Econometrica, Vol. 57, No. 5 (September, 1989), 1027–1057

SIMULATION AND THE ASYMPTOTICS OF OPTIMIZATION ESTIMATORS

BY ARIEL PAKES AND DAVID POLLARD[1]

A general central limit theorem is proved for estimators defined by minimization of the length of a vector-valued, random criterion function. No smoothness assumptions are imposed on the criterion function, in order that the results might apply to a broad class of simulation estimators. Complete analyses of two simulation estimators, one introduced by Pakes and the other by McFadden, illustrate the application of the general theorems.

KEYWORDS: Computationally intractable probabilities, discrete choice, aggregation, simulation estimators, discontinuous objective functions, Vapnik Cervonenkis classes, empirical processes.

1. INTRODUCTION

CONSIDER A MODEL in which the true value, θ_0, of a parameter vector is implicitly defined as the unique solution to an equation $G(\theta) = 0$, for a suitable vector-value function, G. A natural way to estimate θ_0 is to construct a sequence $\{G_n\}$ of random functions that converges to G in some sense, then find the $\hat{\theta}_n$ that makes $G_n(\hat{\theta}_n)$ as close to zero as possible. This paper presents conditions under which such a $\hat{\theta}_n$ converges to θ_0 and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ satisfies a central limit theorem. We avoid smoothness assumptions on G_n, so that G_n can be a discontinuous function of θ.

Our analysis is motivated by a desire to obtain the asymptotic properties of a broad class of simulation estimators: estimators that arise in cases where simulation experiments are used to provide estimates of complicated functions that otherwise could not (or could not easily) be evaluated. As our examples will illustrate, the simulation process often generates a discontinuous $G_n(\cdot)$.

To illustrate the usefulness of simulation estimators, consider a simple econometric model which specifies a set of conditions on population moments

$$G(\theta) = \int h(x, \theta) P(dx),$$

and assumes they equal zero at the true θ_0. Characteristically, estimators of θ_0 would be obtained by drawing a random sample of size n from the population of interest, and then finding that value of θ that makes the sample moment,

$$g_n(\theta) = n^{-1} \sum_{i=1}^{n} h(x_i, \theta),$$

as close as possible to zero.

[1] We are grateful to Daniel McFadden for discussions of his preprint on simulated moment estimators, which persuaded us to make several revisions in our original working paper. We are also grateful to two referees and an editor of this journal for constructive criticism. This research was supported in part by NSF Grants No. SES-8520261 and DMS-8503347.

This can only be done if it is easy, or at least practicable, to evaluate $g_n(\theta)$. For many models of current interest, particularly those whose restrictions involve multi-dimensional integrals, this computational problem is extremely burdensome, or impossible, even with the most sophisticated of computing equipment. Simulation can often be used to circumvent this problem. If simulation experiments can be used to produce a good estimate, $G_n(\theta)$, of $g_n(\theta)$, then one feasible estimator of θ_0 is the value of θ that makes $G_n(\theta)$ as close as possible to zero.

In the moment example, if $h(x,\theta)$ were difficult to calculate but it had an interpretation as a conditional expectation of a tractable function,

$$h(x,\theta) = \int H(x,\zeta,\theta)P(d\zeta|x),$$

with $P(\cdot|x)$ a *known* family of distributions, then a simulation estimator would be easy to construct. One could generate observations $\zeta_{i1},\ldots,\zeta_{is}$ from the distribution $P(\cdot|x_i)$, form the average

$$\hat{h}(x_i,\theta) = s^{-1}\sum_{j=1}^{s} H(x_i,\zeta_{ij},\theta)$$

for each θ, then estimate θ_0 by making

$$G_n(\theta) = n^{-1}\sum_{i=1}^{n} \hat{h}(x_i,\theta)$$

as close as possible to zero.

Section 4 provides a detailed analysis of two examples, one introduced by Pakes (1986), and the other by McFadden (1989), where this method of simulation can be used. The examples illustrate how one can verify the conditions of the general limit theorems that are presented in Section 3. They also show how simulation can be used to circumvent two familiar types of computational problems: evaluating intractable aggregation formulae, and evaluating discrete response probabilities.

In Pakes (1986), the function $H(\cdot,\theta)$ determined an individual's responses to a stimulus conditional on a vector of parameters (θ) from a microeconomic behavioral model, and P provided the distribution of the stimulus in the population of interest. The problem was to estimate the true value of the micro parameters, θ_0, by explicitly aggregating the micro responses into the totals (or economy-wide aggregates) predicted by different values of θ, and then fitting the aggregate predictions to aggregate data. For each different value of the stimulus the individual responses were easy to evaluate. However, the integral required to derive the aggregate implications of θ, that is $h(\cdot,\theta)$, proved intractable. To circumvent this problem, Pakes drew a random sample from P, calculated $H(\cdot,\theta)$ for each draw, and then estimated θ_0 by minimizing a distance between the simulated aggregates and the aggregate data. In Pakes's model $H(\cdot,\theta)$ was a

discontinuous function of θ, and, since the function minimized was a sum of these functions, it was also discontinuous. Pakes's problem is our Example 4.1.

In the discrete response problem studied by McFadden (1989), individual choices were a function of an only partially observed vector of deviates, and a vector of parameters to be estimated. In principle, for any given value of the parameter vector, θ, the probabilities of alternative choices could be evaluated by determining the choice made for every possible realization of the unobserved deviates, and then taking the conditional expectation of the indicator functions for the different choices. In practice, the desired integral is often too complicated to evaluate. Earlier, Lerman and Manski (1981) had proposed simulating the response probabilities for the discrete choice problem, and then finding that value of θ that maximized a likelihood function in which the simulated frequencies replaced the intractable true choice probabilities. Lerman and Manski's (1981) heuristic argument for the limit properties of their estimators required s, the number of simulation draws per observation, to grow large as well as n. However, examples in which both s and n were kept large required an impractical amount of computer time. McFadden (1989) noted that by combining moment conditions in which the theoretical choice probabilities enter linearly with a simulation estimator for those probabilities, one could obtain a simulation estimator for the parameters of the discrete response model that could be expected to have desirable asymptotic properties when the number of simulation draws per sample observation is held fixed, and sample size tends to infinity. This works because linearity allows the errors in the simulation to be averaged out over the sample. McFadden's problem will be our Example 4.2.

Given the insights provided by these articles, it is easy to generate numerous other examples where simulation can be used to solve an otherwise intractable computational problem. Most seem to fit in the moment framework outlined above, or something very similar.

Section 3 of this paper provides conditions under which $\hat{\theta}_n$, the estimator of θ_0 obtained from a random criterion function $G_n(\theta)$, is consistent and asymptotically normal. The theorems of this section are general enough to cover a broad class of simulation estimators. All but one of the conditions underlying each theorem are standard and require little explanation. The difficult (yet critical) condition insures that $G_n(\theta) - G(\theta)$ is small uniformly in θ. For consistency arguments something like a uniform law of large numbers is needed. For the finer asymptotics of the central limit theorem a more stringent bound is needed, but uniform only in small neighborhoods of θ_0.

Section 2 summarizes one method for checking the uniformity conditions of the theorems in Section 3. The method is particularly useful for applications such as the study of simulation estimators, where the criterion function can have discontinuities. It is heavily dependent upon independence assumptions, which makes it unsuitable for many time series applications. In the presence of dependence, methods based on smoothness assumptions are more successful. These methods correspond roughly to the empirical process technique of bracketing (see Section 6 of Pollard (1985), or Pollard (1989b)).

A reader who prefers to get an overview of the limit theorems before plunging into the details of their uniformity conditions could read Section 3 before Section 2.

Notation

Throughout the paper we use the $O_p(\cdot)$, $o_p(\cdot)$ notation of Mann and Wald (1944), as exposited by Chernoff (1956). When applied to vectors and matrices, the symbols should be interpreted entry by entry.

The symbol $\|\cdot\|$ denotes not only the usual Euclidean norm but also a matrix norm: $\|(b_{ij})\| = (\Sigma_{i,j} b_{ij}^2)^{1/2}$. It has the useful property that $\|BX\| \leqslant \|B\| \, \|x\|$ for each vector x and each conformable matrix B.

If x is a $k \times 1$ vector, we will write $\operatorname{diag}(x)$ for the $k \times k$ diagonal matrix with the elements of x along its principal diagonal.

The symbol $\rightsquigarrow$ denotes convergence in distribution.

2. EMPIRICAL PROCESS METHODS

This section describes a specialized technique that is particularly useful for deriving limit theorems for estimators obtained by minimizing random criterion functions with discontinuities.

Let $\xi_1, \xi_2, \ldots$ be independent observations sampled from a distribution P on a set $\mathscr{X}$. The empirical measure P_n is defined as the probability measure that places mass $1/n$ at each of $\xi_1, \ldots, \xi_n$. For each measurable subset D of $\mathscr{X}$, the strong law of large numbers implies that $P_n D$ converges almost surely to PD, and the central limit theorem implies that $\sqrt{n}(P_n D - PD)$ has an asymptotic normal distribution. There is now a large literature devoted to uniform extensions of these results for classes of sets and classes of functions. Some of these extensions provide ready-made ways of checking the uniformity conditions that will underlie the theorems in Section 3.

The simplest uniformity problem was solved most elegantly by Vapnik and Červonenkis (1971). They gave conditions for a class $\mathscr{D}$ of measurable subsets of $\mathscr{X}$ to satisfy a uniform strong law of large numbers:

$$\sup_{\mathscr{D}} |P_n D - PD| \to 0 \quad \text{almost surely.} \tag{2.1}$$

Amongst their results was a very simple combinatorial condition on $\mathscr{D}$ that guarantees (2.1) for every distribution P. An exposition of their approach, modified to take advantage of recent refinements, appears in Section II.4 of Pollard (1984).

Classes of sets that satisfy the combinatorial condition of Vapnik and Červonenkis are called VC classes (or polynomial classes by Pollard (1984)). In the next definition $\#(\cdot)$ denotes cardinality.

(2.2) DEFINITION: A class of sets $\mathscr{D}$ is called a *VC class* if there exist constants A and V such that: if S is a finite subset of $\mathscr{X}$ then

$$\#\{S \cap D : D \in \mathscr{D}\} \leqslant A(\#S)^V.$$

The VC property requires that the number of distinct subsets picked out by $\mathscr{D}$ from an S of size n grows much more slowly than 2^n, the maximum number of distinct subsets of S. Thus the class of all finite subsets of $\mathscr{R}^2$ containing 3 or fewer points is a VC class, because the number of subsets of S that it can pick out grows like n^3. A less obvious example is the VC class of all closed balls in $\mathscr{R}^2$; the number of subsets it picks out also grows like n^3. Notice that subclasses of a VC class are VC classes.

The VC property guarantees a very strong form of (2.1). The proof of the next lemma, and the proofs of other results about VC classes in this section, may be found in Pollard (1984).

(2.3) LEMMA: *If $\mathscr{D}$ is a VC class, then the uniform strong law of large numbers* (2.1) *holds for every P.*

Strictly speaking, the statement of this lemma is incomplete because it omits the necessary measurability qualifications. It would be sufficient to add the assumption that $\mathscr{D}$ be permissible in the sense of Appendix C of Pollard (1984). In practice one checks permissibility by showing that the set of indicator functions of sets in $\mathscr{D}$ can be represented as $\{f(\cdot, t) : t \in T\}$, where $f(x, t)$ is a function jointly measurable in its arguments and T is a Borel subset of a compact metric space. Other assertions in this section could be modified similarly to ensure complete measure theoretic veracity.

