COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1120

A STOPPING RULE FOR THE COMPUTATION OF GENERALIZED METHOD OF MOMENTS ESTIMATORS

Donald W. K. Andrews

April 1996

A STOPPING RULE FOR THE COMPUTATION OF GENERALIZED METHOD OF MOMENTS ESTIMATORS

Donald W. K. Andrews[1]

Cowles Foundation for Research in Economics
Yale University

November 1995

Revised: April 1996

ABSTRACT

To obtain consistency and asymptotic normality, a generalized method of moments (GMM) estimator typically is defined to be an approximate global minimizer of a GMM criterion function. To compute such an estimator, however, can be problematic because of the difficulty of global optimization. In consequence, practitioners usually ignore the problem and take the GMM estimator to be the result of a local optimization algorithm. This yields an estimator that is not necessarily consistent and asymptotically normal. The use of a local optimization algorithm also can run into the problem of instability due to flats or ridges in the criterion function, which makes it difficult to know when to stop the algorithm.

To alleviate these problems of global and local optimization, we propose a stopping-rule (SR) procedure for computing GMM estimators. The SR procedure eliminates the need for global search with high probability. And, it provides an explicit SR for problems of stability that may arise with local optimization problems.

Keywords: Computation, generalized method of moments estimator, global optimization, j-step estimator, stopping rule.

JEL Classification: C13.

1. INTRODUCTION

This paper proposes a stopping-rule (SR) procedure for the computation of GMM estimators. In contrast to local optimization algorithms typically employed in practice, the SR procedure guarantees that the GMM estimator is consistent and asymptotically normal. The SR procedure is quite flexible. In particular, it can be taken to involve similar computations to the local optimization algorithms typically in use. A drawback of the SR procedure is that there is a small probability that the SR criterion is too stringent and one is left with the problem of searching for an approximate global optimum of the GMM criterion function.

The SR procedure can be described briefly as follows. First, one obtains an initial estimator, say $\widehat{\theta}_0$, typically using some local optimization algorithm, perhaps with multiple starting values. Next, one checks to see if $\widehat{\theta}_0$ satisfies the SR $||A_n(\widehat{\theta}_0)G_n(\widehat{\theta}_0)||^2 \leq c_{r,n}/n$, where $||A_n(\theta)G_n(\theta)||^2$ is the GMM criterion function and $c_{r,n}$ is a constant that depends on r, the number of over-identifying restrictions. If $\widehat{\theta}_0$ fails the SR, one needs to look for a new initial estimator, perhaps by considering new starting values for the local optimization algorithm. If $\widehat{\theta}_0$ satisfies the SR, one computes a j-step estimator $\widehat{\theta}_j$ using j Newton–Raphson (NR) iterations, as in Robinson (1988), starting from the initial estimator $\widehat{\theta}_0$. The final GMM estimator is then taken to be the value of θ that minimizes the GMM criterion function over all values considered up to this point, including $\widehat{\theta}_j$. The final estimator is guaranteed to be consistent and asymptotically normal. In fact, it is guaranteed to be an approximate global minimizer of the GMM criterion function.

The remainder of the paper is organized as follows. Section 2 describes the difficulty of finding an approximate optimum of the GMM criterion function by global search. It also describes the problem of terminating a local optimization algorithm when instability, due to flats or ridges, occurs. Section 3 describes the SR procedure in detail. Section 4 provides the asymptotic justification for the SR procedure. The regularity conditions of Pakes and Pollard (1989) are used when establishing these results. An Appendix provides proofs of the results given in Section 4.

2. COMPUTATIONAL PROBLEMS

2.1. Global Optimization

Standard definitions of extremum estimators, including GMM estimators, require that one minimize a criterion function over a parameter space $\Theta \subset R^d$. The best specified definitions in the literature do not require one to *precisely* minimize the criterion function — a task that is usually impossible to carry out in practice. Rather, they require that one finds a value $\widehat{\theta}$ that is close to minimizing the criterion function. For example, Pakes and Pollard (1989) require that $\widehat{\theta}$ yields a value of the criterion function that is within $o_p(1)$ of the minimum to obtain consistency of $\widehat{\theta}$ and within $o_p(n^{-1/2})$ of the minimum to obtain consistency and asymptotic normality of $\widehat{\theta}$.

If the criterion function and parameter space are convex, as occurs in a number of econometric applications, e.g., see Pratt (1981), then the criterion function has a unique local minimum, which also is the global minimum. In this case, an approximate global minimum can be computed by a local optimization algorithm started at any value in the parameter space.

Our concern here is with non-convex problems. For such problems, the most widely used method to compute econometric extremum estimators is an algorithm called *multi-start.* This algorithm is described as follows. One starts with an initial value θ_I, obtained either judiciously or randomly, one uses a local optimization algorithm to converge to a local minimum, and then one repeats the process a number of times with different starting values. The estimator $\widehat{\theta}$ is taken to be the value in Θ that corresponds to the smallest value of the criterion function that is computed during the multi-start process. Often this algorithm is applied somewhat informally with the choice of starting values and number of starts specified vaguely.

The multi-start algorithm has the advantages of being easy to apply (given some local optimization algorithm) and of being tractable (because the number of starts can be chosen with the speed of function and function derivative evaluations in mind, and the local optimization algorithm can be chosen with the nature of the criterion function in mind). A major problem with the multi-start algorithm, however, is that it does not necessarily find the global optimum or an approximate global optimum. In consequence, the estimator it delivers is not necessarily consistent and asymptotically normal.

To ensure that one is close to the minimum of the criterion function, it is necessary to carry out a global optimization of the criterion function. The theory and practice of global optimization of functions is a subfield of computer science that has been increasingly active in the last twenty years. Various approaches have been explored, e.g., see Floudas and Pardalos (1992) and Horst and Tuy (1992). This research area, however, is still in a state of considerable flux. No consensus has emerged as to the best methods for global optimization even for well-studied classes of functions. No widely available and widely used software has been developed.

One reason for the unsatisfactory state of practical global optimization methods is that global optimization is an intrinsically hard problem. This is quantified by results from another subfield of computer science, viz., *information-based complexity theory.* Complexity theory utilizes the concept of ε–cardinality. The (worst case) ε–cardinality of a global optimization problem is defined to be the minimal number of function and first derivative evaluations necessary to find the global optimum of a function to within ε (i.e., to find a value $\widehat{\theta}$ in the domain Θ of a function f such that $f(\widehat{\theta}) \leq \inf_{\theta \in \Theta} f(\theta) + \varepsilon$), given that the function to be optimized could be any function in some specified class.

For example, consider the class $\mathcal{F}$ of functions from Θ to R, where Θ is a nonempty compact subset of R^d, each member of which is s–times continuously differentiable with uniformly bounded s–th derivatives in all directions. The ε–cardinality of the global optimization problem for this class of functions is of order $\varepsilon^{-d/s}$ and it is achieved by searching over a regular grid of points, see Nemirovsky and Yudin (1983) (or, for a summary of their results, see Traub, Wasilkowski, and Wosniakowski (1988, Ch. 5, Sec. 8)). (These results hold even when one allows for random and adaptive selection of the function and derivative evaluation points and random algorithms.) Robinson (1988) discusses and analyzes the use of the grid search method in an econometric context to obtain n^{ξ}–consistent estimators for a given $\xi > 0$.

The use of grid search, however, shows how difficult the global optimization problem is. Consider a function with $s = 1$, first derivatives bounded by M, and parameter space $\Theta = [0, L]^d$.

To obtain an ε–approximate solution, one needs $(LM/\varepsilon)^d$ function evaluations. For example, in econometric applications it would not be unusual to have $L = M = d = 10$, which requires $10^{20}/\varepsilon^{10}$ function evaluations. If $\varepsilon = .01$, then 10^{40} evaluations are required. At one second per evaluation, 3.17×10^{32} years are required to carry out the computation. Alternatively, consider a lower dimensional problem with $d = 5$, $L = M = 10$, and $\varepsilon = .01$. In this case, 10^{20} function evaluations are required. At one second per evaluation, the computation time still is 31.7 million years.

