3.5 Consistency of approximate M-estimators of ψ type. As in Sec. 3.3, let $(X, \mathcal{A}, P)$ be a probability space and Θ a locally compact separable metric space. Let $\psi(\theta, x)$ be a function of x in X and $\theta \in \Theta$ with values in a Euclidean space $\mathbb{R}^m$. Let $X_1, X_2, \ldots$ be independent with values in X and distribution P. A sequence of estimators $T_n := T_n(X_1, \ldots, X_n)$ with values in Θ will be called *approximate M-estimators of ψ type* if

$$\frac{1}{n}\sum_{i=1}^{n} \psi(T_n, X_i) \to 0 \text{ almost uniformly as } n \to \infty. \tag{3.5.1}$$

If ψ is jointly measurable, as will follow from assumptions to be given, then since estimators T_n by definition are assumed to be statistics (measurable functions of the observations), the almost uniform convergence in (3.5.1) will be equivalent to almost sure convergence. Recall that if T_n are M-estimators of ψ type, the expression on the left in (3.5.1) equals 0, at least with probabilities converging to 1. Convergence of T_n to some θ_0 will be proved under some assumptions as follows.

(B-1) For each $\theta \in \Theta$, the function $\psi(\theta, \cdot)$ is $\mathcal{A}$-measurable.

(B-2) For almost all x, $\psi(\cdot, x)$ is continuous on Θ.

(B-3) $\lambda(\theta) := E\psi(\theta, \cdot)$ is defined and finite for all θ, and for some θ_0, $\lambda(\theta_0) = 0$, while $\lambda(\theta) \neq 0$ for all $\theta \neq \theta_0$.

(B-4) There is a continuous, positive function $b(\cdot)$ on Θ, bounded away from 0, so that for some $b_0 > 0$, $b(\theta) \geq b_0$ for all θ, and

(i) $\Psi(x) := \sup_\theta |\psi(\theta, x)|/b(\theta)$ is integrable,

(ii) $\liminf_{\theta \to \infty} |\lambda(\theta)|/b(\theta) \geq 1$, and

(iii) $E\{\limsup_{\theta \to \infty} |\psi(\theta, x) - \lambda(\theta)|/b(\theta)\} < 1$.

A first question about the assumptions is: how are we to verify them, given that the true distribution P of the observations is unknown? (B-1) and (B-2) don't depend on P, so they can be checked. In (B-3), $\psi(\theta, \cdot)$ will be integrable for all P and θ if it is a bounded function of x for each θ. If ψ is bounded uniformly in x and θ, as for the classes of ψ functions with $-A \leq \psi(\theta, x) \leq A < +\infty$ considered in the 1-dimensional location case, so much the better.

To verify that there is unique θ_0 with $\lambda(\theta_0) = 0$ is not as easy, but if ψ has some strict monotonicity property (or multidimensional extensions of such a property) it may be possible to show that this is true for all P. It may also be that existence of a pseudo-true θ_0 is a restriction on P, for example, in the case of the median, that P has to have a unique median (although actually (B-2) doesn't hold for the ψ function corresponding to the median).

For (B-4)(i), $\Psi(x)$ will be integrable for all P if and only if it is bounded. For this, it's sufficient that $\psi(\theta, x)$ be bounded uniformly in θ and x.

For (B-4)(ii), if $\theta \in \mathbb{R}$ and ψ is real-valued, one way to ensure the condition for all P is that for all x,

$$\liminf_{\theta \to -\infty} \psi(\theta, x)/b(\theta) \geq 1 \quad \text{and} \quad \liminf_{\theta \to +\infty} \psi(\theta, x)/b(\theta) \leq -1.$$

In higher dimensions, the situation is more complicated because instead of just two directions for going to infinity there are infinitely many, but on the other hand for ψ vector valued, it will tend to be small or zero less often.

(B-4)(iii) will hold if $\limsup_{\theta\to\infty} |\psi(\theta,x)-\lambda(\theta)|/b(\theta) < 1$ for all x, but this may still not be straightforward to check since $\lambda(\theta)$ depends on the unknown P.

A difficulty about (B-4) is that parts (i) and (iii) require $b(\theta)$ to be not too small, whereas part (ii) requires it to be not too large.

Some other comments on the assumptions: recall that one way ψ functions commonly arise is as the gradients with respect to θ of ρ functions. If so, then (B-2) implies that for almost all x, $\rho(\cdot,x)$ is a C^1 function of θ, as the narrow-sense Huber functions are. (B-1) and (B-2) imply that ψ is separable, as in (A-1) of Sec. 3.3, with S any countable dense subset of Θ and $A := \{x : \psi(\cdot,x) \text{ is not continuous }\}$. (B-1), (B-2) and the integrability in (B-3) are mild regularity conditions. If there were $\theta \neq \phi$ with $\lambda(\theta)=\lambda(\phi)=0$, then (3.5.1) could hold by the law of large numbers when T_n is either near ϕ or near θ, so T_n would not necessarily converge. Thus λ having a unique zero at some θ_0 is a natural assumption for consistency, specifically $T_n \to \theta_0$. (B-4) is the most technical, least intuitive of the assumptions.