There are several very simple methods for constructing VC classes. The most basic of these shows that the VC property is closely related to finite dimensionality. Recall that a class of functions $\mathscr{G}$, is said to be finite dimensional if each g in $\mathscr{G}$ can be expressed as a linear combination of a fixed, finite set of basis functions $g_1, \ldots, g_k$ in $\mathscr{G}$. For example, the class of all polynomials of degree 3 on the real line is finite dimensional; every such polynomial is a linear combination of the basis functions 1, x, x^2, and x^3.

(2.4) LEMMA: *If $\mathscr{G}$ is a finite dimensional vector space of real-valued functions on $\mathscr{X}$, then the class of all sets of the form $\{g \geqslant t\}$ or $\{g > t\}$, with $g \in \mathscr{G}$ and $t \in \mathscr{R}$, is a VC class.*

Typically one constructs VC classes by first generating a basic class using the last lemma, and then combining the basic sets using a fixed finite number of Boolean operations. The second step is justified by the next lemma. (In the fourth assertion the superscript c denotes a complement.)

(2.5) LEMMA: *If $\mathscr{C}$ and $\mathscr{D}$ are VC classes, then so are*

(i) $\mathscr{C} \cup \mathscr{D}$,

(ii) $\{C \cup D : C \in \mathscr{C},\ D \in \mathscr{D}\}$,

(iii) $\{C \cap D : C \in \mathscr{C},\ D \in \mathscr{D}\}$,

(iv) $\{C^c : C \in \mathscr{C}\}$.

Our final lemma for classes of sets proves that the VC property is preserved by the operation of taking inverse images. It gives us a way to prove limit theorems for sequences $\{T(\xi_i)\}$ obtained by applying a fixed transformation to each ξ_i. It also shows why the sets in a VC class need not have smooth boundaries nor have simple connectedness properties: such regularity can be destroyed by a highly irregular map T.

(2.6) LEMMA: *If T is a map from a set $\mathscr{X}$ into a set $\mathscr{Y}$, and if $\mathscr{D}$ is a VC class of subsets of $\mathscr{Y}$, then $\{T^{-1}D : D \in \mathscr{D}\}$ is a VC class of subsets of $\mathscr{X}$.*

PROOF: Let S be a finite subset of $\mathscr{X}$. Suppose D_1 and D_2 are sets in $\mathscr{D}$ whose inverse images pick out different subsets from S:

$$\left(T^{-1}D_1\right) \cap S \neq \left(T^{-1}D_2\right) \cap S.$$

Then D_1 and D_2 pick out different subsets from the image of S under T:

$$(TS) \cap D_1 \neq (TS) \cap D_2.$$

Thus

$$\begin{aligned} &\#\left\{S \cap \left(T^{-1}D\right) : D \in \mathscr{D}\right\} \\ &\quad \leqslant \#\left\{(TS) \cap D : D \in \mathscr{D}\right\} \\ &\quad \leqslant A(\#TS)^V \\ &\quad \leqslant A(\#S)^V. \end{aligned}$$

The results for VC classes of sets admit several generalizations to classes of functions. Nolan and Pollard (1987) have introduced the concept of a Euclidean class as one possibility; Dudley (1987) has studied a multitude of other plausible generalizations. We consider only Euclidean classes in this paper.

Let $\mathscr{F}$ be a class of real-valued functions on $\mathscr{X}$. An envelope for $\mathscr{F}$ is any function F such that $|f| \leqslant F$ for all f in $\mathscr{F}$. If μ is a measure on $\mathscr{X}$ for which F is integrable, it is natural to think of $\mathscr{F}$ as a subset of $\mathscr{L}^1(\mu)$, the space of all μ-integrable functions. This space comes equipped with a distance defined by the $\mathscr{L}^1(\mu)$ norm. The closed ball with center f_0 and radius R consists of all f in $\mathscr{L}^1(\mu)$ for which $\int |f - f_0|\, d\mu \leqslant R$.

(2.7) DEFINITION: Call $\mathscr{F}$ *Euclidean* for the envelope F if there exist positive constants A and V with the following property: if $0 < \varepsilon \leqslant 1$ and if μ is a measure

for which $\int F d\mu < \infty$, then there are functions $f_1, \ldots, f_k$ in $\mathscr{F}$ such that

(i) $k \leq A\varepsilon^{-V}$,

(ii) $\mathscr{F}$ is covered by the union of the closed balls with radius $\varepsilon \int F d\mu$ and centers $f_1, \ldots, f_k$. That is, for each f in $\mathscr{F}$, there is an f_i with $\int |f - f_i| \, d\mu \leq \varepsilon \int F d\mu$.

The constants A and V must not depend on μ.

Implicit in this definition is the assumption that the functions in $\mathscr{F}$ and the envelope F are measurable with respect to a fixed σ-field on $\mathscr{X}$, and that the measure μ lives on this σ-field.

If a class of functions is Euclidean it is necessarily manageable in the sense of Pollard (1989a), which provides the necessary proofs for this section. A class $\mathscr{F}$ that consists of the indicator functions of sets in a class $\mathscr{D}$ is Euclidean (for the envelope $F \equiv 1$) if and only if $\mathscr{D}$ is a VC class. Thus theorems for Euclidean classes will always include results for VC classes as special cases. For example, the next lemma generalizes Lemma 2.3.

(2.8) LEMMA: *If $\mathscr{F}$ is Euclidean for the envelope F and if $\int F dP < \infty$, then*

$$\sup_{\mathscr{F}} \left| \int f dP_n - \int f dP \right| \to 0 \quad \textit{almost surely}.$$

Here are some examples of Euclidean classes. They all involve some sort of smoothness or finite dimensionality, but not necessarily in the way required by traditional proofs of uniform limit theorems.

(2.9) EXAMPLE: Let $\{g_1, \ldots, g_k\}$ be a finite set of functions on $\mathscr{X}$. For each positive, finite M write $\mathscr{F}_M$ for the class of all linear combinations $\Sigma_i \alpha_i g_i(\cdot)$ with $\Sigma_i |\alpha_i| \leq M$. It is Euclidean for the envelope $F = M \max_i |g_i|$. Without the bound on the coefficients the class would still be Euclidean, but only for the trivial reason that an infinite envelope excludes all but trivial μ measures from Definition 2.7. Of course such classes are amenable to traditional techniques.

(2.10) EXAMPLE: Let $K(\cdot)$ be a function of bounded variation (but not necessarily continuous) on the real line. Let $\mathscr{F}$ consist of all rescaled translates of K: that is, functions of the form

$$f_{\sigma, y}(x) = K\left(\frac{y - x}{\sigma}\right),$$

where y ranges over $\mathscr{R}$ and $\sigma > 0$. It is Euclidean for the constant envelope $F \equiv \sup |K|$. Such classes are useful in the study of nonparametric smoothing procedures. See Nolan and Pollard (1987).

(2.11) EXAMPLE: Let $\mathscr{F}_k$ denote the class of all real functions on $\mathscr{R}$ that are bounded in absolute value by a fixed function F and satisfy: for each f in $\mathscr{F}_k$ there is a partition of $\mathscr{R}$ into k intervals on each of which f is linear. The class $\mathscr{F}_k$ is Euclidean for the envelope F. The possibility that f might have discontinuities at the partitioning points makes this class difficult to analyze by traditional methods.

In each of the last three examples the Euclidean property could be verified by application of the lemmas that follow. As with VC classes, the best strategy for identifying Euclidean classes is to combine simpler classes according to the rules that preserve the Euclidean property. The starting point is usually one of the next two lemmas.

To each real valued function on a set $\mathscr{X}$ there corresponds a subset of the higher dimensional set $\mathscr{X} \otimes \mathscr{R}$, its subgraph:

$$\text{subgraph}(f) = \{(x,t) \in \mathscr{X} \otimes \mathscr{R} : 0 < t < f(x) \text{ or } 0 > t > f(x)\}.$$

For example, the subgraph of any of the piecewise linear functions in Example 2.11 is made up of a union of k subsets of $\mathscr{R}^2$, each of which is a quadrilateral or a union of two triangular regions. The next lemma shows that all polynomial classes of functions, in the sense of Pollard (1984), are Euclidean.

(2.12) LEMMA: *If* $\{\text{subgraph}(f) : f \in \mathscr{F}\}$ *is a VC class of sets, then* $\mathscr{F}$ *is Euclidean for every envelope.*

The second basic method deduces the Euclidean property from an analogous property of bounded subsets of the ordinary Euclidean space $\mathscr{R}^d$.

(2.13) LEMMA: *Let* $\mathscr{F} = \{f(\cdot, t) : t \in T\}$ *be a class of functions on* $\mathscr{X}$ *indexed by a bounded subset* T *of* $\mathscr{R}^d$. *If there exists an* $\alpha > 0$ *and a nonnegative function* $\phi(\cdot)$ *such that*

$$|f(x,t) - f(x,t')| \leqslant \phi(x)\|t - t'\|^{\alpha} \quad \text{for } x \in \mathscr{X} \quad \text{and} \quad t, t' \in T,$$

then $\mathscr{F}$ *is Euclidean for the envelope* $|f(\cdot, t_0)| + M\phi(\cdot)$, *where* t_0 *is an arbitrary point of* T *and* $M = (2\sqrt{d}\ \sup_T \|t - t_0\|)^{\alpha}$.

PROOF: For simplicity we consider the case $d = 2$. Write D for $\sup_T \|t - t_0\|$. Enclose T in a square S of side $2D$. Given ε with $0 < \varepsilon \leqslant 1$, choose an integer k with $1 \leqslant k\varepsilon^{1/\alpha} \leqslant 2$. Partition S into k^2 subsquares of side $2D/k$. From each subsquare that intersects T choose, arbitrarily, a point in the intersection. Let $\{t_1, \ldots, t_N\}$ be the set of all such points. Of course $N \leqslant k^2 \leqslant 4\varepsilon^{-2/\alpha}$, which is the right rate of growth for a Euclidean class.

Each t in T belongs to at least one of the subsquares. The corresponding t_i lies a distance no greater than $\Delta = \sqrt{2}\,2D/k$ from t. Write F for the given envelope.

Then, for all x,

$$|f(x,t)-f(x,t_i)| \leqslant \phi(x)\,\Delta^{\alpha} \leqslant \varepsilon F(x).$$

When both sides are integrated with respect to a measure μ this gives the bound required by Definition 2.7.

The closure properties for Euclidean classes are determined by pointwise algebraic operations analogous to the Boolean operations that preserve the VC property.

(2.14) LEMMA: *If $\mathscr{F}$ is Euclidean for an envelope F, and $\mathscr{G}$ is Euclidean for an envelope G, then*

(i) $\{f+g: f\in\mathscr{F},\ g\in\mathscr{G}\}$ *is Euclidean for the envelope* $F+G$;

(ii) $\{fg: f\in\mathscr{F},\ g\in\mathscr{G}\}$ *is Euclidean for the envelope* FG;

(iii) *both* $\{\max(f,g): f\in\mathscr{F},\ g\in G\}$ *and* $\{\min(f,g): f\in F,\ g\in G\}$ *are Euclidean for the envelope* $\max(F,G)$;

(iv) *for each positive M the class* $\{\alpha f: f\in\mathscr{F},\ \alpha\in\mathscr{R},\ |\alpha|\leqslant M\}$ *is Euclidean for the envelope MF.*

(2.15) LEMMA: *If T is a measurable map from $\mathscr{X}$ into $\mathscr{Y}$, and if $\mathscr{F}$ is a class of functions on $\mathscr{Y}$ that is Euclidean for an envelope F, then the class of composed functions* $\{f\circ T: f\in\mathscr{F}\}$ *is Euclidean for the envelope* $F\circ T$.

