Simulation-Based Estimation

Steven Stern*
University of Virginia

March 1997

Abstract

In this paper, I present a number of leading examples in the empirical literature that use simulation-based estimation methods. For each example, I describe the model, why simulation is needed, and how to simulate the relevant object. There is a section on simulation methods and another on simulation-based estimation methods. The paper concludes by considering the significance of each of the examples discussed and commenting on potential future areas of interest.

1. Introduction

Almost all empirical problems in economics involve maximizing a likelihood function or solving a set of moment conditions, and most derivatives of log likelihood functions can be interpreted as a set of moment conditions. There are many problems, especially those with rich specifications of the model, where the relevant moment conditions can not be evaluated analytically or numerically. Up until recently, this problem forced empirical economists to restrict the set of problems they worked on to those simple enough to evaluate analytically or numerically. But, over the last few years, there has been an explosion of empirical work using simulation methods. Simulation allows us to estimate otherwise intractable models by approximating high dimensional integrals. The generic problem for simulation is to evaluate

$$Eh\left(U\right)=\int h\left(u\right)f\left(u\right)du \tag{1.1}$$

where U is a random variable with density $f\left(u\right)$ and $h\left(u\right)$ is a function implied by theory. The basic idea of simulation is to draw random variables from $f\left(\bullet\right)$ and use them to compute a sample mean of $h\left(U\right)$. In particular, if u^r, $r=1,2,...,R$ are R draws from $f\left(\bullet\right)$, then

$$\frac{1}{R}\sum_{r=1}^{R}h\left(u^r\right) \tag{1.2}$$

I would like to thank Mike Brien and especially Frank Wolak for helpful comments and Gayla Gentry for excellent typing. All remaining errors are mine.

is an unbiased estimator of $Eh(U)$ with variance that $\rightarrow 0$ as $R \rightarrow \infty$.

This paper describes a leading set of empirical research papers that use simulation. For each paper, I present a stylized version of the model, discuss why simulation methods are needed, and later the significance of the paper and value of using simulation methods in the paper. There is also a section showing the reader how to use simulation methods. By the end of the paper, the reader should understand the basics of simulation, should have a sense of where to go for more information on specific topics, and should have an appreciation for its value in empirical work.

There are many econometric problems that involve evaluating high dimensional integrals. Consider a model of occupational choice where there are a number of choices available to each individual. This is a special case of the qualitative response models described in, for example, Daniel McFadden (1984). The value of any particular choice will depend upon a set of observed variables (such as sex, race, age, and education) and an error representing the effect of unobserved variables (such as preferences for particular activities and unobserved ability). We assume the individual chooses the occupation with the greatest value. The goal is to estimate how the observed variables affect the value of each occupation. We can do this by observing the choice each individual makes and his observed characteristics. But, because the errors for each occupation are unobserved, our model does not imply particular occupations being chosen; rather, it implies only probabilities of each occupation being chosen. A maximum likelihood estimator of the observed effects would maximize the product of probabilities of the chosen outcomes (assuming each individual was making an independent choice). In order to maximize this product, we need to evaluate the probabilities of the chosen outcomes.

If we assume that the distribution of the errors is independently and identically distributed (*iid*) Extreme Value, the probabilities have multinomial logit functional forms that can be evaluated easily. But the *iid* Extreme value assumption is very restrictive in that it implies that probabilities are not sensitive to irrelevant alternatives. For example, consider a model where two of the choices are policeman and economist and compare it to a model where a third alternative is fireman. The *iid* Extreme value assumption implies that $\Pr[\text{economist}] / \Pr[\text{policeman}]$ is independent of whether being a fireman is also an option. Assuming that being a policeman and fireman are close substitutes (relative to being an economist), one would expect that people who become firemen would have become policemen had being a fireman not been an option; such an assumption requires that the errors associated with being a fireman and policeman are correlated. This implies that the ratio of probabilities would increase when fireman is added because new firemen were policemen before fireman was an option. The lack of change in the ratio is called the independence of irrelevant alternatives problem. One could generalize the *iid* Extreme value assumption to the Generalized Extreme value assumption. This would lead to nested logit probabilities. While more general than the multinomial logit probabilities and not suffering from the independence of irrelevant alternatives problem, the nested logit probabilities are still somewhat restrictive in that they force us to make somewhat arbitrary assumptions about nesting of choices. For example, we might nest occupations according to SIC occupational codes. But instead we might nest them in terms of some other inherent characteristic such as required education or riskiness. Unfortunately, for nested logit, we must decide a priori what nesting structure to use. Also, there are some models where choices do not fit neatly into a nesting structure.

A very flexible functional form assumption for the errors is that they are distributed multivariate normal. But this assumption implies that the choice probabilities are high dimensional integrals with no good numerical approximation except in special cases (Jerry Hausman and David Wise 1978; and J. S. Butler and Robert Moffitt 1982). Simulation provides a feasible way to approximate the choice probability. This approximate choice probability either can be used in a moment condition to get a method of moments estimator of the observed variables' effect, or it can be used in a likelihood function to get a maximum likelihood estimator of the effect. The details of this are discussed below.

Consider a different example where an individual is solving a stochastic, dynamic programming problem. Again, we want to estimate how various observed characteristics affect the choices the individual makes over time. Except under restrictive conditions described by John Rust (1988), part of the process in estimating these effects will involve evaluating high dimensional integrals; usually the dimension of the integral will be the length of the individual's planning horizon. Simulation provides a way to approximate those integrals. For example, consider pricing a European option. There is a stock whose price is changing over time according to some stochastic process such that the joint distribution of stock prices over $t = 1, 2, ..., T$ is F. Denote the marginal distribution of the price at a particular time t as $F_t(p_t)$. Assume it is very difficult to evaluate F or F_t, but it is easy to simulate from the stochastic process associated with F. The European option allows one to purchase the stock at a specified price $\bar{p}_T$ at time T. A straightforward way to evaluate the value of the option at time 0 is to simulate the stochastic process $\{p_t\}_{t=1}^{T}$ and compute $\beta^T (p_T - \bar{p}_T)$ where β is a one-period discount factor . This process can be repeated many times to derive an unbiased, precise estimate of $E\beta^T (p_T - \bar{p}_T)$, the price of the option. We could complicate the problem by changing the option into an American option where one can purchase the stock at the agreed price $\bar{p}$ any time between T_1 and T_2.[1] Now the optimal stategy is a reservation function $\tilde{p}_t$ for $t = T_1, T_1 + 1, ..., T_2$ such that one purchases the stock at price $\bar{p}$ the first period when $p_t > \tilde{p}_t$. The goal is to optimize $E\beta^T (p_T - \bar{p}_T)$over $\{\tilde{p}_t\}_{t=T_1}^{T_2}$ where T is a random variable denoting the time that the option is exercised. One can simulate $E\beta^T (p_T - \bar{p}_T)$ for any guess of $\{\tilde{p}_t\}_{t=T_1}^{T_2}$ and therefore maximize $E\beta^T (p_T - \bar{p}_T)$ for any guess of $\{\tilde{p}_t\}_{t=T_1}^{T_2}$.

Section 2 considers some stylized models resembling more complicated, influential models appearing in the literature. For each example, I show why the model is difficult to estimate because of an integration problem similar to equation (1.1) and how each can be estimated with simulation. Section 3 develops the simulation methods used in these papers. First, I discuss the basics of simulating from specified distributions so one can draw u^r in equation (1.2) for a large class of densities $f(\bullet)$ in equation (1.1); this is the first step in being able

[1] I am abstracting away from other important problems such as how to choose a discount factor, variable discount factors, and dividend processes.

to simulate equation (1.1) using equation (1.2). Then, I show how to simulate integrals like equation (1.1) using simulators like equation (1.2) and better simulators that minimize variance and provide smoothness. Finally, I show how to use the simulator of equation (1.1) in an estimation procedure. Essentially, one replaces the intractable moments in the estimator's optimality condition with their simulated counterparts. This section provides details on how to do this and the simulation-based estimator's asymptotic properties. Section 4 provides some more detail on some of the models in the literature and the significance of their results. Section 5 discusses areas of current research developing better simulation methods and using simulation methods.

2. Examples

2.1. Probit

The first example considered is the probit problem. Probit does not require simulation. But this example will help develop ideas for later examples. Consider a model of whether to own one's home. Let the value of owning a home y_i^* depend upon some observed characteristics (such as age, income, and family size) X_i, and an error u_i:

$$y_i^* = X_i\beta + u_i \tag{2.1}$$

where i indexes people in a sample, $i = 1, 2, ..., N$. Assume that we observe for each person, X_i and an indicator y_i of whether i owns; $y_i = 1$ if i owns, and $y_i = 0$ if i does not own. Further, assume that i owns if and only if the value of owning is positive; $y_i = 1$ if and only if $y_i^* > 0$. Then the probability of observing $y_i = 1$ conditional on X_i is

$$\begin{aligned} P_i &= \Pr\left[y_i = 1 \mid X_i\right] = \Pr\left[y_i^* > 0 \mid X_i\right] \\ &= \Pr\left[X_i\beta + u_i > 0\right] \\ &= \Pr\left[u_i > -X_i\beta\right] \\ &= 1 - F\left(-X_i\beta\right) \end{aligned} \tag{2.2}$$

where $F(\bullet)$ is the distribution function of u_i. If we assume that $u_i \sim iidN\left(0, \sigma^2\right)$, then

$$P_i = 1 - \Phi\left(\frac{-X_i\beta}{\sigma}\right) = \Phi\left(\frac{X_i\beta}{\sigma}\right) \tag{2.3}$$

where $\Phi(\bullet)$ is the standard normal distribution function. The last equality follows from the symmetry of the normal density function. Similarly, the probability that i does not own is

$$\Pr\left[y_i = 0 \mid X_i\right] = \Phi\left(\frac{-X_i\beta}{\sigma}\right) = 1 - \Phi\left(\frac{X_i\beta}{\sigma}\right). \tag{2.4}$$

Note that increases in β and σ such that β/σ remains constant has no effect on either probability. Thus, σ is not identified by the data and is usually set to one. Note that this normalizing procedure does not preclude estimating the effect of a change in X_i on P_i; P_i depends upon X_i only through β/σ. The maximum likelihood (ML) estimator of β maximizes the log likelihood function

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left[y_i \log P_i + (1 - y_i) \log (1 - P_i) \right]. \tag{2.5}$$

The derivative of the log likelihood contribution of observation i is

$$\begin{aligned} \frac{\partial L_i}{\partial \beta} &= y_i \frac{1}{P_i} \frac{\partial P_i}{\partial \beta} - (1 - y_i) \frac{1}{1 - P_i} \frac{\partial P_i}{\partial \beta} \\ &= \frac{1}{P_i} \frac{1}{1 - P_i} \left[y_i (1 - P_i) - (1 - y_i) P_i \right] \frac{\partial P_i}{\partial \beta} \\ &= \frac{1}{P_i} \frac{1}{1 - P_i} \left[y_i - P_i \right] \frac{\partial P_i}{\partial \beta}. \end{aligned} \tag{2.6}$$

More generally, equation (2.6) can be thought of as $Q_i^{'} [y_i - P_i]$ where Q_i is a matrix of instruments for observation i and $y_i - P_i = y_i - Ey_i$. The instruments $Q_i$must be exogenous: plim $\left[\frac{1}{N} \sum_{i=1}^{N} Q_i' (y_i - P_i) \right] = 0$. Assuming that X_i is fixed, both X_i and $\partial P_i / \partial \beta$ evaluated at some value of β (which is a function of X_i) are good instruments. Also, $\sum_i Q_i^{'} \frac{\partial P_i}{\partial \beta} / N$ must converge to a matrix of full rank (equal to the number of parameters being estimated). A method of moments (MOM) estimator of β solves

$$0 = \frac{1}{N} \sum_{i=1}^{N} Q_i^{'} [y_i - P_i]. \tag{2.7}$$

For a fixed Q_i, we can solve equation (2.7) by making an initial guess of β, say β_0. Given β_0, we can evaluate equation (2.7) and its derivative with respect to β in a derivative based optimization routine[2] to make a new guess of β. Using the new guess, we can reevaluate equation (2.7) and its derivative to get another new guess of β. We can continue this process until we converge to a solution of equation (2.7). In the probit example, we can set $Q_i = P_i^{-1} (1 - P_i)^{-1} (\partial P_i / \partial \beta)$ (where P_i and its derivative are evaluated at β_0) to make MOM estimation asymptotically equivalent to maximum likelihood estimation. In order to maximize equation (2.5) over β or solve equation (2.7), we need to be able to evaluate P_i for any value of β. Most computers have library routines to evaluate $\Phi(\bullet)$, so this is not a problem. But in more complicated problems, the function analogous to $\Phi(\bullet)$ (i.e., the Ey) will be infeasible to evaluate. So we will continue with this example assuming one can not evaluate $\Phi(\bullet)$.

Consider writing

$$P_i = \int_{-\infty}^{X_i \beta} \phi(u)\, du = \int_{-\infty}^{\infty} 1 [u < X_i \beta] \phi(u)\, du \tag{2.8}$$

[2] We can use an optimization routine to minimize the square of equation (2.7): $\frac{1}{N} \sum_{i=1}^{N} [y_i - P_i]' Q_i Q_i' [y_i - P_i]$.

where $\phi(\bullet)$ is the standard normal density function and $1[\bullet]$ is equal to one if the argument is true, and it is equal to zero if the argument is not true. The last integral in equation (2.8) is $E1[U < X_i\beta]$. Consider drawing a set of standard normal random variables, $u^r, r = 1, 2, ..., R$, and applying the function $1[u^r < X_i\beta]$ to each one. Let

$$\hat{P}_i = \frac{1}{R}\sum_{r=1}^{R} 1[u^r < X_i\beta] \tag{2.9}$$

be a simulator of P_i. The basic approach is to simulate U from its distribution Φ, evaluate $h(u^r) = 1[u^r < X_i\beta]$, repeat this R times, and average the R draws of $h(U)$. Then, because $P_i = E1[U < X_i\beta]$, $\hat{P}_i$ is an unbiased simulator of P_i. Also,

$$Var\hat{P}_i = \frac{1}{R}P_i(1 - P_i)^{3} \tag{2.10}$$

which goes to zero as $R \rightarrow \infty$. We can write $\hat{P}_i = P_i + \xi_i$ where $E\xi_i = 0$ and $Var\xi_i = Var\hat{P}_i$.