An alternative approach is to employ a global optimization method that may not be optimal in terms of worst case $\varepsilon-$ cardinality, but may be more efficient than a grid search for many of the functions in the class of functions $\mathcal{F}$. This approach has promise, especially as computer hardware and algorithms improve. To date, however, no algorithms that guarantee convergence to the global optimum have become standard in econometrics or elsewhere. When global optimization methods are used in econometrics, algorithms that do not guarantee convergence to an (ε–approximate) global optimum often are employed. An example is the simulated annealing algorithm run for a finite length of time, e.g., see Goffe, Ferrier, and Rogers (1994).

An alternative to global optimization is provided by Veall (1990). He proposes a test of the null hypothesis that a given trial value $\widehat{\theta}$ is a global optimum. The test requires that one draws N^* random variables from a uniform distribution on the parameter space Θ, evaluates the criterion function at those points, and compares a particular function of the resulting criterion function values to the value of the criterion function at $\widehat{\theta}$. For a given significance level α, the test guarantees (for large N^*) that the null hypothesis that $\widehat{\theta}$ is the global optimum will be falsely rejected with probability less than or equal to α.

A serious problem with Veall's test, however, is that the error one wants to control in this testing context is not that of falsely rejecting the null. Rather, the error of foremost concern is that of falsely accepting the null, viz., of falsely concluding that $\widehat{\theta}$ is a global optimum. The latter probability is the power of the test and it depends on N^*. No method is provided with Veall's test to calculate how large N^* must be in order to ensure that the probability that the test falsely accepts the null is less than some specified value α. In consequence, the use of Veall's test does not necessarily produce estimators that are consistent and asymptotically normal.

In sum, the computation of extremum estimators is problematic in practice because they require the solution of a global optimization problem that is often difficult to compute. The standard method of computing the solution, i.e., multi-start, produces a local optimum that is not necessarily global. In consequence, the estimator it produces is not necessarily consistent and asymptotically normal. Global search, which does guarantee an approximate global optimum, is intractable in many problems. Other methods, such as simulated annealing or the use of Veall's testing procedure, may reduce the chance that a local minimum is erroneously chosen to be the estimator. They do not guarantee, however, that a global or approximate global minimum is found. Hence, they do not guarantee that the estimator produced is consistent and asymptotically normal.

2.2. Local Optimization

As discussed above, in current practice the computation of an extremum estimator typically relies on a local optimization algorithm. Especially when the parameter space is of high dimension, a local optimization algorithm may have trouble finding an exact local optimum. It may converge slowly or not at all, due to flatness of the criterion function in certain directions or due to small bumps or ripples in the criterion function. In such cases, it is useful to have a stopping rule that specifies when the algorithm is "close enough" to a local optimum.

In practice, informal stopping rules often are utilized. Usually, such rules are based on how much the criterion function or the parameter estimates change over a number of iterations of the algorithm or on how close the vector of derivatives of the criterion function is to zero. A common problem with such rules is that little guidance is available to indicate when small changes are "small enough" or when close is "close enough."

For the asymptotic results of Huber (1967), Hansen (1982), and others, the vector of derivatives of the criterion function is "close enough" to zero for asymptotic normality of the estimator if it is

$o_p(n^{-1/2})$. Unfortunately, this result is difficult to exploit in practice because it is not clear how one should pick a reasonable sequence of values that is $o_p(n^{-1/2})$ out of the multitude of sequences that have that property. In addition, it is impossible to pick a reasonable sequence without first normalizing the magnitude of the criterion function in some way, since the scale of the criterion function and its derivatives usually is arbitrary.

To conclude, a second problem that arises in the computation of extremum estimators is that of specifying a reasonable stopping rule for a local optimization algorithm.

3. DESCRIPTION OF THE STOPPING-RULE APPROACH

Here we present a computational approach for GMM estimators with over-identifying restrictions that alleviates the problems discussed in Section 2.

3.1. Notation and Introduction

We start by introducing some notation. The GMM criterion function under consideration is

$$\|A_n(\theta)G_n(\theta)\|^2 \, , \tag{3.1}$$

where $G_n(\theta)$ is a k-vector of moment conditions, $A_n(\theta)$ is a nonsingular $k \times k$ weight matrix, $\theta \in \Theta \subset R^d$ is the unknown parameter of interest, and $\|\cdot\|$ is the Euclidean norm. The moment conditions and weight matrix satisfy

$$G_n(\theta) \xrightarrow{p} G(\theta) \ \text{ and } \ A_n(\theta) \xrightarrow{p} A(\theta) \ \ \forall\theta \in \Theta \tag{3.2}$$

for some functions $G(\theta)$ and $A(\theta)$. The latter are assumed to satisfy

$$G(\theta) = 0 \ \text{ iff } \ \theta = \theta_0 \tag{3.3}$$

and $A(\theta)$ is nonsingular for all $\theta \in \Theta$. The limit function $G(\theta)$ is assumed to be differentiable in θ on a neighborhood of θ_0 with $k \times d$ derivative $\Gamma(\theta)$. $\Gamma_n(\theta)$ denotes some consistent estimator of $\Gamma(\theta)$. That is,

$$\Gamma_n(\theta) \xrightarrow{p} \Gamma(\theta) \ \text{ for all } \ \theta \ \text{ in a neighborhood of } \ \theta_0 \, . \tag{3.4}$$

For the case where $G_n(\theta)$ is not everywhere differentiable, one can define $\Gamma_n(\theta)$ using numerical derivatives of $G_n(\theta)$, see Pakes and Pollard (1989, p. 1043).

Typically a GMM estimator is computed by (i) choosing a weight matrix A_n that does not depend on θ, usually $A_n = I_k$, (ii) obtaining an $n^{1/2}$-consistent and asymptotically normal estimator $\widetilde{\theta}_0$ by minimizing (at least as best as one can) the criterion function $||A_n G_n(\theta)||^2$ over Θ, (iii) choosing a weight matrix $A_n(\theta)$ that is more efficient than A_n of step (i) and evaluating it at $\widetilde{\theta}_0$, and (iv) obtaining the final $n^{1/2}$-consistent and asymptotically normal GMM estimator $\widetilde{\theta}$ by minimizing $||A_n(\widetilde{\theta}_0)G_n(\theta)||^2$ over Θ (again, at least as best as one can).

The above procedure requires two global optimizations to guarantee consistency and asymptotic normality of the GMM estimator. The SR procedure we consider, on the other hand, does not require a global investigation of the GMM criterion function. The basic idea behind the SR procedure is that one can exploit knowledge of how large the GMM criterion function should be when evaluated at its minimizing θ value to tell whether a particular value is an approximate minimizer. This allows one to avoid carrying out a global investigation of the criterion function.

In particular, the minimized GMM criterion function (multiplied by n) has a χ_r^2 distribution asymptotically, where $r = k - d > 0$ is the number of over-identifying restrictions — a result well-known from the theory of tests of over-identifying restrictions. In consequence, one can see whether a trial value $\widehat{\theta}_0$ yields a value of the criterion function that is small relative to typical realizations of a χ_r^2 random variable. If it is, then $\widehat{\theta}_0$ must yield a criterion function value that is close to that of the globally minimizing value (though it may not be close enough to yield asymptotic normality of $\widehat{\theta}_0$).

By starting with such a value $\widehat{\theta}_0$ and doing j iterations of an NR procedure, one obtains a j-step estimator $\widehat{\theta}_j$ that is consistent and asymptotically normal. The j-step estimator $\widehat{\theta}_j$ could be taken to be the final GMM estimator. One can show, however, that any estimator $\widehat{\theta}$ that leads to a smaller criterion function value than $\widehat{\theta}_j$ has the same asymptotic distribution as $\widehat{\theta}_j$. Thus, we take the final GMM estimator $\widehat{\theta}$ to be the value which minimizes the criterion function over all values considered, including $\widehat{\theta}_j$. This choice is consistent with the idea that the GMM estimator is an optimization estimator.