PROOF: Given a measure μ on $\mathscr{X}$, write μ_T for its image measure on $\mathscr{Y}$ under the map T. If $f_1,\ldots,f_k$ are the functions for which

$$\min_i \int |f-f_i|\,d\mu_T \leqslant \varepsilon \int F\,d\mu_T,$$

then $f_1\circ T,\ldots,f_k\circ T$ are appropriate for μ, because

$$\int g\,d\mu_T = \int g\circ T\,d\mu$$

for every nonnegative, measurable g on $\mathscr{Y}$.

As an illustration of how these lemmas may be applied, we will prove that the class $\mathscr{F}_k$ from Example 2.11 is Euclidean for its envelope F. From Lemma 2.12, it is good enough to prove that the subgraphs form a VC class. Each subgraph is a union of at most $2k$ triangular regions in $\mathscr{R}^2$. Each triangular region is the intersection of three open or closed halfspaces in $\mathscr{R}^2$. So, with $2k$ application of Lemma 2.5(ii) followed by three applications of Lemma 2.5(iii), the problem is reduced to proving that the class of all halfspaces is a VC class. Every halfspace can be represented as $\{g_{\alpha,\beta}\geqslant t\}$ or $\{g_{\alpha,\beta}> t\}$ for some real numbers α, β, t,

where $g_{\alpha,\beta}(x, y) = \alpha x + \beta y$. The class of all $g_{\alpha,\beta}$ functions is a two dimensional vector space. Lemma 2.4 completes the argument.

Lemma 2.8 is a uniform analogue of the strong law of large numbers. The empirical process literature also contains uniform analogues of the central limit theorem. These are expressed in terms of the standardized empirical process $\nu_n = \sqrt{n}(P_n - P)$. This process acts linearly to produce a standardized sum for each f in $\mathcal{L}^2(P)$,

$$\nu_n(f) = n^{-1/2} \sum_{i=1}^{n} \left[f(\xi_i) - \int f dP \right],$$

which converges in distribution to a normal with variance $\int f^2 dP - [\int f dP]^2$ and zero mean. The empirical central limit theorems give conditions under which the convergence is locally uniform in f, in the sense that small $\mathcal{L}^2(P)$ perturbations of f have only a small effect on $\nu_n(f)$. We do not need a precise statement of the limit theorem (Section VII.5 of Pollard (1984)) in this paper, because it is only the perturbation property that we need in order to check the uniformity conditions of the theorems in Section 3.

(2.16) LEMMA: *Let $\mathcal{F}$ be a Euclidean class with envelope F for which $\int F^2 dP < \infty$. For each $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$\limsup \mathbb{P}\left\{ \sup_{[\delta]} |\nu_n(f_1) - \nu_n(f_2)| > \eta \right\} < \varepsilon,$$

where $[\delta]$ denotes the set of all pairs of functions in $\mathcal{F}$ with $\int (f_1 - f_2)^2 dP < \delta^2$.

The assertion of the lemma translates into a smoothness property for a class $\{f(\cdot, \theta) : \theta \in \Theta\}$ if the parameterization is continuous at θ_0 in the $\mathcal{L}^2(P)$ sense, that is, if

$$\int [f(\cdot, \theta) - f(\cdot, \theta_0)]^2 dP \to 0 \quad \text{as} \quad \theta \to \theta_0.$$

If the envelope F is square-integrable with respect to P, a simple sufficient condition for $\mathcal{L}^2(P)$ continuity at θ_0 is continuity of the function $f(x, \cdot)$ at θ_0, for P almost all x. This follows by the Dominated Convergence Theorem because $[f(x, \theta) - f(x, \theta_0)]^2$ converges almost-surely to zero and it is bounded by the integrable function $4F^2$. When $f(\cdot, \theta)$ is the indicator function of a set $D(\theta)$ the almost-sure convergence is usually verified by showing: (i) each x in the interior of $D(\theta_0)$ belongs to $D(\theta)$ for all θ close enough to θ_0; (ii) each x in the interior of the complement of $D(\theta_0)$ belongs to the complement of $D(\theta)$ for all θ close enough to θ_0; (iii) the boundary of $D(\theta_0)$ has zero P measure. Conditions (i) and (ii) merely restate the definition of continuity of indicator functions at

each x not on the boundary of $D(\theta_0)$. Condition (iii) anticipates that it is only the boundary points where the convergence of the indicator functions might fail.

The combination of almost sure continuity and domination will be familiar to those readers who have studied uniform laws of large numbers, such as the one proved by Hansen (1982). For us the combination plays a completely different role; it is a sufficient condition for translating a uniform central limit theorem, which is a more powerful local result than a law of large numbers, into a parametric form. To get uniform central limit theorems directly from the domination condition one needs more detailed information about rates of convergence of local oscillations of the functions. In empirical process theory, this is made precise by the bracketing method described in Section 6 of Pollard (1985), or in Pollard (1989b).

(2.17) LEMMA: *If* $\{f(\cdot,\theta):\theta\in\Theta\}$ *is a Euclidean class with envelope* F *for which* $\int F^2\,dP<\infty$, *and if the parameterization is* $\mathscr{L}^2(P)$ *continuous at* θ_0, *then, for each sequence of positive numbers* $\{\delta_n\}$ *converging to zero,*

$$\sup_{\|\theta-\theta_0\|<\delta_n}|\nu_n f(\cdot,\theta)-\nu_n f(\cdot,\theta_0)|=o_p(1).$$

PROOF: Fix $\varepsilon>0$ and $\eta>0$. We need to prove that

$$\limsup \mathbb{P}\left\{\sup_{\|\theta-\theta_0\|<\delta_n}|\nu_n f(\cdot,\theta)-\nu_n f(\cdot,\theta_0)|>\eta\right\}<\varepsilon.$$

Choose δ according to Lemma 2.16. When n is large enough,

$$\sup_{\|\theta-\theta_0\|<\delta_n}\int[f(\cdot,\theta)-f(\cdot,\theta_0)]^2\,dP<\delta^2,$$

by virtue of the $\mathscr{L}^2(P)$ continuity of the parameterization. That is, the class $[\delta]$ eventually contains all the pairs $f(\cdot,\theta)$, $f(\cdot,\theta_0)$ for which $\|\theta-\theta_0\|<\delta_n$. The assertion of Lemma 2.16 is then stronger than the requirement of the present lemma.

3. GENERAL LIMIT THEOREMS

In this section we state and prove a consistency theorem and a central limit theorem for an estimator $\hat{\theta}_n$ that comes close enough to minimizing the length $\|G_n(\cdot)\|$ of a random, vector-valued function. This function is defined on a subset Θ of some $\mathscr{R}^d$. It should be thought of as an estimate of a deterministic, vector-valued function $G(\cdot)$ that is also defined on Θ. The true value θ_0 is defined implicitly as the unique point in Θ for which $G(\theta_0)=0$.

The requirements for the theorems usually include a uniformity condition for G_n: a condition that prescribes the rate at which G_n-G must converge to zero uniformly over particular neighborhoods of θ_0. These uniformity conditions are in a form well suited to the application of the uniform limit theorems from Section 2.

The section concludes with two lemmas that state conditions under which the Euclidean norm $\|\cdot\|$ can be replaced by random norms that depend on θ, without disturbing the main limit theorems.

Consistency is a global property. It makes an assertion about an estimator that potentially could be anywhere in the parameter space. The conditions needed to establish consistency are, therefore, necessarily global. Theorem 3.1 spells out one possible set of conditions. The estimator $\hat{\theta}_n$ is taken as any value that comes close enough (condition (i)) to providing a global minimum for $\|G_n(\cdot)\|$. Since θ_0 is included in the set over which the minimum is taken, $\|G_n(\hat{\theta}_n)\|$ cannot be much bigger than $\|G_n(\theta_0)\|$. If $G_n(\theta_0)$ is eventually (condition (ii)) close to zero, the assumed value of $G(\theta_0)$, it follows that $G_n(\hat{\theta}_n)$ must also get close to zero. If small values of $\|G_n(\theta)\|$ can occur only near θ_0 (condition (iii)), this forces $\hat{\theta}_n$ close to θ_0. No direct use is made of the assumption that $G(\theta_0)$ is zero; the theorem applies to any θ_0 in Θ satisfying (ii) and (iii).

(3.1) THEOREM: *Under the following conditions $\hat{\theta}_n$ converges in probability to θ_0.*

(i) $$\|G_n(\hat{\theta}_n)\| \leq o_p(1) + \inf_{\theta \in \Theta} \|G_n(\theta)\|,$$

(ii) $$G_n(\theta_0) = o_p(1),$$

(iii) $$\sup_{\|\theta - \theta_0\| > \delta} \|G_n(\theta)\|^{-1} = O_p(1) \quad \textit{for each} \quad \delta > 0.$$

PROOF: Fix $\varepsilon > 0$ and $\delta > 0$. Condition (iii) means that there exists a finite M for which

$$\limsup \mathbb{P}\left\{ \sup_{\|\theta - \theta_0\| > \delta} \|G_n(\theta)\|^{-1} > M \right\} < \varepsilon.$$

As the range of the infimum on the right-hand side of (i) includes θ_0,

$$\|G_n(\hat{\theta}_n)\| \leq o_p(1) + \|G_n(\theta_0)\| = o_p(1),$$

and hence

$$\mathbb{P}\left\{ \|G_n(\hat{\theta}_n)\|^{-1} > M \right\} \to 1.$$

It follows that, with probability of at least $1 - 2\varepsilon$ for all n large enough,

$$\|G_n(\hat{\theta}_n)\|^{-1} > M \geq \sup_{\|\theta - \theta_0\| > \delta} \|G_n(\theta)\|^{-1}.$$

These inequalities force $\hat{\theta}_n$ to lie within a distance δ of θ_0. That is,

$$\limsup \mathbb{P}\left\{ \|\hat{\theta}_n - \theta_0\| > \delta \right\} \leq 2\varepsilon.$$

As ε and δ can be chosen arbitrarily close to zero, the asserted convergence in probability is established.

Conditions for the strong consistency of $\hat{\theta}_n$ would be obtained by replacing the $o_p(\cdot)$ and $O_p(\cdot)$ quantities in (i), (ii), and (iii) by their almost sure analogues. Since our main goal is a distributional result for $\hat{\theta}_n$, we omit the proof of the stronger theorem.

Condition (iii) says roughly that, outside a neighborhood of θ_0, there is probability close to one that $\|G_n(\theta)\|$ stays bounded away from zero. A sufficient condition for this is that the deterministic $\|G(\theta)\|$ has a similar property and that G_n is everywhere relatively close to G, as shown by the next corollary.

(3.2) COROLLARY: *Under the following conditions $\hat{\theta}_n$ converges in probability to the unique θ_0 in Θ for which $G(\theta_0) = 0$:*

$$\text{(i)} \qquad \|G_n(\hat{\theta}_n)\| \leqslant o_p(1) + \inf_{\theta \in \Theta} \|G_n(\theta)\|,$$

$$\text{(ii)} \qquad \inf_{\|\theta - \theta_0\| > \delta} \|G(\theta)\| > 0 \quad \textit{for each} \quad \delta > 0,$$

$$\text{(iii)} \qquad \sup_{\theta} \frac{\|G_n(\theta) - G(\theta)\|}{1 + \|G_n(\theta)\| + \|G(\theta)\|} = o_p(1).$$

PROOF: The result could be deduced from the previous theorem, but it is just as easy to prove it directly by an argument similar in spirit to Huber's (1967) case B consistency proof.

