10 Simulation-Assisted Estimation

10.1 Motivation

So far we have examined how to simulate choice probabilities but have not investigated the properties of the parameter estimators that are based on these simulated probabilities. In the applications we have presented, we simply inserted the simulated probabilities into the log-likelihood function and maximized this function, the same as if the probabilities were exact. This procedure seems intuitively reasonable. However, we have not actually shown, at least so far, that the resulting estimator has any desirable properties, such as consistency, asymptotic normality, or efficiency. We have also not explored the possibility that other forms of estimation might perhaps be preferable when simulation is used rather than exact probabilities.

The purpose of this chapter is to examine various methods of estimation in the context of simulation. We derive the properties of these estimators and show the conditions under which each estimator is consistent and asymptotically equivalent to the estimator that would arise with exact values rather than simulation. These conditions provide guidance to the researcher on how the simulation needs to be performed to obtain desirable properties of the resultant estimator. The analysis also illuminates the advantages and limitations of each form of estimation, thereby facilitating the researcher's choice among methods.

We consider three methods of estimation:

1. *Maximum Simulated Likelihood: MSL.* This procedure is the same as maximum likelihood (ML) except that simulated probabilities are used in lieu of the exact probabilities. The properties of MSL have been derived by, for example, Gourieroux and Monfort,(1993), Lee (1995), and Hajivassiliou and Ruud (1994).
2. *Method of Simulated Moments: MSM.* This procedure, suggested by McFadden (1989), is a simulated analog to the traditional method of moments (MOM). Under traditional MOM for discrete choice, residuals are defined as the difference between

the 0–1 dependent variable that identifies the chosen alternative and the probability of the alternative. Exogenous variables are identified that are uncorrelated with the model residuals in the population. The estimates are the parameter values that make the variables and residuals uncorrelated in the sample. The simulated version of this procedure calculates residuals with the simulated probabilities rather than the exact probabilities.

3. *Method of Simulated Scores: MSS.* As discussed in Chapter 8, the gradient of the log likelihood of an observation is called the score of the observation. The method of scores finds the parameter values that set the average score to zero. When exact probabilities are used, the method of scores is the same as maximum likelihood, since the log-likelihood function is maximized when the average score is zero. Hajivassiliou and McFadden (1998) suggested using simulated scores instead of the exact ones. They showed that, depending on how the scores are simulated, MSS can differ from MSL and, importantly, can attain consistency and efficiency under more relaxed conditions.

In the next section we define these estimators more formally and relate them to their nonsimulated counterparts. We then describe the properties of each estimator in two stages. First, we derive the properties of the traditional estimator based on exact values. Second, we show how the derivation changes when simulated values are used rather than exact values. We show that the simulation adds extra elements to the sampling distribution of the estimator. The analysis allows us to identify the conditions under which these extra elements disappear asymptotically so that the estimator is asymptotically equivalent to its nonsimulated analog. We also identify more relaxed conditions under which the estimator, though not asymptotically equivalent to its nonsimulated counterpart, is nevertheless consistent.

10.2 Definition of Estimators

10.2.1. Maximum Simulated Likelihood

The log-likelihood function is

$$\mathrm{LL}(\theta) = \sum_n \ln P_n(\theta),$$

where θ is a vector of parameters, $P_n(\theta)$ is the (exact) probability of the observed choice of observation n, and the summation is over a sample

of N independent observations. The ML estimator is the value of θ that maximizes $\text{LL}(\theta)$. Since the gradient of $\text{LL}(\theta)$ is zero at the maximum, the ML estimator can also be defined as the value of θ at which

$$\sum_n s_n(\theta) = 0,$$

where $s_n(\theta) = \partial \ln P_n(\theta)/\partial\theta$ is the score for observation n.

Let $\check{P}_n(\theta)$ be a simulated approximation to $P_n(\theta)$. The simulated log-likelihood function is $\text{SLL}(\theta) = \sum_n \ln \check{P}_n(\theta)$, and the MSL estimator is the value of θ that maximizes $\text{SLL}(\theta)$. Stated equivalently, the estimator is the value of θ at which $\sum_n \check{s}_n(\theta) = 0$, where $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta)/\partial\theta$.

A preview of the properties of MSL can be given now, with a full explanation reserved for the next section. The main issue with MSL arises because of the log transformation. Suppose $\check{P}_n(\theta)$ is an unbiased simulator of $P_n(\theta)$, so that $E_r \check{P}_n(\theta) = P_n(\theta)$, where the expectation is over draws used in the simulation. All of the simulators that we have considered are unbiased for the true probability. However, since the log operation is a nonlinear transformation, $\ln \check{P}_n(\theta)$ is not unbiased for $\ln P_n(\theta)$ even though $\check{P}_n(\theta)$ is unbiased for $P_n(\theta)$. The bias in the simulator of $\ln P_n(\theta)$ translates into bias in the MSL estimator. This bias diminishes as more draws are used in the simulation.

To determine the asymptotic properties of the MSL estimator, the question arises of how the simulation bias behaves when the sample size rises. The answer depends critically on the relationship between the number of draws that are used in the simulation, labeled R, and the sample size, N. If R is considered fixed, then the MSL estimator does not converge to the true parameters, because of the simulation bias in $\ln \check{P}_n(\theta)$. Suppose instead that R rises with N; that is, the number of draws rises with sample size. In this case, the simulation bias disappears as N (and hence R) rises without bound. MSL is consistent in this case. As we will see, if R rises faster than $\sqrt{N}$, MSL is not only consistent but also efficient, asymptotically equivalent to maximum likelihood on the exact probabilities.

In summary, if R is fixed, then MSL is inconsistent. If R rises at any rate with N, then MSL is consistent. If R rises faster than $\sqrt{N}$, then MSL is asymptotically equivalent to ML.

The primary limitation of MSL is that it is inconsistent for fixed R. The other estimators that we consider are motivated by the desire for a simulation-based estimator that is consistent for fixed R. Both MSM and MSS, if structured appropriately, attain this goal. This benefit comes at a price, however, as we see in the following discussion.

10.2.2. Method of Simulated Moments

The traditional MOM is motivated by the recognition that the residuals of a model are necessarily uncorrelated in the population with factors that are exogenous to the behavior being modeled. The MOM estimator is the value of the parameters that make the residuals in the *sample* uncorrelated with the exogenous variables. For discrete choice models, MOM is defined as the parameters that solve the equation

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)] z_{nj} = 0, \tag{10.1}$$

where

d_{nj} is the dependent variable that identifies the chosen alternative: $d_{nj} = 1$ if n chose j, and $= 0$ otherwise, and
z_{nj} is a vector of exogenous variables called instruments.