Robinson (1988, Thms. 2–6) has shown that a j-step estimator is consistent and asymptotically normal if a suitable initial estimator is used. His results apply to differentiable GMM criterion functions. Our results given below extend his to include non-differentiable GMM criterion functions. More importantly, our results differ from his in terms of the choice of the initial estimator. Robinson's initial estimator is obtained by a global search over a regular grid of points with mesh size that is designed to yield an n^{ξ}-consistent estimator for some $0 < \xi \leq 1/2$. As argued in Section 2, such a global search is intractable in many problems. In addition, Robinson (1988) considers a random search to obtain an initial estimator that is n^{ξ}-consistent. Our results also differ from Robinson's in terms of the definition of the final GMM estimator, which in Robinson's case is the j-step estimator, whereas in our case it is the value that minimizes the GMM criterion function over the values already considered including the j-step estimator.

3.2. Description of the SR Approach

We now describe the SR approach in more detail. It entails the following steps:

(i) Starting with a trial estimator, $\widehat{\theta}_0$, one checks to see whether it satisfies the SR. That is, one checks to see if

$$(3.5) \qquad \|A_n(\widehat{\theta}_0)G_n(\widehat{\theta}_0)\|^2 \leq \frac{c_{r,n}}{n} ,$$

where $c_{r,n}$ is a constant defined below. Here, $A_n(\theta)$ must be an asymptotically optimal weight matrix. That is, $A_n(\theta)$ must be such that $A_n(\theta) \xrightarrow{p} A(\theta)$ for all θ in a neighborhood of θ_0 and $A(\theta_0)'A(\theta_0) = V^{-1}$, where V is the asymptotic covariance matrix of $n^{1/2}G_n(\theta_0)$. Typically, $\widehat{\theta}_0$ is the GMM estimator computed via the standard two-step procedure described above (though it need not be). It usually is the result of a multi-start algorithm that utilizes one or more starts. If $\widehat{\theta}_0$ passes the SR, one proceeds to the second step below. If not, one looks for a new trial estimator. For example, one might apply the multi-start algorithm with new starting values.

(ii) Given that $\widehat{\theta}_0$ passes the SR, one computes one-step, two-step, ..., j-step NR iterations starting from $\widehat{\theta}_0$, call them $\widehat{\theta}_1, \widehat{\theta}_2$, ..., $\widehat{\theta}_j$. The number j depends on $c_{r,n}$ (see below), but it is typically three or less. By definition, the j-step estimator is

$$\widehat{\theta}_j = \widehat{\theta}_{j-1} - (\widehat{\Gamma}'_{j-1}\widehat{A}'_{j-1}\widehat{A}_{j-1}\widehat{\Gamma}_{j-1})^{-1}\widehat{\Gamma}'_{j-1}\widehat{A}'_{j-1}\widehat{A}_{j-1}G_n(\widehat{\theta}_{j-1}) \tag{3.6}$$

for $j \geq 1$, where $\widehat{\Gamma}_{j-1} = \Gamma_n(\widehat{\theta}_{j-1})$, $\widehat{A}_{j-1} = A_n(\widehat{\theta}_{j-1})$, and $A_n(\theta)$ is an asymptotically optimal weight matrix.

(iii) One takes the final GMM estimator $\widehat{\theta}$ to be the value that minimizes $||A_n(\theta)G_n(\theta)||^2$ over all the values considered to this point including $\widehat{\theta}_0$, ..., $\widehat{\theta}_j$. (Or, if it is computationally more convenient, one can take $\widehat{\theta}$ to be the value that minimizes $||A_n(\theta)G_n(\theta)||^2$ over a subset of the values considered to this point, provided the subset includes $\widehat{\theta}_j$.) The final GMM estimator is guaranteed to be consistent and asymptotically normal with asymptotically efficient GMM covariance matrix. That is,

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0,\ (\Gamma'V^{-1}\Gamma)^{-1})\ , \tag{3.7}$$

where $\Gamma = \Gamma(\theta_0)$.

3.3. Choice of the Cutoff Value $c_{r,n}$

Here we discuss the choice of the cutoff value $c_{r,n}$. Given the trial value $\widehat{\theta}_0$, we want the SR to be such that if $\widehat{\theta}_0$ passes the SR, then $\widehat{\theta}_0$ is suitable for use as the initial estimator for the j-step estimator. It turns out that we need an SR that tells us whether $\widehat{\theta}_0$ is n^{ξ}-consistent for some $0 < \xi \leq 1/2$, because this guarantees that $\widehat{\theta}_0$ is sufficiently close to θ_0 to provide a good starting value. To this end, we show below that an estimator $\widehat{\theta}_0$ is n^{ξ}-consistent iff

$$||A_n(\widehat{\theta}_0)G_n(\widehat{\theta}_0)||^2 = O_p(n^{-2\xi})\ . \tag{3.8}$$

Thus, we need to determine a suitable sequence of cutoff values $\{c_{r,n}\ :\ n \geq 1\}$ such that $c_{r,n}/n$ is $O_p(n^{-2\xi})$ for some $0 < \xi \leq 1/2$.

A suitable sequence of cutoff values is one that is not too small, because if $c_{r,n}$ is too small there is a high probability that there is no value θ in Θ for which the criterion function is less than or equal to $c_{r,n}$. On the other hand, if the cutoff value $c_{r,n}$ is too large, one may end up with a poor initial estimator from which to start the j iterations.

To determine what is neither "too small" nor "too large," we first need to find an appropriate normalization of the criterion function $||A_n(\theta)G_n(\theta)||^2$, since the criterion function can be multiplied by any scalar constant without changing its relative magnitude for different values of θ. The natural normalization to use is to require the weight matrix $A_n(\theta)$ to be an asymptotically efficient weight matrix, as is done in Step (i) above. This choice of weight matrix yields a scale invariant criterion function. Multiplication of $G_n(\theta)$ by a scalar or any nonsingular matrix leaves $||A_n(\theta)G_n(\theta)||^2$ unchanged, because $A_n(\theta)$ changes to offset the change in $G_n(\theta)$.

This choice of weight matrix has a second advantage that is quite important. In particular, with this choice of weight matrix, the minimum over $\theta \in \Theta$ of the criterion function (multiplied by n) has a known asymptotic distribution that is nuisance parameter free. Its asymptotic distribution is a chi-square distribution with $r = k-d$ degrees of freedom. This result allows us to choose a cutoff value $c_{r,n}$ such that we know that with high probability there exists a value $\theta \in \Theta$ such that the criterion function evaluated at θ is less than the cutoff value.

Based on the asymptotic result referred to above, we approximate the distribution of $\inf_{\theta\in\Theta} n||A_n(\theta)G_n(\theta)||^2$ by a χ^2_r distribution. Using this approximation, if we let $c_{r,n}$ be the $(1-\alpha_n)$-quantile of a χ^2_r distribution, then the probability that

$$\inf_{\theta\in\Theta} ||A_n(\theta)G_n(\theta)||^2 \leq \frac{c_{r,n}}{n} \tag{3.9}$$

is approximately α_n. That is, α_n is the approximate probability that the chosen cutoff value $c_{r,n}$ is such that no value of θ satisfies the SR.

For most sample sizes, the choice of $\alpha_n = .05$ is reasonable. The corresponding $c_{r,n}$ values are given in Table I for $r = 1, ..., 20$. The choice of $\alpha_n = .05$ has the advantage that it yields a fairly small probability that the SR is too stringent, i.e., that no initial estimator satisfies the SR. Furthermore, it has the advantage that if one fails to find an initial estimator that satisfies the SR, then one can conclude that either (i) a test of the over-identifying restrictions rejects the null with the conventional significance level of .05 or (ii) the over-identifying restrictions hold but one is not able to find a suitable initial estimator due to the intractability of global optimization or due to an excessively stringent SR, which occurs with approximate probability .05. With the standard method of computing the GMM estimator, the second possibility is generally ignored when a test of over-identifying restrictions is carried out.

We note that for every paper in the literature that computes a GMM estimator, call it $\widehat{\theta}_0$, and for which the standard test of over-identifying restrictions fails to reject the null at a 5% significance level, the estimator $\widehat{\theta}_0$ satisfies the SR defined with $\alpha_n = .05$. Thus, in a very wide variety of cases, an initial estimator that satisfies the SR is available.

