4314	0.8971
Operating Cost	-1.224	0.1593	0.2069	0.1393	0.2998
Station availability	0.615	0.1452	0.1536	0.1410	0.1757
Sports utility vehicle	0.901	0.1484	0.1486	0.1482	0.1493
Sports car	0.700	0.1625	0.1513	0.1626	0.1518
Station wagon	-1.500	0.0674	0.0645	0.0674	0.0659
Truck	-1.086	0.0556	0.0520	0.0555	0.0556
Van	-0.816	0.0558	0.0468	0.0557	0.0471
Dummy for EV	-1.032	0.4249	0.5022	0.3777	0.6035
Commute < 5 & EV	0.372	0.1660	0.1763	0.1608	0.1927
College & EV	0.766	0.2182	0.2374	0.2073	0.2796
Dummy for CNG	0.626	0.1482	0.1670	0.1391	0.2139
Dummy for methanol	0.415	0.1464	0.1474	0.1440	0.1534
College & methanol	0.313	0.1243	0.1256	0.1223	0.1308
Error Components					
Non-EV	2.464	0.5414	0.7184	0.4428	1.0252
Non-CNG	1.072	0.3773	0.4109	0.2781	0.5711
Size	7.455	1.8194	2.0408	1.5538	2.4734
Luggage Space	5.994	1.2483	1.6617	1.0483	2.7719
Log Likelihood	-7375.34				

Table 6.3. Multinomial Logit Model

	Model 1 Parameter Estimates	Model 1 SE	Model 2 Parameter Estimates	Model 2 SE
Variables				
Price/ln(income)	-0.185	0.027	-0.4240	0.0298
Range	0.350	0.027	0.5036	0.0447
Acceleration	-0.716	0.111	-0.9771	0.1263
Top speed	0.261	0.080	0.3592	0.0814
Pollution	-0.444	0.100	-0.6567	0.1161
Size	0.935	0.311	1.4179	0.3430
"Big Enough"	0.143	0.076	0.2248	0.0845
Luggage Space	0.501	0.188	1.0161	0.2574
Operating Cost	-0.768	0.073	-1.1447	0.0897
Station availability	0.413	0.097	0.6350	0.1074
Sports utility vehicle	0.820	0.144	0.8806	0.1458
Sports car	0.637	0.156	0.6869	0.1580
Station wagon	-1.437	0.065	-1.5229	0.0663
Truck	-1.017	0.055	-1.0776	0.0551
Van	-0.799	0.053	-0.8272	0.0542
Dummy for EV	-0.179	0.169	-0.6979	0.2384
Commute < 5 & EV	0.198	0.082	0.3102	0.0840
College & EV	0.443	0.108	0.6863	0.1145
Dummy for CNG	0.345	0.091	0.4216	0.1056
Dummy for methanol	0.313	0.103	0.4886	0.1105
College & methanol	0.228	0.089	0.3070	0.0903
Artificial Variables				
Price/ln(income)			0.0019	0.0927
Range			-0.0349	0.0551
Acceleration			-1.3728	2.1388
Top speed			-0.2071	0.6383
Pollution			0.0977	0.6764
Size			21.5773*	9.5000
"Big enough"			0.2837	0.3832
Luggage space			3.8731*	3.4638
Operating cost			4.2245*	0.8369
Station availability			0.6741*	0.3781
Dummy for EV			2.3476*	0.5704
Dummy for CNG			1.2364*	0.4798
Log Likelihood	-7391.83		-7356.61	

Notes: Model 1 is from Brownstone & Train (1996)
* denotes the artificial variables with $|T| > 1$