The residuals are $d_{nj} - P_{nj}(\theta)$, and the MOM estimator is the parameter values at which the residuals are uncorrelated with the instruments in the sample.

This MOM estimator is analogous to MOM estimators for standard regression models. A regression model takes the form $y_n = x_n'\beta + \varepsilon_n$. The MOM estimator for this regression is the β at which

$$\sum_n (y_n - x_n'\beta) z_n = 0$$

for a vector of exogenous instruments z_n. When the explanatory variables in the model are exogenous, then they serve as the instruments. The MOM estimator in this case becomes the ordinary least squares estimator:

$$\sum_n (y_n - x_n'\beta) x_n = 0,$$

$$\sum_n x_n y_n = \sum_n x_n x_n'\beta,$$

$$\hat{\beta} = \left(\sum_n x_n x_n' \right)^{-1} \left(\sum_n x_n y_n \right),$$

which is the formula for the least squares estimator. When instruments are specified to be something other than the explanatory variables, the

MOM estimator becomes the standard instrumental variables estimator:

$$\sum_n (y_n - x_n'\beta)z_n = 0,$$
$$\sum_n z_n y_n = \sum_n z_n x_n'\beta,$$
$$\hat{\beta} = \left(\sum_n z_n x_n'\right)^{-1}\left(\sum_n z_n y_n\right),$$

which is the formula for the instrumental variables estimator. This estimator is consistent if the instruments are independent of ε in the population. The estimator is more efficient the more highly correlated the instruments are with the explanatory variables in the model. When the explanatory variables, x_n, are themselves exogenous, then the ideal instruments (i.e., those that give the highest efficiency) are the explanatory variables themselves, $z_n = x_n$.

For discrete choice models, MOM is defined analogously and has a similar relation to other estimators, especially ML. The researcher identifies instruments z_{nj} that are exogenous and hence independent in the population of the residuals $[d_{nj} - P_{nj}(\theta)]$. The MOM estimator is the value of θ at which the sample correlation between instruments and residuals is zero. Unlike the linear case, equation (10.1) cannot be solved explicitly for $\hat{\theta}$. Instead, numerical procedures are used to find the value of θ that solves this equation.

As with regression, ML for a discrete choice model is a special case of MOM. Let the instruments be the scores: $z_{nj} = \partial \ln P_{nj}(\theta)/\partial\theta$. With these instruments, MOM is the same as ML:

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]z_{nj} = 0,$$
$$\sum_n \left(\sum_j d_{nj}\frac{\partial \ln P_{nj}(\theta)}{\partial\beta}\right) - \left(\sum_j P_{nj}(\theta)\frac{\partial \ln P_{nj}(\theta)}{\partial\beta}\right) = 0,$$
$$\sum_n \frac{\partial \ln P_{ni}(\theta)}{\partial\beta} - \sum_n \sum_j P_{nj}(\theta)\frac{1}{P_{nj}(\theta)}\frac{\partial P_{nj}(\theta)}{\partial\theta} = 0,$$
$$\sum_n s_n(\theta) - \sum_n \sum_j \frac{\partial P_{nj}(\theta)}{\partial\theta} = 0,$$
$$\sum_n s_n(\theta) = 0,$$

which is the defining condition for ML. In the third line, i is the chosen alternative, recognizing that $d_{nj} = 0$ for all $j \neq i$. The fourth

line uses the fact that the sum of $\partial P_{nj}/\partial\theta$ over alternatives is zero, since the probabilities must sum to 1 before and after the change in θ.

Since MOM becomes ML and hence is fully efficient when the instruments are the scores, the scores are called the ideal instruments. MOM is consistent whenever the instruments are independent of the model residuals. It is more efficient the higher the correlation between the instruments and the ideal instruments.

An interesting simplification arises with standard logit. For the standard logit model, the ideal instruments are the explanatory variables themselves. As shown in Section 3.7.1, the ML estimator for standard logit is the value of θ that solves $\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]x_{nj} = 0$, where x_{nj} are the explanatory variables. This is a MOM estimator with the explanatory variables as instruments.

A simulated version of MOM, called the method of simulated moments (MSM), is obtained by replacing the exact probabilities $P_{nj}(\theta)$ with simulated probabilities $\check{P}_{nj}(\theta)$. The MSM estimator is the value of θ that solves

$$\sum_n \sum_j [d_{nj} - \check{P}_{nj}(\theta)]z_{nj} = 0$$

for instruments z_{nj}. As with its nonsimulated analog, MSM is consistent if z_{nj} is independent of $d_{nj} - \check{P}_{nj}(\theta)$.

The important feature of this estimator is that $\check{P}_{nj}(\theta)$ enters the equation linearly. As a result, if $\check{P}_{nj}(\theta)$ is unbiased for $P_{nj}(\theta)$, then $[d_{nj} - \check{P}_{nj}(\theta)]z_{nj}$ is unbiased for $[d_{nj} - P_{nj}(\theta)]z_{nj}$. Since there is no simulation bias in the estimation condition, the MSM estimator is consistent, even when the number of draws R is fixed. In contrast, MSL contains simulation bias due to the log transformation of the simulated probabilities. By not taking a nonlinear transformation of the simulated probabilities, MSM avoids simulation bias.

MSM still contains simulation noise (variance due to simulation). This noise becomes smaller as R rises and disappears when R rises without bound. As a result, MSM is asymptotically equivalent to MOM if R rises with N.

Just like its unsimulated analog, MSM is less efficient than MSL unless the ideal instruments are used. However, the ideal instruments are functions of $\ln P_{nj}$. They cannot be calculated exactly for any but the simplest models, and, if they are simulated using the simulated probability, simulation bias is introduced by the log operation. MSM is usually applied with nonideal weights, which means that there is a loss of

efficiency. MSM with ideal weights that are simulated without bias becomes MSS, which we discuss in the next section.

In summary, MSM has the advantage over MSL of being consistent with a fixed number of draws. However, there is no free lunch, and the cost of this advantage is a loss of efficiency when nonideal weights are used.

10.2.3. Method of Simulated Scores

MSS provides a possibility of attaining consistency without a loss of efficiency. The cost of this double advantage is numerical: the versions of MSS that provide efficiency have fairly poor numerical properties such that calculation of the estimator can be difficult.