Fix $\delta > 0$. Write ε for the corresponding infimum on the left-hand side of (ii). Then

$$\mathbb{P}\{\|\hat{\theta}_n - \theta_0\| > \delta\} \leqslant \mathbb{P}\{\|G(\hat{\theta}_n)\| \geqslant \varepsilon\},$$

so it will suffice to show that $\|G(\hat{\theta}_n)\| = o_p(1)$. To do this, invoke the triangle inequality and (iii) to get

$$\begin{aligned} \|G(\hat{\theta}_n)\| &\leqslant \|G_n(\hat{\theta}_n)\| + \|G(\hat{\theta}_n) - G_n(\hat{\theta}_n)\| \\ &\leqslant \|G_n(\hat{\theta}_n)\| + o_p(1)\big[1 + \|G_n(\hat{\theta}_n)\| + \|G(\hat{\theta}_n)\|\big], \end{aligned}$$

which rearranges to

$$\|G(\hat{\theta}_n)\|\big[1 - o_p(1)\big] \leqslant o_p(1) + \|G_n(\hat{\theta}_n)\|\big[1 + o_p(1)\big].$$

The right-hand side is of order $o_p(1)$ because, from (i) and (iii) and the requirement $G(\theta_0) = 0$,

$$\|G_n(\hat{\theta}_n)\| \leqslant o_p(1) + \|G_n(\theta_0)\| = o_p(1).$$

The assertion of the theorem follows.

For the purposes of direct verification, the slightly more stringent requirements of Corollary 3.2 are often more convenient than the general condition in

Theorem 3.1. The assumptions of Hansen's (1982) Theorem 2.2 imply the (almost sure analogues of the) conditions assumed for our corollary; but Huber's (1967) case B assumptions correspond to a generality somewhere between our theorem and its corollary. As the discussion that will follow our Lemma 3.4 will show, we sometimes do need the greater generality of Theorem 3.1.

Once $\hat{\theta}_n$ is known to converge to θ_0, further limiting properties of the estimator require only local assumptions on the behavior of G_n and G in small neighborhoods of θ_0. Not only does this relieve local limit theorems of the burden of the global conditions in the preceding theorem and corollary, but it also leaves open the possibility that consistency might be established by some other ad hoc argument.

The next theorem gives conditions under which $\hat{\theta}_n$, which is now assumed to converge in probability to θ_0, satisfies a central limit theorem. The argument breaks naturally into two steps. First we establish $\sqrt{n}$-consistency by means of a comparison between $\|G_n(\hat{\theta}_n)\|$ and $\|G_n(\theta_0)\|$. Informally stated, the new equicontinuity condition (iii) implies that

$$\|G(\theta)\| \leq O_p\big(\|G_n(\theta)\|\big) + O_p\big(\|G_n(\theta_0)\|\big) + o_p(n^{-1/2})$$

uniformly near θ_0. Since $\hat{\theta}_n$ comes close to minimizing $\|G_n(\cdot)\|$, the quantity $\|G_n(\hat{\theta}_n)\|$ cannot be much larger than $\|G_n(\theta_0)\|$, which is of order $O_p(n^{-1/2})$. Approximate linearity of G near θ_0 transfers the same rate of convergence to $\hat{\theta}_n - \theta_0$. The argument for the second step need concern only values of θ in a $O_p(n^{-1/2})$ neighborhood of θ_0. There conditions (ii) and (iii) combine to show G_n is uniformly well approximated by a linear function L_n. The θ_n^* that minimizes $\|L_n(\cdot)\|$ has an explicit form, from which asymptotic normality of $\sqrt{n}\,(\theta_n^* - \theta_0)$ is easily established. A comparison between $\|G_n(\hat{\theta}_n)\|$ and $\|G_n(\theta_n^*)\|$ shows that $\hat{\theta}_n$ must lie within $o_p(n^{-1/2})$ of θ_n^*, which implies the desired central limit theorem.

(3.3) THEOREM: *Let $\hat{\theta}_n$ be a consistent estimator of θ_0, the unique point of Θ for which $G(\theta_0) = 0$. If:*

(i) $\|G_n(\hat{\theta}_n)\| \leq o_p(n^{-1/2}) + \inf_{\theta} \|G_n(\theta)\|$;

(ii) *$G(\cdot)$ is differentiable at θ_0 with a derivative matrix Γ of full rank;*

(iii) *for every sequence $\{\delta_n\}$ of positive numbers that converges to zero,*

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

(iv) $\sqrt{n}\, G_n(\theta_0) \rightsquigarrow N(0, V)$;

(v) *θ_0 is an interior point of Θ;*

then

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \rightsquigarrow N\big(0, (\Gamma'\Gamma)^{-1}\Gamma' V \Gamma (\Gamma'\Gamma)^{-1}\big).$$

PROOF: First we prove $\sqrt{n}$-consistency. The assumed consistency allows us to choose a sequence $\{\delta_n\}$ that converges to zero slowly enough to ensure

$$\mathbb{P}\{\|\hat{\theta}_n - \theta_0\| \geq \delta_n\} \to 0.$$

With probability tending to one for this sequence, the supremum in condition (iii) runs over a range that includes the random value $\hat{\theta}_n$. Thus

$$\begin{aligned}&\|G_n(\hat{\theta}_n) - G(\hat{\theta}_n) - G_n(\theta_0)\| \\ &\quad \leq o_p(n^{-1/2}) + o_p(\|G_n(\hat{\theta}_n)\|) + o_p(\|G(\hat{\theta}_n)\|).\end{aligned}$$

By the triangle inequality, the left-hand side is larger than

$$\|G(\hat{\theta}_n)\| - \|G_n(\hat{\theta}_n)\| - \|G_n(\theta_0)\|.$$

Thus

$$\|G(\hat{\theta}_n)\|[1 - o_p(1)] \leq o_p(n^{-1/2}) + \|G_n(\hat{\theta}_n)\|[1 + o_p(1)] + \|G_n(\theta_0)\|.$$

From conditions (i) and (iv)

$$\|G_n(\hat{\theta}_n)\| \leq \|G_n(\theta_0)\| + o_p(n^{-1/2}) = O_p(n^{-1/2}).$$

It follows that

$$\|G(\hat{\theta}_n)\| = O_p(n^{-1/2}).$$

The differentiability condition (ii) implies the existence of a positive constant C for which (remember that $G(\theta_0) = 0$)

$$\|G(\theta)\| \geq C\|\theta - \theta_0\| \quad \text{near } \theta_0.$$

In particular, $\|\hat{\theta}_n - \theta_0\| = O_p(\|G(\hat{\theta}_n)\|) = O_p(n^{-1/2})$.

Next we establish asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, by arguing that $G_n(\cdot)$ is very well approximated by the linear function

$$L_n(\theta) = \Gamma(\theta - \theta_0) + G_n(\theta_0)$$

within a $O_p(n^{-1/2})$ neighborhood of θ_0. More precisely, we need the approximation error to be of order $o_p(n^{-1/2})$ at $\hat{\theta}_n$ and at the θ_n^* that minimizes $\|L_n(\cdot)\|$ globally. For $\hat{\theta}_n$ this follows directly from (ii) and (iii) together with the $\sqrt{n}$-consistency results already established:

$$\begin{aligned}\|G_n(\hat{\theta}_n) - L_n(\hat{\theta}_n)\| &\leq \|G_n(\hat{\theta}_n) - G(\hat{\theta}_n) - G_n(\theta_0)\| \\ &\quad + \|G(\hat{\theta}_n) - \Gamma(\hat{\theta}_n - \theta_0)\| \\ &\leq o_p(n^{-1/2}) + o_p(\|G_n(\hat{\theta}_n)\|) + o_p(\|G(\hat{\theta}_n)\|) \\ &\quad + o_p(\|\hat{\theta}_n - \theta_0\|) \\ &= o_p(n^{-1/2}).\end{aligned}$$

To correspond to a minimum of $\|L_n(\cdot)\|$, the vector $\Gamma(\hat{\theta}_n^* - \theta_0)$ must be equal to the projection of $-G_n(\theta_0)$ onto the column space Γ. Hence

$$\sqrt{n}\,(\theta_n^* - \theta_0) = -\sqrt{n}\,(\Gamma'\Gamma)^{-1}\Gamma' G_n(\theta_0).$$

From (iv), the right-hand side has the asymptotic normal distribution specified in the statement of the theorem. Consequently $\theta_n^* = \theta_0 + O_p(n^{-1/2})$, and the $\{\delta_n\}$ sequence can be assumed to satisfy

$$\mathbb{P}\{\|\theta_n^* - \theta_0\| \geq \delta_n\} \to 0.$$

Because θ_0 is an interior point of θ this implies that θ_n^* lies in Θ with probability tending to one. To simplify the argument slightly we shall act as if $\|\theta_n^* - \theta_0\| < \delta_n$ and θ_n^* belongs to Θ always. A more precise treatment would show that the contributions from those values of θ_n^* not satisfying these two requirements are eventually absorbed into an $o_p(1)$ error term.

From the differentiability condition (ii) we get

$$\|G(\theta_n^*)\| \leq \|\Gamma(\theta_n^* - \theta_0)\| + o(\|\theta_n^* - \theta_0\|) = O_p(n^{-1/2}).$$

From (iii) we get

$$\begin{aligned}\|G_n(\theta_n^*)\| - \|G(\theta_n^*)\| - \|G_n(\theta_0)\| \\ \leq o_p(n^{-1/2}) + o_p(\|G_n(\theta_n^*)\|) + o_p(\|G(\theta_n^*)\|),\end{aligned}$$

which rearranges to give $\|G_n(\theta_n^*)\| = O_p(n^{-1/2})$. Then we can argue as for $\hat{\theta}_n$ to deduce that

$$\|G_n(\theta_n^*) - L_n(\theta_n^*)\| = o_p(n^{-1/2}).$$

We now know that G_n and L_n are close at both $\hat{\theta}_n$, which almost minimizes $\|G_n\|$, and θ_n^*, which minimizes $\|L_n\|$. This forces $\hat{\theta}_n$ to come close to minimizing $\|L_n\|$:

$$\begin{aligned}\|L_n(\hat{\theta}_n)\| - o_p(n^{-1/2}) &\leq \|G_n(\hat{\theta}_n)\| \\ &\leq \|G_n(\theta_n^*)\| + o_p(n^{-1/2}) \\ &\leq \|L_n(\theta_n^*)\| + o_p(n^{-1/2}).\end{aligned}$$

That is,

$$\|L_n(\hat{\theta}_n)\| = \|L_n(\theta_n^*)\| + o_p(n^{-1/2}).$$

Squaring both sides we get

$$\|L_n(\hat{\theta}_n)\|^2 = \|L_n(\theta_n^*)\|^2 + o_p(n^{-1}),$$

the cross product term being absorbed into the $o_p(n^{-1})$ because $\|L_n(\theta_n^*)\|$ is of order $O_p(n^{-1/2})$. The quadratic form $\|L_n(\theta)\|^2$ has the simple expansion

$$\|L_n(\theta)\|^2 = \|L_n(\theta_n^*)\|^2 + \|\Gamma(\theta - \theta_n^*)\|^2,$$

about its global minimum. (The cross-product term vanishes because the residual vector, $L_n(\theta_n^*)$, must be orthogonal to the columns of Γ.) Put θ equal to $\hat{\theta}_n$, then

equate the two expressions for $\|L_n(\hat{\theta}_n)\|^2$ to deduce that

$$\left\|\Gamma\left(\hat{\theta}_n - \theta_n^*\right)\right\| = o_p(n^{-1/2}).$$

As Γ has full rank, this is equivalent to

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \sqrt{n}\left(\theta_n^* - \theta_0\right) + o_p(1),$$

from which the asserted central limit theorem follows.