Assumption (B-4(i)) gives $|\psi(\theta,x)| \le \Psi(x)b(\theta)$ for an integrable function Ψ. If U is a neighborhood of θ whose closure is compact, then $b(\cdot)$, being continuous, is bounded on U. It follows that

$$\sup\{|\psi(\theta,x)-\psi(\phi,x)| : \phi \in U\}) \le 2\Psi(x)\sup\{b(\phi) : \phi \in U\},$$

an integrable function. This, assumption (B-2), and dominated convergence imply:

(B-2′) For any θ, as a neighborhood U of θ converges to $\{\theta\}$,

$$E(\sup\{|\psi(\theta,x)-\psi(\phi,x)| : \phi \in U\}) \to 0.$$

Then, it follows that $\lambda(\cdot)$ as defined in (B-3) is continuous.

The next fact is not needed for the proof of consistency (Theorem 3.5.4 below) but it may be useful in checking hypothesis (B-4) by suggesting what function(s) to use for $b(\theta)$, if we can control $\lambda(\theta)$ well enough without knowing P.

3.5.2 Proposition. If (B-1) through (B-4) all hold, for some $b(\theta)$ and b_0, then (B-4) also holds for $B(\theta) := \max(|\lambda(\theta)|, b_0)$ or $B_1(\theta) := \max(|\lambda(\theta)|, b_0')$ where $b_0' := \liminf_{\phi\to\infty} |\lambda(\phi)|$ in place of $b(\theta)$.

Proof. Clearly, $B(\cdot)$ is continuous and $\ge b_0$. From (ii) for $b(\theta)$ and $b(\theta) \ge b_0$, we have $b_0' \ge b_0$, so (ii) holds for $B(\cdot)$. Also, (ii) for $b(\cdot)$ implies that for any $\varepsilon > 0$, there is a compact K such that for $\theta \notin K$, $b(\theta) \le (1+\varepsilon)|\lambda(\theta)|$. For ε small enough, this implies (iii) for $B(\cdot)$, and also (i) for the supremum over the complement of K. For the supremum over K, (i) is equivalent for any two positive continuous functions, such as $b(\cdot)$ and $B(\cdot)$, both bounded away from 0.

Since $B_1 \ge B$, clearly (i) and (iii) hold for B_1. Also, from the definitions, (ii) holds for B_1. □

Assuming (B-1), (B-2) and (B-3), we have that (B-4) holds for some function $b(\cdot)$ if and only if both $b_0' > 0$ and (B-4) holds for $B_1(\theta)$ in place of $b(\theta)$ by Proposition 3.5.2.

Still, the function $\lambda(\cdot)$ depends on the law P which usually is unknown to the statistician. Thus the assumptions would usually need to be checked for all P in some class (which may or may not be parametrized by Θ).

3.5.3 Lemma. If (B-1) and (B-4) hold, then for any sequence $\{T_n\}$ of approximate M-estimators of ψ type there is a compact set $C \subset \Theta$ such that $T_n \in C$ eventually a.s., specifically (3.3.11) holds.

Proof. For a compact set C, let

$$w_C(x) \ := \ \sup\{|\psi(\theta, x) - \lambda(\theta)|/b(\theta) : \ \theta \notin C\}.$$

By (B-4) (i) and (iii) and dominated convergence, since $0 \le w_C(x) \le \Psi(x) + E\Psi$ for any C, we can take C large enough so that $Ew_C(x) < 1$. Then we can take $\varepsilon > 0$ small enough so that $Ew_C < 1 - 3\varepsilon$. Note that if $C \subset D$ for another compact set D, we have $w_D(x) \le w_C(x)$ for all x and so $Ew_D \le Ew_C$. Thus by (B-4(ii)), we can take the compact C large enough so that $|\lambda(\theta)| > (1-\varepsilon)b(\theta)$ for $\theta \notin C$, and $Ew_C < 1 - 3\varepsilon$ still holds.

By the strong law of large numbers for $w(\cdot)$, a.s. for n large enough

$$sup\{\frac{1}{n}\sum_{i=1}^{n} |\psi(\theta, X_i) - \lambda(\theta)|/b(\theta) : \ \theta \notin C\} \ \le \ \frac{1}{n}\sum_{i=1}^{n} w(X_i) \ \le \ 1 - 2\varepsilon,$$

so for θ not in C,

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} |\psi(\theta, X_i) - \lambda(\theta)| \ \le \ (1-2\varepsilon)b(\theta) \ \le \ (1-2\varepsilon)|\lambda(\theta)|/(1-\varepsilon) \ \le \ (1-\varepsilon)|\lambda(\theta)|,$$
$$\text{so} \quad |\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \psi(\theta, X_i)| \ \ge \ \varepsilon|\lambda(\theta)| \ \ge \ \varepsilon(1-\varepsilon)b_0 \ > \ 0.$$

This implies (3.3.11). □

3.5.4 Theorem. Let $\{T_n\}$ be a sequence of approximate M-estimators of ψ type. If
(a) (B-1), (B-2), (B-3) and (B-4) hold,
or if
(b) (B-1), (B-2′) and (B-3) hold, and (3.3.11) holds for some compact C,
then $T_n \to \theta_0$ almost uniformly.