The method of scores is defined by the condition

$$\sum_n s_n(\theta) = 0,$$

where $s_n(\theta) = \partial \ln P_n(\theta)/\partial\theta$ is the score for observation n. This is the same defining condition as ML: when exact probabilities are used, the method of scores is simply ML.

The method of simulated scores replaces the exact score with a simulated counterpart. The MSS estimator is the value of θ that solves

$$\sum_n \check{s}_n(\theta) = 0,$$

where $\check{s}_n(\theta)$ is a simulator of the score. If $\check{s}_n(\theta)$ is calculated as the derivative of the log of the simulated probability; that is, $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta)/\partial\theta$, then MSS is the same as MSL. However, the score can be simulated in other ways. When the score is simulated in other ways, MSS differs from MSL and has different properties.

Suppose that an unbiased simulator of the score can be constructed. With this simulator, the defining equation $\sum_n \check{s}_n(\theta) = 0$ does not incorporate any simulation bias, since the simulator enters the equation linearly. MSS is therefore consistent with a fixed R. The simulation noise decreases as R rises, such that MSS is asymptotically efficient, equivalent to MSL, when R rises with N. In contrast, MSL uses the biased score simulator $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta)/\partial\theta$, which is biased due to the log operator. MSS with an unbiased score simulator is therefore better than MSL with its biased score simulator in two regards: it is consistent under less stringent conditions (with fixed R rather than R rising with N) and is efficient under less stringent conditions (R rising at any rate with N rather than R rising faster than $\sqrt{N}$).

The difficulty with MSS comes in finding an unbiased score simulator. The score can be rewritten as

$$s_n(\theta) = \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} = \frac{1}{P_{nj}(\theta)} \frac{\partial P_{nj}}{\partial \theta}.$$

An unbiased simulator for the second term $\partial P_{nj}/\partial\theta$ is easily obtained by taking the derivative of the simulated probability. Since differentiation is a linear operation, $\partial \check{P}_{nj}/\partial\theta$ is unbiased for $\partial P_{nj}/\partial\theta$ if $\check{P}_{nj}(\theta)$ is unbiased for $P_{nj}(\theta)$. Since the second term in the score can be simulated without bias, the difficulty arises in finding an unbiased simulator for the first term $1/P_{nj}(\theta)$. Of course, simply taking the inverse of the simulated probability does not provide an unbiased simulator, since $E_r(1/\check{P}_{nj}(\theta)) \neq 1/P_{nj}(\theta)$. Like the log operation, an inverse introduces bias.

One proposal is based on the recognition that $1/P_{nj}(\theta)$ is the expected number of draws of the random terms that are needed before an "accept" is obtained. Consider drawing balls from an urn that contains many balls of different colors. Suppose the probability of obtaining a red ball is 0.20. That is, one-fifth of the balls are red. How many draws would it take, on average, to obtain a red ball? The answer is $1/0.2 = 5$. The same idea can be applied to choice probabilities. $P_{nj}(\theta)$ is the probability that a draw of the random terms of the model will result in alternative j having the highest utility. The inverse $1/P_{nj}(\theta)$ can be simulated as follows:

1. Take a draw of the random terms from their density.
2. Calculate the utility of each alternative with this draw.
3. Determine whether alternative j has the highest utility.
4. If so, call the draw an *accept*. If not, then call the draw a *reject* and repeat steps 1 to 3 with a new draw. Define B^r as the number of draws that are taken until the first *accept* is obtained.
5. Perform steps 1 to 4 R times, obtaining B^r for $r = 1, \ldots, R$. The simulator of $1/P_{nj}(\theta)$ is $(1/R)\sum_{r=1}^{R} B^r$.

This simulator is unbiased for $1/P_{nj}(\theta)$. The product of this simulator with the simulator $\partial \check{P}_{nj}/\partial\theta$ provides an unbiased simulator of the score. MSS based on this unbiased score simulator is consistent for fixed R and asymptotically efficient when R rises with N.

Unfortunately, the simulator of $1/P_{nj}(\theta)$ has the same difficulties as the accept–reject simulators that we discussed in Section 5.6. There is no guarantee than an accept will be obtained within any given number of draws. Also, the simulator is not continuous in parameters. The

discontinuity hinders the numerical procedures that are used to locate the parameters that solve the MSS equation.

In summary, there are advantages and disadvantages of MSS relative to MSL, just as there are of MSM. Understanding the capabilities of each estimator allows the researcher to make an informed choice among them.

10.3 The Central Limit Theorem

Prior to deriving the properties of our estimators, it is useful to review the central limit theorem. This theorem provides the basis for the distributions of the estimators.

One of the most basic results in statistics is that, if we take draws from a distribution with mean μ and variance σ, the mean of these draws will be normally distributed with mean μ and variance σ/N, where N is a large number of draws. This result is the central limit theorem, stated intuitively rather than precisely. We will provide a more complete and precise expression of these ideas.

Let $t = (1/N)\sum_n t_n$, where each t_n is a draw a from a distribution with mean μ and variance σ. The realization of the draws are called the sample, and t is the sample mean. If we take a different sample (i.e., obtain different values for the draws of each t_n), then we will get a different value for the statistic t. Our goal is to derive the sampling distribution of t.

For most statistics, we cannot determine the sampling distribution exactly for a given sample size. Instead, we examine how the sampling distribution behaves as sample size rises without bound. A distinction is made between the limiting distribution and the asymptotic distribution of a statistic. Suppose that, as sample size rises, the sampling distribution of statistic t converges to a fixed distribution. For example, the sampling distribution of t might become arbitrarily close to a normal with mean t^* and variance σ. In this case, we say that $N(t^*, \sigma)$ is the limiting distribution of t and that t converges in distribution to $N(t^*, \sigma)$. We denote this situation as $t \xrightarrow{d} N(t^*, \sigma)$.

In many cases, a statistic will not have a limiting distribution. As N rises, the sampling distribution keeps changing. The mean of a sample of draws is an example of a statistic without a limiting distribution. As stated, if t is the mean of a sample of draws from a distribution with mean μ and variance σ, then t is normally distributed with mean μ and variance σ/N. The variance decreases as N rises. The distribution changes as N rises, becoming more and more tightly dispersed around

the mean. If a limiting distribution were to be defined in this case, it would have to be the degenerate distribution at μ: as N rises without bound, the distribution of t collapses on μ. This limiting distribution is useless in understanding the variance of the statistic, since the variance of this limiting distribution is zero. What do we do in this case to understand the properties of the statistic?