Notice that the equicontinuity condition (iii) is in a form to which Huber's (1967) Lemma 3 (or Lemma 4 of Pollard (1985)) can be applied if $G_n(\cdot)$ happens to be an average of the form $P_n h(\cdot, \theta)$. For if $\|G_n(\theta)\| + \|G(\theta)\|$ is reduced to $C\|\theta - \theta_0\|$, the quantity within the supremum in condition (iii) increases to

$$\frac{\left\|\sqrt{n}\,(P_n - P)\left[h(\cdot, \theta) - h(\cdot, \theta_0)\right]\right\|}{1 + \sqrt{n}\,C\|\theta - \theta_0\|}.$$

The limiting normal distribution involves the matrices Γ and V, which depend implicitly upon the unknown θ_0. In practice one would need consistent estimators of these matrices before the limit distribution could be used as an approximation. For Γ, its interpretation as a derivative of G suggests an estimator

$$\hat{\Gamma}_{ni} = \varepsilon_n^{-1}\left[G_n\left(\hat{\theta}_n + \varepsilon_n u_i\right) - G_n\left(\hat{\theta}_n\right)\right]$$

for the ith column of Γ, where u_i is the unit vector with 1 in its ith place and $\{\varepsilon_n\}$ is a sequence that converges in probability to zero. The uniformity condition (iii) of Theorem 3.3 implies that this equals

$$\varepsilon_n^{-1}\left[G\left(\hat{\theta}_n + \varepsilon_n u_i\right) - G\left(\hat{\theta}_n\right)\right] + \varepsilon_n^{-1}\left[o_p(n^{-1/2}) + o_p(\varepsilon_n)\right],$$

which converges in probability to Γu_i provided $n^{-1/2}\varepsilon_n^{-1} = O_p(1)$. For example, $\varepsilon_n = n^{-\delta}$ would lead to a consistent estimator for Γ provided $\delta \leqslant 1/2$.

The statement of the theorem gives little explicit information about the dependence of V on θ_0. For the moment example in the introduction there is, however, a natural estimator for V. For notational simplicity consider the case where $s = 1$, so that

$$G_n(\theta) = P_n H(\cdot, \cdot, \theta),$$

where P_n is the empirical measure for the vectors $(x_i, \zeta_{i,1})$. Lemma 2.17, which we use to establish the uniformity condition (iii) of the asymptotic normality argument (Theorem 3.3), requires the class

$$\mathscr{H} = \{H(\cdot, \cdot, \theta) : \theta \in \Theta\}$$

to be Euclidean with square integrable envelope and $H(\cdot, \cdot, \theta)$ to be $\mathscr{L}^2(P)$ continuous at $\theta = \theta_0$. This makes

$$\mathscr{V} = \{H(\cdot, \cdot, \theta)H(\cdot, \cdot, \theta)' : \theta \in \Theta\}$$

Euclidean with an integrable envelope (Lemma 2.14), and

$$V(\theta) = PH(\cdot,\cdot,\theta)H(\cdot,\cdot,\theta)'$$

continuous at $\theta = \theta_0$ with $V = V(\theta_0)$. Consequently, if we define

$$V_n(\theta) = P_n H(\cdot,\cdot,\theta)H(\cdot,\cdot,\theta)',$$

Lemma 2.8 insures that

$$\left\| V_n\left(\hat{\theta}_n\right) - V \right\| \leq \sup_{\theta} \| V_n(\theta) - V(\theta) \| + \left\| V\left(\hat{\theta}_n\right) - V(\theta_0) \right\| = o_p(1).$$

That is, $V_n(\hat{\theta}_n)$ is a consistent estimator of V.

The asymptotic distribution in Theorem 3.3 is determined by both the behavior of $\sqrt{n}\, G_n(\theta_0)$ and the solution of a minimization problem for the Euclidean norm $\|\cdot\|$. If a different norm is used the asymptotic variance matrix is changed. With the proper choice of norm the asymptotic efficiency of $\hat{\theta}_n$ can be improved—the discussion for the multinomial problem of Example 4.1 will elaborate. The next two lemmas specify appropriate constraints on the choice of the norm.

For each nonsingular matrix A a new norm $\|\cdot\|_A$ is defined by setting $\|x\|_A = \|Ax\|$. The choice of A for the limit theorems in this section could depend on both θ and the data from which the random $G_n(\cdot)$ is constructed. That is, the norms could be defined by matrices $\{A_n(\theta)\}$ whose elements are random variables that depend on θ. A typical example is the method of minimum chi-square for the classical multinomial model, where the difference between observed and expected cell counts is weighted using a diagonal matrix with elements inversely proportional to the square root of estimated cell counts.

The first lemma gives conditions on the random matrices that preserve the consistency conditions of Theorem 3.1. If $A_n(\theta)$ became too nearly singular for values of θ not near θ_0, the norms $\|A_n(\theta)G_n(\theta)\|$ could be close to zero outside neighborhoods of θ_0. Condition (b) of the lemma prevents this degeneracy by placing a bound on the matrix norm of the inverse of $A_n(\theta)$. Condition (a) ensures that $A_n(\theta_0)G_n(\theta_0)$ converges in probability to zero.

(3.4) LEMMA: *Let* $\{A_n(\theta): \theta \in \Theta\}$ *be a family of sequences of nonsingular, random matrices for which*

(a) $$\|A_n(\theta_0)\| = O_p(1),$$

(b) $$\sup_{\theta \in \Theta} \left\| A_n(\theta)^{-1} \right\| = O_p(1).$$

If $G_n(\cdot)$ *satisfies conditions* (ii) *and* (iii) *of Theorem* 3.1 *then these conditions also hold with* $G_n(\theta)$ *replaced by* $A_n(\theta)G_n(\theta)$.

PROOF: From (a) and condition (ii) of the theorem:

$$\| A_n(\theta_0)G_n(\theta_0) \| \leq \| A_n(\theta_0) \| \, \| G_n(\theta_0) \| = O_p(1)o_p(1) = o_p(1).$$

For the analogue of (iii), first notice that, from the definition of the matrix norm,

$$\left\| A_n(\theta)^{-1}x \right\| \leq \left\| A_n(\theta)^{-1} \right\| \|x\| \quad \text{for all } x.$$

Put x equal to $A_n(\theta)G_n(\theta)$, then rearrange to get

$$\| A_n(\theta)G_n(\theta) \|^{-1} \leq \left\| A_n(\theta)^{-1} \right\| \| G_n(\theta) \|^{-1}.$$

Thus, for each $\delta > 0$,

$$\sup_{\|\theta-\theta_0\|>\delta} \| A_n(\theta)G_n(\theta) \|^{-1} \leq \sup_{\theta\in\Theta} \left\| A_n(\theta)^{-1} \right\| \sup_{\|\theta-\theta_0\|>\delta} \| G_n(\theta_n) \|^{-1}.$$

On the right-hand side, both factors are of the order $O_p(1)$.

Notice that the lemma imposes no uniform upper bound on $\|A_n(\theta)\|$. If G_n were replaced by A_nG_n in condition (iii) of Corollary 3.2 the ratio on the left-hand side could get close to 1 if $\|A_n(\theta)\|$ were unbounded. Corollary 3.2 would suffice if we were to restrict ourselves to bounded A_n, but in some cases that would be an unnatural restriction. For example, with the method of minimum chi-square in the multinomial problem, it would amount to an assumption that all cell probabilities were bounded away from zero. We discuss this further in Example 4.1.

Once consistency has been established, only the behavior of $\{A_n(\theta)\}$ in shrinking neighborhoods is relevant. The final lemma requires that $A_n(\theta)$ be close to a fixed nonsingular matrix A uniformly over these neighborhoods. It is this matrix A that will be passed through to the limiting variance matrix.

(3.5) LEMMA: *Let* $\{A_n(\theta{:}\theta \in \Theta)\}$ *be a family of sequences of nonsingular, random matrices for which there exists a nonsingular, nonrandom matrix* A *such that*

$$\sup_{\|\theta-\theta_0\|<\delta_n} \| A_n(\theta) - A \| = o_p(1)$$

whenever $\{\delta_n\}$ *is a sequence of positive numbers that converges to zero. If conditions* (ii), (iii), *and* (iv) *of Theorem* 3.3 *are satisfied by* $G_n(\cdot)$ *and* $G(\cdot)$, *then they are also satisfied if the* $G_n(\theta)$ *is replaced by* $A_n(\theta)G_n(\theta)$, *the* $G(\theta)$ *by* $AG(\theta)$, *the* V *by* AVA', *and the* Γ *by* $A\Gamma$.

PROOF: The convergence in distribution of the pair $(A_n(\theta_0), \sqrt{n}\, G_n(\theta_0))$ to the pair $(A, N(0, V))$ implies that $\sqrt{n}\, A_n(\theta_0)G_n(\theta_0) \rightsquigarrow N(0, AVA')$. (A formal argument would use Theorem 4.4 of Billingsley (1968) and the Continuous Mapping Theorem.) Existence of a derivative with full rank for $AG(\theta)$ at θ_0 is a trivial consequence of the nonsingularity of A.

For the uniformity condition, subtract and add terms $A_n(\theta)G(\theta)$ and $A_n(\theta)G_n(\theta_0)$, then invoke the triangle inequality to get the bound

$$\begin{aligned}
&\|A_n(\theta)G_n(\theta) - AG(\theta) - A_n(\theta_0)G_n(\theta_0)\| \\
&\quad \leq \|A_n(\theta)\|\,\|G_n(\theta) - G(\theta) - G_n(\theta_0)\| + \|A_n(\theta) - A\|\,\|G(\theta)\| \\
&\qquad + \|A_n(\theta) - A_n(\theta_0)\|\,\|G_n(\theta_0)\| \\
&\quad \leq O_p(1)\|G_n(\theta) - G(\theta) - G_n(\theta_0)\| \\
&\qquad + o_p(1)\|G(\theta)\| + o_p(1)O_p(n^{-1/2}) \\
&\quad = o_p(n^{-1/2}) + o_p\big(\|G_n(\theta)\|\big) + o_p\big(\|G(\theta)\|\big)
\end{aligned}$$

uniformly over the neighborhood $\{\|\theta - \theta_0\| < \delta_n\}$. We need this to be less than $o_p(1)$ times $n^{-1/2} + \|A_nG_n(\theta)\| + \|AG(\theta)\|$, which is greater than

$$n^{-1/2} + \|AG_n(\theta)\| - \|A_n(\theta) - A\|\,\|G_n(\theta)\| + \|AG(\theta)\|,$$

again uniformly over the neighborhood. Because A is nonsingular, there exists a positive C for which this last expression is greater than

$$n^{-1/2} + \big[C - o_p(1)\big]\|G_n(\theta)\| + C\|G(\theta)\|.$$

The asserted analogue of the uniformity condition (iii) follows immediately.