In equation (2.7), we constructed a MOM estimator of β that depended upon evaluating P_i. Again, pretending that we can not evaluate P_i, consider substituting $\hat{P}_i$ for P_i. Then, we can write the orthogonality condition as

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N} Q_i'\left[y_i - \hat{P}_i\right] = \frac{1}{N}\sum_{i=1}^{N} Q_i'[y_i - P_i - \xi_i] \\ &= \frac{1}{N}\sum_{i=1}^{N} Q_i'[y_i - P_i] - \frac{1}{N}\sum_{i=1}^{N} Q_i'\xi_i. \end{aligned} \tag{2.11}$$

The β that solves the orthogonality condition in equation (2.11) is the method of simulated moments estimator (MSM) of β. The last term in the second line disappears asymptotically in N even for fixed R by a law of large numbers because ξ_i is independently distributed with zero mean. The first term in the second line is the orthogonality condition for the MOM estimator. Thus, the effect of simulating P_i with $\hat{P}_i$ disappears asymptotically in N. Because the MOM estimator (with P_i) is consistent and the only difference between the MOM estimator and MSM estimator disappears asymptotically, the MSM estimator (with $\hat{P}_i$) converges to the MOM estimator asymptotically and is therefore also consistent. The asymptotic covariance matrix of the MOM estimator $\tilde{\beta}$ is

$$D\left(\tilde{\beta}\right) = \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i'\frac{\partial P_i}{\partial\beta}\right]^{-1} \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Var(y_i)\, Q_i'Q_i\right] \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial P_i'}{\partial\beta}Q_i\right]^{-1} \tag{2.12}$$

where $Var(y_i) = P_i(1 - P_i)$. The asymptotic covariance matrix of the MSM estimator $\hat{\beta}$ is

$$D\left(\hat{\beta}\right) = \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i'\frac{\partial P_i}{\partial\beta}\right]^{-1} \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Var(y_i - \xi_i)\, Q_i'Q_i\right] \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial P_i'}{\partial\beta}Q_i\right]^{-1} \tag{2.13}$$

[3] Each random variable $1[U < X_i\beta]$ is a Bernoulli random variable with probability P_i. This implies the result.

where

$$\begin{aligned} Var\left(y_i - \xi_i\right) &= Var\left(y_i\right) + Var\left(\xi_i\right) \\ &= Var\left(y_i\right) + \frac{1}{R} Var\left(y_i\right) \\ &= \left[1 + \frac{1}{R}\right] P_i \left(1 - P_i\right). \end{aligned} \tag{2.14}$$

The $Var\left(y_i\right)$ term in $D\left(\tilde{\beta}\right)$ is replaced with $Var\left(y_i - \xi_i\right)$ in $D\left(\hat{\beta}\right)$ because we have added randomness ξ_i to each residual in equation (2.11). The first equality in equation (2.14) follows because y_i and ξ_i are independent. The second equality follows from equation (2.10). Note that simulation increases the asymptotic covariance matrix by a factor of $\left[1 + \frac{1}{R}\right]$, and that as $R \to \infty$, the loss in efficiency associated with simulation disappears: $\left[1 + \frac{1}{R}\right] \to 1$. A more rigorous argument for the asymptotic properties of the MSM estimator can be found in seminal papers by McFadden (1989) and Ariel Pakes and David Pollard (1989) or in other surveys such as Vassilis Hajivassiliou (1993), Michael Keane (1993), Hajivassiliou, McFadden, and Paul Ruud (1994), Hajivassiliou and Ruud (1994), and McFadden and Ruud (1994).

In equation (2.5), we constructed the log likelihood function; the β that maximizes the equation is the ML estimator of β. Consider substituting $\hat{P}_i$ for P_i (again assuming that we can not evaluate P_i analytically). Then, we can write the log likelihood function as

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left[y_i \log \hat{P}_i + \left(1 - y_i\right) \log\left(1 - \hat{P}_i\right)\right]. \tag{2.15}$$

The value of β that maximizes equation (2.15) is the maximum simulated likelihood (MSL) estimator $\hat{\beta}$ of β. The consistency argument used for the MSM estimator relied upon $\frac{1}{N}\sum_{i=1}^{N} Q_i' \xi_i \to 0$ as $N \to \infty$ which required that the first order condition be linear in the simulation error ξ_i. For MSL estimation, the simulation error is not linear because we are taking logs of $\hat{P}_i$ and $1 - \hat{P}_i$. Because $E \log \hat{P}_i \neq \log E\hat{P}_i$, there is a bias in simulating $\log P_i$. However, as $R \to \infty$, $\hat{P}_i \to P_i$ by equation (2.10). Thus, we can construct a consistent MSL estimator of β by letting $R \to \infty$ as $N \to \infty$. In theory, the need to let $R \to \infty$ for MSL estimation is a significant disadvantage of MSL estimation relative to MSM. But much work has been done on constructing better simulators of P_i; these are discussed in Section 3. These better simulators are smooth in the parameters, they are usually strictly bounded between zero and one (so that $\log \hat{P}_i$ always exists), and they frequently have small variances. Axel Börsch-Supan and Hajivassiliou (1993) show in some Monte Carlo experiments that if a "better" simulator for P_i is used , then, even for fixed R, MSL estimators have small asymptotic bias and root mean squared errors.

In fact, the MSM and MSL estimators discussed above perform poorly in practice because the simulator defined in equation (2.9) has the following shortcomings relative to P_i. First, for most values of β, a small change in β changes none of the values of $1\left[u^r < X_i\beta\right]$, and for the remaining values of β, a small change in β changes one of the $1\left[u^r < X_i\beta\right]$ terms from 0 to 1 or from 1 to 0. Thus, the simulator is a step function in β. Step functions are very difficult to optimize over because derivatives are equal to zero everywhere but points with discrete steps and do not exist at such points. Second, especially for small P_i, $\Pr\left[\hat{P}_i = 0\right]$ is large; this is a significant problem for MSL estimation which needs to simulate $\log P_i$. Third, the $Var\left(\hat{P}_i\right)$ is unnecessarily large. In Section 3, I discuss better simulators that have none of these problems and that allow MSM estimators and MSL estimators to behave nicely. Importance sampling methods use a well chosen distribution to rewrite the integral in equation (1.1) as the expected value of a function of a new random variable whose simulator is smooth (it is also frequently strictly bounded and has smaller variance). Decomposition methods break up the random variable to be simulated into two random variables only one of which needs to be simulated; the expected value of $h(U)$ conditional on the simulated random variable is smooth, strictly bounded, and has smaller variance. Antithetic acceleration methods choose negatively correlated simulators to reduce variance.

2.2. Multinomial Probit

The next example considered is the multinomial probit problem. The multinomial probit problem is a generalization of the probit problem, and, unlike the probit problem, it does require simulation. It is also the leading problem in developing simulation methods (e.g., McFadden 1989; Börsch-Supan and Hajivassiliou 1993; and Monte Carlo studies such as Hajivassiliou, McFadden, and Ruud 1996; and John Geweke, Keane, and David Runkle 1994, 1996). Consider a model of occupational choice, and let occupations be indexed by $j = 1, 2, .., J$. Let the value to person i of choosing occupation j be y^*_{ij}, and assume that y^*_{ij} depends upon some observed characteristics (such as age, sex, education, and occupation specific characteristics) X_{ij}, and an error u_{ij}:

$$y^*_{ij} = X_{ij}\beta + u_{ij}. \tag{2.16}$$

For observed characteristics that do not vary over choices (such as age, sex, and education), we can construct occupation-specific dummy variables and interact them with those characteristics. Assume that we observe for each person, $X_i = (X_{i1}, X_{i2}, ..., X_{iJ})'$ and a vector of indicators $y_i = (y_{i1}, y_{i2}, ..., y_{iJ})'$ where $y_{ij} = 1$ if i chooses occupation j, and $y_{ij} = 0$ if i does not choose occupation j. Assume that i chooses one occupation: $\sum_{j=1}^{J} y_{ij} = 1$. Further, assume that i chooses j if and only if the value of j is greater than the value of any other choice: $y_{ij} = 1$ if and only if $y^*_{ij} > y^*_{ik}$ for all $k \neq j$. Note that if $J = 2$, then we can define $\tilde{y}_i = y^*_{i1} - y^*_{i2}$, $\tilde{X}_i = X_{i1} - X_{i2}$, and $\tilde{u}_i = u_{i1} - u_{i2}$, assume $\tilde{u}_i \sim N(0, \sigma^2)$, and treat the problem as a probit problem.

In general, the probability of observing $y_{ij} = 1$ is

$$\begin{aligned} P_{ij} &= \Pr\left[y_{ij} = 1 \mid X_i\right] = \Pr\left[y^*_{ij} > y^*_{ik} \quad \forall k \neq j \mid X_i\right] \\ &= \Pr\left[X_{ij}\beta + u_{ij} > X_{ik}\beta + u_{ik} \quad \forall k \neq j\right] \\ &= \Pr\left[u_{ik} - u_{ij} < (X_{ij} - X_{ik})\beta \quad \forall k \neq j\right]. \end{aligned} \tag{2.17}$$

Let $u^*_{ijk} = u_{ik} - u_{ij}$ and $X^*_{ijk} = X_{ij} - X_{ik}$ for all $k \neq j$, and let $F(\bullet)$ be the distribution function of $u^*_{ij} = \left(u^*_{ij1}, u^*_{ij2}, ..., u^*_{ijJ}\right)'$, the $(J-1)$-dimensional vector of error differences. Then, equation (2.17) becomes

$$\begin{aligned} P_{ij} &= \Pr\left[u^*_{ijk} < X^*_{ijk}\beta \quad \forall k \neq j\right] \qquad (2.18) \\ &= \int_{-\infty}^{X^*_{ijJ}\beta} \cdots \int_{-\infty}^{X^*_{ij1}\beta} dF\left(u^*_{ij}\right) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1\left[u^*_{ijk} < X^*_{ijk}\beta \quad \forall k \neq j\right] dF\left(u^*_{ij}\right). \end{aligned}$$

In general, equation (2.18) is a $(J-1)$-dimensional integral and can not be integrated analytically or numerically with any precision for $J > 4$ (see Hausman and Wise 1978 for a numerical solution when $J = 4$). If we assume that $u_{ik} \sim iid$ Extreme Value,[4] then equation (2.18) becomes the multinomial logit probability

$$P_{ij} = \frac{\exp\{X_{ij}\beta\}}{\sum_{k=1}^{J}\exp\{X_{ik}\beta\}}. \qquad (2.19)$$

But the multinomial logit problem suffers from the independence of irrelevant alternatives problem discussed in Section 1: P_{ij}/P_{ik} is independent of whether any other choice l is available.

Instead, assume that $u_i = (u_{i1}, u_{i2}, ..., u_{iJ})' \sim iidN(0, \Omega)$ which implies that $u^*_{ij} \sim iidN\left(0, \Omega^*_j\right)$ where the kl-element of Ω^*_j is

$$\Omega^*_{jkl} = E\left(u^*_{ijk}u^*_{ijl}\right) = E(u_{ik} - u_{ij})(u_{il} - u_{ij}) = \Omega_{kl} + \Omega_{jj} - \Omega_{jk} - \Omega_{jl} \qquad (2.20)$$

where Ω_{kl} is the kl-element of Ω. This assumption allows for general correlations between errors across choices subject to the usual restriction that Ω is positive definite. Then P_{ij} in equation (2.18) becomes

$$P_{ij} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1\left[u^*_{ijk} < X^*_{ijk}\beta \quad \forall k \neq j\right] d\Phi\left(u^*_{ij} \mid \Omega^*_j\right) \qquad (2.21)$$

where $\Phi\left(u^*_{ij} \mid \Omega^*_j\right)$ is the multivariate normal distribution function (with covariance matrix Ω^*_j) for u^*_{ij}. There are some significant identification restrictions: a) only Ω^*_j is identified by the data and b) β is proportional to $\sqrt{\Omega^*_{jkk}}$ for some chosen diagonal element Ω^*_{jkk} of Ω^*_j (David Bunch 1991).

The maximum likelihood (ML) estimator of $\theta = (\beta, \Omega)$ maximizes the log likelihood function

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log P_{ij} \qquad (2.22)$$

[4] $F(u_{ik}) = \exp\{-\exp(-u_{ik})\}$.

where $L_i = \sum_{j=1}^{J} y_{ij} \log P_{ij}$ is the log likelihood contribution of observation i. The derivative of L_i with respect to θ is

$$\begin{aligned} \frac{\partial L_i}{\partial \theta} &= \sum_{j=1}^{J} y_{ij} \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta} \qquad (2.23) \\ &= \sum_{j=1}^{J} \left(y_{ij} - P_{ij} + P_{ij}\right) \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta} \\ &= \sum_{j=1}^{J} \left(y_{ij} - P_{ij}\right) \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta} + \sum_{j=1}^{J} \frac{\partial P_{ij}}{\partial \theta} \\ &= \sum_{j=1}^{J} \left(y_{ij} - P_{ij}\right) \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta}. \end{aligned}$$

The last step follows because $\sum_{j=1}^{J} P_{ij} = 1$ which implies that $\sum_{j=1}^{J} \partial P_{ij} / \partial \theta = 0$. As was true for the probit problem, $Ey_{ij} = P_{ij}$ which suggests that equation (2.23) can be thought of as $Q_i^{'} \left[y_i - P_i\right]$ where Q_i is a matrix of instruments for observation i, y_i is the vector of choice indicators, and P_i is the associated vector of choice probabilities. Note that $Ey_i = P_i$. A MOM estimator of θ solves

$$0 = \frac{1}{N} \sum_{i=1}^{N} Q_i^{'} \left[y_i - P_i\right]. \qquad (2.24)$$

In order to maximize the log likelihood function in equation (2.22) over θ or solve the moment conditions in equation (2.24), we need to be able to evaluate P_i for any value of θ. This is a significant problem because each element of P_i is a $(J-1)$-dimensional integral similar to equation (2.21).

Analogous to the probit problem, we can think of the integral in equation (2.21) as $E1\left[u_{ijk}^{*} < X_{ijk}^{*} \beta \quad \forall k \neq j\right]$. Consider drawing a set of $N\left(0, \Omega_j^{*}\right)$ random variables, $u_{ij}^{*r}, r = 1, 2, ..., R$, and applying $1\left[u_{ijk}^{*} < X_{ijk}^{*} \beta \quad \forall k \neq j\right]$ to each one. Let

$$\hat{P}_{ij} = \frac{1}{R} \sum_{r=1}^{R} 1\left[u_{ijk}^{*r} < X_{ijk}^{*} \beta \quad \forall k \neq j\right] \qquad (2.25)$$

be a simulator of P_{ij}. As in the probit problem, the basic approach is to simulate U_{ij}^{*} from its $N\left(0, \Omega_j^{*}\right)$ distribution, evaluate $h\left(U_{ij}^{*}\right) = 1\left[U_{ijk}^{*} < X_{ijk}^{*} \beta \quad \forall k \neq j\right]$, repeat this R times, and average the R draws of $h\left(U_{ij}^{*}\right)$. Then, because $P_{ij} = E1\left[U_{ijk}^{*} < X_{ijk}^{*} \beta \quad \forall k \neq j\right]$, $\hat{P}_{ij}$ is an unbiased simulator of P_{ij}. Also,

$$Var \hat{P}_{ij} = \frac{1}{R} P_{ij} \left(1 - P_{ij}\right) \qquad (2.26)$$

which goes to zero as $R \to \infty$.[5] We can write $\hat{P}_{ij} = P_{ij} + \xi_{ij}$ where $E\xi_{ij} = 0$ and $Var \xi_{ij} = Var \hat{P}_{ij}$.

[5] I am ignoring the $Cov\left(\hat{P}_{ij}, \hat{P}_{ik}\right)$ terms because they depend upon how the $\hat{P}_{ij}$'s are simulated. If they are simulated independently, then the covariance terms are zero.

In equation (2.24), we constructed a MOM estimator of θ that required evaluating P_i. Because we can not evaluate P_i, consider substituting $\hat{P}_i$ for P_i as we did in the probit problem. Then, we can write the orthogonality condition as

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N} Q_i^{'}\left[y_i - \hat{P}_i\right] = \frac{1}{N}\sum_{i=1}^{N} Q_i^{'}\left[y_i - P_i - \xi_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} Q_i^{'}\left[y_i - P_i\right] - \frac{1}{N}\sum_{i=1}^{N} Q_i^{'}\xi_i \end{aligned} \tag{2.27}$$

where $\xi_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{iJ})'$. The θ that solves the orthogonality condition in equation (2.27) is the MSM of θ. As in the probit problem, the last term in the second line disappears asymptotically in N even for fixed R by a law of large numbers because ξ_i is independently distributed with zero mean. The first term in the second line is the orthogonality condition for the MOM estimator. Thus, the effect of simulating P_i with $\hat{P}_i$ disappears asymptotically in N. Because the MOM estimator (with P_i) is consistent, the MSM estimator (with $\hat{P}_i$) is also consistent. The asymptotic covariance matrix of the MSM estimator $\hat{\beta}$ is

$$D\left(\hat{\beta}\right) = \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i^{'}\frac{\partial P_i}{\partial \beta}\right]^{-1} \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i^{'} D\left(y_i - \xi_i\right) Q_i\right] \text{plim}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial P_i}{\partial \beta}^{\prime} Q_i\right]^{-1} \tag{2.28}$$

where $D(\bullet)$ is a covariance matrix and

$$\begin{aligned} D\left(y_i - \xi_i\right) &= D\left(y_i\right) + D\left(\xi_i\right) \\ &= D\left(y_i\right) + \frac{1}{R} D\left(y_i\right) \\ &= \left[1 + \frac{1}{R}\right] D\left(y_i\right). \end{aligned} \tag{2.29}$$

As in the probit problem, the $D\left(y_i - \xi_i\right)$ term in $D\left(\hat{\beta}\right)$ occurs because we have added randomness ξ_i to each residual in equation (2.27). Again, note that simulation increases the asymptotic covariance matrix by a factor of $\left[1 + \frac{1}{R}\right]$, and that, as $R \to \infty$, the loss in efficiency associated with simulation disappears: $\left[1 + \frac{1}{R}\right] \to 1$.