Table 6.4. Mixed Multinomial Logit Model

	Parameter Estimates	SE
Variables		
Price/ln(income)	-0.3622	0.0669
Range	0.6753	0.0965
Acceleration	-1.2688	0.2591
Top speed	0.4027	0.1553
Pollution	-0.7929	0.1980
Size	1.7351	0.6694
"Big Enough"	0.2695	0.1468
Luggage Space	2.2631	0.6426
Operating Cost	-1.8056	0.2912
Station availability	0.7029	0.1896
Sports utility vehicle	0.9234	0.1498
Sports car	0.7270	0.1645
Station wagon	-1.5246	0.0681
Truck	-1.1195	0.0559
Van	-0.8191	0.0564
Dummy for EV	-1.5733	0.5819
Commute < 5 & EV	0.4793	0.2242
College & EV	1.0534	0.3114
Dummy for CNG	0.7709	0.2018
Dummy for methanol	0.5435	0.1922
College & methanol	0.3849	0.1542
Error Components		
Non-EV	3.3802	0.7647
Non-CNG	1.1042	0.4990
Size	8.0788	2.7021
Luggage space	7.6220	1.7153
Operating cost	4.4532	0.8014
Station availability	1.3987	0.5730
Log Likelihood	-7358.93	

REFERENCES

Amemiya, Takashi (1985) *Advanced Econometrics*, Cambridge: Harvard University Press.

Becker, Walter; Eymann, Angelika; McFadden, Daniel (1994) "Multinomial Approximation and Sequential Simulation," University of California Working Paper, Berkeley.

Beggs, John (1988) "A Simple Model for Heterogeneity in Binary Logit Models," *Economics Letters*; 27(3),245 49.

Ben-Akiva; Moshe; Bolduc, D. (1996) "Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure," MIT Working Paper.

Billingsley, Patrick (1986) *Probability and Measure*. New York: Wiley.

Borsch Supan, Axel (1990) "Recent Developments in Flexible Discrete Choice Models: Nested Logit Analysis versus Simulated Moments Probit Analysis," in Fischer, M. et al, eds. *Spatial choices and processes. Studies in Regional Science and Urban Economics*, 21, Amsterdam: North Holland, 203-17.

Brownstone, David; Bunch, D.; Golob, T.; Ren, W. (1996) "Transactions Choice Model for Forecasting Demand for Alternative-Fueled Vehicles," in S. McMullen, ed, *Research in Transportation Economics*, 4, 87-129.

Brownstone, David; Train, Kenneth (1996) "Forecasting New Product Penetration with flexible Substitution Patterns," University of California Working Paper, Berkeley.

Cencov, N.N. (1982) Statistical Decision Rules and Optimal Inference, Translations of Mathematical Monographs, Vol. 53, American Mathematical Society.

Chan, K. (1993) "On the central limit theorem for an ergodic Markov chain," *Stochastic Processes and their Applications* 47, 113-117.

Chavas, Jean; Segerson, Kathleen (1986) "Singularity and Auotregressive Disturbances in Linear Logit Models," *Journal of Business and Economic Statistics*; 4(2), 161-69.

Chesher, Andrew; Santos Silva, Joao (1995) "Taste Variation in Discrete Choice Models," University of Bristol Working Paper

Chib, S.; Greenberg, E. (1995) "Markov Chain Monte Carlo simulation methods in econometrics," *Econometric Theory*, forthcoming.

Chintagunta, Pradeep (1992) "Estimating a Multinomial Probit Model of Brand Choice Using the Method of Simulated Moments," *Marketing Science*; 11(4), 386-407.

Devroye, L. (1986) *Non-Uniform Random Variate Generation* Springer, New York.

Dubin, Jeffrey; Zeng, Langche (1991) "The Heterogeneous Logit," Model Caltech Social Science Working Paper: 759.

Enberg, John; Gottschalk, Peter; Wolf, Douglas (1990) "A Random Effects Logit Model of Work Welfare Transitions," *Journal of Econometrics*; 43(1 2),63-75.

Follman, David; Lambert, D. (1989) "Generalized Logistic Regression by Nonparametric Mixing," *Journal of the American Statistical Association*, 84, 295-300.

Formann, A. (1992) "Linear Logistic Latent Class Analysis for Polytomous Data," *Journal of the American Statistical Association*, 87, 476-486.

Fuss; Mel; McFadden, Daniel; Mundlak, Yair (1978) "A Survey of Functional Forms in the Economic Analysis of Production," in M. Fuss and D. McFadden, eds, *Production Economics: A Dual Approach to Theory and Applications*, Amsterdam: North Holland, 219-68.