4. ANALYSES OF THE EXAMPLES

This section provides a detailed examination of the asymptotic behavior of the estimators introduced by Pakes (1986) and McFadden (1989). Both examples illustrate the effect of replacing an intractable function by a random function generated from a simulation sample s times as large as the original data sample. The linearity in the estimating equations makes the randomness from the simulation act like an extra additive source of randomness in the data, but scaled down by a factor of s^{-1}. This is seen clearly in the form of the limit distribution for the simulation estimators.

The analysis of both examples proceeds as follows. We begin by outlining the model and deriving the objective function to be minimized. Assuming the estimator is obtained by minimizing the objective function up to a term of order $o_p(1/\sqrt{n})$, we check for all but the uniformity conditions of the consistency and asymptotic normality theorems in Section 3. Finally we show that the required uniformity conditions are also satisfied.

4.1. *Example*

Pakes (1986) fit an optimal stopping model for the renewal of patents. Each year patentees had to decide whether to pay a renewal fee in order to keep their patents in force. The renewal decision was based on the expected discounted value of future returns from holding the patent. Since the stochastic process

generating those returns was assumed to be Markovian, the renewal decision depended only on current returns. The stopping rule specified a value that current returns had to exceed in order for the patent to be renewed. For any given value of the parameter vector determining the distribution of initial returns and the Markov process generating subsequent returns, say θ, the model determined a vector, $\pi(\theta)$, of the expected proportion dropping out at each age. The data contained the observed dropout proportions, p_n. If $\pi(\cdot)$ had been an easily calculable function of θ, any of the usual estimation procedures for the multinomial distribution could have been used to estimate θ. The elements of $\pi(\theta)$ were, however, determined by the proportion of current returns greater than the stopping value, and the Markov and stopping processes combined to produce a distribution of current returns which was not tractable. This led Pakes to substitute a simulation estimator, $\hat{\pi}_s(\theta)$, for $\pi(\theta)$ in the likelihood equations used to estimate θ_0. For a fixed θ, the simulation estimate was obtained by taking ns random draws from the implied initial distribution, passing each through the process determined by the model, and then simply counting up the proportions that dropped out at each age. (Note that s need not be an integer in this example.)

We discuss the asymptotic properties of simulated minimum distance estimators for this problem. (Minor modifications, along the lines of Pollard (1979), provide the properties of Pakes's simulated maximum likelihood estimator.) Our discussion begins by checking the consistency and asymptotic normality conditions of Theorems 3.1 and 3.3 for the estimator which minimizes

$$\| G_n(\theta) \| = \| p_n - \hat{\pi}_s(\theta) \|.$$

This is the simulated analogue of the estimator which minimizes $\| g_n(\theta) \| = \| p_n - \pi(\theta) \|$, an estimator which satisfies the conditions of Theorems 3.1 and 3.3 by virtue of the standard limit properties of $g_n(\theta_0)$ and the differentiability of $\pi(\theta)$ at $\theta = \theta_0$ (see below). Later we invoke Lemmas 3.4 and 3.5 to insure consistency and asymptotic normality when we minimize instead $A_n(\theta)G_n(\theta)$, with $A_n(\theta) = \operatorname{diag}[\hat{\pi}_s(\theta)^{-1/2}]$ and $A_n(\theta) = \operatorname{diag}[p_n^{-1/2}]$ (thus producing the simulated analogues of the traditional minimum chi-square and modified minimum chi-square estimators). Since

$$\| G_n(\theta) \| \leqslant \| p_n - \pi(\theta) \| + \| \pi(\theta) - \hat{\pi}_s(\theta) \|,$$

the law of large numbers ensures that $\| G_n(\theta_0) \| = o_p(1)$, which is condition (ii) of Theorem 3.1. To obtain (iii), and hence consistency, we assume, as did Pakes (1986), the identification condition that

$$\inf_{\| \theta - \theta_0 \| > \delta} \| G(\theta) \| = \inf_{\| \theta - \theta_0 \| > \delta} \| \pi(\theta) - \pi(\theta_0) \| > 0 \quad \text{for all } \delta > 0.$$

Now note that

$$\inf_{\| \theta - \theta_0 \| > \delta} \| G_n(\theta) \| \geqslant \inf_{\| \theta - \theta_0 \| > \delta} \| G(\theta) \| - \sup_{\theta} \| G_n(\theta) - G(\theta) \|.$$

The right-hand side of this expression will be bounded away from zero with probability tending to one, and (iii) will be satisfied, provided $\sup_\theta \|G_n(\theta) - G(\theta)\| = o_p(1)$. This is the uniform law we verify below.

To ensure asymptotic normality we check the conditions of Theorem 3.3. Pakes proved that $\pi(\theta)$ was differentiable at $\theta = \theta_0$ and assumed that its derivative matrix Γ was of full column rank. Also θ was specified to be an open subset of Euclidean space so that θ_0 was trivially in its interior. Left to check are the limit properties of the objective function evaluated at $\theta = \theta_0$, and the stochastic equicontinuity condition, or (iii) and (iv) of Theorem 3.3.

The multivariate central limit theorem (Theorem 11.10 of Breiman (1968)) guarantees that

$$\sqrt{n}\,[\,p_n - \pi(\theta_0)] \rightsquigarrow N(0, V),$$

where $V = \operatorname{diag}[\pi(\theta_0)] - \pi(\theta_0)\pi(\theta_0)'$. Now recall that at the true θ_0 the simulation mimics the data generation process for a sample of size sn. Consequently $\sqrt{n}\,[\hat{\pi}_s(\theta_0) - \pi(\theta_0)]$ has the same limit distribution as $\sqrt{n}\,[p_{sn} - \pi(\theta_0)]$: normal with mean zero and variance $s^{-1}V$. Moreover, since the data generation and simulation processes are independent,

$$\begin{aligned}\sqrt{n}\,G_n(\theta_0) &= \sqrt{n}\,[\,p_n - \pi(\theta_0)] - \sqrt{n}\,[\hat{\pi}_s(\theta_0) - \pi(\theta_0)] \\ &\rightsquigarrow N\big(0, (1 + s^{-1})V\big).\end{aligned}$$

So the limit distribution of $\sqrt{n}\,G_n(\theta_0)$ differs from that of $\sqrt{n}\,g_n(\theta_0) = \sqrt{n}\,[p_n - \pi(\theta_0)]$ only through the presence of the scalar $(1 + s^{-1})$, which reflects the extra independent source of randomness generated by the simulation process. Theorem 3.3 then insures that the limit distribution of the simulated minimum distance estimator of θ differs from that of the estimator which minimizes $\|p_n - \pi(\theta)\|$ (the estimator that would be obtained were we able to calculate $\pi(\cdot)$) only by the fact that the covariance matrix of the former is $(1 + s^{-1})$ times that of the latter.

Since

$$\frac{\|G_n(\theta) - G(\theta) - G_n(\theta_0)\|}{n^{-1/2} + \|G_n(\theta)\| + \|G(\theta)\|} \leq \|n^{1/2}[G_n(\theta) - G(\theta) - G_n(\theta_0)]\|,$$

condition (iii) of Theorem 3.3 will be satisfied provided

$$\sup_{\|\theta - \theta_0\| < \delta_n} \|\sqrt{n}\,[\hat{\pi}_s(\theta) - \pi(\theta)] - \sqrt{n}\,[\hat{\pi}_s(\theta_0) - \pi(\theta_0)]\| = o_p(1)$$

for every sequence $\{\delta_n\}$ converging to zero. If independent simulation draws are used to evaluate $\hat{\pi}_s(\theta)$ for each different θ then this condition will *never* be satisfied, since the left-hand side will always be more variable than $\sqrt{n}\,[\hat{\pi}_s(\theta_0) - \pi(\theta_0)]$. We show below, however, that the condition will be satisfied if the same simulation draws are used to evaluate $\hat{\pi}_s(\theta)$ for different values of θ.

The conditions discussed thus far also ensure the consistency and asymptotic normality of the estimator formed by minimizing $\|A_n(\theta)G_n(\theta)\|$, where $\{A_n(\theta): \theta \in \Theta\}$ is a family of nonsingular random matrices satisfying the

conditions of Lemmas 3.4 and 3.5. That is, if $\hat{\theta}_n$ is such an estimator and A is the nonsingular probability limit of $A_n(\theta_0)$, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightsquigarrow N\left(0, (1+s^{-1})M(A)\right), \quad \text{where}$$

$$M(A) = (\Gamma' A' A \Gamma)^{-1} \Gamma' A' A V A' A \Gamma (\Gamma' A' A \Gamma)^{-1}.$$

We check the conditions of Lemmas 3.4 and 3.5 for $A_n(\theta) = \text{diag}[\hat{\pi}_s(\theta)^{-1/2}]$ and for $A_n(\theta) = \text{diag}(p_n^{-1/2})$. These are the weighting matrices that provide the simulated analogues of the method of minimum chi-square and modified minimum chi-square, respectively, and in both cases $A = \text{diag}[\pi(\theta_0)^{-1/2}]$. A modification of Aitken's theorem (Theil (1984)) shows that $M(A) - M(\text{diag}[\pi(\theta_0)^{-1/2}])$ is positive semi-definite for every nonsingular A, so the use of an $A = \text{diag}[\pi(\theta_0)^{-1/2}]$ leads to an asymptotically efficient estimator for any fixed value of s.

Pakes (1986) assumed that $\pi(\theta_0) > 0$. Thus the law of large numbers ensures that $\|A_n(\theta_0)\| = O_p(1)$. Moreover, since all the elements $A_n(\theta)^{-1}$ are bounded by one, $\sup_\theta \|A_n(\theta)^{-1}\| = O_p(1)$, and we have verified the conditions of Lemma 3.4. Note that we have not required $\sup_\theta \|A_n(\theta)\|$ to be stochastically bounded. Thus, the minimum chi-square estimator can contend with elements in the parameter space that generate cell probabilities that get arbitrarily small. This is a possibility we would have difficulty excluding a priori, and it generates a need for the generality of Theorem 3.1 that is not available in Corollary 3.2 (see the discussion following Lemma 3.4). Finally, since $\pi(\theta_0) > 0$, the continuity of the map from θ to $\pi(\theta)$ at $\theta = \theta_0$ together with the condition that $\sup_\theta \|\pi(\theta) - \hat{\pi}_s(\theta)\| = o_p(1)$, both of which are verified below, suffice for Lemma 3.5.

We now come back to the problem of verifying the uniformity conditions (iii) of Theorems 3.1 and 3.3. More detail on the underlying model for patent renewals is required for this. That model assumes that the sequence of returns earned from holding a patent, should that patent be kept in force, is determined by a random draw of the vector of independent random variables

$$\xi = (Z, X_1, \ldots, X_L, Y_1, \ldots, Y_L)$$

which has distribution P on $\mathscr{X} = \mathscr{R} \otimes (0, \infty)^{2L}$. Here Z has a standard normal distribution and the X_i and Y_i have exponential distributions with unit means $(i = 1, \ldots, L)$. For a given value of θ, where $\theta = (\mu, \sigma, \lambda, \delta, \beta, \phi, \gamma)$ is in $\Theta = \mathscr{R} \otimes (0, \infty)^6$, the returns in year j, say R_j, are generated from ξ by putting

$$R_1 = \exp(\mu + \sigma Z), \text{ and } R_j = \{\lambda R_{j-1} \geqslant Y_j\} \max\left[\delta R_{j-1}, \beta_j X_j - \gamma\right],$$

for $2 \leqslant j \leqslant L$, where $\beta_j = \phi^j \beta$, and $\{\cdot\}$ is notation for the indicator function which takes the value of one if the logical condition inside it is satisfied, and zero elsewhere. A patent is renewed in year j if it was renewed in all previous years, and R_j is greater than the stopping value, $\tau_j(\theta)$. Pakes proved that the $\tau_j(\cdot)$ are differentiable functions of $\theta (1 \leqslant j \leqslant L)$.