To find an "optimal" set of instruments, consider the relationship between equations (2.23) and (2.24). They imply that choosing $Q_{ij} = P_{ij}^{-1}\left(\partial P_{ij} / \partial \theta\right)$ makes MOM estimation asymptotically equivalent to ML estimation. We can simulate the proposed instruments conditional on an initial guess of θ. Any instruments that satisfy the classic conditions for instruments will provide a consistent MSM estimator of θ. If we construct $Q_{ij} = P_{ij}^{-1}\left(\partial P_{ij} / \partial \theta\right)$ using a consistent estimator of θ, then the only loss in efficiency relative to ML estimation is the loss due to simulation, and that can be made significantly smaller by increasing R, by using good simulation methods, or both.

Below is a roadmap for using MSM to estimate multinomial probit parameters:

A. Choose an identifiable parameterization for Ω and initial values for $\theta = (\beta, \Omega)$. Make sure that the initial guess results in probabilities reasonably far from zero or one so that instruments will be well behaved. In particular, the initial values for θ should result in instruments that are highly correlated with $P_{ij}^{-1}(\partial P_{ij}/\partial\theta)$ evaluated at the true value of θ and should not have significant multicollinearity problems. Sometimes it may be better to start with crude instruments such as X and then to update them with a consistent estimate of θ. [6]

B. Choose a simulator:

$$\hat{P}_{ij} = \frac{1}{R}\sum_{r=1}^{R}\tilde{P}_{ij}^{r} \tag{2.30}$$

where $\tilde{P}_{ij}^{r}$ is an unbiased simulator of P_{ij} using the rth draw of random variables (good choices for $\tilde{P}_{ij}^{r}$ are discussed in Section 3.2).

C. Simulate $2NJR$ standard normal random variables.[7] Store NJR of them in an instruments random number file and NJR in an estimation random number file. These random numbers will be used throughout the estimation process and never changed.

D. Given the initial guess of θ and the instruments random number file, simulate the instruments as

$$Q_{ij} = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{P}_{ij}^{r}\right)^{-1}\left(\partial\tilde{P}_{ij}^{r}/\partial\theta\right). \tag{2.31}$$

In particular, for each observation i and choice j, use the ijrth draw of the standard normals to generate a normal vector with covariance matrix Ω_j^* (Section 3.1 describes how to do this). Use these to construct $\tilde{P}_{ij}^{r}$ and $\partial\tilde{P}_{ij}^{r}/\partial\theta$ and plug these into equation (2.31). Store the simulated instruments.

E. Given the initial guess of θ, the simulated instruments, and the estimation random number file to construct $\hat{P}_{ij}$ in equation (2.30), solve equation (2.27) for θ. This is a MSM estimator of θ. It is consistent but not efficient (because the instruments were constructed using an arbitrary choice of θ).

F. Given the initial MSM estimator, reperform steps (D) and (E) once using the consistent estimate of θ from step E. This provides an (almost)[8] efficient estimator (relative to the class of simulation estimators).

[6] See Hajivassiliou (1992) for a discussion on "reasonable" instrument selection.

[7] Remember that N = sample size, J = number of alternatives, and R = number of draws.

[8] Efficiency is not reached because $\left(\tilde{P}_{ij}^{r}\right)^{-1}$ is not a consistent estimator of $(P_{ij})^{-1}$. However, papers such as McFadden and Ruud (1994) and Keane (1994) suggest the loss of efficiency is small.

Solving equation (2.27) requires using an optimization algorithm to find the θ that minimizes

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{P}_i\right]' Q_i Q_i' \left[y_i - \hat{P}_i\right].^{9} \tag{2.32}$$

The derivatives of $\hat{P}_i$ are well behaved for good simulators, so derivative based optimization routines should be used. In step (C), it was suggested not to change the random numbers used for simulation over the course of estimation. First of all, there is no need to change them. Even as Ω is changing, it is straightforward to use the original standard normal random variables to create $N(0,\Omega)$ random variables (see Section 3.1.1). Second, if the random variables do change, then the value of θ that solves equation (2.32) will change as the random variables change; there is no guarantee that the optimization algorithm will ever converge.

In equation (2.22), we constructed the log likelihood function; the θ that maximizes the equation is the ML estimator of θ. Consider substituting $\hat{P}_i$ for P_i. Then, we can write the log likelihood function as

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\log \hat{P}_{ij}. \tag{2.33}$$

The value of θ that maximizes equation (2.33) is the MSL estimator $\hat{\theta}$ of θ. As in the probit problem, consistency of the MSL estimator relies upon $R \to \infty$ as $N \to \infty$. But, again, Börsch-Supan and Hajivassiliou (1993) suggest this is not a very important restriction on the use of MSL, at least for the multinomial probit problem. There are two significant advantages of MSL estimation relative to MSM estimation. First, ML estimation is efficient relative to MOM. Second, for MSM, we have to simulate $\hat{P}_{ij}$ for all $j = 1, 2, ..., J$, while for MSL estimation, we need only simulate $\hat{P}_{ij}$ for the chosen alternative j.

2.3. Dynamic Programming Models

Dynamic programming models have become popular models for describing dynamic behavior. Examples of such models are the retirement model of James Berkovec and Stern (1991), the child spacing model of V. Joseph Hotz and Robert Miller (1988), and the international currency and asset choice model of Ravi Bansal et al. (1995). An early example, Pakes (1986), is a model of patent renewal. The firm must decide each year whether to renew a patent conditional on a set of information I_t. If it renews the patent, it pays a patent renewal cost c_t, receives a random return ρ_t, and has the option to renew the patent again the following year. If it does not renew, it receives zero forever after. The value of renewing a patent of age t is

$$V(t) = \max\left\{0, \rho_t + \beta E\left[V(t+1) \mid I_t\right] - c_t\right\} \tag{2.34}$$

where β is a discount factor. Pakes uses a rich specification for the stochastic process generating returns $\{\rho_t\}_{t=1}^{T}$. Significantly simplifying his model, let $\rho_{t+1} = \max\left[\delta\rho_t, Z_t\right]$ where

[9] The first order condition for this optimization problem has form equivalent to the moment condition in equation (2.24).

$\log\rho_1 \sim N\left(\mu, \sigma_\rho^2\right)$ and the density of Z_t is $g_t(z)$. The law of motion for ρ_{t+1} implies that there is some probability that the firm will discover a new application of the patent with value Z_t. The distribution of Z_t is changing over time such that the variance of Z_t is decreasing over time (firms are more likely to find new applications early in an invention's life). The goal is to estimate a set of parameters θ that determine the joint density of returns $\{\rho_t\}_{t=1}^T$. It is not possible to write down the joint density of $\{\rho_t\}_{t=1}^T$. Thus, it is not possible to use maximum likelihood estimation or method of moments estimation. But I show in Section 3.1.2 how to simulate sequences of $\{\rho_t\}_{t=1}^T$.

Pakes uses these simulated sequences to estimate θ using maximum simulated likelihood estimation. In particular, the model implies a sequence of reservation returns $\{\bar{\rho}_t\}_{t=1}^T$ where the firm pays the renewal fee c_t if and only if $\rho_t \geq \bar{\rho}_t$.[10] Let $\pi_j(t)$ be the probability that a firm from cohort j renews its patent up until t and then does not renew its patent in year t. Pakes simulates $\pi_j(t)$ using a frequency simulator (described in Section 3).

The log likelihood function can be written in terms of the π's and aggregate data on the proportion of patents renewed conditional on age and cohort. Let $n(t,j)$ be the number of patents not renewed at age t from cohort j. Then the log likelihood function is

$$L(\theta) = \sum_j \sum_{t=1}^T n(t,j) \log \pi(t,j). \tag{2.35}$$

The features common to all of these stochastic, dynamic models are that a) they allow for complex, rational behavior in a stochastic, dynamic environment, b) they usually allow for much more interesting policy analyses than in simpler models (see Kenneth Wolpin 1996), but c) it is very difficult to evaluate analytically high-dimensional integrals associated with estimation. Simulation provides a solution for the last problem.

2.4. Market Entry

Another area of expanding literature is empirical models of game theory. Timothy Bresnahan and Peter Reiss (1991) describe a general structure for problems when the game's agents make discrete choices, and they describe some of the hurdles that must be overcome to implement the model empirically. Frequently, there are problems of nonexistence of equilibria, potential multiple equilibria, and difficult integrals to evaluate. An example of this type of problem, Steven Berry (1992), is a model of entry into specific airline route markets. Each city pair is a separate observation. Entry into a market depends upon the solution of an oligopoly equilibrium game. There is firm heterogeneity which leads to integration problems with

[10] Computing the sequence of reservation values is difficult and usually requires evaluating a high dimensional integral. Pakes uses a symbolic mathematics program called MACSYMA to do this, and the particular method is specific to his assumptions. But, in general, there is a solution which can be found numerically. Also, it needs to be solved only once for each guess of the parameters. Hotz et al. (1994) suggest a different solution to the problem. Other researchers (e.g., Miller 1984; and Berkovec and Stern 1991) make strong assumptions concerning information sets and error distributions to provide an analytical solution to the relevant integral.

boundaries of integration that do not lend themselves to standard solution methods (i.e., the support is not a rectangle). The model has multiple equilibria. But, under reasonable conditions, all of the equilibria have the same number, N_i, of entrants in market i. The likelihood function for N_i is very hard to evaluate or simulate, but EN_i is easy to simulate.

Each market i has K potential entrants. Let $s_{ik} = 1$ if and only if firm k enters market i, and let $s_i = (s_{i1}, s_{i2}, ..., s_{iK})$ be the vector of strategies for all potential entrants. Let $\pi_{ik}(s_i)$ be profits for firm k in market i conditional on s_i. A pure strategy equilibrium requires that each entrant makes nonnegative profits and that each nonentrant would have lost profits had it entered the market. Berry assumes that

$$\pi_{ik}(s_i) = v[N(s_i)] + \varepsilon_{ik} \tag{2.36}$$

where $N(s_i)$ is the number of entrants implied by s_i $\left(N(s_i) = \sum_k s_{ik}\right)$ and that $\partial v / \partial N < 0$; i.e., profits depend upon strategies s_i only through their effect on the number of entrants N_i. He shows that there is a unique equilibrium number of entrants N_i^*; for all equilibrium strategies s_i^*, $N(s_i^*) = N_i^*$. Thus, he can avoid multiple equilibrium problems by focusing on N instead of s.

He assumes that

$$\begin{aligned} v[N] &= X_i\beta - \delta \ln N + \rho e_i \\ \varepsilon_{ik} &= W_{ik}\alpha + \sigma u_{ik} \end{aligned} \tag{2.37}$$

where X_i is a set of market specific variables, ρe_i is a market specific error with variance ρ^2, W_{ik} is a set of characteristics specific to firm k in market i, and σu_{ik} is an error specific to firm k in market i with variance σ^2. The goal is to estimate $\theta = (\beta, \delta, \rho, \alpha, \sigma)$ using data on market characteristics X_i, firm-market characteristics W_{ik}, and the number of entrants in the market N_i for a large number of US airline markets. This can be accomplished using simulated values of EN_i with the method of simulated moments (described in Section 3).

The Berry model is useful to consider in that it is an early example of estimating a game theoretic model. It is likely that empirical applications of game theory will become popular over time and that simulation methods will play a critical role in their development. Also, the Berry model is an example of a problem where MSM is used because the associated likelihood function is too hard to evaluate or simulate.

2.5. Unobserved Heterogeneity

The problem of unobserved heterogeneity has received some attention in the literature (e.g., James Heckman 1981). The problem is that individuals have person-specific unobserved characteristics that affect their behavior. These person-specific effects significantly complicate estimation because one needs to integrate over the distribution of the effects. I consider two papers, Berkovec and Stern (1991) and Berry, James Levinsohn, and Pakes (1995), to illustrate the issues involved.

Berkovec and Stern (1991) write down a straightforward dynamic programming model of retirement decisions. Let $d_{tj} = 1$ if and only if choice j is chosen at time t, $d_t =$

$(d_{t1}, d_{t2}, ..., d_{tJ})$, and H_t be the history of choices and states up to and including time t, $H_t = (d_1, d_2, ..., d_t)$. In Berkovec and Stern, the choices are to retire (not work), stay in one's job, take a new full-time job, or take a new part-time job ($J = 4$). One's relevant history is the type of job and tenure in that job at time t. The utility one gets from choice j at t is

$$u_{tj} = u_j^* (H_t) + \varepsilon_{tj}. \tag{2.38}$$

In Berkovec and Stern, u_j^* represents pecuniary returns (wages and pensions) and nonpecuniary returns (job conditions and leisure) from the various choices. Using Bellman's equation, we can write the value of choice j given H_{t-1} as

$$V_j (H_{t-1}) = u_j^* (H_{t-1}) + \varepsilon_{tj} + \beta E [V_j (H_t) \mid d_t, H_{t-1}] \tag{2.39}$$

where β is a discount factor.

Berkovec and Stern specify the error in equation (2.38) as

$$\varepsilon_{tj} = \eta + \alpha_j + v_{tj} \tag{2.40}$$

where η is a person-specific effect, α_j is a "type of job" specific effect, and v_{tj} is a time specific effect with two components, one specific to the particular job and one independent over time. The α terms allow for different people have different preferences for leisure, and the job-specific terms allow for the kind of matching effects prevalent in the labor literature (Boyan Jovanovic 1979). The inclusion of α terms and job specific terms mean that relevant choice probabilities involve evaluating T-dimensional integrals where T is the length of the individual's planning horizon (in Berkovec and Stern, $T = 32$). Yet exclusion of these terms leads to somewhat unrealistic models.

Consider a dynamic model of contraceptive choice (Hotz and Miller 1988). In this example, the choices are methods of contraception, and histories are number and ages of children. The error terms represent unobserved preferences for particular contraceptive methods.[11] If there are no choice-specific components to these errors over time, then one should observe that contraceptive choices are independent over time after conditioning on observed characteristics. This is clearly not consistent with the data. Yet inclusion of choice specific errors makes estimation impossible without simulation methods because evaluating choice probabilities involves solving high dimensional integrals (the dimension is at least the number of contraceptive methods minus one).

Many dynamic programming models are simple enough so that they do not require simulation methods (see, for example, Miller 1984; Rust 1987; and Wolpin 1984). But this is only because the model assumes away (probably important) unobserved heterogeneity. Berkovec and Stern find that unobserved heterogeneity plays an important role in their model. I expect it to play an important role in other models as its inclusion becomes more feasible through simulation methods.

[11] There is an added source of randonmess not found in Berkovec and Stern. In particular, because fertility can not be controlled completely by contraception, the history attained next period conditional on the contraceptive choice made this period is random.