In some cases the standard test of over-identifying restrictions may reject the null at 5% for all parameter values $\widehat{\theta}_0$ that one has computational time to consider. It still may be helpful to use the criterion function $||A_n(\theta)G_n(\theta)||^2$ with an asymptotically efficient weight matrix $A_n(\theta)$ to compare different parameter estimates in such cases, because of the natural normalization that is provided by this weight matrix.

3.4. Choice of the Number of Iterations j

We now discuss the value of j (for the j-step estimator) that is required for given $c_{r,n}$ values. If $c_{r,n}/n = O_p(n^{-2\xi})$, then it turns out that one needs

$$j > -\log(2\xi)/\log 2\ . \tag{3.10}$$

For example, if $\xi \in (1/4, 1/2]$, (1/8,1/4], or (1/16,1/8], then it suffices to take j = 1, 2, or 3 respectively.

Note that $c_{r,n}$ need not depend on n. If it does not, then an initial estimator that satisfies the SR is n^{ξ}-consistent with $\xi = 1/2$. In this case, the above condition (3.10) just requires $j = 1$.

If $c_{r,n}$ grows with n, however, then the probability α_n that no value satisfies the SR goes to zero as $n \to \infty$, as may be deemed desirable. In fact, the relationship between α_n and $c_{r,n}$ as $c_{r,n} \to \infty$ with n is

$$\alpha_n \approx K c_{r,n}^{r/2} e^{-c_{r,n}/2} \tag{3.11}$$

for some constant K (see Johnson and Kotz (1970, p. 179)). In consequence, α_n declines to zero very quickly unless $c_{r,n}$ increases very slowly. For example, if one wants α_n to decline geometrically fast, say $\alpha_n \approx K_1 n^{-\lambda}$ for some $K_1 > 0$ and $\lambda > 0$, then $c_{r,n}$ must increase at a logarithmic rate and $c_{r,n}/n = O(n^{-2\xi})$ $\forall \xi < 1/2$. In this case, one can take $j = 1$.

Alternatively, if one wants α_n to decline exponentially fast, say $\alpha_n \approx K_2 e^{-\lambda n^\gamma}$ for some $K_2 > 0$, $\lambda > 0$, and $\gamma > 0$, then $c_{r,n}$ must satisfy $c_{r,n} \approx K_3 n^\gamma$ for some $K_3 > 0$. In this case, if $\gamma < 7/8$, then $c_{r,n} = O(n^{-2\xi})$ with $\xi > 1/16$ and one can take $j = 3$. Note that one would probably not want to choose $c_{r,n}$ such that it corresponds to $\xi \leq 1/16$ because this could yield an initial estimator that passes the SR but is not very close to θ_0.

In sum, for any choice of α_n (and corresponding $c_{r,n}$) that ranges between α_n (and $c_{r,n}$) being independent of n to α_n declining at the very quick rate of $\alpha_n \approx K_2 e^{-\lambda n^\gamma}$ for $\gamma < 7/8$, it suffices to take $j = 3$. Thus, a conservative strategy is to choose $j = 3$. This covers a very wide range of α_n values and is not unduly burdensome computationally since it involves at most two iterations more than may be strictly necessary.

4. RESULTS

In this section, we provide the results that establish that the SR procedure yields a consistent, asymptotically normal, and asymptotically efficient GMM estimator. The regularity conditions that we use are very close to those of Pakes and Pollard (1989), who provide a discussion of many of the assumptions. A feature of these assumptions is that they do not require that the GMM criterion function is differentiable in θ.

The conditions outlined in the paragraph containing equations (3.1)–(3.4) are NOT assumed to be in force throughout this section. We let "$\wedge$" denote the min operator, i.e., $a \wedge b = \min\{a, b\}$. We let $||\cdot||$ denote the Euclidean norm of vectors and matrices. Thus, for a matrix, $||B|| = (\text{tr}(B'B))^{1/2}$, where "tr" is the trace operator.

The first result establishes the basic asymptotic properties of the one-step estimator $\widehat{\theta}_1$ (i.e., $\widehat{\theta}_j$ with $j = 1$) based on the initial estimator $\widehat{\theta}_0$.

Assumption RC (Rate of Convergence): *$n^\xi(\widehat{\theta}_0 - \theta_0) = O_p(1)$ for some $0 < \xi \leq \frac{1}{2}$.*

ASSUMPTION 1: $G_n(\theta) \xrightarrow{p} G(\theta)$ for all θ in some neighborhood of θ_0 for some function $G(\cdot)$ that satisfies $G(\theta_0) = 0$, *$G(\cdot)$ is differentiable on a neighborhood of θ_0 with derivative $\Gamma(\cdot)$ that is Lipschitz at θ_0, and $\Gamma = \Gamma(\theta_0)$ is full rank d.*

ASSUMPTION 2: *Either* (a) *For every sequence $\{\delta_n : n \geq 1\}$ of numbers that converges to zero,*

$$n^{1/2} \sup_{\|\theta-\theta_0\|<\delta_n} \|G_n(\theta) - G(\theta) - G_n(\theta_0)\| = o_p(1)$$

or (b) *$G_n(\cdot)$ is differentiable on a neighborhood of θ_0 for all n with derivative that satisfies*

$$n^{1/4} \sup_{\|\theta-\theta_0\|<\delta_n} \left\| \frac{\partial}{\partial\theta'} G_n(\theta) - \Gamma(\theta) \right\| = o_p(1)$$

for every sequence $\{\delta_n : n \geq 1\}$ as in part (a).

ASSUMPTION 3: $\sqrt{n} G_n(\theta_0) \xrightarrow{d} N(0, V)$ *for some $k \times k$ covariance matrix V.*

ASSUMPTION 4: *For every sequence $\{\delta_n : n \geq 1\}$ of numbers that converges to zero, $\sup_{\|\theta-\theta_0\|<\delta_n} \|A_n(\theta) - A\| = o_p(1)$ for some $k \times k$ nonsingular matrix A.*

ASSUMPTION 5: *For all $0 < \gamma \leq 1/4$ and all $0 < M < \infty$,*

$$n^{\gamma} \sup_{\|\theta-\theta_0\|\leq Mn^{-\gamma}} \|\Gamma_n(\theta) - \Gamma(\theta)\| = O_p(1) \ .$$

ASSUMPTION 6: *V is nonsingular and $A'A = V^{-1}$.*

The above assumptions are fairly standard with the possible exception of Assumption 2(a). The latter is employed when $G_n(\theta)$ is not differentiable in θ. It can be verified using the stochastic equicontinuity results given in Pakes and Pollard (1989) or Andrews (1993, 1994).

THEOREM 1: *Under Assumptions* RC *and* 1–5,

(a) $n^{(2\xi)\wedge\frac{1}{2}}(\widehat{\theta}_1 - \theta_0) = O_p(1)$,

(b) *if $\xi > \frac{1}{4}$ in Assumption* RC, *then*

$$n^{1/2}(\widehat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, (\Gamma'A'A\Gamma)^{-1}\Gamma'A'AVA'A\Gamma(\Gamma'A'A\Gamma)^{-1}) \ ,$$

(c) *if* $\xi > \frac{1}{4}$ *in Assumption* RC *and Assumption* 6 *also holds, then*

$$n^{1/2}(\widehat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, (\Gamma' V^{-1}\Gamma)^{-1}) \ \ \textit{and} \ \ n\|A_n(\widehat{\theta}_1)G_n(\widehat{\theta}_1)\|^2 \xrightarrow{d} \chi^2_r \ .$$

By taking the results of Theorem 1 for the one-step estimator and iterating j times, we obtain the desired results for the j-step estimator.

Corollary 1: (a) *Under Assumptions* RC *and* 1–5, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' A' A\Gamma)^{-1}\Gamma' A' AVA' A\Gamma(\Gamma' A' A\Gamma)^{-1}) \ .$$

(b) *Under Assumptions* RC *and* 1–6, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' V^{-1}\Gamma)^{-1}) \ \ \textit{and} \ \ n\|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j)\|^2 \xrightarrow{d} \chi^2_r \ .$$

Next, we determine necessary and sufficient conditions for the initial estimator $\widehat{\theta}_0$ to be n^{ξ}-consistent.