If our original statistic does not have a limiting distribution, then we often can transform the statistic in such a way that the transformed statistic has a limiting distribution. Suppose, as in in our example of a sample mean, that the statistic we are interested in does not have a limiting distribution because its variance decreases as N rises. In that case, we can consider a transformation of the statistic that normalizes for sample size. In particular, we can consider $\sqrt{N}(t-\mu)$. Suppose that this statistic does indeed have a limiting distribution, for example $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$. In this case, we can derive the properties of our original statistic from the limiting distribution of the transformed statistic. Recall from basic principles of probabilities that, for fixed a and b, if $a(t-b)$ is distributed normal with zero mean and variance σ, then t itself is distributed normal with mean b and variance σ/a^2. This statement can be applied to our limiting distribution. For large enough N, $\sqrt{N}(t-\mu)$ is distributed approximately $N(0,\sigma)$. Therefore, for large enough N, t is distributed approximately $N(\mu,\sigma/N)$. We denote this as $t \overset{a}{\sim} N(\mu,\sigma/N)$. Note that this is not the limiting distribution of t, since t has no nondegenerate limiting distribution. Rather, it is called the asymptotic distribution of t, derived from the limiting distribution of $\sqrt{N}(t-\mu)$.

We can now restate precisely our concepts about the sampling distribution of a sample mean. The central limit theorem states the following. Suppose t is the mean of a sample of N draws from a distribution with mean μ and variance σ. Then $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$. With this limiting distribution, we can say that $t \overset{a}{\sim} N(\mu,\sigma/N)$.

There is another, more general version of the central limit theorem. In the version just stated, each t_n is a draw from the same distribution. Suppose t_n is a draw from a distribution with mean μ and variance σ_n, for $n = 1,\ldots,N$. That is, each t_n is from a different distribution; the distributions have the same mean but different variances. The generalized version of the central limit theorem states that $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$, where σ is now the average variance: $\sigma = (1/N)\sum_n \sigma_n$. Given this limiting distribution, we can say that $t \overset{a}{\sim} N(\mu,\sigma/N)$. We will use both versions of the central limit theorem when deriving the distributions of our estimators.

10.4 Properties of Traditional Estimators

In this section, we review the procedure for deriving the properties of estimators and apply that procedure to the traditional, non-simulation-based estimators. This discussion provides the basis for analyzing the properties of the simulation-based estimators in the next section.

The true value of the parameters is denoted θ^*. The ML and MOM estimators are roots of an equation that takes the form

$$\sum_n g_n(\hat{\theta})/N = 0. \tag{10.2}$$

That is, the estimator $\hat{\theta}$ is the value of the parameters that solve this equation. We divide by N, even though this division does not affect the root of the equation, because doing so facilitates our derivation of the properties of the estimators. The condition states that the average value of $g_n(\theta)$ in the sample is zero at the parameter estimates. For ML, $g_n(\theta)$ is the score $\partial \ln P_n(\theta)/\partial\theta$. For MOM, $g_n(\theta)$ is the set of first moments of residuals with a vector of instruments, $\sum_j (d_{nj} - P_{nj})z_{nj}$. Equation (10.2) is often called the moment condition. In its nonsimulated form, the method of scores is the same as ML and therefore need not be considered separately in this section. Note that we call (10.2) an equation even though it is actually a set of equations, since $g_n(\theta)$ is a vector. The parameters that solve these equations are the estimators.

At any particular value of θ, the mean and variance of $g_n(\theta)$ can be calculated for the sample. Label the mean as $g(\theta)$ and the variance as $W(\theta)$. We are particularly interested in the sample mean and variance of $g_n(\theta)$ at the true parameters, θ^*, since our goal is to estimate these parameters.

The key to understanding the properties of an estimator comes in realizing that each $g_n(\theta^*)$ is a draw from a distribution of $g_n(\theta^*)$'s in the population. We do not know the true parameters, but we know that each observation has a value of $g_n(\theta^*)$ at the true parameters. The value of $g_n(\theta^*)$ varies over people in the population. So, by drawing a person into our sample, we are essentially drawing a value of $g_n(\theta^*)$ from its distribution in the population.

The distribution of $g_n(\theta^*)$ in the population has a mean and variance. Label the mean of $g_n(\theta^*)$ in the population as $\mathbf{g}$ and its variance in the population as $\mathbf{W}$. The sample mean and variance at the true parameters, $g(\theta^*)$ and $W(\theta^*)$, are the sample counterparts to the population mean and variance, $\mathbf{g}$ and $\mathbf{W}$.

We assume that $\mathbf{g} = 0$. That is, we assume that the average of $g_n(\theta^*)$ in the population is zero at the true parameters. Under this assumption, the estimator provides a sample analog to this population expectation: $\hat{\theta}$ is the value of the parameters at which the sample average of $g_n(\theta)$ equals zero, as given in the defining condition (10.2). For ML, the assumption that $\mathbf{g} = 0$ simply states that the average score in the population is zero, when evaluated at the true parameters. In a sense, this can be considered the definition of the true parameters, namely, θ^* are the parameters at which the log-likelihood function for the entire population obtains its maximum and hence has zero slope. The estimated parameters are the values that make the slope of the likelihood function in the sample zero. For MOM, the assumption is satisfied if the instruments are independent of the residuals. In a sense, the assumption with MOM is simply a reiteration that the instruments are exogenous. The estimated parameters are the values that make the instruments and residuals uncorrelated in the sample.

We now consider the population variance of $g_n(\theta^*)$, which we have denoted $\mathbf{W}$. When $g_n(\theta)$ is the score, as in ML, this variance takes a special meaning. As shown in Section 8.7, the information identity states that $\mathbf{V} = -\mathbf{H}$, where

$$-\mathbf{H} = -E\left(\frac{\partial^2 \ln P_n(\theta^*)}{\partial\theta\,\partial\theta'}\right)$$

is the information matrix and $\mathbf{V}$ is the variance of the scores evaluated at the true parameters: $\mathbf{V} = \text{Var}(\partial \ln P_n(\theta^*)/\partial\theta)$. When $g_n(\theta)$ is the score, we have $\mathbf{W} = \mathbf{V}$ by definition and hence $\mathbf{W} = -\mathbf{H}$ by the information identity. That is, when $g_n(\theta)$ is the score, $\mathbf{W}$ is the information matrix. For MOM with nonideal instruments, $\mathbf{W} \neq -\mathbf{H}$, so that $\mathbf{W}$ does not equal the information matrix.