The original data are generated by this mechanism with $\theta = \theta_0$ from independent vectors $\xi_1, \ldots, \xi_n$. The simulation is constructed from a further sample of

size sn from the ξ distribution. Note that each θ in Θ partitions the set $\mathscr{X}$ into $L+1$ subsets, the first L corresponding to those realizations of ξ for which R_j is less than τ_j for the first time $[1 \leqslant j \leqslant L]$, and the last set corresponding to those values for which $R_j \geqslant \tau_j$, for all L years. As we vary θ over Θ these partitions generate a class of subsets of $\mathscr{X}$. Our proof of the uniformity conditions consists of verifying that this class is a VC class of subsets of $\mathscr{X}$ and then applying the results on VC classes listed in Section 2.

This procedure is described best in empirical process terms. Let

$$x = (z, x_1, \ldots, x_L, y_1, \ldots, y_L)$$

be the generic point in $\mathscr{X}$, and put

$$r_1(x, \theta) = \exp(\mu + \sigma z)$$

and

$$r_j(x, \theta) = \{\lambda r_{j-1}(x, \theta) \geqslant y_j\} \max[\delta r_{j-1}(x, \theta), \beta_j x_j - \gamma] \quad \text{for} \quad 2 \leqslant j \leqslant L.$$

For each θ define $L+1$ subsets of $\mathscr{X}$ by

$$D_j(\theta) = \bigcap_{i<j} \{r_i(x, \theta) \geqslant \tau_i(\theta)\} \cap \{r_j(x, \theta) < \tau_j(\theta)\}$$

and

$$D_{L+1}(\theta) = \left[\bigcup_{j \leqslant L} D_j(\theta)\right]^c.$$

Finally set

$$\mathscr{D}_j = \{D_j(\theta) : \theta \in \Theta\}, \quad \text{for} \quad 1 \leqslant j \leqslant L+1.$$

Then the class of all subsets of $\mathscr{X}$ generated by varying θ is

$$\mathscr{D} = \bigcup_j \mathscr{D}_j.$$

Let P_n be the empirical measure of the original sample, and Q_n be the empirical measure of the simulation sample. Then p_n has jth component $P_n D_j(\theta_0)$, and $\hat{\pi}_s(\theta)$ has jth component $Q_n D_j(\theta)$. Note that

$$\sup_\theta \|\pi(\theta) - \hat{\pi}_s(\theta)\| \leqslant (L+1) \max_j \sup_{\theta \in \Theta} |\pi_j(\theta) - \hat{\pi}_{s,j}(\theta)|.$$

Thus, to prove the uniform law required for consistency it will suffice to show that $\sup_\theta |\pi_j(\theta) - \hat{\pi}_{s,j}(\theta)| = o_p(1)$ for each j. But, from the definition of $\mathscr{D}$,

$$\sup_\theta |\pi_j(\theta) - \hat{\pi}_{s,j}(\theta)| \leqslant \sup_{D \in \mathscr{D}} |PD - Q_n D|.$$

Lemma 2.3 guarantees this last term goes to zero almost surely if $\mathscr{D}$ is a VC class, a condition we establish below.

The argument for the equicontinuity condition required to complete the asymptotic normality proof is similar. Provided the parameterization for the class $\mathscr{D}$ is $\mathscr{L}^2(P)$ continuous at θ_0, and $\mathscr{D}$ is indeed a VC class, Lemma 2.17 implies that for each j

$$\sup_{\|\theta-\theta_0\|<\delta_n} \sqrt{ns}\left|\left[\pi_j(\theta)-\hat{\pi}_{s_j}(\theta)\right]-\left[\pi_j(\theta_0)-\hat{\pi}_{s_j}(\theta_0)\right]\right| = o_p(1),$$

for every sequence $\{\delta_n\}$ that converges to zero. We sketch a proof of the $\mathscr{L}^2(P)$ continuity of the θ-parameterization of $\mathscr{D}$ below. Pakes (1986) provides an alternative proof of continuity.

Verifying that $\mathscr{D}$ is indeed a VC class is an exercise in applying the criteria of Section 2. Let

$$r_j^*(x,\theta) = \max\left\{\max_{2\leqslant q\leqslant j}\left[\delta^{j-q}\left(x_q\beta_q-\gamma\right)\right], \delta^{j-1}r_1\right\}.$$

$r_j^*(x,\theta)$ is the maximum current return a patent with a draw of x could earn and, as can be verified by repeated substitution into the formulae given above, if this patent is still in force in year j, its returns will be $r_j^*(x,\theta)$. A patent in force in $j-1$ will be in force in year j if $r_j^* \geqslant \tau_j$ and $\lambda r_{j-1}^* \leqslant y_j$, so, omitting the dependence of the return function on θ and x we have

$$D_j(\theta) = \left[\left\{r_j^* < \tau_j\right\} \cup \left\{\lambda r_{j-1}^* < y_j\right\}\right] \cap \bigcap_{i=1}^{j-1}\left[\left\{\lambda r_{j-1-i}^* \geqslant y_{j-i}\right\} \cap \left\{r_{j-i}^* \geqslant y_{j-i}\right\}\right]$$

where

$$\left\{r_j^* \geqslant \tau_j\right\} = \bigcup_{2\leqslant q\leqslant j}\left\{\delta^{j-q}\left(\beta_q x_q-\gamma\right) \geqslant \tau_j\right\} \cup \left\{\delta^{j-1}r_1 \geqslant \tau_1\right\},$$

and an analogous expression can be written for $\{\lambda r_{j-1}^* \geqslant y_j\}$, for $j=1,\ldots,L$. Since Lemma 2.5 ensures that classes of sets formed from the intersection (or union) of the elements of one VC class with those from another are VC classes, it will suffice to show that each generating set of the form $\{a_1(\theta)x_i + a_2(\theta)y_i + a_3(\theta) \geqslant 0\}$, or $\{b_1(\theta)r_1(z,\theta) + b_2(\theta) \geqslant 0\}$, or $\{b_3(\theta)r_1(z,\theta) \geqslant y_i\}$, traces out a VC class of subsets of $\mathscr{X}$ as θ ranges over Θ. Note that the $a_j(\cdot)$ and $b_j(\cdot)$ are continuous functions of θ $(j=1,2,3)$. Also because the class of functions generated from $r_1(z,\theta) = \exp(\sigma z+\mu)$ by varying θ is not a finite dimensional vector space, use of Lemma 2.4 by itself does not complete the proof.

Write T for the map from $\mathscr{X}$ into $\mathscr{R}^{3L+1}$ that takes x onto the vector $(x,\log y_1,\ldots,\log y_L)$. Let $\mathscr{G}$ be the vector space of all real-valued functions on $\mathscr{R}^{3L+1}$, and recall that Lemma 2.4 ensures that the class $\mathscr{C}$ of all subsets of $\mathscr{R}^{3L+1}$ of the form $\{g \geqslant t\}$, with $g \in \mathscr{G}$ and $t \in \mathscr{R}$, is a VC class. Then each of the generating sets can be written as

$$\left\{d_1(\theta)x_i + d_2(\theta)y_i + d_3(\theta)z + d_4(\theta)\log y_i + d_5(\theta) \geqslant 0\right\},$$

with the $d_j(\cdot)$ continuous functions of θ. As θ ranges over Θ these sets trace out a subset of $\mathscr{G}$, and since subclasses of a VC class are a VC class, each forms a VC class. But then Lemma 2.6 ensures that the inverse images of these sets trace out a VC class of subsets of $\mathscr{X}$, which completes the proof that $\mathscr{D}$ is a VC class.

The proof of continuity of $\theta \to D_j(\theta)$, as a map from θ into $\mathscr{L}^2(P)$, can be built up in a similar fashion. Since continuity is preserved by the formation of intersections and unions, it suffices to prove the continuity of the map for the generating sets. Since each of the generating sets is a closed halfspace whose boundary has zero P measure, the argument after Lemma 2.16 establishes their $\mathscr{L}^2(P)$ continuity with respect to θ at $\theta = \theta_0$.

4.2. *Example*

McFadden (1989) proposed a simulation method for estimating the parameters of a multinomial probit model. We will show that his estimator fits into the framework outlined in this paper. To simplify the analysis needed to verify the uniformity conditions, we will substitute combinatorial assumptions for the various smoothness assumptions of McFadden. Our methods will depend on the empirical process theory described in Section 2, whereas McFadden's methods allowed him to deduce the asymptotic distribution of his estimator by means of an elegant limit theorem due to Huber (1967).

An individual has m alternatives to choose between. His choice is determined by a set of m vector covariates $z_1, \ldots, z_m$ and a random vector α of weights. Alternative i is chosen if $z_i'\alpha$ is bigger than all the other $z_j'\alpha$. The vector α is generated as a $k \times 1$ vector function $h(\eta, \theta_0)$ of an r-dimensional random vector η with known distribution; the unknown value θ_0 is an interior point in a k-dimensional parameter space Θ. If the covariates are stacked as the rows of an $m \times k$ matrix Z, the choice is specified by the response vector

$$d = D[Zh(\eta, \theta_0)],$$

where $D(\cdot)$ maps $\mathscr{R}^m$ into $\{0,1\}^m$, putting a one in the position of the largest component and zeros elsewhere. The choice corresponds to the position in the vector d that contains the one. Ties would be indicated by a one in multiple positions of d. Following McFadden we assume a zero probability for ties.

The choices of n individuals are determined in this fashion from random pairs (Z_i, η_i) for $i = 1, \ldots, n$. These are assumed independent and identically distributed. From the observed response vectors d_i and matrices of covariates Z_i, we must estimate the unknown θ_0.

Write $\pi(Z, \theta)$ for the conditional expectation of $D[Zh(\eta, \theta)]$ given Z. For a $k \times m$ matrix $W(Z, \theta)$ of instruments, define

$$\begin{aligned} G(\theta) &= \int W(Z,\theta)[d - \pi(Z,\theta)]\, d\mathbb{P} \\ &= \int W(Z,\theta)[\pi(Z,\theta_0) - \pi(Z,\theta)]\, d\mathbb{P}. \end{aligned}$$

Clearly, $G(\theta_0) = 0$.