Assumption D (Definition of Initial Estimator): $\|A_n(\widehat{\theta}_0)G_n(\widehat{\theta}_0)\|^2 = O_p(n^{-2\xi})$ *for some* $0 < \xi \leq \frac{1}{2}$.

Assumption C (Consistency): $\widehat{\theta}_0 \xrightarrow{p} \theta_0$.

Note that an initial estimator $\widehat{\theta}_0$ that satisfies the SR (3.5) also satisfies Assumption D, because $c_{r,n}$ is chosen in Section 3 such that $c_{r,n}/n = O_p(n^{-2\xi})$ for some $0 < \xi \leq \frac{1}{2}$.

Theorem 2: *Under Assumptions* 1–4, *Assumption* RC *holds if and only if Assumptions* D *and* C *hold.*

Combining Theorem 2 and Corollary 1 gives the following result.

Corollary 2: (a) *Under Assumptions* D, C, *and* 1– 5, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' A' A\Gamma)^{-1}\Gamma' A' AVA' A\Gamma(\Gamma' A' A\Gamma)^{-1}) \ .$$

(b) *Under Assumptions* D, C, *and* 1–6, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' V^{-1}\Gamma)^{-1}) \ \ and \ \ n\|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j)\|^2 \xrightarrow{d} \chi^2_r \ .$$

We now specify two sets of sufficient conditions for consistency of $\widehat{\theta}_0$ (Assumption C).

Assumption D1: $\|A_n(\widehat{\theta}_0)G_n(\widehat{\theta}_0)\| = o_p(1)$.

Assumption 7: $G_n(\theta_0) = o_p(1)$.

Assumption 8: $\sup_{\|\theta-\theta_0\|>\delta} \|G_n(\theta)\|^{-1} = O_p(1) \ \ \forall\delta > 0$.

Assumption 9: $\|A_n(\theta_0)\| = O_p(1)$ and $\sup_{\theta\in\Theta} \|A_n^{-1}(\theta)\| = O_p(1)$.

Assumption 10: $\sup_{\theta\in\Theta} \|G_n(\theta) - G(\theta)\| = o_p(1)$ *for some function* $G(\cdot)$ *on* Θ.

Assumption 11: $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| > 0 \ \ \forall\delta > 0$.

Assumption 12: $\sup_{\theta\in\Theta} \|A_n(\theta) - A(\theta)\| = o_p(1)$, $\sup_{\theta\in\Theta} \|A(\theta)\| < \infty$, *and* $\inf_{\theta\in\Theta} \lambda_{\min}(A(\theta)) > 0$ *for some function* $A(\cdot)$ *on* Θ.

Proposition 1: *Under Assumptions* D1 *and either* 7–9 *or* 10–12, $\widehat{\theta}_0 \xrightarrow{p} \theta_0$.

Proposition 1 is due to Pakes and Pollard (1989).

The following Corollary is similar to Theorem 2 and Corollary 2. It replaces the consistency assumption (Assumption C) by sufficient conditions for consistency. Part (a) is established by combining Proposition 1 and Theorem 2. Parts (b) and (c) are obtained by combining Proposition 1 and Corollary 1.

Corollary 3: (a) *Under Assumptions* 1–4 *and either* 8–9 *or* 10–12, *Assumption* RC *holds if and only if Assumption* D *holds.*

(b) *Under Assumptions* D, 1–5, *and either* 8–9 *or* 10–12, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' A' A\Gamma)^{-1}\Gamma' A' A V A' A\Gamma(\Gamma' A' A\Gamma)^{-1}) \ .$$

(c) *Under Assumptions* D, 1–6, *and either* 8–9 *or* 10–12, *for all* $j > -\log(2\xi)/\log 2$,

$$n^{1/2}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, (\Gamma' V^{-1}\Gamma)^{-1}) \ \ \textit{and} \ \ n\|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j)\|^2 \xrightarrow{d} \chi^2_r \ .$$

Next, we show that the value of the GMM criterion function minimized over $\theta \in \Theta$ is asymptotically χ^2_r.

Assumption 13: *θ_0 is in the interior of Θ.*

Theorem 3: *Under Assumptions* 1–4, 6, 13, *and either* 8–9 *or* 10–12,

$$n \inf_{\theta\in\Theta} \|A_n(\theta)G_n(\theta)\|^2 \xrightarrow{d} \chi^2_r \ .$$

We now provide results that show that the final GMM estimator $\widehat{\theta}$ is consistent and asymptotically normal.

Let $\widehat{\widehat{\theta}}$ denote some estimator. We say $\widehat{\widehat{\theta}}$ satisfies Condition (i) if

Condition (i): $\widehat{\widehat{\theta}} \xrightarrow{p} \theta_0$ and $\|A_n(\widehat{\widehat{\theta}})G_n(\widehat{\widehat{\theta}})\| \le \inf_{\theta\in\Theta} \|A_n(\theta)G_n(\theta)\| + o_p(n^{-1/2})$.

Theorem 4: *Under Assumptions* D, 1–5, 13, *and either* 8–9 *or* 10–12, $\widehat{\theta}_j$ *satisfies Condition* (i) *for all* $j > -\log(2\xi)/\log 2$.

Proposition 2: *Under Assumptions* 1–4 *and* 13, *if an estimator* $\widehat{\widehat{\theta}}$ *satisfies Condition* (i), *then*

$$n^{1/2}(\widehat{\widehat{\theta}} - \theta_0) \xrightarrow{d} N(0, \ (\Gamma' A' A \Gamma)^{-1}\Gamma' A' A V A' A \Gamma(\Gamma' A' A \Gamma)^{-1}) \ .$$

Proposition 2 is due to Pakes and Pollard (1989).

The definition of the final GMM estimator $\widehat{\theta}$ and Theorem 4 imply that $\widehat{\theta}$ satisfies Condition (i). Proposition 2 then gives the desired asymptotic normality result for $\widehat{\theta}$.

Corollary 4: *Let $\widehat{\theta}$ be an estimator that satisfies*

$$\|A_n(\widehat{\theta})G_n(\widehat{\theta})\|^2 \le \|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j)\|^2 \ .$$

Then, under Assumptions D, 1–5, 13, *and either* 8–9 *or* 10–12,

(a) $n^{1/2}(\widehat{\theta}-\theta_0) \xrightarrow{d} N(0,\ (\Gamma' A' A \Gamma)^{-1}\Gamma' A' A V A' A \Gamma(\Gamma' A' A \Gamma)^{-1})$ *provided* $j > -\log(2\xi)/\log 2$,

(b) *if Assumption* 6 *also holds*

$$n^{1/2}(\widehat{\theta}-\theta_0) \xrightarrow{d} N(0,\ (\Gamma' V^{-1}\Gamma)^{-1})$$

provided $j > -\log(2\xi)/\log 2$.

APPENDIX

PROOF OF THEOREM 1: First, we consider the case where Assumption 2(a) holds. Using Assumptions RC and 1, element by element mean–value expansions yield

$$G(\widehat{\theta}_0) = \Gamma(\theta^*)(\widehat{\theta}_0 - \theta_0) \ , \tag{A.1}$$

where θ^* lies on the line segment joining $\widehat{\theta}_0$ and θ_0 and may differ across the rows of $\Gamma(\theta^*)$. This result and Assumption 2(a) give

$$\begin{aligned} & \|G_n(\widehat{\theta}_0) - \Gamma(\theta^*)(\widehat{\theta}_0 - \theta_0) - G_n(\theta_0)\| \\ = \ & \|G_n(\widehat{\theta}_0) - G(\widehat{\theta}_0) - G_n(\theta_0)\| = o_p(n^{-1/2}) \ . \end{aligned} \tag{A.2}$$

(Assumption 2(a) is applicable here because consistency of $\widehat{\theta}_0$ implies the existence of a sequence of constants $\{\delta_n : n \geq 1\}$ for which $\delta_n \to 0$ and $P(\|\widehat{\theta}_0 - \theta_0\| > \delta_n) \to 0$.)