Proof. Hypotheses (a) imply (B-2′) as noted at its statement, and (3.3.11) by Lemma 3.5.3, so we can assume hypotheses (b). Then, we can assume that Θ is the compact set C in (3.3.11). By (B-2′), $\lambda(\cdot)$ is continuous. Let U be any open neighborhood of θ_0. Then on the compact set $C \setminus U$, λ is strictly positive by (B-3) and attains its infimum, which is $\ge 5\delta$ for some $\delta > 0$. For each $\theta \in C \setminus U$, take a neighborhood U_θ by (B-2′) such that

$$E(\sup\{|\psi(\phi, x) - \psi(\theta, x)| : \ \phi \in U_\theta\}) \ \le \ \delta. \tag{3.5.5}$$

Then $|\lambda(\phi) - \lambda(\theta)| \ \le \ \delta$ for $\phi \in U_\theta$. Since $C \setminus U$ is compact, take a finite subcover $U_j \ := \ U_{\theta_j}$ for some $M < \infty$ and $j = 1, \ldots, M$. Then

$$S \ := \ \sup\{\frac{1}{n}|\sum_{i=1}^{n} \psi(\phi, X_i) - \lambda(\phi)| : \ \phi \in C \setminus U\} \ \le \ T_1 + T_2 + T_3$$

where $T_1 := \max_{1\le j\le M} \frac{1}{n}\sum_{i=1}^n \sup\{|\psi(\phi,X_i)-\psi(\theta_j,X_i)| : \phi\in U_j\}$,

$$T_2 := \max_{1\le j\le M} \frac{1}{n}|\sum_{i=1}^n \psi(\theta_j,X_i)-\lambda(\theta_j)|,$$

and $T_3 := \max_j \sup\{|\lambda(\phi)-\lambda(\theta_j)| : \phi\in U_j\}$. Then $T_3\le\delta$ by (3.5.5) since for each $\phi\in U_j$,

$$|\lambda(\phi)-\lambda(\theta_j)| = |E\psi(\phi,x)-E\psi(\theta_j,x)| \le E|\psi(\phi,x)-\psi(\theta_j,x)|.$$

Almost surely for n large enough, $T_1\le 2\delta$ by (3.5.5) applied to $\theta=\theta_j$ and the strong law of large numbers M times; also, $T_2\le\delta$ by (B-3) and the strong law of large numbers M times, once for each θ_j. Then $S\le 2\delta+\delta+\delta = 4\delta$. But $|\lambda(\phi)|\ge 5\delta$ for $\phi\in C\setminus U$ implies that $\frac{1}{n}|\sum_{i=1}^n\psi(\phi,X_i)|\ge\delta$ for $\phi\in C\setminus U$ and n large enough, which implies $T_n\in U$ eventually a.s., specifically $1_{T_n\in U}\to 1$ almost uniformly. So $T_n\to\theta_0$ almost uniformly. □

PROBLEMS

1. Consider $\psi(\theta,x)=\rho'(x-\theta)$ for ρ equal to wide-sense Huber function (b) on p. 6 of section 3.4, $\rho(x)=(c^2+x^2)^{1/2}$ for some $c>0$. Take $c=1$. Verify that in this section, conditions (B-1) through (B-4) all hold for any law P. *Hints*: for (B-3), show that $\lambda'(\theta)<0$ for all θ, and find limits of $\lambda(\theta)$ as $\theta\to-\infty$ or $+\infty$.

2. Consider the narrow-sense Huber functions, (c) on p. 6 of section 3.4, where for some $b>0$, $\rho(x)=x^2$ for $|x|\le b$ and $c|x|+d$ otherwise, where c and d are chosen to make ρ a C^1 function. Show that in this case there is always a θ_0 such that $\lambda(\theta_0)=0$ (by the intermediate value theorem: show that $\lambda(\cdot)$ is continuous, positive at some θ and negative and some other θ). Show however that if a law P has an interval of medians longer than $2b$, in other words its distribution function $F(x)=P((-\infty,x])$ is equal to $1/2$ on such an interval, then $\lambda(\theta)$ is not 0 at a unique point θ_0 but is zero on some interval (u,v) with $u<v$.

NOTES

This section is based on the paper by Huber (1967), pp. 224-226.

REFERENCE

Huber, P. J. (1967). See sec. 3.3.