Why does this distinction matter? We will see that knowing whether $\mathbf{W}$ equals the information matrix allows us to determine whether the estimator is efficient. The lowest variance that any estimator can achieve is $-\mathbf{H}^{-1}/N$. For a proof, see, for example, Greene (2000) or Ruud (2000). An estimator is efficient if its variance attains this lower bound. As we will see, this lower bound is achieved when $\mathbf{W} = -\mathbf{H}$ but not when $\mathbf{W} \neq -\mathbf{H}$.

Our goal is to determine the properties of $\hat{\theta}$. We derive these properties in a two-step process. First, we examine the distribution of $g(\theta^*)$, which, as stated earlier, is the sample mean of $g_n(\theta^*)$. Second, the distribution of $\hat{\theta}$ is derived from the distribution of $g(\theta^*)$. This two-step process is not necessarily the most direct way of examining traditional estimators.

However, as we will see in the next section, it provides a very convenient way for generalizing to simulation-based estimators.

Step 1: The Distribution of $g(\theta^*)$
Recall that the value of $g_n(\theta^*)$ varies over decision makers in the population. When taking a sample, the researcher is drawing values of $g_n(\theta^*)$ from its distribution in the population. This distribution has zero mean by assumption and variance denoted $\mathbf{W}$. The researcher calculates the sample mean of these draws, $g(\theta^*)$. By the central limit theorem, $\sqrt{N}(g(\theta^*) - 0) \xrightarrow{d} N(0, \mathbf{W})$ such that the sample mean has distribution $g(\theta^*) \overset{a}{\sim} N(0, \mathbf{W}/N)$.

Step 2: Derive the Distribution of $\hat{\theta}$ from the Distribution of $g(\theta^*)$
We can relate the estimator $\hat{\theta}$ to its defining term $g(\theta)$ as follows. Take a first-order Taylor's expansion of $g(\hat{\theta})$ around $g(\theta^*)$:

$$g(\hat{\theta}) = g(\theta^*) + D[\hat{\theta} - \theta^*], \tag{10.3}$$

where $D = \partial g(\theta^*)/\partial\theta'$. By definition of $\hat{\theta}$ (that is, by defining condition (10.2)), $g(\hat{\theta}) = 0$ so that the right-hand side of this expansion is 0. Then

$$\begin{aligned} 0 &= g(\theta^*) + D[\hat{\theta} - \theta^*], \\ \hat{\theta} - \theta^* &= -D^{-1}g(\theta^*), \\ \sqrt{N}(\hat{\theta} - \theta^*) &= \sqrt{N}(-D^{-1})g(\theta^*). \end{aligned} \tag{10.4}$$

Denote the mean of $\partial g_n(\theta^*)/\partial\theta'$ in the population as $\mathbf{D}$. The sample mean of $\partial g_n(\theta^*)/\partial\theta'$ is D, as defined for equation (10.3). The sample mean D converges to the population mean $\mathbf{D}$ as the sample size rises. We know from step 1 that $\sqrt{N}g(\theta^*) \xrightarrow{d} N(0, \mathbf{W})$. Using this fact in (10.4), we have

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}). \tag{10.5}$$

This limiting distribution tells us that $\hat{\theta} \overset{a}{\sim} N(\theta^*, \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}/N)$.

We can now observe the properties of the estimator. The asymptotic distribution of $\hat{\theta}$ is centered on the true value, and its variance decreases as the sample size rises. As a result, $\hat{\theta}$ converges in probability to θ^* as the sample sise rises without bound: $\hat{\theta} \xrightarrow{p} \theta$. The estimator is therefore consistent. The estimator is asymptotically normal. And its variance is $\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}/N$, which can be compared with the lowest possible variance, $-\mathbf{H}^{-1}/N$, to determine whether it is efficient.

For ML, $g_n(\cdot)$ is the score, so that the variance of $g_n(\theta^*)$ is the variance of the scores: $\mathbf{W} = \mathbf{V}$. Also, the mean derivative of $g_n(\theta^*)$

is the mean derivative of the scores: $\mathbf{D} = \mathbf{H} = E(\partial^2 \ln P_n(\theta^*)/\partial\theta\,\partial\theta')$, where the expectation is over the population. By the information identity, $\mathbf{V} = -\mathbf{H}$. The asymptotic variance of $\hat{\theta}$ becomes $\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}/N = \mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}/N = \mathbf{H}^{-1}(-\mathbf{H})\mathbf{H}^{-1}/N = -\mathbf{H}^{-1}/N$, which is the lowest possible variance of any estimator. ML is therefore efficient. Since $\mathbf{V} = -\mathbf{H}$, the variance of the ML estimator can also be expressed as $\mathbf{V}^{-1}/N$, which has a readily interpretable meaning: the variance of the estimator is equal to the variance of the scores evaluated at the true parameters, divided by sample size.

For MOM, $g_n(\cdot)$ is a set of moments. If the ideal instruments are used, then MOM becomes ML and is efficient. If any other instruments are used, then MOM is not ML. In this case, $\mathbf{W}$ is the population variance of the moments and $\mathbf{D}$ is the mean derivatives of the moments, rather than the variance and mean derivatives of the scores. The asymptotic variance of $\hat{\theta}$ does not equal $-\mathbf{H}^{-1}/N$. MOM without ideal weights is therefore not efficient.

10.5 Properties of Simulation-Based Estimators

Suppose that the terms that enter the defining equation for an estimator are simulated rather than calculated exactly. Let $\check{g}_n(\theta)$ denote the simulated value of $g_n(\theta)$, and $\check{g}(\theta)$ the sample mean of these simulated values, so that $\check{g}(\theta)$ is the simulated version of $g(\theta)$. Denote the number of draws used in simulation for each n as R, and assume that independent draws are used for each n (e.g., separate draws are taken for each n). Assume further that the same draws are used for each value of θ when calculating $\check{g}_n(\theta)$. This procedure prevents *chatter* in the simulation: the difference between $\check{g}(\theta_1)$ and $\check{g}(\theta_2)$ for two different values of θ is not due to different draws.

These assumptions on the simulation draws are easy for the researcher to implement and simplify our analysis considerably. For interested readers, Lee (1992) examines the situation when the same draws are used for all observations. Pakes and Pollard (1989) provide a way to characterize an equicontinuity condition that, when satisfied, facilitates analysis of simulation-based estimators. McFadden (1989) characterizes this condition in a different way and shows that it can be met by using the same draws for each value of θ, which is the assumption that we make. McFadden (1996) provides a helpful synthesis that includes a discussion of the need to prevent chatter.

The estimator is defined by the condition $\check{g}(\hat{\theta}) = 0$. We derive the properties of $\hat{\theta}$ in the same two steps as for the traditional estimators.