Berry, Levinsohn, and Pakes (1995) allow for unobserved heterogeneity in a static model of automobile demand. Let the utility to person i of choosing brand j be

$$u_{ij} = \alpha \log (y_i - p_j) + X_j\beta + \xi_j + \sum_{k=1}^{K} \sigma_k X_{jk} v_{ik} + \varepsilon_{ij} \tag{2.41}$$

where y_i is income, p_j is the price of j, X_j is a set of observed characteristics of choice j, ξ_j is a set of unobserved characteristics of j, and v_{ik} is the preference for characteristics k specific to person i, $k = 1, 2, .., K$. Note that this model allows for brand characteristics to interact with unobserved personal characteristics. This allows for effects such as big families preferring big cars or wealthy families having a strong preference for air conditioning. It also allows for unobserved brand characteristics ξ_j to be correlated with price by treating ξ_j as a fixed effect. Assuming otherwise is unreasonable in that ξ_j is common across people and observed by people; thus it must affect prices. Berry, Levinsohn, and Pakes use aggregate panel data on the market share of each brand and the distribution of income to estimate α, β, and the σ's (along with other parameters not relevant for this discussion). In particular, let $v_i = (v_{i1}, v_{i2}, ..., v_{iK})'$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iJ})'$. Then, given a joint distribution F for y_i, v_i, and ε_i, one can compute a brand's market share as

$$\int_{Aj} dF (y_i, v_i, \varepsilon_i) \tag{2.42}$$

where A_j is that part of the support of y_i, v_i, and ε_i such that brand j is preferred to all other brands. Berry, Levinsohn, and Pakes use Current Population Survey data to specify a marginal distribution for y_i and estimate the parameters of the joint distribution of v_i and ε_i. In another example of this type, Benjamin Scafidi (1996) uses Census data to estimate the joint distribution of income, race, and education, and then uses this distribution with block group Census data and methodology similar to Berry, Levinsohn, and Pakes to estimate parameters associated with choice of school and neighborhood. Both of these suggest ways to deal with the unobserved heterogeneity inherent in aggregate data that must be integrated out. Simulation provides an alternative to looking for a restrictive functional form whose only value is its nice aggregation properties.

3. Methods[12]

This section discusses various simulation methods. First it provides some of the basics of simulation. Then it discusses integration by simulation. Finally, it discusses using simulation in estimation and compares different estimators.

[12] Significant portions of this section are very similar to Sections 2 and 3 in Stern (forthcoming). A well-written discussion of the same topics appears in Geweke (forthcoming).

3.1. Simulators

Throughout this section, I will assume that the goal is to simulate $Eh(U)$ where U is a vector of random variables with known distribution F and that h and F depend upon a set of parameters θ to be estimated. A basic approach is to simulate U from its distribution F, evaluate $h(U)$, repeat this R times, and average the R draws of $h(U)$. As noted in Section 2, a special case that has received much attention is $\Pr[y_{ij} = 1 \mid X_i]$ from the multinomial probit example.

3.1.1. Simulation of Random Variables

The first goal is to simulate a pseudorandom variable[13] U from its distribution F. In general, if $Z \sim$ Uniform $(0,1)$, then $F^{-1}(Z) \sim F$.[14] For example, Pakes (1986) uses a generalized exponential distribution to simulate new innovations. Consider another example: the exponential distribution function with $F(z) = 1 - \exp\{-\lambda z\}$. Thus, $F^{-1}(z) = -\ln(1-z)/\lambda$, implying $-\ln(1-Z)/\lambda \sim F$. Therefore, we can simulate a standard uniform random variable, $Z \sim$ Uniform $(0,1)$, and then compute $U = -\ln(1-Z)/\lambda$, and U will be distributed exponential(λ). In general, this method is called the inversion method. Consider the example where $U \sim N[\mu, \sigma^2]$; then F^{-1} has no closed form. But most computers have library routines to approximate Φ^{-1}, the inverse standard normal distribution function, and $U = \mu + \sigma\Phi^{-1}(Z) \sim N[\mu, \sigma^2]$. The intuition for the general result is seen in Figure 1. The curve F is some general distribution function. The inversion method essentially picks a random point on the vertical axis and then finds the random variable on the horizontal axis associated with it. This is equivalent to evaluating $F^{-1}(Z)$. A proof of the general result is as follows: Let $U = F^{-1}(Z)$. Then $\Pr[U < u] = \Pr[F^{-1}(Z) < u] = \Pr[Z < F(u)] = F(u)$. The second to last step follows because F is a monotone function, and the last step follows because $Z \sim$ Uniform $(0,1)$ with distribution function $G(z) = z$.

FIGURE 1
Random Sampling

Truncated random variables can be simulated in a similar way. For example, assume $U \sim N[\mu, \sigma^2]$ but let it be truncated with truncation points a and b. Then, because

$$F(u) = \left[\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right] / \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right], \tag{3.1}$$

[13] This is called pseudorandom because the computer process used to generate U is actually deterministic; it only appears to be random. There is a large literature on the properties of various pseudorandom number generators.

[14] Most computers have a library routine to generate standard uniform random variables. See, for example, Brian Ripley (1987) for a discussion of standard uniform random number generators.

Figure 3.1:

U can be simulated by solving equation (3.1) for u as

$$U = \sigma \Phi^{-1} \left\{ Z \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \right] + \Phi\left(\frac{a-\mu}{\sigma}\right) \right\} + \mu. \tag{3.2}$$

An alternative would be to draw U from the untruncated distribution and keep it if and only if $a < U < b$. This is the simplest version of an acceptance-rejection (AR) method (Luc Devroye 1986 or Ripley 1987). But AR methods frequently have poor properties. First, they can take a long time to simulate especially if $\Pr[a < U < b]$ is small. Second, they are usually not continuous in some set of parameters θ that might affect for example μ. In particular, consider a case where at some value θ, the rth draw is accepted, but at $\theta + \delta$, the rth draw is not accepted (because θ affects a, b, or the parameters of the distribution of U). Because the draws used for θ would be different than those used for $\theta + \delta$, the simulator would differ by a discrete amount. For these reasons, AR methods have limited application.

In Hotz et al. (1994), it is necessary to draw errors from discrete distributions corresponding to discrete choices and discrete random events. A small generalization of the inversion method works here. Assume $U = j$ with probability p_j for $j = 1, 2, ..., n$. Let $P_j = \Pr[U \leq j] = \sum_{k=1}^{j} p_k$. Then, letting $U = j$ if and only if $P_{j-1} < Z \leq P_j$ (where $P_0 = 0$) leads to the desired distribution for U. This is equivalent to inserting a distribution for a discrete random variable in Figure 1.

Random variables frequently can be simulated by using a composition formula. For example, because a binomial random variable is the sum of *iid* Bernoulli random variables, we can simulate a binomial random variable by simulating *iid* Bernoulli's and then adding them up. A more useful example is simulating multivariate $U \sim N[\mu, \Omega]$ such as in the multinomial probit example in Section 2.2. Let $Z \sim N[0, I]$, and let C be any matrix such that $CC' = \Omega$ (e.g., the Choleski decomposition of Ω). Then $U = CZ + \mu \sim N[\mu, \Omega]$.[15] So we can simulate U by simulating Z and then transforming it. In general, if $U = h(Z)$ for some specified function h, we can simulate U by simulating Z and then evaluating $h(Z)$.

3.1.2. Simulating $Eh(U)$ Directly

The most straightforward simulator for $Eh(U)$ is the direct simulator

$$\hat{E}h(U) = \frac{1}{R} \sum_{r=1}^{R} h(u^r) \tag{3.3}$$

where $u^r, r = 1, 2, ..., R$ are R *iid* pseudorandom draws of U. If a) h is continuous and differentiable with respect to θ and b) the simulator for U is continuous and differentiable with respect to θ, then $\hat{E}h(U)$ will be continuous and differentiable with respect to θ. Equation (3.3) is an unbiased simulator of $Eh(U)$, and its variance is $Var[h(U)]/R$. Note that as $R \to \infty$, the variance of the simulator → zero.

An early, often cited simulator of this type is used in Steven Lerman and Charles Manski (1981) to simulate the probabilities necessary for multinomial probit. Section 2.2 describes how to simulate $P_{ij} = \Pr[y_{ij} = 1 \mid X_i]$ which is represented in equation (2.21). But the simulator defined in equation (2.25) is neither continuous nor differentiable in θ. In particular, for most values of θ, small changes in θ change none of the conditions for j to be chosen ($u^*_{ijk} < X^*_{ijk}\beta \quad \forall k \neq j$), while for some values of θ, a small change in θ causes one of the conditions to change by a discrete amount (for some $k \neq j$, $u^*_{ijk} > X^*_{ijk}\beta$ at θ, but $u^*_{ijk} < X^*_{ijk}\beta$ at $\theta + \Delta$ for some small Δ). Therefore, $\hat{P}_{ij}$ is a step function in θ. The step function nature of the simulator makes it difficult to solve an orthogonality condition or maximize a likelihood function, and it makes it difficult to approximate the covariance matrix of the estimator (see Section 3.2). Also, especially for small P_{ij}, R must be large to guarantee that $\hat{P}_{ij}$ will be positive which is necessary for simulation of the log likelihood function. Thus, this simulator does not work well for the multinomial probit problem especially when there are many alternatives (J is large); better simulators are discussed below.

Most of the applications discussed in Section 2 use a special case of the direct simulator defined in equation (3.3). Berkovec and Stern (1991) and Berry, Levinsohn, and Pakes (1995) need to simulate probabilities of choices conditional on some unobserved heterogeneity components. Both assume that the conditional choice probabilities are multinomial logit:

$$P_j(u) = \exp\left\{\bar{V}_j + u_j\right\} / \sum_k \exp\left\{\bar{V}_k + u_k\right\} \tag{3.4}$$

where $\bar{V}_k$, $k = 1, 2, ..., J$, are the deterministic components of the values of the choices available. Given draws of $\{u^r_k\}_{r=1}^{R}$, one can simulate $EP_j(U)$ as

$$\hat{E}P_j(U) = \frac{1}{R} \sum_{r=1}^{R} \left[\exp\left\{\bar{V}_j + u^r_j\right\} / \sum_k \exp\left\{\bar{V}_k + u^r_k\right\}\right]. \tag{3.5}$$

[15] U is normal because Z is normal. $EU = E(CZ + \mu) = CEZ + \mu = \mu$. $E(U - \mu)(U - \mu)' = ECZZ'C' = CEZZ'C' = CC' = \Omega$.

Pakes (1986) directly simulates the patent nonrenewal probabilities $\pi\left(t,j\right)$ discussed in Section 2.3. Pakes simulates draws of $\{\rho_t^r\}_{t=1}^T$ and observes the first period t where $\rho_t^r < \bar{\rho}_t$; this is the nonrenewal period for draw r. Pakes constructs a frequency simulator by counting the proportion of draws where nonrenewal occurs at t for each t. In particular, he uses the following algorithm to simulate $\pi\left(t,j\right)$:

```
A. Initialize the simulated patent nonrenewal probabilities $\hat{\pi}_j\left(t\right)$ to 0 for $t = 1, 2, ..., T$.
B. Do for $r = 1, 2, ..., R$:
    1. Simulate the first period realization of the patent return $\rho_1^r \sim N\left(\mu, \sigma_r\right)$.
    2. For each $t = 2, 3, ..., T$ simulate $\rho_t^r$ by:
        a)(simulate a new potential use with value $Z_t^r$ and pick the best use)
            Simulate $Z_t^r \sim G_t\left(z\right)$ introduced in Section 2.3 and set $\rho_t^r = \max\left[\delta\rho_{t-1}^r, Z_t^r\right]$.
        b)(check to see if a nonrenewal occurs) If $\rho_t^r < \bar{\rho}_t$ or $t = T$,
            then, for this draw, nonrenewal occurs at $t$, so increment $\hat{\pi}_j\left(t\right)$
            by one and do next $r$.
C. Divide $\hat{\pi}_j\left(t\right)$ by $R$ for $t = 1, 2, ..., T$.
```

Once Pakes can simulate $\pi\left(t,j\right)$, he can substitute the simulated $\pi\left(t,j\right)$ for the true $\pi\left(t,j\right)$ in his log likelihood function in equation (2.35) and maximize it over the parameters of the model. This is an application of maximum simulated likelihood discussed in general in Section 3.2.2.

Berry (1992) directly simulates the expected number of entrants in market i. Using the notation in Section 2.4, he simulates e_i^r and u_{ik}^r for each market i and firm k, computes $\nu_i(N)$ in equation (2.37), and $\pi_{ik}\left(s_i\right)$ in equation (2.36), and then the equilibrium strategy of each firm, implying N_i for each market. This provides an unbiased estimator of EN_i whose variance can be reduced by repeating the same experiment R independent times (the variance is proportional to R^{-1}). Unfortunately, this simulator is not continuous in the parameters θ.

3.1.3. Importance Sampling

Several methods allow us to improve the performance of a simulator significantly either in terms of reduced variance, better smoothness properties, and/or better computation time properties. One of these methods, importance sampling, uses a well-chosen distribution to simulate $Eh(U)$ when it is either difficult to draw U from its distribution F or when h is not smooth. In some cases, one can write equation (1.1) as

$$Eh(U) = \int \frac{h(u) f(u)}{g(u)} g(u) \, du \tag{3.6}$$

where $g(u)$ is a density such that a) it is easy to draw U from g; b) f and g have the same support; c) it is easy to evaluate $h(u)f(u)/g(u)$ given u; and d) $h(u)f(u)/g(u)$ is bounded and smooth in the parameters over the support of U. Note that equation (3.6) is $E[h(U)f(U)/g(U)]$ where $U \sim g$. The importance sampling simulator for $Eh(U)$ is

$$\hat{E}h(U) = \frac{1}{R} \sum_{r=1}^{R} \frac{h(u^r) f(u^r)}{g(u^r)} \tag{3.7}$$

where u^r, $r = 1, 2, ..., R$ are R draws from g. The purpose of conditions (a) and (c) are to increase computational speed. The purpose of condition (d) is variance bounding and smoothness. Smoothness makes optimization over parameters in $h(\bullet)$ and $f(\bullet)$ much more efficient; derivative based optimization algorithms can be used only if the simulator is smooth. Essentially, we are oversampling the "important" parts of the support of U by choosing the importance sampling density g carefully; thus the name importance sampling. This is seen more clearly in Figure 2. Assume we want to simulate equation (1.1) with $h(u) f(u)$ drawn in the figure. The optimal importance sampling density would be a density with the same support as f but proportional to $h(u) f(u)$. If such a density $\tilde{g}$ could be found, there would be no variation in $h(u) f(u) / \tilde{g}(u)$ over u, and so the simulator would have a zero variance. Deriving such a density involves computing $Eh(U)$ which we assumed was not feasible. But in Figure 2, $g(u) = \exp\{-u\}$ has a shape similar to $h(u) f(u)$, so the variation in $h(u) f(u) / g(u)$ over u is significantly smaller than the variation in $h(u)$.

(INSERT FIGURE 2 APPROXIMATELY HERE)

Consider simulating $\Pr[y_{ij} = 1 \mid X_i]$ for the multinomial probit problem. Let $\phi\left(\bullet \mid \Omega_j^*\right)$ be the multivariate density function (with covariance matrix Ω_j^*) for u_{ij}^*. Equation (2.21) can be written as

$$P_{ij} = \int_{-\infty}^{X_{ijJ}^*\beta} \cdots \int_{-\infty}^{X_{ij1}^*\beta} \left[\phi\left(u_j^* \mid \Omega_j^*\right) / g\left(u_j^*\right)\right] g\left(u_j^*\right) du_{j1}^* \cdots du_{jJ}^* \tag{3.8}$$

for some multivariate density g satisfying Conditions (a) through (d). Consider g where the kth element of u_j^* is distributed independently truncated normal with upper truncation point $X_{ijk}^*\beta$ and variance Ω_{jkk}^* for each k. The candidate g satisfies Conditions (a), (b), and (c), and is smooth over the support. But $h(u)f(u)/g(u)$ is not bounded especially when Ω_j^* has large off-diagonal terms. Thus, this choice of g may be problematic in that if a draw of u is chosen where $\mid h(u)f(u)/g(u) \mid$ is very large, then the simulator will explode. In fact, in general, it is the boundedness condition that is difficult to satisfy. For the multinomial probit problem, the Geweke-Keane-Hajivassiliou (GHK) and decomposition simulators discussed below both can be thought of as importance sampling simulators that satisfy all four of the conditions.

