Statistics 582

Revision of Section 4.4: Consistency of Maximum Likelihood Estimates

Wellner; 1/9/98

Some Uniform Strong Laws of Large Numbers

Suppose that:

A. $X, X_1, \ldots, X_n$ are i.i.d. P on the measurable space $(\mathcal{X}, \mathcal{A})$.

B. For each $\theta \in \Theta$, $f(x, \theta)$ is a measurable, real-valued function of x, $f(\cdot, \theta) \in L_1(P)$.

Let $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$. Since $f(\cdot, \theta) \in L_1(P)$ for each θ,

$$\mu(\theta) = Ef(X, \theta) = \int f(x, \theta) dP(x) \equiv Pf(\cdot, \theta)$$

exists and is finite. Moreover, by the strong law of large numbers,

$$\begin{aligned} \mathbb{P}_n f(\cdot, \theta) &\equiv \int f(x, \theta) d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \\ &\rightarrow_{a.s.} Ef(X, \theta) = Pf(\cdot, \theta) = \mu(\theta). \end{aligned} \tag{0.1}$$

It is often useful and important to strengthen (0.1) to hold uniformly in $\theta \in \Theta$:

$$\sup_{\theta \in \Theta} |\mathbb{P}_n f(\cdot, \theta) - Pf(\cdot, \theta)| \rightarrow_{a.s.} 0\,. \tag{0.2}$$

Note that the left side in (0.2) is equal to

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf|\,.$$

Here is how (0.2) can be used: suppose that we have a sequence $\widehat{\theta}_n$ of estimators, possibly dependent on $X_1, \ldots, X_n$, such that $\widehat{\theta}_n \rightarrow_{a.s.} \theta_0$. Suppose that $\mu(\theta)$ is continuous at θ_0. We would like to conclude that

$$\mathbb{P}_n f(\cdot, \widehat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, \widehat{\theta}_n) \rightarrow_{a.s.} \mu(\theta_0). \tag{0.3}$$

The convergence (0.3) *does not follow* from (0.1); but (0.3) *does follow* from (0.2):

$$\begin{aligned} \left|\mathbb{P}_n f(\cdot, \widehat{\theta}_n) - \mu(\theta_0)\right| &\leq \left|\mathbb{P}_n f(\cdot, \widehat{\theta}_n) - \mu(\widehat{\theta}_n)\right| + \left|\mu(\widehat{\theta}_n) - \mu(\theta_0)\right| \\ &\leq \sup_{\theta \in \Theta} \left|\mathbb{P}_n f(\cdot, \theta) - \mu(\theta)\right| + \left|\mu(\widehat{\theta}_n) - \mu(\theta_0)\right| \\ &= \|\mathbb{P}_n - P\|_{\mathcal{F}} + \left|\mu(\widehat{\theta}_n) - \mu(\theta_0)\right| \\ &\rightarrow_{a.s.} 0 + 0 = 0\,. \end{aligned}$$

The following theorems, due to Le Cam, give conditions on f and P under which (2) holds. The first theorem is a prototype for what are now known in empirical process theory as "Glivenko-Cantelli theorems".

Theorem 1. Suppose that:
(a) Θ is compact.
(b) $f(x, \cdot)$ is continuous in θ for all x.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and $|f(x, \theta)| \le F(x)$ for all $x \in \mathcal{X}$, $\theta \in \Theta$.
Then (0.2) holds; i.e.

$$\sup_{\theta \in \Theta} |\mathbb{P}_n f(\cdot, \theta) - Pf(\cdot, \theta)| \rightarrow_{a.s.} 0\,.$$

The second theorem is a "one-sided" version of theorem 1 which is useful for the theory of maximum likelihood estimation.

Theorem 2. Suppose that:
(a) Θ is compact.
(b) $f(x, \cdot)$ is upper semicontinuous in θ for all x.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and $f(x, \theta) \le F(x)$ for all $x \in \mathcal{X}$, $\theta \in \Theta$.
(d) For all θ and all sufficiently small $\rho > 0$

$$\sup_{|\theta' - \theta| < \rho} f(x, \theta')$$

is measurable in x.
Then

$$\text{limsup}_{n \to \infty} \sup_{\theta \in \Theta} \mathbb{P}_n f(\cdot, \theta) \le_{a.s.} \sup_{\theta \in \Theta} Pf(\cdot, \theta) = \sup_{\theta \in \Theta} \mu(\theta).$$

We proceed by first proving Theorem 2. Then Theorem 1 will follow as a consequence of Theorem 2.