Step 1: The Distribution of $\check{g}(\theta^*)$

To identify the various components of this distribution, let us reexpress $\check{g}(\theta^*)$ by adding and subtracting some terms and rearranging:

$$\begin{aligned}\check{g}(\theta^*) &= \check{g}(\theta^*) + g(\theta^*) - g(\theta^*) + E_r\check{g}(\theta^*) - E_r\check{g}(\theta^*) \\ &= g(\theta^*) + [E_r\check{g}(\theta^*) - g(\theta^*)] + [\check{g}(\theta^*) - E_r\check{g}(\theta^*)],\end{aligned}$$

where $g(\theta^*)$ is the nonsimulated value and $E_r\check{g}(\theta^*)$ is the expectation of the simulated value over the draws used in the simulation. Adding and subtracting terms obviously does not change $\check{g}(\theta^*)$. Yet, the subsequent rearrangement of the terms allows us to identify components that have intuitive meaning.

The first term $g(\theta^*)$ is the same as arises for the traditional estimator. The other two terms are extra elements that arise because of the simulation. The term $E_r\check{g}(\theta^*) - g(\theta^*)$ captures the bias, if any, in the simulator of $g(\theta^*)$. It is the difference between the true value of $g(\theta^*)$ and the expectation of the simulated value. If the simulator is unbiased for $g(\theta^*)$, then $E_r\check{g}(\theta^*) = g(\theta^*)$ and this term drops out. Often, however, the simulator will not be unbiased for $g(\theta^*)$. For example, with MSL, $\check{g}_n(\theta) = \partial \ln \check{P}_n(\theta)/\partial\theta$, where $\check{P}_n(\theta)$ is an unbiased simulator of $P_n(\theta)$. Since $\check{P}_n(\theta)$ enters nonlinearly via the log operator, $\check{g}_n(\theta)$ is not unbiased. The third term, $\check{g}(\theta^*) - E_r\check{g}(\theta^*)$, captures simulation noise, that is, the deviation of the simulator for given draws from its expectation over all possible draws.

Combining these concepts, we have

$$\check{g}(\theta) = A + B + C, \tag{10.6}$$

where

A is the same as in the traditional estimator,
B is simulation bias,
C is simulation noise.

To see how the simulation-based estimators differ from their traditional counterparts, we examine the simulation bias B and noise C.

Consider first the noise. This term can be reexpressed as

$$\begin{aligned}C &= \check{g}(\theta^*) - E_r\check{g}(\theta^*) \\ &= \frac{1}{N}\sum_n \check{g}_n(\theta^*) - E_r\check{g}_n(\theta^*) \\ &= \sum_n d_n/N,\end{aligned}$$

where d_n is the deviation of the simulated value for observation n from its expectation. The key to understanding the behavior of the simulation noise comes in noting that d_n is simply a statistic for observation n. The sample constitutes N draws of this statistic, one for each observation: $d_n, n = 1, \ldots, N$. The simulation noise C is the average of these N draws. Thus, the central limit theorem gives us the distribution of C.

In particular, for a given observation, the draws that are used in simulation provide a particular value of d_n. If different draws had been obtained, then a different value of d_n would have been obtained. There is a distribution of values of d_n over the possible realizations of the draws used in simulation. The distribution has zero mean, since the expectation over draws is subtracted out when creating d_n. Label the variance of the distribution as $\mathcal{S}_n/R$, where $\mathcal{S}_n$ is the variance when one draw is used in simulation. There are two things to note about this variance. First, $\mathcal{S}_n/R$ is inversely related to R, the number of draws that are used in simulation. Second, the variance is different for different n. Since $g_n(\theta^*)$ is different for different n, the variance of the simulation deviation also differs.

We take a draw of d_n for each of N observations; the overall simulation noise, C, is the average of these N draws of observation-specific simulation noise. As just stated, each d_n is a draw from a distribution with zero mean and variance $\mathcal{S}_n/R$. The generalized version of the central limit theorem tells us the distribution of a sample average of draws from distributions that have the same mean but different variances. In our case,

$$\sqrt{N}C \xrightarrow{d} N(0, \mathbf{S}/R),$$

where $\mathbf{S}$ is the population mean of $\mathcal{S}_n$. Then $C \overset{a}{\sim} N(0, \mathbf{S}/NR)$.

The most relevant characteristic of the asymptotic distribution of C is that it decreases as N increases, even when R is fixed. Simulation noise disappears as sample size increases, even without increasing the number of draws used in simulation. This is a very important and powerful fact. It means that increasing the sample size is a way to decrease the effects of simulation on the estimator. The result is intuitively meaningful. Essentially, simulation noise cancels out over observations. The simulation for one observation might, by chance, make that observation's $\check{g}_n(\theta)$ too large. However, the simulation for another observation is likely, by chance, to be too small. By averaging the simulations over observations, the errors tend to cancel each other. As sample size rises, this canceling out property becomes more powerful until, with large enough samples, simulation noise is negligible.

Consider now the bias. If the simulator $\check{g}(\theta)$ is unbiased for $g(\theta)$, then the bias term B in (10.6) is zero. However, if the simulator is biased, as with MSL, then the effect of this bias on the distribution of $\check{g}(\theta^*)$ must be considered.

Usually, the defining term $g_n(\theta)$ is a function of a statistic, ℓ_n, that can be simulated without bias. For example, with MSL, $g_n(\theta)$ is a function of the choice probability, which can be simulated without bias; in this case ℓ_n is the probability. More generally, ℓ_n can be any statistic that is simulated without bias and serves to define $g_n(\theta)$. We can write the dependence in general as $g_n(\theta) = g(\ell_n(\theta))$ and the unbiased simulator of $\ell_n(\theta)$ as $\check{\ell}_n(\theta)$ where $E_r\check{\ell}_n(\theta) = \ell_n(\theta)$.