3.1.4. GHK Simulator

The GHK simulator, developed by Geweke (1991), Hajivassiliou (1990), and Keane (1994), has been found to perform very well in Monte Carlo studies for simulating multinomial probit probabilities. (Börsch-Supan and Hajivassilliou 1993; Geweke, Keane, and Runkle 1994, 1996; Hajivassilliou, McFadden, and Ruud 1996; Keane 1994; and Stern forthcoming.) The GHK algorithm switches back and forth among a) computing univariate, truncated normal probabilities; b) simulating draws from univariate, truncated normal distributions; and c) computing normal distributions conditional on previously drawn truncated normal random variables. Because each step is straightforward and fast, the algorithm can decompose the more difficult problem into a series of feasible steps. First, I illustrate the GHK algorithm in terms of simulating a bivariate normal probability. Consider the "isodensity" curves drawn in Figure 3. Each of these curves represents the locus of points with the same density where the density is bivariate normal with negative means and a positive covariance: $U \sim N[\mu, \Omega]$. Consider simulating the $\Pr[U > 0]$. This is the volume under the density curve in the upper right quadrant. Using the GHK algorithm, we compute $\hat{p}_1 = \Pr[U_1 > 0]$ which can be done analytically: $\hat{p}_1^r = \Phi\left[\mu_1/\sqrt{\Omega_{11}}\right]$. Then we simulate u_1^r from a left truncated normal using equation (3.2) where the truncation point is C in Figure 3. Let u_1^r in Figure 3 be such a draw. Next, we can compute the distribution of $U_2 \mid u_1^r$: $U_2 \mid u_1^r \sim N[\mu_2 + (\Omega_{12}/\Omega_{11})(u_1^r - \mu_1), \Omega_{11}(1-\rho^2)]$ where $\rho = \Omega_{12}/\sqrt{\Omega_{11}\Omega_{22}}$ is the correlation between U_1 and U_2. Given this distribution, we can compute $\hat{p}_2^r = \Pr[U_2 > 0 \mid u_1^r] = \Phi\left[(\mu_2 + (\Omega_{12}/\Omega_{11})(u_1^r - \mu_1))/\sqrt{\Omega_{11}(1-\rho^2)}\right]$. The GHK simulator of $\Pr[U > 0]$ is $\frac{1}{R}\sum_{r=1}^{R} \hat{p}_1^r \hat{p}_2^r$.

(INSERT FIGURE 3 APPROXIMATELY HERE)

This simulator is of no significant value when dealing with a bivariate normal (because there are accurate numerical methods for approximating bivariate normal probabilities). But the method generalizes for problems of higher dimension. Define k to be an index of which condition $u_{jk}^* < X_{ijk}^*\beta$ is being evaluated, μ to be the conditional mean, σ^2 to be the conditional variance, and $\hat{P}_{ij}$ to be the simulator of P_{ij}. The algorithm to simulate $\Pr\left[u_j^* < X_{ij}^*\beta\right]$ is:

A. Set $k = 1$, $\mu = 0$, $\sigma^2 = \Omega_{jkk}^*$, and $\hat{P}_{ij} = 1$.

B. Compute $p = \Pr\left[u_{jk}^* < X_{ijk}^*\beta\right] = \Phi\left[\left(X_{ijk}^*\beta - \mu\right)/\sigma\right]$ analytically, and increment $\hat{P}_{ij} = \hat{P}_{ij}p$.

C. Draw u_{jk}^* from a truncated normal distribution with mean μ, variance σ^2, and upper truncation point $X_{ijk}^*\beta$.

D. If $k < J - 1$, increment k by 1; otherwise go to (G).

E. Compute (analytically) the distribution of u^*_{jk} conditional on $u^*_{j1}, u^*_{j2}, ..., u^*_{jk-1}$. Note that this is normal with an analytically computable mean vector μ and variance σ^2.

F. Go to (B).

G. $\hat{P}_{ij}$ is the simulator.

The algorithm relies upon the fact that normal random variables conditional on other normal random variables are still normal. The GHK simulator is strictly bounded between $(0, 1)$ because each increment to $\hat{P}_{ij}$ is strictly bounded between $(0, 1)$. It is continuous and differentiable with respect to θ because each increment to $\hat{P}_{ij}$ is continuous and differentiable. Its variance is smaller than the frequency simulator described in Section 2.2 because each draw of $\hat{P}_{ij}$ is strictly bounded between $(0, 1)$ while each draw of the frequency simulator is either 0 or 1. A minor modification of the algorithm provides draws of normal random variables u^*_j conditional on j being chosen.[16] Other minor modifications are useful for related problems.

Some intuition for the GHK algorithm is that one is decomposing the joint density of the errors into a product of conditional densities. Such intuition might lead one to think that the GHK algorithm provides a "good" simulator of draws from the joint density. This is wrong because, when simulating from the first marginal density (when $k = 1$ in the algorithm), one ignores the information that the later conditional densities provide about the first marginal density. This is seen in Figure 3 for a bivariate normal problem. As before, assume that $U \sim N[\mu, \Omega]$ and again consider simulating $\Pr[U > 0]$. The GHK simulator would first compute $\hat{p}^r_1 = 1 - \Phi\left[-\mu_1/\sqrt{\Omega_{11}}\right]$, then simulate u^r_1 from its marginal, left truncated distribution, then compute $\hat{p}^r_2 = \Pr[U_2 > 0 \mid u^r_1] = 1 - \Phi\left[-\left(\mu_2 + \left(\Omega_{12}/\Omega_{11}\right)\left(u^r_1 - \mu_1\right)\right)/\sqrt{\Omega_{11}\left(1 - \rho^2\right)}\right]$, and then multiply $\hat{p}^r_1$ and $\hat{p}^r_2$ for the rth draw. This algorithm would oversample points near A in Figure 3 relative to points near B even though points near B have higher density. This occurs because u^r_1 is simulated from its marginal distribution ignoring correlation between U_1 and U_2. Even though the GHK algorithm does not simulate from the joint density, it still provides an unbiased simulator of P_{ij} because it is an importance sampling algorithm.

3.1.5. Decomposition Simulators

Next, two decomposition simulators are described. The Stern (1992) simulator uses the property that the sum of two normal random vectors is also normal. The goal is to simulate $\Pr\left[u^*_j < V^*_j\right]$. For the multinomial probit problem, V^*_j is a vector whose kth element is $X^*_{ijk}\beta$. Decompose $u^*_j = Z_1 + Z_2$ where $Z_1 \sim N[0, \lambda]$, $Z_2 \sim N\left[0, \Omega^*_j - \lambda\right]$, Z_1 and Z_2 are independent, and λ is chosen to be a diagonal matrix as large as possible such that $\Omega^*_j - \lambda$

[16] One needs to be careful how these draws are used. See the next paragraph.

is positive definite.[17] Then equation (2.21) can be written as

$$P_j = \int \Pr\left[Z_1 < V_j^* - z_2\right] g\left(z_2\right) dz_2 = \int \prod_k \Phi\left(\frac{V_{jk}^* - z_{2k}}{\lambda_k}\right) g\left(z_2\right) dz_2 \tag{3.9}$$

where g is the joint normal density of Z_2. The equality in equation (3.9) holds because the elements of Z_1 are independent. Equation (3.9) can be simulated as

$$\frac{1}{R}\sum_{r=1}^{R}\prod_k \Phi\left(\frac{V_{jk}^* - z_{2k}^r}{\lambda_k}\right) \tag{3.10}$$

where $z_{2k}^r, k = 1, 2, ..., J-1$, are pseudorandom draws of Z_2. The Stern simulator has all of the theoretical properties of the GHK simulator that are discussed above. So which one performs better is an empirical matter. Most Monte Carlo studies (Börsch-Supan and Hajivassiliou 1993; and Hajivassiliou, McFadden, and Ruud 1996) suggest that the GHK simulator provides simulators of P_j with smaller bias and variance (and therefore parameter estimators with smaller bias and variance), but Stern (1994, forthcoming) suggest cases where the Stern simulator has smaller bias and variance or where the simulation error is second order relative to small sample bias.

Another decomposition simulator for the multinomial probit problem, suggested by McFadden (1989), changes the specification of equation (2.16) to

$$y_{ij}^* = X_{ij}\beta + u_{ij} + \tau e_{ij} \tag{3.11}$$

where τ is a small number and $e_{ij} \sim iid$ Extreme Value. In the limit, as $\tau \to 0$, $\Pr\left[y_{ij} = 1 \mid X_i\right]$ converges to a multinomial probit probability. But for any $\tau > 0$,

$$\Pr\left[y_{ij} = 1 \mid X_i\right] = \int \left[\exp\left\{\frac{X_{ij}\beta + u_{ij}}{\tau}\right\} / \sum_k \exp\left\{\frac{X_{ik}\beta + u_{ik}}{\tau}\right\}\right] \phi\left(u_i \mid \Omega\right) du_i \tag{3.12}$$

where $\phi\left(\bullet \mid \Omega\right)$ is the joint multivariate normal density (with covariance matrix Ω) of the u_i's. The term $\exp\left\{\frac{X_{ij}\beta+u_{ij}}{\tau}\right\} / \sum_k \exp\left\{\frac{X_{ik}\beta+u_{ik}}{\tau}\right\}$ in equation (3.12) is the $\Pr\left[y_i = 1 \mid X_i, u_i\right]$ (because once we condition on u_i, the only randomness left is the extreme value error e_{ij} terms). Equation (3.12) is the expected value (over draws of U) of the multinomial logit probability and can be simulated directly. The idea in McFadden (1989) is to think of equation (3.11) as a kernel-type approximation of equation (2.16) for small "bandwidth" τ. This idea is used in Maxim Engers and Stern (1996). However, assuming that equation (3.11) is the true structure (where τ is a parameter that sometimes can be estimated) takes away no flexibility and frequently eases simulation because multinomial logit probabilities are easy to evaluate. Multivariate normality is a desirable assumption because of its flexible covariance matrix. But there are very few applications where theory dictates that the error in equation (2.16) should be multivariate normal. Berkovec and Stern (1991), Stern (1993), and Berry, Levinsohn, and Pakes (1995) all use the McFadden specification as the "true" specification because it is computationally convenient.

[17] An easy way to pick λ is to set each diagonal element of λ equal to the smallest eigenvalue of Ω_j^* minus a small amount.

3.1.6. Antithetic Acceleration[18]

Antithetic acceleration is a powerful variance reduction method (see Geweke 1988) that can be used in combination with other simulation methods. In any simulation method, there is some probability that the pseudorandom draws will be unusually large (or small). Antithetic acceleration prevents such events from increasing the simulator variance. Consider the general problem of simulating $Eh(U)$ where $U \sim F$. Let $Z \sim$ Uniform$(0,1)$. Then $h\left(F^{-1}(Z)\right)$ is a simulator of $Eh(U)$ because $F^{-1}(Z) \sim U$. But $h\left(F^{-1}(1-Z)\right)$ is also a simulator of $Eh(U)$ (because $1 - Z \sim$ Uniform$(0,1)$ also). The antithetic acceleration simulator of $Eh(U)$ is

$$\hat{E}h(U) = \frac{1}{2R}\sum_{r=1}^{R}\left[h\left(F^{-1}(z^r)\right) + h\left(F^{-1}(1-z^r)\right)\right] \tag{3.13}$$

where z^r is a pseudorandom draw of Z. Note that when $F^{-1}(z^r)$ is unusually large, $F^{-1}(1-z^r)$ is unusually small, and the associated element of the sum in equation (3.13) will average out to $\left[h\left(F^{-1}(z^r)\right) + h\left(F^{-1}(1-z^r)\right)\right]/2$ which is neither unusually large nor small. When F is $N[0,\sigma^2]$, equation (3.13) becomes

$$\hat{E}h(U) = \frac{1}{2R}\sum_{r=1}^{R}\left[h(u^r) + h(-u^r)\right] \tag{3.14}$$

where u^r is a pseudorandom draw of U. Note that antithetic acceleration requires $2R$ evaluations of h for R independent draws of z^r. Thus we should compare the precision of antithetic acceleration with R draws to the precision of simulation without antithetic acceleration with $2R$ draws; both require the same number of evaluations of h. Still, for any F with symmetric density, if h is linear, the variance of $\hat{E}h(U)$ using antithetic acceleration is zero. For any F and monotone h, the variance of $\hat{E}h(U)$ with R draws and antithetic acceleration is smaller than the variance of $\hat{E}h(U)$ with $2R$ draws and no antithetic acceleration. Note that, with antithetic acceleration and R draws, the variance of the simulator is

$$Var\left[\hat{E}h\left(F^{-1}(Z)\right)\right] = \frac{Var\left[h\left(F^{-1}(Z)\right)\right] + Cov\left[h\left(F^{-1}(Z)\right), h\left(F^{-1}(1-Z)\right)\right]}{2R}. \tag{3.15}$$

In general, the covariance term in equation (3.15) needs to be negative for antithetic acceleration to reduce variance. A sufficient condition for this is that h is monotone. If $Eh(U)$ is being simulated to estimate a parameter θ with N observations and h is monotone, then the increase in $Var\left(\hat{\theta}\right)$ due to simulation when antithetic acceleration is used is of order $(1/N)$ times the increase in $Var\left(\hat{\theta}\right)$ due to simulation when antithetic acceleration is not used. The value of this result is discussed in Section 3.2.

There are simulation problems where antithetic acceleration does not help. For example, let $U \sim N[0,\sigma^2]$, and let $h(U) = U^2$. Then the $Var\left[\hat{E}h(U)\right]$ with antithetic acceleration

[18] The reader might be interested in the related idea of control variates. See David Hendry (1984) for a description and Brian Brown and Whitney Newey (forthcoming) for a nice application.

and R draws is greater than that without antithetic acceleration and $2R$ draws. This is because $h(-U) = h(U)$ which means that the variance is twice as great as with no antithetic acceleration and $2R$ draws. In general, deviations from monotone h will diminish the relative performance of antithetic acceleration. But John Hammersley and David Handscomb (1964) suggest generalizations of antithetic acceleration that will reduce variance for more general h.

Consider applying antithetic acceleration to the entry problem in Berry (1992). The errors in Berry are the e_i's and u_{ik}'s in equation (2.37); let $\upsilon \sim N\left(0, \Psi\right)$ be a set of complete errors for the model, and let $\hat{\upsilon}$ be a simulator for υ. Conditional on $\hat{\upsilon}$, one can evaluate $\pi_{ik}\left(s_i\right)$ in equation (2.36) and determine the number of entrants N. But if $\hat{\upsilon}$ is unusually large (small), then profits conditional on entry will be large (small) leading to unusually high (low) entry. The effect of such a draw can be minimized by using $-\hat{\upsilon}$ as a draw of υ to average out deviations from EN.

3.2. Estimation Methods

In this section, we use the simulators developed in Section 3.1 in some estimation methods. Two different estimation methods are discussed: method of simulated moments (MSM) and maximum simulated likelihood (MSL).[19] Each method is described, and its theoretical properties are discussed. The section ends with a comparison of the different estimation procedures.

3.2.1. Method of Simulated Moments

Many estimation problems involve finding a parameter vector θ that solves a set of orthogonality conditions

$$0 = Q'h\left(y, X \mid \theta\right) \tag{3.16}$$

where Q is a set of instruments. Such estimators are called method of moments (MOM) estimators. All least squares methods are special cases of equation (3.16), and many maximum likelihood score equations can be thought of as moment conditions. For example, Robert Avery, Lars Hansen, and Hotz (1983) suggest how to recast the multinomial probit problem as a MOM problem where $h(y, X|\theta)$ is the vector $y - E(y|X)$.