Geweke, John (1989) "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica* 57, 1317-39.

Gilks, W; Richardson, S.; Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.

Gonul, Fusun; Srinivasan, Kannan (1993) "Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models," *Marketing Science*; 12(3), 213-29.

Gourieroux, Christian; Monfort, Alain (1992) "Simulation Based Inference in Models with Heterogeneity," *Annales d'Economie et de Statistique*; 0(20-21), 69-107.

Gourieroux, Christian; Monfort, A.; Renault, E. (1993) "Indirect Inference," *Journal of Applied Econometrics*; 8(0), S85-118.

Gourieroux, Christian; Monfort, Alain (1993) "Simulation-Based Inference: A Survey with Special Reference to Panel Data Models," *Journal of Econometrics*; 59(1-2), 5-33.

Gourieroux, Christian; Monfort, Alain (1994) "Testing Non-Nested Hypotheses," in R. Engle and D. McFadden, eds., *Handbook of Econometrics*, IV, 2585-2637.

Gourieroux, Christian; Monfort, Alain (1996) *Simulation Based Methods in Econometrics*. Oxford University Press: Oxford.

Hajivassiliou, Vassilis; Ruud, Paul (1994) "Classical Estimation Methods for LDV Models using Simulation," in R. Engle and D. McFadden, eds *Handbook of Econometrics* IV, 2384-2441.

Hajivassiliou, Vassilis; McFadden, Daniel; Ruud, Paul (1996) "Simulation of Multivariate Normal Rectangle Probabilities and their Derivatives," *Journal of Econometrics*; 72, 85-134.

Hall, Peter; Heyde, C. (1980) *Martingale Limit Theory and Its Applications*, Academic Press.

Hendry, David (1984) "Monte Carlo experimentation in econometrics" in Z. Griliches and M. Intriligator (ed) *Handbook of Econometrics*, v.2, North Holland, Amsterdam, 937-976.

Hendry, David (1988) "Encompassing," *National Institute Economic Review*; 0(125), 88-92.

Huber, Peter (1967) "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in L. LeCam and J. Neyman *Proceedings of the Fifth Berkeley Symposium on Statistics and Probability*, vol 1, Berkeley: University of California Press, 129-156.

Ibragimov, I.; Has'minskii, R. (1981) *Statistical Estimation*, Springer-Verlag: New York.

Jain, Dipak C.; Vilcassim, Naufel; Chintagunta, Pradeep (1994) "A Random Coefficients Logit Brand Choice Model Applied to Panel Data," *Journal of Business and Economic Statistics*; 12(3), 317-28.

Keane, Michael (1994) "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62, 95-116.

Keane, Michael; Wolpin, Kenneth (1994) "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *Review of Economics and Statistics*; 76(4), 648-72.

Lee, L.F.; Cheshire, A. (1986) "Specification Testing when Score Test Statistics are Identically Zero," *Journal of Econometrics*, 31, 121-49.

Matzkin, Rosa (1994) "Restrictions of Economic Theory in Econometric Methods," in R. Engle and D. McFadden, eds, *Handbook of Econometrics*, IV, Amsterdam: North Holland, 2524-2558.

McFadden, Daniel (1984) "Econometric Analysis of Qualitative Response Models," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, II, North Holland: Amsterdam.

McFadden, Daniel (1987) "Regression-Based Specification Tests for the Multinomial Logit Model," *Journal of Econometrics*, 34, 63-82.

McFadden, Daniel (1989) "A Method of Simulated Moments for Estimation of Discrete Choice Models without Numerical Integration", *Econometrica*, 57, 995-1026.

McFadden, Daniel; Ruud, Paul (1994) "Estimation by Simulation," *Review of Economics and Statistics*; 76(4), 591-608.

McFadden, Daniel (1996) "Willingness to Pay for Natural Resources," University of California Working Paper, Berkeley.

McFadden, Daniel; Train, Kenneth (1996) "Mixed Multinomial Logit Models for Discrete Response," University of California Working Paper, Berkeley.