We can now reexpress $\check{g}_n(\theta)$ by taking a Taylor's expansion around the unsimulated value $g_n(\theta)$:

$$\begin{aligned}\check{g}_n(\theta) = g_n(\theta) &+ \frac{\partial g(\ell_n(\theta))}{\partial \ell_n}[\check{\ell}_n(\theta) - \ell_n(\theta)] \\ &+ \tfrac{1}{2}\frac{\partial^2 g(\ell_n(\theta))}{\partial \ell_n^2}[\check{\ell}_n(\theta) - \ell_n(\theta]^2,\end{aligned}$$

$$\check{g}_n(\theta) - g_n(\theta) = g_n'[\check{\ell}_n(\theta) - \ell_n(\theta)] + \tfrac{1}{2}g_n''[\check{\ell}_n(\theta) - \ell_n(\theta)]^2,$$

where g_n' and g_n'' are simply shorthand ways to denote the first and second derivatives of $g_n(\ell(\cdot))$ with respect to ℓ. Since $\check{\ell}_n(\theta)$ is unbiased for $\ell_n(\theta)$, we know $E_r g_n'[\check{\ell}_n(\theta) - \ell_n(\theta)] = g_n'[E_r\check{\ell}_n(\theta) - \ell_n(\theta)] = 0$. As a result, only the variance term remains in the expectation:

$$\begin{aligned}E_r\check{g}_n(\theta) - g_n(\theta) &= \tfrac{1}{2}g_n''E_r[\check{\ell}_n(\theta) - \ell_n(\theta)]^2 \\ &= \tfrac{1}{2}g_n''\operatorname{Var}_r\check{\ell}_n(\theta).\end{aligned}$$

Denote $\operatorname{Var}_r\check{\ell}_n(\theta) = \mathcal{Q}_n/R$ to reflect the fact that the variance is inversely proportional to the number of draws used in the simulation. The simulation bias is then

$$\begin{aligned}E_r\check{g}(\theta) - g(\theta) &= \frac{1}{N}\sum_n E_r\check{g}_n(\theta) - g_n(\theta) \\ &= \frac{1}{N}\sum_n g_n''\frac{\mathcal{Q}_n}{2R} \\ &= \frac{\mathcal{Z}}{R},\end{aligned}$$

where $\mathcal{Z}$ is the sample average of $g_n''\mathcal{Q}_n/2$.

Since $B = \mathcal{Z}/R$, the value of this statistic normalized for sample size is

$$\sqrt{N}B = \frac{\sqrt{N}}{R}\mathcal{Z}. \tag{10.7}$$

If R is fixed, then B is nonzero. Even worse, $\sqrt{N}B$ rises with N, in such a way that it has no limiting value. Suppose that R is considered to rise with N. The bias term then disappears asymptotically: $B = \mathcal{Z}/R \xrightarrow{p} 0$. However, the normalized bias term does not necessarily disappear. Since $\sqrt{N}$ enters the numerator of this term, $\sqrt{N}B = (\sqrt{N}/R)\mathcal{Z} \xrightarrow{p} 0$ only if R rises faster than $\sqrt{N}$, so that the ratio $\sqrt{N}/R$ approaches zero as N increases. If R rises slower than $\sqrt{N}$, the ratio $\sqrt{N}/R$ rises, such that the normalized bias term does not disappear but in fact gets larger and larger as sample size increases.

We can now collect our results for the distribution of the defining term normalized by sample size:

$$\sqrt{N}\check{g}(\theta^*) = \sqrt{N}(A + B + C), \tag{10.8}$$

where

$\sqrt{N}A \xrightarrow{d} N(0, \mathbf{W})$,	the same as in the traditional estimator,
$\sqrt{N}B = \frac{\sqrt{N}}{R}\mathcal{Z}$,	capturing simulation bias,
$\sqrt{N}C \xrightarrow{d} N(0, \mathbf{S}/R)$,	capturing simulation noise.

Step 2: Derive Distribution of $\hat{\theta}$ from Distribution of $\check{g}(\theta^*)$
As with the traditional estimators, the distribution of $\hat{\theta}$ is directly related to the distribution of $\check{g}(\theta^*)$. Using the same Taylor's expansion as in (10.3), we have

$$\sqrt{N}(\hat{\theta} - \theta^*) = -\check{D}^{-1}\sqrt{N}\check{g}(\theta^*) = -\check{D}^{-1}\sqrt{N}(A + B + C), \tag{10.9}$$

where $\check{D}$ is the derivative of $\check{g}(\theta^*)$ with respect to the parameters, which converges to its expectation $\check{\mathbf{D}}$ as sample size rises. The estimator itself is expressed as

$$\hat{\theta} = \theta^* - \check{D}^{-1}(A + B + C). \tag{10.10}$$

We can now examine the properties of our estimators.

10.5.1. Maximum Simulated Likelihood

For MSL, $\breve{g}_n(\theta)$ is not unbiased for $g_n(\theta)$. The bias term in (10.9) is $\sqrt{N}B = (\sqrt{N}/R)\mathcal{Z}$. Suppose R rises with N. If R rises faster than $\sqrt{N}$, then

$$\sqrt{N}B = (\sqrt{N}/R)\mathcal{Z} \xrightarrow{p} 0,$$

since the ratio $\sqrt{N}/R$ falls to zero. Consider now the third term in (10.9), which captures simulation noise: $\sqrt{N}C \xrightarrow{d} N(0, \mathbf{S}/R)$. Since $\mathbf{S}/R$ decreases as R rises, we have $\mathbf{S}/R \xrightarrow{p} 0$ as $N \to \infty$ when R rises with N. The second and third terms disappear, leaving only the first term. This first term is the same as appears for the nonsimulated estimator. We have

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta^*) = -\mathbf{D}^{-1}\sqrt{N}A &\xrightarrow{d} N(0, \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}) \\ &= N(0, \mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}) \\ &= N(0, -\mathbf{H}^{-1}),\end{aligned}$$

where the next-to-last equality occurs because $g_n(\theta)$ is the score, and the last equality is due to the information identity. The estimator is distributed

$$\hat{\theta} \overset{a}{\sim} N(\theta^*, -\mathbf{H}^{-1}/N).$$

This is the same asymptotic distribution as ML. When R rises faster than $\sqrt{N}$, MSL is consistent, asymptotically normal and efficient, and asymptotically equivalent to ML.

Suppose that R rises with N but at a rate that is slower than $\sqrt{N}$. In this case, the ratio $\sqrt{N}/R$ grows larger as N rises. There is no limiting distribution for $\sqrt{N}(\hat{\theta} - \theta^*)$, because the bias term, $(\sqrt{N}/R)\mathcal{Z}$, rises with N. However, the estimator itself converges on the true value. $\hat{\theta}$ depends on $(1/R)\mathcal{Z}$, not multiplied by $\sqrt{N}$. This bias term disappears when R rises at any rate. Therefore, the estimator converges on the true value, just like its nonsimulated counterpart, which means that $\hat{\theta}$ is consistent. However, the estimator is not asymptotically normal, since $\sqrt{N}(\hat{\theta} - \theta^*)$ has no limiting distribution. Standard errors cannot be calculated, and confidence intervals cannot be constructed.