In many MOM problems, the orthogonality condition can not be evaluated analytically. For example, in the multinomial probit problem, evaluating $E[y|X]$ involves evaluating equation (2.21). MSM replaces $h(y, X|\theta)$ with an unbiased simulator $\hat{h}(y, X|\theta)$ and then finds

[19] Two other methods receiving significant attention in the literature are the method of simulated scores (Hajivassiliou and McFadden 1990; and Hajivassiliou 1992) and Gibbs sampling (Stuart Geman and Donald Geman 1984; Martin Tanner and Wing Wong 1987; Alan Gelfand and Adrian Smith 1990; George Casella and Edward George 1992; James Albert and Siddhartha Chib 1993; Robert McCulloch and Peter Rossi 1994, 1995; and Geweke and Keane forthcoming). The discussion of these two methods are limited here because of space limitations and because, as they are relatively new estimators, knowledge about them is still changing very rapidly.

the θ that solves

$$0 = Q'\hat{h}\left(y, X \mid \theta\right). \tag{3.17}$$

The θ that solves equation (3.17) is the MSM estimator of θ, $\hat{\theta}$. McFadden (1989) and Pakes and Pollard (1989) show, using an argument similar to the one in Sections 2.2 and 2.3, that, as long as $\hat{h}(y, X|\theta)$ is an unbiased simulator of $h(y, X|\theta)$, deviations between $\hat{h}$ and h will wash out by the Law of Large Numbers (because equation (3.17) is linear in $\hat{h}$) and $plim(\hat{\theta}) = \theta$ as the sample size $N \to \infty$ even for small R.[20] The asymptotic covariance matrix for $\hat{\theta}$ is

$$D\left(\hat{\theta}\right) = \Xi\,\mathrm{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i' D\left(y_i - \xi_i\right) Q_i\right]\Xi' \tag{3.18}$$

where

$$\Xi = \mathrm{plim}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i' \frac{\partial h(y, X|\theta)}{\partial \theta}\right]^{-1}, \tag{3.19}$$

$\xi = \hat{h}(y, X|\theta) - h(y, X|\theta)$ is the residual caused by simulation, and $D\left(y_i - \xi_i\right) = D\left(y_i\right) + D\left(\xi_i\right)$ (McFadden 1989, p. 1006). If $\hat{h}(y, X|\theta)$ is simulated directly, then $D\left(\xi_i\right) = D\left(y_i\right)/R$ as in the multinomial probit discussion in Section 2.2. Again, note that for this case simulation increases the asymptotic covariance matrix by a factor of $\left[1 + \frac{1}{R}\right]$, and that, as $R \to \infty$, the loss in efficiency associated with simulation disappears: $\left[1 + \frac{1}{R}\right] \to 1$. If a better simulator is used, $D\left(\xi_i\right) < D\left(y_i\right)/R$. For example, if antithetic acceleration is used, then the loss in precision due to simulation may become of order N^{-1} which requires no adjustment to the asymptotic covariance matrix. Most researchers do not adjust standard errors for simulation(see Berkovec and Stern 1991; Berry 1992; or Hotz et al. 1994 for exceptions); instead they use methods where simulation variance is known to be small a priori (Berry, Levinsohn, and Pakes 1995; and Stern 1993), use a large number of draws R (Pakes 1986), or measure sensitivity of estimates to R (Börsch-Supan et al. 1992). Berkovec and Stern (1991), Bong-Soo Lee and Beth Ingram (1991), Berry, Levinsohn, and Pakes (1995), Berry (1992), Darrell Duffie and Kenneth Singleton (1993), Hotz et al. (1994), Keane (1994), Bansal et al. (1995), and Berry, Levinsohn, and Pakes (1995) all use MSM and rely upon the asymptotic properties of MSM to provide consistent estimates.

There are two approaches to choosing good instruments. The first is to represent the score statistic in the form of equation (3.16) and then choose Q accordingly. This was done in the multinomial probit discussion in Section 2.2. If this can be accomplished, the MOM estimator that the MSM estimator converges to as $R \to \infty$ is asymptotically equivalent to a ML estimator. The second is to pick $Q = D^{-1}\left(y_i\right)\left(\partial h(y, X|\theta)/\partial\theta\right)$. This is essentially an application of the GLS idea applied to the general MOM estimation problem.

Below is a roadmap for using MSM (this is a generalization of the multinomial probit MSM algorithm described in Section 2):

A. Choose an initial values for θ. Make sure that the initial guess results in reasonable moments.

[20] Extra technical conditions are found in McFadden (1989) and Pakes and Pollard (1989).

B. Choose a simulator.

C. Simulate $2NKR$ random variables where K is the number of random variables needed per observation. Store NKR of them in an instruments random number file and NKR in an estimation random number file. These random numbers will be used throughout the estimation process and never changed.

D. Given the initial guess of θ and the instruments random number file, simulate the instruments. Store the simulated instruments.

E. Given the initial guess of θ, the simulated instruments, and the estimation random number file, solve equation (3.17) for θ. This is an MSM estimator of θ.

F. Given the initial MSM estimator, reperform steps (D) and (E) once.

Consider applying this algorithm to the estimation problem in Berry (1992). Berry needs to simulate EN_i where N_i is the number of entrants in market i. First, we make a guess of $\theta = (\beta, \delta, \rho, \alpha, \sigma)$. Next, we need a simulator for EN_i; this is described in Section 3.1.2. Berry uses exogenous variables as instruments and thus does not need to simulate them. The MSM estimator of θ minimizes $n^{-1} \sum_{i=1}^{n} \left(N_i - \hat{E}N_i\right)' W_i A_i W_i' \left(N_i - \hat{E}N_i\right)$ over θ where W_i is a matrix of instruments and A_i is a weighting matrix. Initially, A_i is set equal to an identity matrix; this provides a consistent estimator of θ. Then, given that estimator, A_i is an estimate of $D^{-1}\left[W_i'\left(N_i - \hat{E}N_i\right)\right]$ (which depends on θ). Using the updated A_i's provides a more efficient estimator of θ. An even more efficient estimor would use instruments associated with $\partial EN_i/\partial\theta$. But, Berry's $\hat{E}N_i$ is not differentiable. Therefore an optimization algorithm that does not require derivatives (such as a simplex algorithm) must be used; such algorithms are an order of magnitude slower than derivative-based optimization algorithms.

3.2.2. Maximum Simulated Likelihood

The basic idea in ML estimation is to maximize the log likelihood of the observed data over the vector of parameters to be estimated. ML estimators are consistent and efficient for a very large class of problems. Their asymptotic distribution is normal for a slightly smaller class of problems. However, there are many likelihood functions that can not be evaluated analytically. In many cases, they can be thought of as expected values of some random function that can be simulated. In general, let $L = \sum_{i=1}^{N} L_i(\theta \mid y_i, X_i)$ be the log likelihood function where $L_i(\theta \mid y_i, X_i)$ is the log likelihood contribution for observation i. Assume L_i can not be evaluated analytically or numerically but it can be simulated. Let

$$\hat{L}_i(\theta \mid y_i, X_i) = \frac{1}{R} \sum_{r=1}^{R} \hat{L}_i^r(\theta \mid y_i, X_i) \tag{3.20}$$

be a simulator of L_i in the sense that $\hat{L}_i(\theta \mid y_i, X_i) \to L_i(\theta \mid y_i, X_i)$ as $R \to \infty$. Then the value of θ that maximizes equation (3.20) is the MSL estimator of θ. The MSL estimator of the multinomial probit parameters, described in Section 2.2, is used by Börsch-Supan et al. (1992) in a model of care for elderly parents. Pakes (1986) uses MSL to estimate his model of patent renewal. The log likelihood function for Pakes (1986) is given in equation (2.33), and it can be simulated given the direct method described in Section 3.1.1.

A property of maximum likelihood is that the score statistic, the derivative of the log likelihood function, should have an expected value of zero at the true value of θ.[21] This idea is the motivation behind the method of simulated scores (MSS) described in Hajivassiliou and McFadden (1990), Hajivassiliou (1992), and Hajivassiliou and Ruud (1994). The potential advantage of MSS is to use an estimator with the efficiency properties of ML and the consistency properties of MSM. The difficulty in this method is to construct an unbiased simulator of the score statistic.

3.3. Comparison of Methods

At this point, there are four simulation based estimation methods: MSM, MSL, MSS, and Monte Carlo Markov Chain (MCMC) methods. I have discussed MSM and MSL in some detail and have alluded to the other two methods. Of the four, MSS is the least developed. While it holds significant promise, there is not much practical experience with it, and its use requires more innovation on the part of the user than the other methods. Gibbs sampling and other MCMC methods have received significant attention, especially among Bayesians. Monte Carlo experiments in Geweke, Keane, and Runkle (1994, 1996) and empirical work in McCulloch and Rossi (1994) suggest that Gibbs sampling methods provide precise estimators in multinomial probit-type problems. But my experience (Stern forthcoming) suggests that Gibbs sampling methods are very expensive relative to MSM and MSL.

Prior to Börsch-Supan and Hajivassiliou (1993), MSM was the preferred method because MSL provides consistent estimators only if $R \to \infty$ as $N \to \infty$. However, Börsch-Supan and Hajivassiliou (1993) and supporting Monte Carlo studies (see, for example, Hajivassiliou, McFadden, and Ruud 1996) show that, even for small, fixed R, MSL provides precise estimators (at least in multinomial probit-type problems), in most cases studied more precise than MSM estimators. While these results rely upon the use of good simulators such as the GHK simulator, such simulators are easy to use. So, in terms of statistical performance, MSL is preferred in multinomial probit-type problems.

In other problems, one should use two criteria: a) statistical performance and b) ease of use and speed. For criterion (a), we know that MSM provides consistent estimators. The performance of MSL for fixed R can be verified only with experimentation or a Monte Carlo study. Such experimentation is straightforward to perform. Criterion (b) varies by application. For the multinomial probit problem, MSL is much easier than MSM because MSL requires simulation of only the choice probability corresponding to the observed alternative while MSM requires simulation of all choice probabilities. In other problems, it is very difficult to simulate the likelihood function (Berry 1992) or one is using an instrumental variables approach to avoid specifying exact functional forms for equations corresponding to

[21] An exception occurs when the support of the data depends upon θ.

endogenous regressors (Berkovec and Stern 1991; Duffie and Singleton 1993).

4. Significance of Results in the Literature Dependent on Simulation

In this section, for each of the applications described in Section 2, I reconsider the role of simulation in the paper, the results dependent upon simulation, and potential extensions of the paper. Then I consider some other important applications that were not discussed in Section 2.

4.1. Multinomial Probit

The first application discussed is the multinomial probit model. Börsch-Supan et al. (1992) have a model of living arrangements of elderly parents with 243 choices. Each period $t = 1, 2, ..., T$, the family must choose among 3 alternatives: the parent living alone, with children, or in a nursing home. The value of each alternative is

$$y_{it}^* = X_{it}\beta + \varepsilon_{it} \tag{4.1}$$

where $\varepsilon = (\varepsilon_{11}, \varepsilon_{21}, \varepsilon_{31}, \varepsilon_{12}, ..., \varepsilon_{3T}) \sim N[0, \Omega]$, i.e., errors are correlated over possibly choices and time. They need to simulate the probability of choosing the sequence of choices $\{y_{it}\}_{t=1}^{T}$ where $T = 5$. Thus there are $3^5 = 243$ alternatives. Börsch-Supan et al. consider eight different specifications for Ω. The first assumes $\Omega = I$, and the second allows for choice-specific random effects that are constant over time. Both of these specifications are simple enough to require no simulation methods. The other six specifications involve various combinations of AR(1) errors, choice-specific random effects, and contemporaneously correlated random effects. All of the specifications involve restrictions on Ω thus significantly reducing the number of parameters in Ω to estimate. Börsch-Supan et al. estimate the model for the first two specifications using ML estimation and the last six using MSL estimation (which is described in Sections 2.2 and 3.2.2). Without simulation, the authors would be restricted to very simple error structures. Börsch-Supan et al. show that more flexible error structures explain the data significantly better.[22] In particular, adding random effects significantly improves the likelihood function, allowing for an AR(1) structure does better yet, and allowing for both an AR(1) structure and random effects does even better. Within period correlation also has significant effects. There results show also that using small R (=3) provides estimates with small simulation error.[23] From a policy perspective, when better covariance structures are used, the time-invariant variables become less important and the time varying variables become more important. This result occurs probably because the better covariance structure captures some of the time-invariant preferences that can be captured only by time-invariant variables when very restricted covariance structures are used.

[22] The pseudo-R^2 increases from 40% to 60%.

[23] $R = 9$ does not change results much.

Simulation tempts one to estimate many parameters of the covariance matrix Ω. Yet identification of Ω is somewhat weak except in applications such as Börsch-Supan et al. where one can use panel data. Hausman and Wise (1978) use a very parsimonious error structure. McCulloch and Rossi (1994) estimate relatively flexible Ω structures using Gibbs sampling (see the footnote at the beginning of Section 3 for references on Gibbs sampling). But Geweke, Keane, and Runkle (1994) find that precision of parameter estimators declines significantly when many elements of Ω need to be estimated. Simulation provides no relief from identification problems.

4.2. Dynamic Programming

The second example, described in Section 2.3, Pakes (1986), uses simulation to evaluate the likelihood function for a complicated stochastic process explaining patent returns. Without simulation, Pakes' problem would be infeasible except for the simplest of stochastic processes. The empirical results in Pakes support the richness of his specification. First, his model fits the data significantly better than earlier deterministic models. Second, the model suggests that there is significant learning about the market value of inventions in the life of a patent and significant obsolescence later ($g_t(z)$ introduced below equation (2.34) has a large variance, and δ is significantly less than 1). Third, there are large returns to patenting. This result is probably biased in Pakes because the data really provides little information about the shape of the high tail of the distribution (the only data available is when a patent is not worth renewing). It relies heavily on a functional form assumption about an unobservable part of the return distribution.

Hotz et al. (1994) use simulation to evaluate expected values of value functions in a stochastic, discrete choice, dynamic programming model. They then use those simulated expected values in an orthogonality condition inside a MSM estimation procedure to estimate the parameters of the model. Part of their estimation procedure involves nonparametrically estimating future choice probabilities and state transition probabilities (see Hotz and Miller 1993) and using them to simplify evaluation of the value functions. They use Rust (1987) to evaluate their method and find that estimators are somewhat sensitive to the method used to nonparametrically estimate future choice probabilities and state transition probabilities. In many cases, biases are large and significant. Also, the method does not handle unobserved heterogeneity. The method relies upon the assumption that individuals (and the econometrician) can observe the behavior of others similar to them and therefore infer the relevant choice and transition probabilities. This is a valid assumption only if there is no significant unobserved heterogeneity. In Rust's GMC bus problem, the assumption of no unobserved heterogeneity may be reasonable;[24] in many other problems it is not. Given these problems, it remains to be seen how influential the method in Hotz et al. will become.

[24] It would not be true if the decision-maker has better information about particular engines than Rust does.

4.3. Entry

Berry (1992) uses simulation to evaluate the expected number of entrants in an airline market. Without simulation, his problem can be solved only for markets where there is a small number of potential entrants or for models with severe restrictions on the value of entering a market, equation (2.36). Berry estimates three models. One model makes the restrictive assumptions necessary to avoid the need for simulation. This model performs significantly worse than the other two models. The second model assumes an order of entry based on the ranking of profits, and the third lets the incumbent move first. Berry finds that the "incumbents" model performs the best. Results vary somewhat over the three specifications. Berry can use the model to predict actual entrants in a market although it is not clear how he deals with multiple equilibria with respect to this issue.