In consequence, using the definition of $\widehat{\theta}_1$, we obtain

$$\begin{aligned} n^{(2\xi)\wedge\frac{1}{2}}(\widehat{\theta}_1 - \theta_0) &= n^{(2\xi)\wedge\frac{1}{2}}(\widehat{\theta}_0 - \theta_0) - (\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0\widehat{\Gamma}_0)^{-1}\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0 n^{(2\xi)\wedge\frac{1}{2}} G_n(\widehat{\theta}_0) \\ &= (\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0\widehat{\Gamma}_0)^{-1}\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0 n^{\xi\wedge\frac{1}{4}}(\widehat{\Gamma}_0 - \Gamma(\theta^*)) n^{\xi\wedge\frac{1}{4}}(\widehat{\theta}_0 - \theta_0) \\ &\quad - (\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0\widehat{\Gamma}_0)^{-1}\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0 n^{(2\xi)\wedge\frac{1}{2}} G_n(\theta_0) + o_p(1) \ . \end{aligned} \tag{A.3}$$

The components of the right-hand side of (A.3) exhibit the following asymptotic behavior. By Assumptions RC, 1, 4, and 5, $(\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0\widehat{\Gamma}_0)^{-1}\widehat{\Gamma}_0'\widehat{A}_0'\widehat{A}_0 = (\Gamma'A'A\Gamma)^{-1}\Gamma'A'A + o_p(1)$. By Assumptions RC, 1, and 5,

$$n^{\xi\wedge\frac{1}{4}}\|\widehat{\Gamma}_0 - \Gamma(\theta^*)\| \leq n^{\xi\wedge\frac{1}{4}}\|\Gamma_n(\widehat{\theta}_0) - \Gamma(\widehat{\theta}_0)\| + n^{\xi\wedge\frac{1}{4}}\|\Gamma(\widehat{\theta}_0) - \Gamma(\theta^*)\| = O_p(1) \ , \tag{A.4}$$

where the equality utilizes the Lipschitz condition on $\Gamma(\cdot)$. By Assumption RC, $n^{\xi\wedge\frac{1}{4}}(\widehat{\theta}_0 - \theta_0)$ is $O_p(1)$ for $0 < \xi \leq \frac{1}{4}$ and $o_p(1)$ for $\frac{1}{4} < \xi \leq \frac{1}{2}$. By Assumption 3, $n^{(2\xi)\wedge\frac{1}{2}}G_n(\theta_0)$ is $o_p(1)$ for $0 < \xi < \frac{1}{4}$ and asymptotically $N(0, V)$ for $\xi \geq \frac{1}{4}$. Substitution of these results in (A.3) gives parts (a) and (b) of the Theorem.

The first result of part (c) of the Theorem holds by part (b) by substituting V^{-1} for $A'A$. The second result of part (c) is obtained under Assumption 2(a) as follows. Equations (A.1) and

(A.2) hold with $\widehat{\theta}_0$ replaced by $\widehat{\theta}_1$, since $\widehat{\theta}_1$ is consistent by part (b). In consequence, (A.2) revised and (A.3) yield

$$\begin{aligned}
\text{(A.5)}\qquad n^{1/2}G_n(\widehat{\theta}_1) &= \Gamma(\theta^*)n^{1/2}(\widehat{\theta}_1-\theta_0)+n^{1/2}G_n(\theta_0)\\
&= -[\Gamma+o_p(1)][(\Gamma'V^{-1}\Gamma)^{-1}\Gamma'V^{-1}n^{1/2}G_n(\theta_0)+o_p(1)]+n^{1/2}G_n(\theta_0)\ \text{and}\\
n^{1/2}A_n(\widehat{\theta}_1)G_n(\widehat{\theta}_1) &= A(I_k-\Gamma(\Gamma'V^{-1}\Gamma)^{-1}\Gamma'V^{-1})n^{1/2}G_n(\theta_0)+o_p(1)\\
&\xrightarrow{d} N(0,\, I_k-A\Gamma(\Gamma'V^{-1}\Gamma)^{-1}\Gamma'A')\ .
\end{aligned}$$

Since $A\Gamma(\Gamma'V^{-1}\Gamma)^{-1}\Gamma'A'$ is symmetric, idempotent, and of rank d, $I_k - A\Gamma(\Gamma'V^{-1}\Gamma)^{-1}\Gamma'A'$ is a projection matrix onto a $(k-d)$-dimensional space. The second result of part (c) now follows by (A.5), the continuous mapping theorem, and the definition of the χ^2_r distribution.

Next, we consider the case where Assumption 2(b) holds. Using Assumptions RC and 2(b), element by element mean–value expansions give

$$\text{(A.6)}\qquad G_n(\widehat{\theta}_0) = G_n(\theta_0)+\frac{\partial}{\partial\theta'}G_n(\theta^*)(\widehat{\theta}_0-\theta_0)\ ,$$

where θ^* lies on the line segment joining $\widehat{\theta}_0$ and θ_0 and may differ across rows of $\frac{\partial}{\partial\theta'}G_n(\theta^*)$. Equation (A.6) replaces (A.2). The proof is now the same as above from equation (A.3) on with $\Gamma(\theta^*)$ replaced by $\frac{\partial}{\partial\theta'}G_n(\theta^*)$. On the right-hand side of (A.4), $n^{\xi\wedge\frac{1}{4}}\left\|\Gamma(\widehat{\theta}_0)-\frac{\partial}{\partial\theta'}G_n(\theta^*)\right\|$ is $O_p(1)$, because it is bounded by

$$\text{(A.7)}\qquad n^{\xi\wedge\frac{1}{4}}\|\Gamma(\widehat{\theta}_0)-\Gamma(\theta^*)\|+n^{\xi\wedge\frac{1}{4}}\left\|\frac{\partial}{\partial\theta'}G_n(\theta^*)-\Gamma(\theta^*)\right\|\ ,$$

which is $O_p(1)$ by Assumptions RC, 1, and 2(b). □

Proof of Corollary 1: Corollary 1 follows from Theorem 1 by applying Theorem 1 part (a) recursively with the initial estimator $\widehat{\theta}_0$ given by $\widehat{\theta}_0$, $\widehat{\theta}_1$, ..., $\widehat{\theta}_{j-2}$ to obtain $n^{(2\xi)\wedge\frac{1}{2}}(\widehat{\theta}_1-\theta_0)=O_p(1)$, $n^{(4\xi)\wedge\frac{1}{2}}(\widehat{\theta}_2-\theta_0)=O_p(1)$, ..., $n^{(2^{j-1}\xi)\wedge\frac{1}{2}}(\widehat{\theta}_{j-1}-\theta_0)=O_p(1)$. For $j>-\log(2\xi)/\log 2$, we have $2^{j-1}\xi>\frac{1}{4}$ and Theorem 1 parts (b) and (c) applied with the initial estimator $\widehat{\theta}_0$ set equal to $\widehat{\theta}_{j-1}$ give parts (a) and (b) of the Corollary respectively. □

PROOF OF THEOREM 2: Assumption RC obviously implies Assumption C. First we show Assumption RC implies Assumption D when Assumption 2(a) holds. Using Assumptions RC, 1, 2(a), and 3, we obtain

$$\begin{aligned} \|G_n(\widehat{\theta}_0)\| &\leq \|G_n(\widehat{\theta}_0) - G(\widehat{\theta}_0) - G_n(\theta_0)\| + \|G(\widehat{\theta}_0)\| + \|G_n(\theta_0)\| \\ &= o_p(n^{-1/2}) + C_1\|\widehat{\theta}_0 - \theta_0\| + o_p(n^{-1/2}) = O_p(n^{-\xi}) \ , \end{aligned} \tag{A.8}$$

where the first equality uses the fact that Assumption 1 implies the existence of a positive constant C_1 for which

$$\|G(\theta)\| \leq C_1\|\theta - \theta_0\| \text{ for } \theta \text{ near } \theta_0 \ . \tag{A.9}$$

By (A.8) and Assumption 4, we obtain the desired result:

$$\|\widehat{A}_0 G_n(\widehat{\theta}_0)\| = \|(A + o_p(1))G_n(\widehat{\theta}_0)\| = O_p(n^{-\xi}) \ . \tag{A.10}$$