When R is fixed, the bias rises as N rises. $\sqrt{N}(\hat{\theta} - \theta^*)$ does not have a limiting distribution. Moreover, the estimator itself, $\hat{\theta}$, contains a bias $B = (1/R)\mathcal{Z}$ that does not disappear as sample size rises with fixed R. The MSL estimator is neither consistent nor asymptotically normal when R is fixed.

The properties of MSL can be summarized as follows:

1. If R is fixed, MSL is inconsistent.
2. If R rises slower than $\sqrt{N}$, MSL is consistent but not asymptotically normal.
3. If R rises faster than $\sqrt{N}$, MSL is consistent, asymptotically normal and efficient, and equivalent to ML.

10.5.2. Method of Simulated Moments

For MSM with fixed instruments, $\check{g}_n(\theta) = \sum_j [d_{nj} - \check{P}_{nj}(\theta)]z_{nj}$, which is unbiased for $g_n(\theta)$, since the simulated probability enters linearly. The bias term is zero. The distribution of the estimator is determined only by term A, which is the same as in the traditional MOM without simulation, and term C, which reflects simulation noise:

$$\sqrt{N}(\hat{\theta} - \theta^*) = -\check{D}^{-1}\sqrt{N}(A + C).$$

Suppose that R is fixed. Since $\check{D}$ converges to its expectation $\mathbf{D}$, we have $-\sqrt{N}\check{D}^{-1}A \xrightarrow{d} N(0, \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1})$ and $-\sqrt{N}\check{D}^{-1}C \xrightarrow{d} N(0, \mathbf{D}^{-1}(\mathbf{S}/R)\mathbf{D}^{-1})$, so that

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathbf{D}^{-1}[\mathbf{W} + \mathbf{S}/R]\mathbf{D}^{-1}).$$

The asymptotic distribution of the estimator is then

$$\hat{\theta} \overset{a}{\sim} N(\theta^*, \mathbf{D}^{-1}[\mathbf{W} + \mathbf{S}/R]\mathbf{D}^{-1}/N).$$

The estimator is consistent and asymptotically normal. Its variance is greater than its nonsimulated counterpart by $\mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}/RN$, reflecting simulation noise.

Suppose now that R rises with N at any rate. The extra variance due to simulation noise disappears, so that $\hat{\theta} \overset{a}{\sim} N(\theta^*, \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}/N)$, the same as its nonsimulated counterpart. When nonideal instruments are used, $\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1} \neq -\mathbf{H}^{-1}$ and so the estimator (in either its simulated or nonsimulated form) is less efficient than ML.

If simulated instruments are used in MSM, then the properties of the estimator depend on how the instruments are simulated. If the instruments are simulated without bias and independently of the probability that enters the residual, then this MSM has the same properties as MSM with fixed weights. If the instruments are simulated with bias and the instruments are not ideal, then the estimator has the same properties as MSL except that it is not asymptotically efficient, since the information

identity does not apply. MSM with simulated ideal instruments is MSS, which we discuss next.

10.5.3. Method of Simulated Scores

With MSS using unbiased score simulators, $\check{g}_n(\theta)$ is unbiased for $g_n(\theta)$, and, moreover, $g_n(\theta)$ is the score such that the information identity applies. The analysis is the same as for MSM except that the information identity makes the estimator efficient when R rises with N. As with MSM, we have

$$\hat{\theta} \overset{a}{\sim} N(\theta^*, \mathbf{D}^{-1}[\mathbf{W} + \mathbf{S}/R]\mathbf{D}^{-1}/N),$$

which, since $g_n(\theta)$ is the score, becomes

$$\hat{\theta} \overset{a}{\sim} N\left(\theta^*, \frac{\mathbf{H}^{-1}[\mathbf{V} + \mathbf{S}/R]\mathbf{H}^{-1}}{N}\right) = N\left(\theta^*, -\frac{\mathbf{H}^{-1}}{N} + \frac{\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}}{RN}\right).$$

When R is fixed, the estimator is consistent and asymptotically normal, but its covariance is larger than with ML because of simulation noise. If R rises at any rate with N, then we have

$$\hat{\theta} \overset{a}{\sim} N(0, -\mathbf{H}^{-1}/N).$$

MSS with unbiased score simulators is asymptotically equivalent to ML when R rises at any rate with N.

This analysis shows that MSS with unbiased score simulators has better properties than MSL in two regards. First, for fixed R, MSS is consistent and asymptotically normal, while MSL is neither. Second, for R rising with N, MSS is equivalent to ML no matter how fast R is rising, while MSL is equivalent to ML only if the rate is faster than $\sqrt{N}$.

As we discussed in Section 10.2.3, finding unbiased score simulators with good numerical properties is difficult. MSS is sometimes applied with biased score simulators. In this case, the properties of the estimator are the same as with MSL: the bias in the simulated scores translates into bias in the estimator, which disappears from the limiting distribution only if R rises faster than $\sqrt{N}$.

10.6 Numerical Solution

The estimators are defined as the value of θ that solves $\check{g}(\theta) = 0$, where $\check{g}(\theta) = \sum_n \check{g}_n(\theta)/N$ is the sample average of a simulated statistic $\check{g}_n(\theta)$. Since $\check{g}_n(\theta)$ is a vector, we need to solve the set of equations for the

parameters. The question arises: how are these equations solved numerically to obtain the estimates?

Chapter 8 describes numerical methods for maximizing a function. These procedures can also be used for solving a set of equations. Let T be the negative of the inner product of the defining term for an estimator: $T = -\check{g}(\theta)'\check{g}(\theta) = -(\sum_n \check{g}_n(\theta))'(\sum_n \check{g}_n(\theta))/N^2$. T is necessarily less than or equal to zero, since it is the negative of a sum of squares. T has a highest value of 0, which is attained only when the squared terms that compose it are all 0. That is, the maximum of T is attained when $\check{g}(\theta) = 0$. Maximizing T is equivalent to solving the equation $\check{g}(\theta) = 0$. The approaches described in Chapter 8, with the exception of BHHH, can be used for this maximization. BHHH cannot be used, because that method assumes that the function being maximized is a sum of observation-specific terms, whereas T takes the square of each sum of observation-specific terms. The other approaches, especially BFGS and DFP, have proven very effective at locating the parameters at which $\check{g}(\theta) = 0$.

With MSL, it is usually easier to maximize the simulated likelihood function rather than T. BHHH can be used in this case, as well as the other methods.