There are a number of interesting extensions one might consider to Berry. First, it is straightforward to allow for firm effects within the MSM estimation strategy (as Berry suggests). Second, it would be worthwhile to find a better simulator for his EN_i's so that more efficient optimization algorithms could be used. There are two problems with his simulator associated with it being a step function in the parameters of the model. The first is the points of discontinuity (where a small change in a parameter leads to one more entrant in one market). The second is the flats in parameter space. It is not obvious how to solve either problem.

4.4. Unobserved Heterogeneity

Berry, Levinsohn, and Pakes (1995) allow for unobserved heterogeneity because they are using only aggregate share data (along with information about the distribution of some personal characteristics in the market). Theoretically, there is some concern that the unobserved heterogeneity parameters in their model are not identified well (i.e., the statistical objective function might be relatively flat with respect to those parameters). However, they find significant interaction effects between unobserved personal characteristics and brand characteristics implying a high value to allowing for unobserved heterogeneity. One could solve their problem by using a good cross-section data set (Pinelopi Goldberg 1995). But one would still need simulation methods to solve multinomial probit type problems or they would have to make a restrictive error distribution assumption such as the nested logit assumption in Goldberg. Also, there are many problems for which good cross section data does not exist (Scafidi 1996).

Berkovec and Stern (1991) is a second example of using simulation to allow for unobserved heterogeneity. They use simulation to allow for unobserved heterogeneity in each individual's preferences over different types of jobs and returns to particular jobs. Their error specification, described in equation (2.40), allows for person-specific, type-of-job-specific, job-specific effects, and independent, idiosyncratic, time-specific effects. All but the last of these effects leads to unobserved heterogeneity that must be integrated out. The results suggest that the standard deviations of all of the effects leading to unobserved heterogeneity are very large relative to the standard deviation of the independent time-specific error (the only error that could be allowed in the model if simulation methods were not available). The inclusion of the unobserved heterogeneity effects significantly reduces the effect of job tenure variables on returns to a job (as would be predicted by matching models such as Jovanovic 1979), and they significantly affect the interpretation of hazard rates into retirement. The latter effect is the well-known result of ignored unobserved heterogeneity in hazard models causing negative duration dependence in the baseline hazard (Christopher Flinn and Heckman 1982).

The method of allowing for unobserved heterogeneity in Berkovec and Stern would be even more useful in other applications where unobserved heterogeneity is probably more important. A good example of this is models of contraception or child spacing (see Wolpin 1984 or Hotz and Miller 1988). Here the choices are which method of contraception to use. Whatever unobserved factors affecting choice exist are highly correlated over time. Ignoring this serial correlation leads to restricting the model to have choices that are independent (after conditioning on observed characteristics). Such an assumption would lead to possibly significant bias on coefficients associated with observed characteristics[25] and unnecessarily poor goodness-of-fit statistics. Whether unobserved heterogeneity is important in dynamic models is an empirical question. Possibly one could estimate a model without unobserved heterogeneity and test for its existence using a Lagrange Multiplier test and thus avoid simulation (if the null hypothesis of no unobserved heterogeneity is accepted). But frequently unobserved heterogeneity will exist, and then it needs to be controlled for.

5. Directions for New Applied Work

To date, the development of structural estimation has been limited by a number of factors including computational cost, data limitations, and skepticism among many economists that structural estimation has much to offer. Simulation will provide little help with data problems.[26] But it greatly expands the set of models that can be evaluated and estimated and at reasonable computer cost. Keane and Wolpin (1994) and Rust (1995) both suggest approximation methods relying to some degree on simulation to deal with the computational issues associated with evaluating dynamic programming models. See, for example, Pakes (1994) and Rust (1994) for surveys of dynamic structural estimation; both provide prominent roles for simulation methods. As more structural models are estimated, tested, and shown to be of value in explaining economic behavior, they will become more widely accepted.

Bayesian econometrics has always been hampered by the large computation costs associated with evaluating posterior distributions. To some degree, all economists and econometricians are Bayesians; it just has not been practical to use the concepts associated with Bayesian econometrics until recently. Simulation (and rapidly improving computer technology) knock down this computation hurdle. Geweke (1989, 1994, 1996) provides many ideas

[25] In nonlinear models, misspecification of the covariance matrix of the errors may lead to inconsistent estimates of all of the parameters.

[26] An exception is that simulation can be used to impute missing observations such as in Victor Lavy, Michael Palumbo, and Stern (1996).

about how simulation can be used to make Bayesian econometrics accessible to anyone with a reasonably good computer. Angelo Melino and Stuart Turnbull (1990) and Neil Shephard (1993) use Bayesian methods to estimate stochastic volatility models. Most of the theoretical development in simulation methods is associated with “Bayesian” methods such as Gibbs sampling and more generally Markov chain Monte Carlo methods (see Chib and Edward Greenberg 1996).

Other areas where simulation offers great potential include missing variables and errors in variables (Newey 1993; and Lavy, Palumbo, and Stern 1996), disequilibrium models, and duration models. Disequilibrium models were one of the first applications of simulation methods (Hajivassiliou 1987; and Guy Laroque and Bernard Salanie 1989). It has seen more work such as Hajivassiliou and Yannis Ioannides (1991). Duration analysis is moving toward dealing with more complicated unobserved heterogeneity structures (Lee Lillard 1993), and already some work (Keun Huh and Robin Sickles 1994; and Fabrizia Mealli and Stephen Pudney 1996) has used simulation.

Simulation brings countless models into the realm of feasibility. We have only begun to understand its potential. As we develop better simulation methods and more powerful computers, we will be able to deal with larger and richer problems.

References

[1] Albert, James H. and Chib, Siddhartha. “Bayesian Analysis of Binary and Polychotomous Data,” J. Amer. Statist. Assoc., June 1993, 88(422), pp. 669-79.

[2] Avery, Robert B.; Hansen, Lars Peter and Hotz V. Joseph . “Multiperiod Probit Models and Orthogonality Condition Estimation,” Int. Econ. Rev., Feb. 1983, 24(1), pp. 21-35.

[3] Bansal, Ravi et al. “Nonparametric Estimation of Structural Models for High-Frequency Currency Market Data,” J. Econometrics, Mar./Apr. 1995, 66(1/2), pp. 251-87.

[4] Berkovec, James and Stern, Steven. “Job Exit Behavior of Older Men,” Econometrica, Jan. 1991, 59(1), pp. 189-210.

[5] Berry, Steven T. “Estimation of a Model of Entry in the Airline Industry,” Econometrica, July 1992, 60(4), pp. 889-917.

[6] Berry, Steven; Levinsohn, James and Pakes, Ariel. “Automobile Prices in Market Equilibrium,” Econometrica, July 1995, 63(4), pp. 841-890.

[7] Börsch-Supan, Axel and Hajivassiliou, Vassilis A. “Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models,” J. Econometrics, Aug. 1993, 58(3), pp. 347-68.

[8] Börsch-Supan, Axel et al. “Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors,” in Topics in the economics of aging. Ed.: David Wise. Chicago and London: U. of Chicago Press, 1992, pp. 79-104.

[9] Bresnahan, Timothy F. and Reiss, Peter C. “Empirical Models of Discrete Games,” J. Econometrics, Apr./May 1991, 48(1/2), pp. 57-81.

[10] Brown, Brian and Newey, Whitney. “Simulation-Based Inference in Semiparametric Procedures,” in Simulation-based inference in econometrics: methods and applications. Ed.: Roberto S. Mariano, Melvyn Weeks, and Til Schuermann, Cambridge: Cambridge University Press, forthcoming.

[11] Bunch, David. “Estimability in the Multinomial Probit Model,” Transportation Research, Part B, Methodological, Feb. 1991, 25B(1), pp. 1- 12.

[12] Butler, J. S. and Moffitt, Robert. “A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model,” Econometrica, May 1982, 50(3), pp. 761-64.

[13] Casella, George and George, Edward. “Explaining the Gibbs Sampler,” American Statistician , Aug. 1992, 46(3), pp. 167-174.

[14] Chib, Siddhartha and Greenberg, Edward. "Markov Chain Monte Carlo Simulation Methods in Econometrics," Econ. Theory, Aug. 1996, 12(3),pp. 409-31.

[15] Devroye, Luc. Non-uniform random variate generation. New York: Springer-Verlag, 1986.

[16] Duffie, Darrell and Singleton, Kenneth J. "Simulated Moments Estimation of Markov Models of Asset Prices," Econometrica, July 1993, 61(4), pp. 929-52.

[17] Engers, Maxim and Stern, Steven. "Long-Term Care and Family Bargaining" Unpub. msc., U. of Virginia, 1996.

[18] Flinn, Christopher J. and Heckman, James J. "New Methods for Analyzing Structural Models of Labor Force Dynamics," J. Econometrics, Jan. 1982, 18(1), pp. 115-68.

[19] Gelfand, Alan and Smith, Adrian. "Sampling-Based Approaches to Calculating Marginal Densities," J. Amer. Statist. Assoc., June 1990, 85(410), pp. 398-409.

[20] Geman, Stuart and Geman, Donald. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Nov. 1984, 6(6), pp. 721- 41.

[21] Geweke, John. "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," J. Econometrics, May/June 1988, 38(1/2), pp. 73-89.

[22] Geweke, John. "Bayesian Inference in Econometric Models Using Monte Carlo Integration," Econometrica, Nov. 1989, 57(6), pp. 1317-39.

[23] Geweke, John. "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, 1991, pp. 571-78.

[24] Geweke, John. "Bayesian Comparison of Econometric Models," Unpub. msc., Federal Reserve Bank of Minneapolis, 1994.

[25] Geweke, John. "Simulation-Based Bayesian Inference for Economic Time Series," in Simulation- based inference in econometrics: methods and applications. Ed.: Roberto S. Mariano, Melvyn Weeks, and Til Schuermann, Cambridge: Cambridge University Press, forthcoming.

[26] Geweke, John. "Monte Carlo Simulation and Numerical Integration," in Handbook of computational economics. Eds.: Hans M. Amman, David A. Kendrick, and John Rust. Amsterdam and New York: Elsevier, 1996, pp. 731-800.

[27] Geweke, John and Keane, Michael. "Bayesian Inference for Dynamic Discrete Choice Models Without the Need for Dynamic Programming," in Simulation-based inference in econometrics: methods and applications. Ed.: Roberto S. Mariano, Melvyn Weeks, and Til Schuermann, Cambridge: Cambridge University Press, forthcoming.

[28] Geweke, John; Keane, Michael and Runkle, David. "Alternative Computational Approaches to Inference in the Multinomial Probit Model," Rev. Econ. Statist., Nov. 1994, 76(4), pp. 609-32.

[29] Geweke, John; Keane, Michael and Runkle, David. "Statistical Inference in the Multinomial Multiperiod Probit Model." Fed. Res. Bank of Minneapolis, Staff Report 177, 1996.

[30] Goldberg, Pinelope K. "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry," Econometrica, July 1995, 63(4), pp. 891-951.

[31] Hajivassiliou, Vassilis A. "The External Debt Repayments Problem of LDC's: An Econometric Model Based on Panel Data," J. Econometrics, Sept./Oct. 1987, 36(1/2), pp. 205-30.

[32] Hajivassiliou, Vassilis A. "Smooth Simulation Estimation of Panel Data LDV Models." Unpub. msc. Yale U., 1990.

[33] Hajivassiliou, Vassilis A. "The Method of Simulated Scores: A Presentation and Comparative Evaluation," Unpub. msc., Yale U., 1992.

[34] Hajivassiliou, Vassilis A. "Simulation Estimation Methods for Limited Dependent Variable Models," in Handbook of statistics. Eds: G. S. Maddala, C. R. Rao, and H. D. Vinod, Amsterdam and New York: North-Holland, 1993, pp. 519-44.

[35] Hajivassiliou, Vassilis and Ioannides, Yannis. "Switching Regressions Models of the Euler Equation: Consumption, Labor Supply, and Liquidity Constraints," Unpub. msc., Yale U., 1991.

[36] Hajivassiliou, Vassilis A. and McFadden, Daniel L. "The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises," Yale Cowles Foundation Discussion Paper No. 967, Dec. 1990.

[37] Hajivassiliou, Vassilis; McFadden, Daniel, and Ruud, Paul. "Simulation of Multivariate Normal Rectangle Probabilities and their Derivatives: Theoretical and Computational Results," J. Econometrics, May/June 1996, 72(2), pp. 85-134.

[38] Hajivassiliou, Vassilis A. and Ruud, Paul A.. "Classical Estimation Methods for LDV Models Using Simulation," in Handbook of econometrics. Eds: Robert F. Engle and Daniel L. McFadden, Amsterdam and New Yourk: North-Holland, 1994, pp. 2384-2441.

[39] Hammersley, John M. and Handscomb, David C. Monte Carlo methods. London: Methuen, 1964.

[40] Hausman, Jerry A. and Wise, David A. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," Econometrica, Mar. 1978, 46(2), pp. 403-26.

[41] Heckman, James. "Statistical Models for Discrete Panel Data," in Structural analysis of discrete data with econometric applications. Eds.:Charles F. Manski and Daniel McFadden. Cambridge, MA: MIT Press, 1981, pp. 114-78.

[42] Hendry, David. "Monte Carlo Experimentation in Econometrics," in Handbook of econometrics. Eds: Zvi Griliches and Michael Intriligator. Amsterdam and New York: North-Holland, 1984, pp. 939-76.

[43] Hotz, V. Joseph and Miller, Robert A. "An Empirical Analysis of Life Cycle Fertility and Female Labor Supply," Econometrica, Jan. 1988, 56(1), pp. 91-118.

[44] Hotz, V. Joseph and Miller, Robert A. "Conditional Choice Probabilities and the Estimation of Dynamic Models," Rev. Econ. Stud., July 1993, 60(3), pp. 497-529.

[45] Hotz, V. Joseph et al. "A Simulation Estimator for Dynamic Models of Discrete Choice," Rev. Econ. Stud., Apr. 1994, 61(2), pp. 265-89.

[46] Huh, Keun and Sickles, Robin C.. "Estimation of the Duration Model by Nonparametric Maximum Likelihood, Maximum Penalized Likelihood, and Probability Simulators," Rev. Econ. Statist., Nov. 1994, 76(4), pp. 683-94.

[47] Jovanovic, Boyan. "Job Matching and the Theory of Turnover," J. Polit. Econ., Oct. 1979, 87(5), pp. 972-90.

[48] Keane, Michael P. "Simulation Estimation for Panel Data Models with Limited Dependent Variables," in Handbook of statistics. Eds: G. S. Maddala, C. R. Rao, and H. D. Vinod. Amsterdam and New York: North-Holland, 1993, pp. 545-572.

[49] Keane, Michael P. "A Computationally Practical Simulation Estimator for Panel Data," Econometrica, Jan. 1994, 62(1), pp. 95-116.

[50] Keane, Michael P. and Wolpin, Kenneth I. "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," Rev. Econ. Statist., Nov. 1994, 76(4), pp. 648-72.

[51] Laroque, Guy and Salanie, Bernard. "Estimation of Multi-Market Fix-Price Models: An Application of Pseudo Maximum Likelihood Methods," Econometrica, July 1989, 57(4), pp. 831-60.

[52] Lavy, Victor, Palumbo, Michael and Stern, Steven. "Simulation of Multinomial Probit Probabilities and Imputation of Missing Data" Unpub. msc., U. of Virginia, 1996.

[53] Lee, Bong-Soo and Ingram, Beth Fisher. "Simulation Estimation of Time-Series Models," J. Econometrics, Feb./Mar. 1991, 47(2/3), pp. 197-205.

[54] Lerman, Steven and Manski, Charles. "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in Structural analsis of discrete data with econometric applications. Eds: Charles Manski and Daniel McFadden. Cambridge: MIT Press, 1981, pp. 305-19.

[55] Lillard, Lee A. "Simultaneous Equations for Hazards: Marriage Duration and Fertility Timing," J. Econometrics, Mar. 1993 56(1/2), pp. 189-217.