If $\pi(Z,\theta)$ were easily calculable, a reasonable estimator of θ_0 would be the value that minimized $\|g_n(\theta)\|$, where

$$g_n(\theta) = n^{-1}\sum_{i=1}^{n} W(Z_i,\theta)\left[d_i - \pi(Z_i,\theta)\right].$$

If the $\pi(Z,\theta)$ were intractable it could be replaced by a simulation estimator. For each individual generate s new random variables $\eta_{i1},\ldots,\eta_{is}$, then replace $\pi(Z_i,\theta)$ by

$$\hat{\pi}_s(Z_i,\theta) = s^{-1}\sum_{j=1}^{s} D\left[Z_i h(\eta_{ij},\theta)\right].$$

When η_i is independent of Z_i the simulation is carried out by generating s new independent observations from the same distribution. However, such independence is not required for the application of the limit theorems. Writing

$$\xi_i = (Z_i, \eta_i, \eta_{i1},\ldots,\eta_{is})$$

for the $(mk + r + rs)$-dimensional vector of data on the ith individual, we assume only that: the $\{\xi_i\}$ are independent and identically distributed; and the conditional expectation of $D[Z_i h(\eta_{ij},\theta)]$ given Z_i is $\pi(Z_i,\theta)$. By permitting dependence between the components of ξ_i we leave open the possibility of using variance reduction techniques in the generation of the simulation sample. With these assumptions it becomes plausible to use

$$G_n(\theta) = n^{-1}\sum_{i=1}^{n} W(Z_i,\theta)\left[d_i - \hat{\pi}_s(Z_i,\theta)\right]$$

to replace $g_n(\theta)$ as the estimator for $G(\theta)$. We define $\hat{\theta}_n$ to be any estimator for which $\|G_n(\hat{\theta}_n)\|$ comes within $o_p(n^{-1/2})$ of minimizing $\|G_n(\cdot)\|$.

We assume the identification condition:

$$\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta)\| > 0 \quad \text{for each} \quad \delta > 0.$$

Following McFadden, we could deduce this from lower level assumptions such as: Θ is compact, $G(\cdot)$ is continuous, and θ_0 is the unique point at which $G(\theta_0) = 0$. We will also require, as did McFadden, that:

$G(\cdot)$ has a nonsingular derivative matrix, say R, at θ_0.

These assumptions take care of all but condition (iii) of Corollary 3.2 (for consistency), and conditions (iii) and (iv) of Theorem 3.3 (for asymptotic normality). To ensure that these conditions are satisfied we will impose additional regularity conditions on the instruments. We assume first that

$$\int \sup_{\theta\in\Theta} \|W(Z,\theta)\|\, d\mathbb{P} < \infty.$$

As $\|W(Z,\theta)\|$ is bounded by $(km)^{1/2}\max|W_{ij}(Z,\theta)|$, this is equivalent to an assumption that the components of $W(Z,\theta)$ are dominated by a function that

does not depend on θ and has a finite first moment. This assumption is used in the proof of consistency. (It also guarantees that $G(\theta)$ is well-defined for every θ.) For the central limit theorem we will need an analogous second moment condition near θ_0. Assume that for some positive δ:

$$\int \sup_{\|\theta-\theta_0\|<\delta} \|W(Z,\theta)\|^2 \, d\mathbb{P} < \infty.$$

Certainly McFadden's uniformly bounded instruments satisfy these moment conditions.

To check the remaining requirements of Corollary 3.2 and Theorem 3.3 we recast $G_n(\cdot)$ as an empirical process indexed by a class of functions, upon which we impose further regularity conditions. Write $x = (X, y, y_1, \ldots, y_s)$ for the generic point in $(mk + r + rs)$-dimensional Euclidean space, the first coordinates being rearranged into the $m \times k$ matrix X, and the other coordinates being partitioned into the $r \times 1$ vectors y and y_j. Define

$$f(x,\theta) = W(X,\theta)\left[D\left[Xh(y,\theta_0)\right] - s^{-1}\sum_{j=1}^{s} D\left[Xh(y_j,\theta)\right]\right].$$

If P_n denotes the empirical measure of the $\{\xi_i\}$ and P denotes their common distribution, we have

$$G_n(\theta) = \int f(x,\theta)\, dP_n, \quad \text{and} \quad G(\theta) = \int f(x,\theta)\, dP.$$

Notice how the assumption about the conditional expectations for the simulation sample is used to get the representation for $G(\theta)$.

The asymptotic normality of

$$\sqrt{n}\, G_n(\theta_0) = \sqrt{n}\left[\int f(x,\theta_0)\, dP_n - \int f(x,\theta_0)\, dP\right]$$

follows from the multivariate central limit theorem for standardized sums of independent random vectors with finite second order moments (Theorem 11.10 of Breiman (1968)). The asymptotic variance matrix, V, equals

$$\operatorname{var}\left[f(\xi_1,\theta_0)\right] = \int f(x,\theta_0) f(x,\theta_0)'\, dP.$$

In the special case where η_i and the η_{ij} are conditionally independent given Z_i, the expression for V simplifies to $(1 + s^{-1})$ times

$$\int W(X,\theta_0)\left[\operatorname{diag}\left[\pi(X,\theta_0)\right] - \pi(X,\theta_0)\pi(X,\theta_0)'\right] W(X,\theta_0)'\, dP.$$

It remains only to check the two uniformity conditions. We will give two combinatorial conditions that will make the class of vector-valued functions $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta\}$ Euclidean, in the sense that the class of all the component real-valued functions satisfies Definition 2.7. Application of Lemma 2.8 to each

component then implies that

$$\sup_{\theta} \| G_n(\theta) - G(\theta) \| \to 0 \quad \text{almost surely,}$$

from which condition (iii) of Corollary 3.2, and hence consistency, follows. Lemma 2.17, applied to the components of $f(\cdot, \theta)$ for θ in a small enough neighborhood of θ_0, will imply a stronger result than the equicontinuity condition (iii) of Theorem 3.3 if $f(\cdot, \theta)$ is $\mathscr{L}^2(P)$ continuous at θ_0. For the $\mathscr{L}^2(P)$ continuity, assume that for P almost all X, each component of $W(X, \theta)$ is continuous in θ at θ_0, and assume that there is zero probability of a tie at $\theta = \theta_0$ for each of the simulations. Because $D[Xh(y_j, \theta)]$ is continuous in θ except at those (X, y_j) pairs for which there is a tie, it follows that $f(x, \theta)$ is almost surely continuous in θ at θ_0. The second moment condition on the instruments converts almost-sure continuity to $\mathscr{L}^2(P)$ continuity, as explained at the end of Section 2.

The Euclidean property for $\mathscr{F}$ is a consequence of more basic assumptions about $h(\cdot, \theta)$ and the instruments $W(\cdot, \theta)$. For each θ in Θ define a subset of $\mathscr{R}^k \otimes \mathscr{R}^r$ by

$$B(\theta) = \{(z, y) \in \mathscr{R}^k \otimes \mathscr{R}^r : z'h(y, \theta) \geqslant 0\}.$$

Assume that $\{B(\theta) : \theta \in \Theta\}$ is a VC class in the sense of Definition 2.2. In special cases this assumption is readily checkable. For example, McFadden checked his regularity conditions for the function

$$h(y, \theta) = \beta(\theta) + A(\theta) y,$$

when each component of $\beta(\theta)$ and $A(\theta)$ depended smoothly on θ. With our assumption the smoothness is irrelevant. As θ ranges over Θ the functions

$$g_\theta(z, y) = z'\beta(\theta) + z'A(\theta) y$$

range over a subset of a finite dimensional vector space $\mathscr{G}$ of real-valued functions on $\mathscr{R}^k \otimes \mathscr{R}^r$. Lemma 2.4 establishes the VC property for the class of all sets of the form $\{g \geqslant 0\}$, with g in $\mathscr{G}$; hence the subclass $\{B(\theta) : \theta \in \Theta\}$ is also a VC class. We assume also that $\{W(\cdot, \theta) : \theta \in \Theta\}$ is Euclidean in the sense that the class of all component functions satisfies Definition 2.7. This assumption can be further reduced by means of the methods of Section 2. For example, if Θ is bounded and if each $W(X, \theta)$ satisfies a Lipschitz condition of the type described in Lemma 2.13, then the Euclidean assumption holds.

Now the criteria from Section 2 lead us directly to the Euclidean property for $\mathscr{F}$. Each component of $f(x, \theta)$ is a bounded linear combination of products involving the components of $W(X, \theta)$ and the components of the two choice functions. By Lemma 2.14, it suffices to establish the Euclidean property for each individual component. For $W(X, \theta)$ the property holds by assumption. The components of the choice functions are indicator functions for sets, so it will suffice to show that these sets generate a VC class as θ ranges over Θ. Consider, for example, the first component of $D[Xh(y, \theta)]$. Write $x_1', \ldots, x_m'$ for the rows

of X. Then the set that corresponds to an individual choosing alternative 1 is the intersection of the sets

$$\Delta_{1j}(\theta) = \{(X, y) : (x_1' - x_j')h(y, \theta) \geqslant 0\}, \quad \text{for} \quad 1 < j \leqslant m.$$

By Lemma 2.5, it suffices to show that the class of all $\Delta_{1j}(\theta)$ sets is a VC class.

Define a map T_{1j} from $\mathscr{R}^{mk} \otimes \mathscr{R}^r$ into $\mathscr{R}^k \otimes \mathscr{R}^r$ by putting $T_{1j}(X, y) = (x_1 - x_j, y)$. Then $\Delta_{1j}(\theta) = T_{1j}^{-1}B(\theta)$. Lemma 2.6 shows that the class of all such inverse images, as $B(\theta)$ ranges over its VC class of sets, is also a VC class. A similar argument could be invoked for the sets corresponding to the choice of any of the other alternatives. The proof that $\mathscr{F}$ is Euclidean is complete.

Department of Economics, Yale University, New Haven, CT 06520, U.S.A.
and
Department of Statistics, Yale University, New Haven, CT 06520, U.S.A.

Manuscript received October, 1987; final revision received March, 1989.

REFERENCES

BILLINGSLEY, P. (1968): *Convergence of Probability Measures*. New York: Wiley.
BREIMAN, L. (1968): *Probability*. Reading, Mass.: Addison-Wesley.
CHERNOFF, H. (1956): "Large Sample Theory: Parametric Case," *Annals of Mathematical Statistics*, 27, 1–22.
DUDLEY, R. M. (1978): "Central Limit Theorems for Empirical Measures," *Annals of Probability*, 6, 899–929 (correction, *ibid* 7 (1979), 909–911).
——— (1987): "Universal Donsker Classes and Metric Entropy," *Annals of Probability*, 15, 1306–1326.
HANSEN, L. (1982): "Large Sample Properties of Method of Moments Estimators," *Econometrica*, 50, 1029–1054.
HUBER, P. J. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233. Berkeley, CA: University of California.
LERMAN, S., AND C. MANSKI (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, Cambridge, Mass.: MIT Press.
MANN, H. B., AND A. WALD (1943): "On Stochastic Limit and Order Relationships," *Annals of Mathematical Statistics*, 14, 217–226.
MCFADDEN, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995–1026.
NOLAN, D., AND D. POLLARD (1987): "U-Processes: Rates of Convergence," *Annals of Statistics*, 15, 780–799.
PAKES, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–785.
POLLARD, D. (1979): "General Chi-Square Goodness-of-Fit Tests with Data-Dependent Cells," *Z. Wahrscheinlichkeitstheore Verw. Geb.*, 50, 317–333.
——— (1984): *Convergence of Stochastic Processes*. New York: Springer-Verlag.
——— (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295–314.
——— (1989a): "Asymptotics via Empirical Processes," *Statistical Science* (forthcoming).
——— (1989b): "Bracketing Methods in Statistics and Econometrics," forthcoming in the Proceedings of the Conference on Nonparametric and Semiparametric Methods in Econometrics and Statistics, Duke University, May, 1988.

THEIL, H. (1984): *Handbook of Econometrics*, Chapter 1, Volume 1, ed. by Z. Griliches and M. Intrilligator. Amsterdam: North Holland.

VAPNIK, V. N., AND A. YA. ČERVONENKIS (1971): "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probability and Its Application*, 16, 264–280.