Mizon, Grayham; Richard, Jean (1986) "The Encompassing Principle and Its Application to Testing Non-nested Hypotheses," *Econometrica*; 54(3), 657-78.

Mizon, Grayham (1984) "The Encompassing Approach in Econometrics," in Hendry, David F., ed.; Wallis, Kenneth F., ed. *Econometrics and Quantitative Economics*. Oxford and

New York: Blackwell, 135-72.

Montgomery, Mark; Richards, Toni; Braun, Henry (1986) "Child Health, Breast Feeding, and Survival in Malaysia: A Random Effects Logit Approach," *Journal of the American Statistical Association*; 81(394), 297-309.

Newey, Whitney; McFadden, Daniel (1994) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden, eds., *Handbook of Econometrics*, IV, North Holland: Amsterdam, 2111-2245.

Neveu, Jacques (1965) *Mathematical Foundations of the Calculus of Probability*, Holden-Day: San Francisco.

Nummelin, E. (1984) *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press: Cambridge.

Orey, S. (1971) *Limit Theory for Markov Chain Transition Probabilities*, van Nostrand: New York.

Pakes, Ariel; Pollard, David (1989) "The Asymptotic Distribution of Simulation Experiments", *Econometrica*, 57, 1027-1057.

Pollard, David (1991) "Bracketing Methods in Statistics and Econometrics," in W. Barnett et al, eds, *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.

Press, William *et al* (1986) *Numerical Recipes*; Cambridge University Press: Cambridge.

Reader, S. (1993) "Unobserved Heterogeneity in Dynamic Discrete Choice Models," *Environment and Planning* A; 25(4), 495-519.

Revelt, David; Train, Kenneth (1996) "Incentives for Appliance Efficiency," University of California Working Paper, Berkeley.

Robert, C. (1996) *Methodes de Monte Carlo par Chaines de Markov*. Economica: Paris.

Roberts, G. (1992) "Convergence diagnostics of the Gibbs sampler," in J. Bernardo *et al* (eds) *Bayesian Statistics* 4, 775-782. Oxford University Press: Oxford.

Roberts, G.; Smith, A. (1994) "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms," *Stochastic Processes and their Applications* **49**, 207-216.

Rubinstein, Reuven (1981) *Simulation and the Monte Carlo Method*, Wiley: New York.

Ryu, Keunkwan (1993) "Monotonicity of the Fisher Information and the Kullback-Leibler Divergence Measure," *Economics Letters*; 42(2-3), 121-28.

Steckel, Joel; Vanhonacker, Wilfried (1988) "A Heterogeneous Conditional Logit Model of Choice," *Journal of Business and Economic Statistics*; 6(3), July 1988, pages 391 98.

Stern, Steven (1994) "Two Dynamic Discrete Choice Estimation Problems and Simulation Method Solutions," *Review of Economics and Statistics*; 76(4), 695-702.

Stout, W. (1974) *Almost Sure Convergence*. New York: Wiley.

Talvitie, Antti (1972) "Comparison of Probabilistic Modal-Choice Models," *Highway Research Record*, 392, 111-120.

Thisted, Ronald (1988) *Elements of Statistical Computing*; Chapman and Hall: New York.

Tierney, L. (1994) "Markov chains for exploring posterior distributions," *Annals of Statistics* **22**, 1701-1762.

Train, Kenneth (1996) "Unobserved Taste Variation in Recreation Demand Models," University of California Working Paper, Berkeley.

Van Praag, B.; Hopf, A. (1987) "Estimation of Continuous Models on the Basis of Set Valued Observations", Working Paper.

Westin, R. (1974) "Predictions from binary choice models," *Journal of Econometrics* **2**, 1-16.

Westin, Richard; Gillen, David (1978) "Parking Location and Transit Demand: A Case Study of Endogenous Attributes in Disaggregate Mode Choice Models," *Journal of Econometrics*; 8(1), 75-101.

White, Halbert (1982) "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*; 50(1), 1-25.

Zellner, Arnold; Min, C. (1995) "Gibbs Sampler Convergence Criteria," *Journal of the American Statistical Association* **90**, 921-927.