[56] Mariano, Roberto S.; Weeks, Melvyn and Schuermann, Til. Simulation-based inference in econometrics: methods and applications. Cambridge: Cambridge University Press, forthcoming.

[57] McCulloch, Robert E. and Rossi, Peter E. "An Exact Likelihood Analysis of the Multinomial Probit Model," J. Econometrics, Sept./Oct. 1994, 64(1/2), pp. 207-40.

[58] McCulloch, Robert E. and Rossi, Peter E. "Value of Household Purchase History Information" Unpub. msc. U. of Chicago, 1995.

[59] McFadden, Daniel. "Econometric Analysis of Qualitative Response Models," in Handbook of econometrics. Eds: Zvi Griliches and Michael Intriligator. Amsterdam and New York: North-Holland, 1984, pp. 1396-1457.

[60] McFadden, Daniel. "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," Econometrica, Sept. 1989, 57(5), pp. 995-1026.

[61] McFadden, Daniel and Ruud, Paul A. "Estimation by Simulation," Rev. Econ. Statist., Nov. 1994, 76(4), pp. 591-608.

[62] Mealli, Fabrizia and Pudney, Stephen. "Occupational Pensions and Job Mobility in Britain: Estimation of a Random-Effects Competing Risks Model," J. Applied Econometrics, May/June 1996, 11(3), pp. 293-320.

[63] Melino, Angelo and Turnbull, Stuart M. "Pricing Foreign Currency Options with Stochastic Volatility," J. Econometrics, July/Aug. 1990, 45(1/2), pp. 239-65.

[64] Miller, Robert A. "Job Matching and Occupational Choice," J. Polit. Econ., Dec. 1984, 92(6), pp. 1086-1120.

[65] Newey, Whitney K. “Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models” Unpub. msc., Massachusetts Institute of Technology, Dept. of Economics Working Paper: 93-18, Nov. 1993.

[66] Pakes, Ariel S. “Patents as Options: Some Estimates of the Value of Holding European Patent Stocks,” Econometrica, July 1986, 54(4), pp. 755-84.

[67] Pakes, Ariel S. “Dynamic Structural Models; Problems and Prospects: Mixed Continuous Discrete Controls and Market Interactions,” in Proceedings of the 6th world congress of the Econometric Society, Barcelona, Spain. Ed.: Christopher A. Sims. Cambridge: Cambridge University Press, 1994, pp. 171-259.

[68] Pakes, Ariel and Pollard, David. “Simulation and the Asymptotics of Optimization Estimators,” Econometrica, Sept. 1989, 57(5), pp. 1027-57.

[69] Ripley, Brian. Stochastic simulation. New York: John Wiley and Sons, 1987.

[70] Rust, John. “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher,” Econometrica, Sept. 1987, 55(5), pp. 999-1033.

[71] Rust, John. “Structural Estimation of Markov Decision Processes” in Handbook of econometrics. Eds.: Robert Engle and Daniel McFadden. Amsterdam and New York: North Holland, 1994, pp. 3082-146.

[72] Rust, John. “Maximum Likelihood Estimation of Discrete Control Processes,” SIAM Journal on Control and Optimization, Sept. 1988, 26(5), pp. 1006-23.

[73] Rust, John. “Using Randomization to Break the Curse of Dimensionality,” Econometrica, forthcoming.

[74] Scafidi, Benjamin. “Neighborhood, Housing, and School Choice,” Unpub. msc., U. of Virginia, 1996.

[75] Shephard, Neil. “Fitting Nonlinear Time-Series Models with Applications to Stochastic Variance Models,” J. Applied Econometrics, Dec. 1993, 8, pp. S135-52.

[76] Stern, Steven. “A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models,” Econometrica, July 1992, 60(4), pp. 943-52.

[77] Stern, Steven. “A Disaggregate Discrete Choice Model of Transportation Demand by Elderly and Disabled People in Rural Virginia,” Transportation Research, July 1993, 27A(4), pp. 315-27.

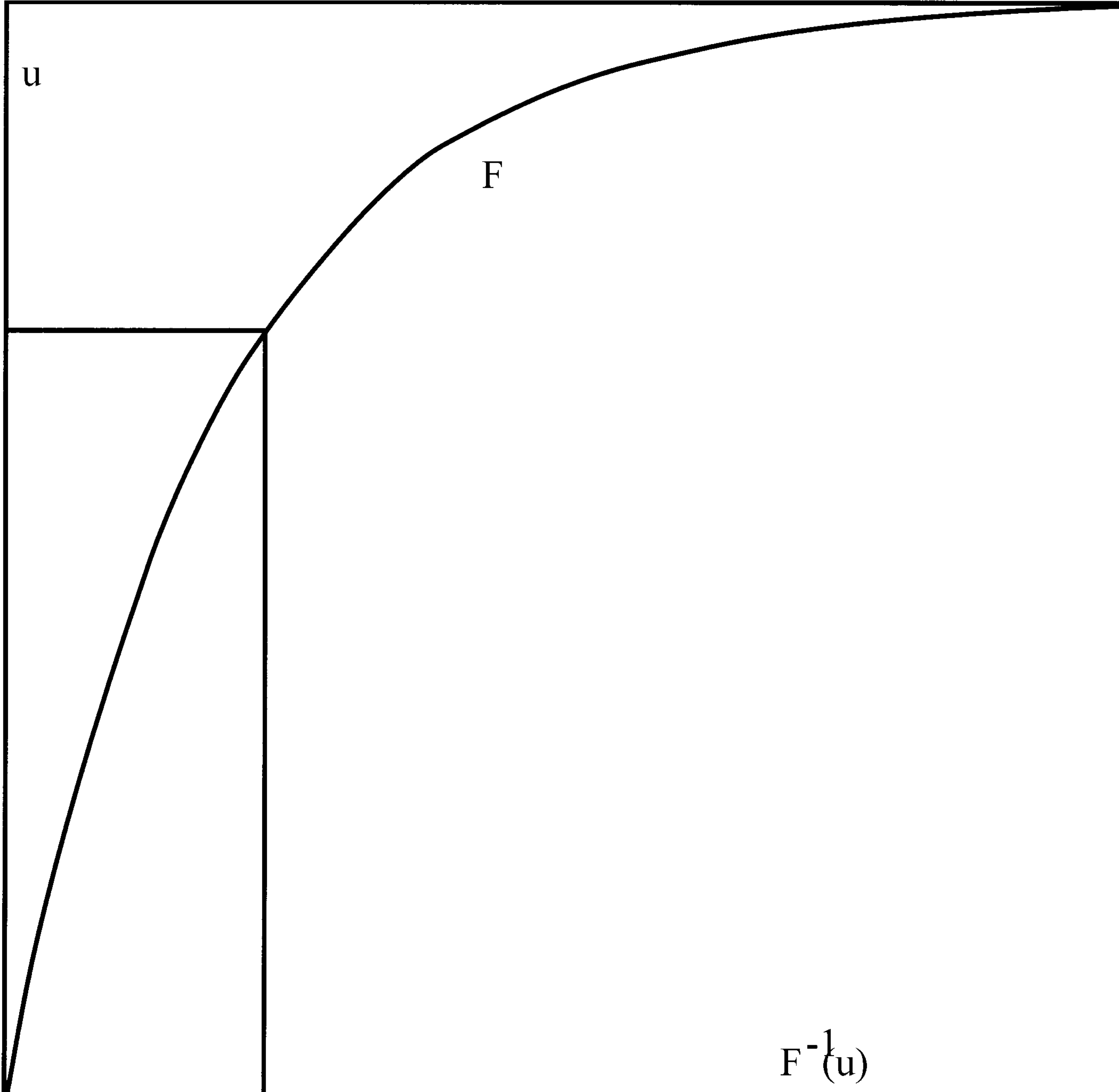
[78] Stern, Steven. “Two Dynamic Discrete Choice Estimation Problems and Simulation Method Solutions,” Rev. Econ. Statist., Nov. 1994, 76(4), pp. 695-702.

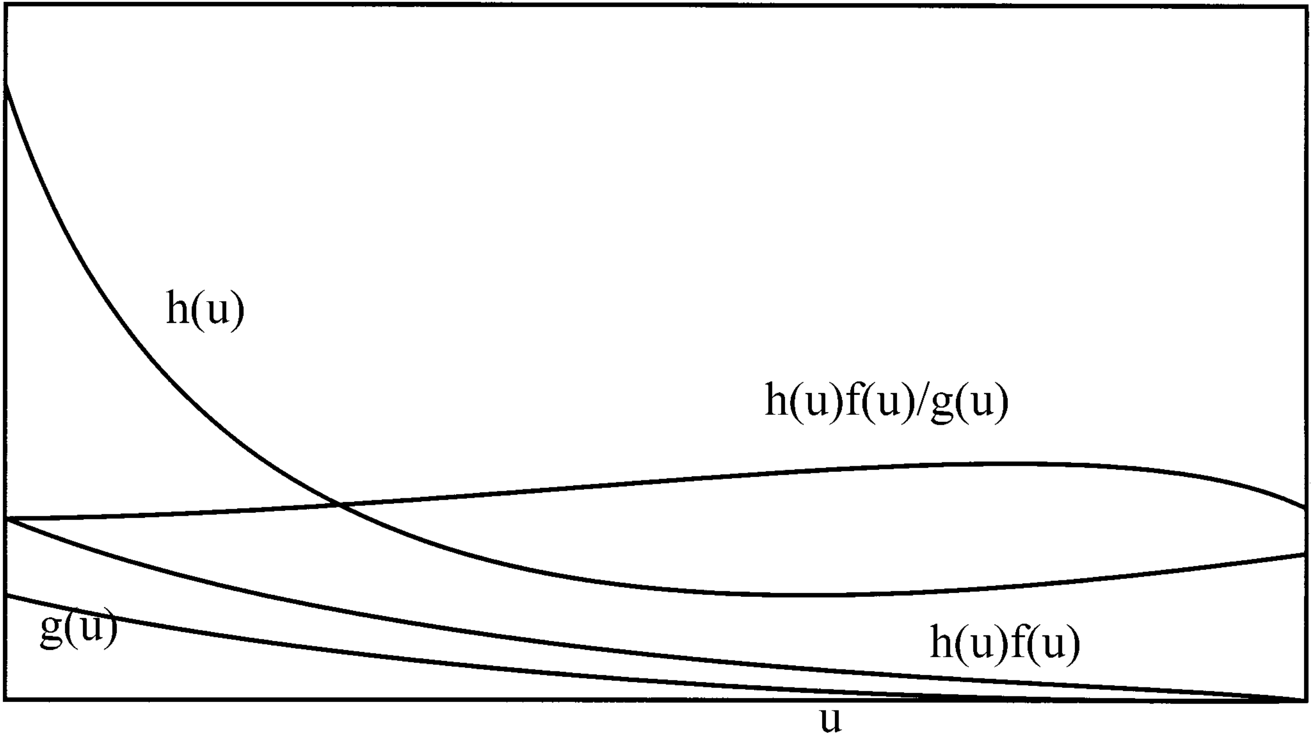
[79] Stern, Steven. “Simulation and Estimation: Motivation and Methods,” in Simulation-based inference in econometrics: methods and applications. Eds: Roberto S. Mariano, Melvyn Weeks, and Til Schuermann. Cambridge: Cambridge University Press, forthcoming.

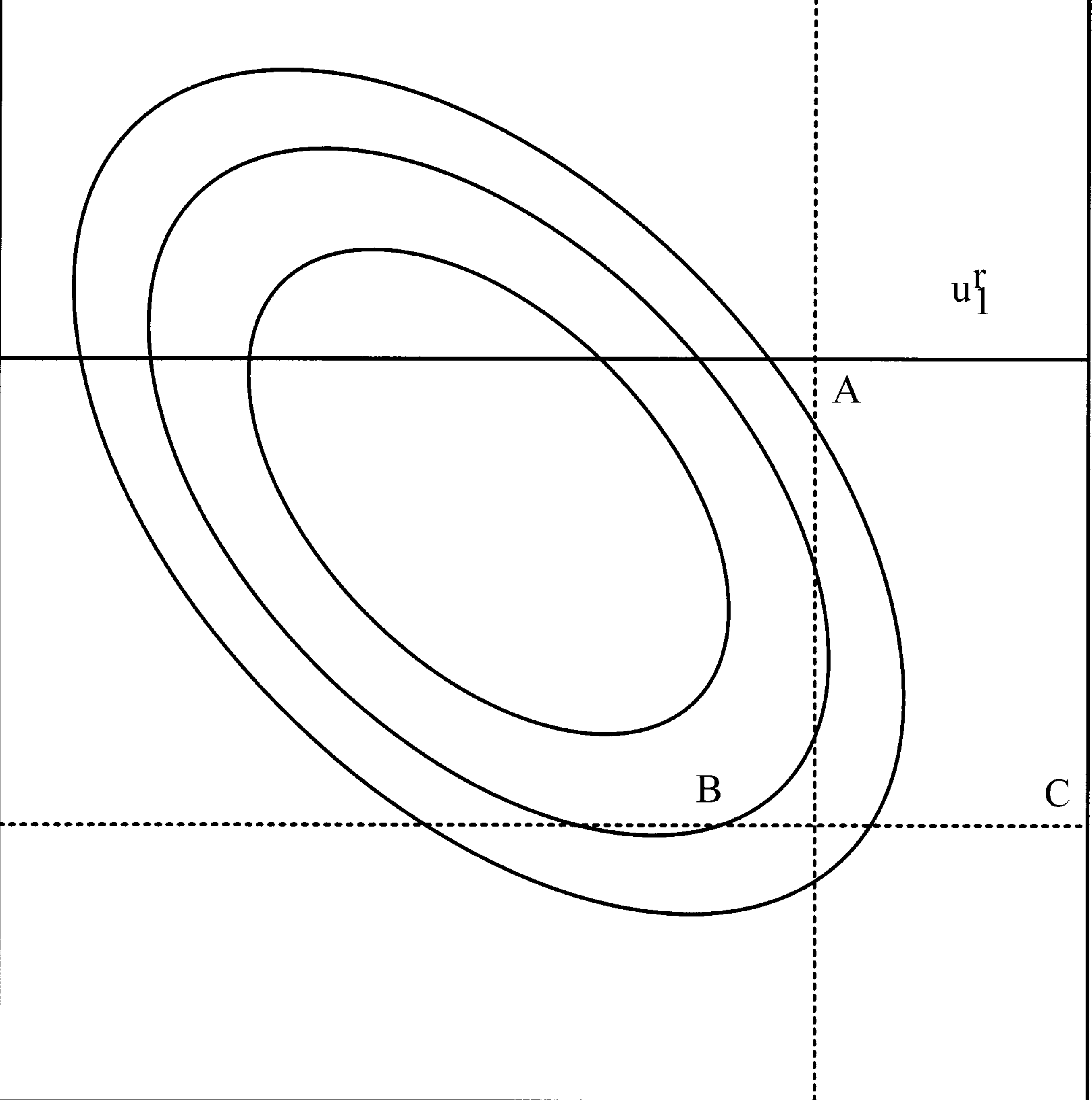
[80] Tanner, Martin A. and Wong, Wing Huing. “The Calculation of Posterior Distributions by Data Augmentation,” J. Amer. Statist. Assoc., June 1987, 82(398), pp. 528-50.

[81] Wolpin, Kenneth. “An Estimable Dynamic Stochastic Model of Fertility and Child Mortality,” J. Polit. Econ., Oct. 1984, 92(5), pp. 852-74.

[82] Wolpin, Kenneth. “Public-Policy Uses of Discrete-Choice Dynamic Programming Models,” Amer. Econ. Rev., May 1996, 86(2), pp. 427-32.


u
F
F⁻¹(u)


h(u)
h(u)f(u)/g(u)
g(u)
h(u)f(u)
u


u^r_l
A
B
C