To show Assumptions D and C imply RC under Assumption 2(a), we write

$$\begin{aligned} \|\widehat{A}_0 G(\widehat{\theta}_0)\| &\leq \|\widehat{A}_0(G_n(\widehat{\theta}_0) - G(\widehat{\theta}_0) - G_n(\theta_0))\| + \|\widehat{A}_0 G_n(\widehat{\theta}_0)\| + \|\widehat{A}_0 G_n(\theta_0)\| \\ &= o_p(n^{-1/2}) + O_p(n^{-\xi}) + o_p(n^{-1/2}) = O_p(n^{-\xi}) \end{aligned} \tag{A.11}$$

using Assumptions D, C, 2(a), 3, and 4. This result and the triangle inequality give

$$\begin{aligned} O_p(n^{-\xi}) &= \|\widehat{A}_0 G(\widehat{\theta}_0)\| \geq \|AG(\widehat{\theta}_0)\| + \|(\widehat{A}_0 A^{-1} - I_k)AG(\widehat{\theta}_0)\| \\ &= (1 + o_p(1))\|AG(\widehat{\theta}_0)\| \geq (1 + o_p(1))C_2\|\widehat{\theta}_0 - \theta_0\| \ , \end{aligned} \tag{A.12}$$

where the second inequality holds because Assumptions 1 and 4 imply that there is a positive constant C_2 for which $\|AG(\theta)\| \geq C_2\|\theta - \theta_0\|$ for θ near θ_0.

Next, we show Assumption RC implies Assumption D under Assumption 2(b). Using (A.6) (which relies on Assumption 2(b)), we obtain

$$\|\widehat{A}_0 G_n(\widehat{\theta}_0)\| \leq \|\widehat{A}_0 G_n(\theta_0)\| + \left\|\widehat{A}_0 \frac{\partial}{\partial\theta'} G_n(\theta^*)(\widehat{\theta}_0 - \theta_0)\right\| = O_p(n^{-\xi}) \ , \tag{A.13}$$

where the equality uses Assumptions RC, 2(b), 3, and 4.

For the converse, we pre-multiply (A.6) by $\widehat{A}_0 = A + o_p(1)$ to get

$$\text{(A.14)} \qquad \|(A + o_p(1))(\Gamma + o_p(1))(\widehat{\theta}_0 - \theta_0)\| \leq \|\widehat{A}_0 G_n(\widehat{\theta}_0)\| + \|\widehat{A}_0 G_n(\theta_0)\| = O_p(n^{-\xi})$$

using Assumptions D, C, 2(b), 3, and 4. Since $A\Gamma$ is full rank d, this implies $\|\widehat{\theta}_0 - \theta_0\| = O_p(n^{-\xi})$. □

Proof of Proposition 1: To establish consistency of $\widehat{\theta}_0$ under Assumptions D1 and 7–9, we apply Pakes and Pollard's (1989) Theorem 3.1 with their $G_n(\cdot)$ replaced by our $A_n(\cdot)G_n(\cdot)$. Assumption D1 is equivalent to Pakes and Pollard's altered condition (i) of their Theorem 3.1. Assumptions 7, 8, and 9 are equivalent to Pakes and Pollard's unaltered condition (ii) of Theorem 3.1, unaltered condition (iii) of Theorem 3.1, and conditions (a) and (b) of their Lemma 3.4, respectively. By Lemma 3.4, conditions (ii) and (iii) of Theorem 3.1 then hold with $G_n(\cdot)$ replaced by $A_n(\cdot)G_n(\cdot)$. Consistency follows by Theorem 3.1.

To establish consistency of $\widehat{\theta}_0$ under Assumptions D1 and 10–12, we apply Pakes and Pollard's Corollary 3.2 with their $G_n(\cdot)$ and $G(\cdot)$ replaced by our $A_n(\cdot)G_n(\cdot)$ and $A(\cdot)G(\cdot)$ respectively. The altered conditions (i) and (ii) of their Corollary 3.2 hold given Assumptions D1, 11, and 12. The altered condition (iii) of Corollary 3.2 holds by the following inequality:

$$\text{(A.15)} \qquad \begin{aligned} &\sup_{\theta \in \Theta} \frac{\|A_n(\theta)G_n(\theta) - A(\theta)G(\theta)\|}{1 + \|A_n(\theta)G_n(\theta)\| + \|A(\theta)G(\theta)\|} \\ &\leq \sup_{\theta \in \Theta} \|A_n(\theta)(G_n(\theta) - G(\theta))\| + \sup_{\theta \in \Theta} \frac{\|(A_n(\theta) - A(\theta))G(\theta)\|}{1 + \|A(\theta)G(\theta)\|} = o_p(1) \end{aligned}$$

using Assumptions 10 and 12. □

Proof of Theorem 3: First we consider the case where Assumption 2(a) holds. Let $\widetilde{\theta}$ be an estimator that satisfies

$$\text{(A.16)} \qquad \|A_n(\widetilde{\theta})G_n(\widetilde{\theta})\| \leq \inf_{\theta \in \Theta} \|A_n(\theta)G_n(\theta)\| + o_p(n^{-1/2}) \; .$$

We show that the conditions of Pakes and Pollard's (1989) Theorem 3.3 hold with the alterations that $\widetilde{\theta}_n$ is replaced by $\widetilde{\theta}$, $G_n(\cdot)$ is replaced by $A_n(\cdot)G_n(\cdot)$, and $G(\cdot)$ is replaced by $AG(\cdot)$. Since

$\inf_{\theta \in \Theta} ||A_n(\theta)G_n(\theta)|| \leq A_n(\theta_0)G_n(\theta_0) = O_p(n^{-1/2})$ by Assumptions 3 and either 9 or 12, $\widetilde{\theta}$ satisfies Assumption D. In consequence, $\widetilde{\theta} \xrightarrow{p} \theta_0$ by Proposition 1. Furthermore, by (A.16), the altered condition (i) of Theorem 3.3 holds. By Assumptions 1, 2(a), 3, and 13, the unaltered conditions (ii), (iii), (iv), and (v) of Theorem 3.3 hold respectively. By Assumption 4, Pakes and Pollard's Lemma 3.5 applies. It implies that the altered conditions (ii)-(v) of Theorem 3.3 hold. Hence, all the conditions of Theorem 3.3 hold. By Pakes and Pollard's proof of Theorem 3.3 with the alterations listed above, we get

$$\text{(A.17)} \quad \begin{aligned} ||A_n(\widetilde{\theta})G_n(\widetilde{\theta})|| &= ||A_n(\theta_n^*)L_n(\theta_n^*)|| + o_p(n^{-1/2}) \\ &= ||[A + o_p(1)][-\Gamma(\Gamma'A'A\Gamma)^{-1}\Gamma'A'AG_n(\theta_0) + G_n(\theta_0)]|| + o_p(n^{-1/2}) \ , \end{aligned}$$

where $L_n(\theta) = \Gamma(\theta - \theta_0) + G_n(\theta_0)$ and θ_n^* minimizes $||A_n(\theta)L_n(\theta)||$ over Θ. Now, by the same argument as used in and below (A.5), $n||A_n(\widetilde{\theta})G_n(\widetilde{\theta})||^2 \xrightarrow{d} \chi_r^2$. By (A.16) and the inequality $\inf_{\theta \in \Theta} ||A_n(\theta)G_n(\theta)|| \leq ||A_n(\widetilde{\theta})G_n(\widetilde{\theta})||$, we see that $n \inf_{\theta \in \Theta} ||A_n(\theta)G_n(\theta)||^2$ has the same asymptotic distribution as $n||A_n(\widetilde{\theta})G_n(\widetilde{\theta})||^2$.