Proof of Theorem 2. Let

$$\psi(x, \theta, \rho) \equiv \sup_{|\theta' - \theta| < \rho} f(x, \theta').$$

Then ψ is measurable (for ρ sufficiently small), bounded by an integrable function F, and

$$\psi(x, \theta, \rho) \searrow f(x, \theta) \qquad \text{as} \qquad \rho \searrow 0 \qquad \text{by} \quad (b).$$

Thus by the monotone convergence theorem

$$\int \psi(x, \theta, \rho) dP(x) \searrow \int f(x, \theta) dP(x) = \mu(\theta).$$

Let $\epsilon > 0$. For each θ, find ρ_θ so that

$$\int \psi(x, \theta, \rho) dP(x) < \mu(\theta) + \epsilon\,.$$

The spheres

$$S(\theta, \rho_\theta) = \{\theta' : |\theta' - \theta| < \rho_\theta\}$$

cover Θ, so by (a) there exists a finite subcover: $\Theta \subset \cup_{j=1}^m S(\theta_j, \rho_{\theta_j})$. for each $\theta \in \Theta$ there is some j, $1 \le j \le m$, such that $\theta \in S(\theta_j, \rho_{\theta_j})$; hence from the definition of ψ it follows that

$$f(x, \theta) \le \psi(x, \theta_j, \rho_{\theta_j})$$

for all x. Therefore

$$\mathbb{P}_n f(\cdot, \theta) \le \mathbb{P}_n \psi(\cdot, \theta_j, \rho_{\theta_j}),$$

and hence

$$\begin{aligned}
\sup_{\theta \in \Theta} \mathbb{P}_n f(\cdot, \theta) &\le \sup_{1 \le j \le m} \mathbb{P}_n \psi(\cdot, \theta_j, \rho_{\theta_j}) \\
&\to_{a.s.} \sup_{1 \le j \le m} P\psi(\cdot, \theta_j, \rho_{\theta_j}) \\
&\le \sup_{1 \le j \le m} \mu(\theta_j) + \epsilon \\
&\le \sup_{\theta \in \Theta} \mu(\theta) + \epsilon\,.
\end{aligned}$$

Hence

$$\text{limsup}_{n\to\infty} \sup_{\theta \in \Theta} \mathbb{P}_n f(\cdot, \theta) \le_{a.s.} \sup_{\theta \in \Theta} \mu(\theta) + \epsilon\,.$$

Letting $\epsilon \downarrow 0$ completes the proof.

□

Proof of Theorem 1. Since f is continuous in θ, condition (d) of Theorem 2 is satisfied: for any countable set D dense in $\{\theta' : |\theta' - \theta| < \rho\}$,

$$\sup_{|\theta' - \theta| < \rho} f(x, \theta') = \sup_{\theta' \in D} f(x, \theta')$$

where the right side is measurable since it is a countable supremum of measurable functions. Furthermore, $\mu(\theta)$ is continuous in θ:

$$\lim_{\theta' \to \theta} \mu(\theta) = \lim_{\theta' \to \theta} \int f(x, \theta') dP(x) = \int f(x, \theta) dP(x)$$

by the dominated convergence theorem. Now Theorem 1 follows from Theorem 2 applied to the functions $g(x, \theta) \equiv f(x, \theta) - \mu(\theta)$ and $-g(x, \theta)$: by Theorem 2 applied to $\{g(x, \theta) : \theta \in \Theta\}$,

$$\text{limsup}_{n\to\infty} \sup_{\theta \in \Theta} (\mathbb{P}_n f(\cdot, \theta) - \mu(\theta)) \le 0 \qquad \text{a.s.}$$

By Theorem 2 applied to $\{-g(x,\theta) : \theta \in \Theta\}$,

$$\text{limsup}_{n\to\infty} \sup_{\theta\in\Theta}(\mu(\theta)) - \mathbb{P}_n f(\cdot,\theta)) \le 0 \qquad \text{a.s.}$$

The conclusion of Theorem 1 follows since

$$\begin{aligned} 0 &\le \sup_{\theta\in\Theta} |\mathbb{P}_n f(\cdot,\theta) - \mu(\theta)| \\ &= \sup_{\theta\in\Theta}(\mathbb{P}_n f(\cdot,\theta) - \mu(\theta)) \vee \sup_{\theta\in\Theta}(\mu(\theta) - \mathbb{P}_n f(\cdot,\theta)) \,. \end{aligned}$$

$\square$

For our application of Theorem 2 to consistency of maximum likelihood, the following Lemma will be useful.