Next, for the case where Assumption 2(b) holds, the above proof goes through provided (A.17) can be established. Equation (A.17) holds by the proof of Theorem 3.3 of Pakes and Pollard with $\sqrt{n}$-consistency established via Theorem 2 above and with the argument of (A.6) used in place of Pakes and Pollard's condition (iii) wherever the latter is used. □

Proof of Theorem 4: Define $\widetilde{\theta}$, $L_n(\theta)$, and θ_n^* as in the proof of Theorem 3. By the proof of Theorem 3, under Assumption 2(a) or 2(b),

$$\text{(A.18)} \qquad ||A_n(\widetilde{\theta})G_n(\widetilde{\theta})|| = ||A_n(\theta_n^*)L_n(\theta_n^*)|| + o_p(n^{-1/2}) \ .$$

By the proof of Theorem 1 above,

$$\text{(A.19)} \qquad \begin{aligned} n^{1/2}(\widehat{\theta}_j - \theta_0) &= -(\Gamma'A'A\Gamma)^{-1}\Gamma'A'An^{1/2}G_n(\theta_0) + o_p(1) \\ &= n^{1/2}(\theta_n^* - \theta_0) + o_p(1) \ . \end{aligned}$$

By the definition of $L_n(\cdot)$, this yields

$$\text{(A.20)} \qquad ||A_n(\widehat{\theta}_j)L_n(\widehat{\theta}_j) - A_n(\theta_n^*)L_n(\theta_n^*)|| = o_p(n^{-1/2}) \ .$$

Next, by the proof of (A.2), under Assumption 2(a) or 2(b), we have

$$\|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j) - A_n(\widehat{\theta}_j)L_n(\widehat{\theta}_j)\| = o_p(n^{-1/2}) \ . \tag{A.21}$$

Equations (A.18), (A.20), and (A.21) combine to give

$$\begin{aligned} \|A_n(\widehat{\theta}_j)G_n(\widehat{\theta}_j)\| &= \|A_n(\widetilde{\theta})G_n(\widetilde{\theta})\| + o_p(n^{-1/2}) \\ &\leq \inf_{\theta\in\Theta} \|A_n(\theta)G_n(\theta)\| + o_p(n^{-1/2}) \end{aligned} \tag{A.22}$$

using the definition of $\widetilde{\theta}$. □

Proof of Proposition 2: The proof when Assumption 2(a) holds is given by Theorem 3.3 and Lemma 3.5 of Pakes and Pollard (1989).

Now suppose Assumption 2(b) holds. First, we show that $\widehat{\widehat{\theta}}$ is $n^{1/2}$-consistent. We have

$$n^{1/2}\|A_n(\widehat{\widehat{\theta}})G_n(\widehat{\widehat{\theta}})\| \leq n^{1/2}\|A_n(\theta_0)G_n(\theta_0)\| + o_p(1) = O_p(1) \ , \tag{A.23}$$

where the inequality holds by Condition (i) and the equality by Assumptions 3 and 4.

Element by element mean value expansions give

$$n^{1/2}G_n(\widehat{\widehat{\theta}}) = n^{1/2}G_n(\theta_0) + \frac{\partial}{\partial\theta'}G_n(\overline{\theta})n^{1/2}(\widehat{\widehat{\theta}} - \theta_0) \ , \tag{A.24}$$

where $\overline{\theta}$ lies between $\widehat{\widehat{\theta}}$ and θ_0 and may differ across rows of $\frac{\partial}{\partial\theta'}G_n(\overline{\theta})$, using Assumption 2(b). Pre-multiplication by $A_n(\widehat{\widehat{\theta}})$ yields:

$$\begin{aligned} n^{1/2}A_n(\widehat{\widehat{\theta}})G_n(\widehat{\widehat{\theta}}) &= A_n(\widehat{\widehat{\theta}})n^{1/2}G_n(\theta_0) + A_n(\widehat{\widehat{\theta}})\frac{\partial}{\partial\theta'}G_n(\overline{\theta})n^{1/2}(\widehat{\widehat{\theta}} - \theta_0) \quad \text{and} \\ O_p(1) &= O_p(1) + (A\Gamma + o_p(1))n^{1/2}(\widehat{\widehat{\theta}} - \theta_0) \ , \end{aligned} \tag{A.25}$$

where the second equation follows from the first, (A.23), consistency of $\widehat{\widehat{\theta}}$, and Assumptions 1, 2(b), 3, and 4. Equation (A.25) implies that $n^{1/2}(\widehat{\widehat{\theta}} - \theta_0) = O_p(1)$ since $A\Gamma$ is full rank d $(< k)$.

We now follow the proof of Pakes and Pollard's (1989) Theorem 3.3 and point out the alterations that need to be made given Assumption 2(b) holds rather than Assumption 2(a). Let

$$L_n(\theta) = \Gamma(\theta - \theta_0) + G_n(\theta_0) \ . \tag{A.26}$$

Pakes and Pollard's (1989) last equation on p. 1041 is replaced by

$$\|A_n(\hat{\hat{\theta}})G_n(\hat{\hat{\theta}}) - AL_n(\hat{\hat{\theta}})\| = o_p(n^{-1/2}) \; , \tag{A.27}$$

which holds by (A.25), consistency of $\hat{\hat{\theta}}$, and Assumptions 1, 2(b), and 4.

Let θ^* be the value that minimizes $\|AL_n(\theta)\|^2$ over Θ. By Pakes ad Pollard's argument on p. 1042, θ^* is consistent. Now, an analogous argument to that given above yields

$$\|A_n(\theta^*)G_n(\theta^*) - AL_n(\theta^*)\| = o_p(n^{-1/2}) \; . \tag{A.28}$$

The remainder of the proof is as in Pakes and Pollard (1989). □

Proof of Corollary 4: By assumption and Theorem 4, we have

$$\begin{aligned} \|A_n(\hat{\theta})G_n(\hat{\theta})\| &\leq \|A_n(\hat{\theta}_j)G_n(\hat{\theta}_j)\| \\ &\leq \inf_{\theta \in \Theta} \|A_n(\theta)G_n(\theta)\| + o_p(n^{-1/2}) \end{aligned} \tag{A.29}$$

provided $j > -\log(2\xi)/\log 2$. Now, the right-hand side is $O_p(n^{-1/2})$ by Theorem 3. In consequence, $\hat{\theta}$ is consistent for θ_0 by applying Proposition 1 with $\hat{\theta}_0$ equal to $\hat{\theta}$. Consistency of $\hat{\theta}$ plus (A.29) imply that $\hat{\theta}$ satisfies Condition (i). Proposition 2 now yields the results of Corollary 4. □

FOOTNOTE

[1] I thank Anne, Brett, Alex, and many friends for their support over the past year. I thank Ariel Pakes for helpful comments. I gratefully acknowledge the research support of the National Science Foundation via grant numbers SBR–9121914 and SBR–9410675.

TABLE I

Cutoff Values $c_{r,n}$ that Correspond to α_n = .05

r	1	2	3	4	5	6	7	8	9	10
$c_{r,n}$	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31

r	11	12	13	14	15	16	17	18	19	20
$c_{r,n}$	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41

REFERENCES

Andrews, D. W. K. (1993): "An Introduction to Econometric Applications of Empirical Process Theory for Dependent Random Variables," *Econometric Reviews* 12, 183–216.

——— (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Vol. 4*, ed. by R. F. Engle and D. McFadden, pp. 2247–2294. New York: North Holland.

Floudas, C. A. and P. M. Pardalos (1992): *Recent Advances in Global Optimization.* Princeton, NJ: Princeton University Press.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994): "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics* 60, 65–99.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029–1054.

Horst, R. and H. Tuy (1992): *Global Optimization*, 2nd ed. New York: Springer–Verlag.

Huber, P. J. (1967): "The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, ed. by L. M. LeCam and J. Neyman. Berkeley: University of California Press.

Johnson, N. L. and S. Kotz (1970): *Continuous Univariate Distributions — 1.* New York: Wiley.

Nemirovsky, A. S. and D. B. Yudin (1983): *Problem Complexity and Method Efficiency in Optimization.* New York: Wiley–Interscience.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57, 1027–1057.

Pratt, J. W. (1981): "Concavity of the Log-Likelihood," *Journal of the American Statistical Association* 76, 137–159.

Robinson, P. M. (1988): "The Stochastic Difference Between Econometric Statistics," *Econometrica* 56, 531–548.

Traub, J. F., G. W. Wasilkowski, and H. Wosniakowski (1988): *Information-Based Complexity.* New York: Academic Press.

Veall, M. R. (1990): "Testing for a Global Maximum in an Econometric Context," *Econometrica* 58, 1459–1465.