Lemma 1. If the conditions of Theorem 2 hold, then $\mu(\theta)$ is upper semicontinuous: i.e.

$$\text{limsup}_{\theta'\to\theta}\mu(\theta') \le \mu(\theta).$$

Proof. Since $f(x,\theta)$ is upper semicontinuous,

$$\text{limsup}_{\theta'\to\theta} f(x,\theta') \le f(x,\theta) \qquad \text{for all } x\,;$$

i.e.

$$\text{liminf}_{\theta'\to\theta}\{f(x,\theta) - f(x,\theta')\} \ge 0 \qquad \text{for all } x\,.$$

Hence it follows by Fatou's lemma that

$$\begin{aligned} 0 &\le E\,\text{liminf}_{\theta'\to\theta}\{f(X,\theta) - f(X,\theta')\} \\ &\le \text{liminf}_{\theta'\to\theta} E\{f(X,\theta) - f(X,\theta')\} \\ &= Ef(X,\theta) - \text{limsup}_{\theta'\to\theta} Ef(X,\theta')\,; \end{aligned}$$

i.e.

$$\text{limsup}_{\theta'\to\theta} Ef(X,\theta') \le Ef(X,\theta) = \mu(\theta)\,.$$

$\square$

Now we are prepared to tackle consistency of maximum likelihood estimates.

Theorem 3. Suppose that $X, X_1, \ldots, X_n$ are i.i.d. P_{θ_0}, $\theta_0 \in \Theta$ with density $p(x,\theta_0)$ with respect to the dominating measure ν, and that:
(a) Θ is compact.
(b) $p(x,\cdot)$ is upper semicontinuous in θ for all x.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and

$$f(x,\theta) \equiv \log p(x,\theta) - \log p(x,\theta_0) \le F(x)$$

for all $x \in \mathcal{X}$, $\theta \in \Theta$.
(d) For all θ and all sufficiently small $\rho > 0$

$$\sup_{|\theta' - \theta| < \rho} p(x, \theta')$$

is measurable in x.
(e) $p(x, \theta) = p(x, \theta_0)$ a.e. $d\nu$ implies that $\theta = \theta_0$.
Then for any sequence of maximum likelihood estimates $\widehat{\theta}_n$ of θ_0,

$$\widehat{\theta}_n \rightarrow_{a.s.} \theta_0 \, .$$

Proof. Let $\rho > 0$. The functions $\{f(x, \theta) : \theta \in \Theta\}$ satisfy the conditions of theorem 2. But we will apply Theorem 2 with Θ replaced by the subset

$$S \equiv \{\theta : \, |\theta - \theta_0| \geq \rho\} \; \subset \Theta \, .$$

Then S is compact, and by Theorem 2

$$P_{\theta_0}\left(\text{limsup}_{n \to \infty} \sup_{\theta \in S} \mathbb{P}_n f(\cdot, \theta) \leq \sup_{\theta \in S} \mu(\theta) \right) = 1$$

where

$$\begin{aligned} \mu(\theta) = E_{\theta_0} f(X, \theta) &= E_{\theta_0} \left\{ \log \frac{p(X, \theta)}{p(X, \theta_0)} \right\} \\ &= -K(P_{\theta_0}, P_\theta) < 0 \qquad \text{for} \;\; \theta \in S. \end{aligned}$$

Furthermore by the Lemma, $\mu(\theta)$ is upper semicontinuous and hence achieves its supremum on the compact set S. Let $\delta = \sup_{\theta \in S} \mu(\theta)$. Then by Lemma 4.1.2 it follows that $\delta < 0$ and we have

$$P_{\theta_0}\left(\text{limsup}_{n \to \infty} \sup_{\theta \in S} \mathbb{P}_n f(\cdot, \theta) \leq \delta \right) = 1 \, .$$

Thus with probability 1 there exists an N such that for all $n > N$

$$\sup_{\theta \in S} \mathbb{P}_n f(\cdot, \theta) \leq \delta / 2 < 0 \, .$$

But

$$\begin{aligned} \mathbb{P}_n f(\cdot, \widehat{\theta}_n) &= \sup_{\theta \in \Theta} \mathbb{P}_n f(\cdot, \theta) \\ &= \sup_{\theta \in \Theta} \frac{1}{n} \{ l_n(\theta) - l_n(\theta_0) \} \geq 0 \, . \end{aligned}$$

Hence $\widehat{\theta}_n \notin S$ for $n > N$; that is, $|\widehat{\theta}_n - \theta_0| < \rho$ with probability 1. Since ρ was arbitrary, $\widehat{\theta}_n$ is a.s. consistent.

Remark 3. Theorem 3 is due to Wald (1949). The present writeup is an adaptation of Chapters 16 and 17 of Ferguson (1996). For further Glivenko - Cantelli theorems, see chapter 2.4 of Van der Vaart and Wellner (1996).

References:

- Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related estimates. *Univ. Calif. Publ. in Statist.* **1**, 277 - 330.
- Ferguson, T. (1996). *A Course in Large Sample Theory.* Chapman and Hall, London.
- Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.
- Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595 - 601.