Chapter 2
Consistency

Preliminary draft. Please do not distribute.

SECTION *1 explains why the Chapter will be mostly concerned with M-estimators, even though the techniques are more widely applicable.*

SECTION *2 presents a variation on Wald's original argument for consistency of maximum likelihood estimators.*

SECTION *3 considers some of the problems that can arise for estimators that range over noncompact parameter spaces, or for models that cannot be successfully compactified.*

SECTION *4 establishes some basic comparison principles for deterministic functions, which lead to limit theorems when applied to sample paths of random criterion functions.*

SECTION **??** *presents a method that typically succeeds because of existence of a suitable uniform law of large numbers.*

SECTION **??** *presents a high-level result for consistency, in terms of uniformity assumptions about the behaviour of the criterion function.*

SECTION *5: consistency for Z-estimators.*

SECTION *6: consistency for* MLE *of monotone density on* $\mathbb{R}^+$.

2.1 Estimators defined by minimization

Consistency::minimization

The statistics and econometrics literatures contain a huge number of theorems that establish consistency of different types of estimators, that is, theorems that prove convergence in some probabilistic sense of an estimator to some desired limiting value.

The variety of different consistency theorems can be overwhelming for a new researcher. In my opinion, any attempt at cataloging all the variations on the consistency theme is a not particularly fruitful activity. I think it is much better to concentrate attention on a smaller number of general principles, accepting that new applications might require some tweaking of the standard methods in order to accommodate the peculiarites of particular examples.

Accordingly, this chapter will focus mainly on a particular type of estimator, namely those estimators defined by minimization (or maximization) of a some stochastic process. Maximum likelihood estimators and

the M-estimators from Chapter 1 are of this type. More generally, suppose $G_n(\theta) = G_n(\omega, \theta)$ is a random variable for each θ in an index set Θ. Suppose also that an estimator $\widehat{\theta}_n = \widehat{\theta}_n(\omega)$ is defined by minimization of $G_n(\cdot)$, or at least is required to come close to minimizing $G_n(\cdot)$ over Θ,

$$G_n(\widehat{\theta}_n) \approx \inf_{\theta \in \Theta} G_n(\theta),$$

in a sense soon to be made more precise. What asymptotic properties must $\widehat{\theta}_n$ have, as a consequence of the minimization?

If $G_n(\theta)$ is a smooth function of θ with Θ a subset of some $\mathbb{R}^k$ then it is sometimes technically simpler to define $\widehat{\theta}_n$ as the zero of the derivative $L_n(\theta) := \partial G_n(\theta)/\partial\theta$. Of course this approach will encounter difficulties if minima are achieved at boundary points, or if $\widehat{\theta}_n$ comes only close to minimizing G_n. As explained in Chapter 5, there are even cases where perfectly straightforward minimizations are made to seem more complicated by a search for a zero derivative. Moreover, at the cost of some artificiality, we could always recast the problem of finding a zero of a process $L_n(\theta)$ as a problem of minimizing a process such as $\|L_n(\theta)\|^2$. See Section 5.

The prototype for rigorous consistency arguments is a theorem due to Wald (1949), for maximum likelihood estimators under independent sampling, with Θ equal to a closed subset of $\mathbb{R}^k$. (He also noted that his argument would work in more general settings.) Section 2 will present a variation on Wald's theorem.

ref for bracketing?

In more modern terminology, the central idea in Wald's method is a form of one-sided bracketing argument. In recent decades it has become more common for proofs to involve uniform two-sided bounds, probably because the methods used to establish the bounds (such as the symmetrization methods from empirical process theory) usually deliver two-sided inequalities. The argument leading from the bounds to the desired concentration of the estimator in a small region of the parameter space typically depend on comparisons involving a single sample path $\theta \mapsto G_n(\omega, \theta)$, for fixed ω and n. To stress this point, I will first present the general arguments first (Section 4) as comparisons for deterministic functions, leaving you to write out their stochastic analogs.

2.2 Wald's method

Consistency::wald

The following result captures the main idea of Wald (1949).

Wald <1>

Theorem. *Let $G_n(\theta) = \sum_{i \le n} g(X_i(\omega), \theta)/n$, with the $\{X_i\}$ independently distributed as P and $g(\cdot, \theta) \in \mathcal{L}^1(P)$ for each θ in Θ. Suppose:*

W1 *(i)* Θ *is a compact metric space*

W2 *(ii)* θ_0 *is a point of* Θ *such that* $Pg(\cdot,\theta) > Pg(\cdot,\theta_0) > -\infty$ *for all* $\theta \neq \theta_0$.

W3 *(iii) For each* $\theta \in \Theta$, *the function* $t \mapsto g(x,t)$ *is lower semi-continuous at* θ *for* P *almost all* x.

> **Remark.** *That is, there is a* P*-negligible set* $\mathcal{N}_\theta$ *for which* $\liminf_{i\to\infty} g(x,t_i) \geq g(x,\theta)$ *if* $t_i \to \theta$ *and* $x \notin \mathcal{N}_\theta$.

W4 *(iv)* $P^x \inf_\theta g(x,\theta) > -\infty$,

W5 *(v)* $\widehat{\theta}_n$ *is a random element of* Θ *for which* $G_n(\widehat{\theta}_n) \leq \epsilon_n + \inf_\theta G_n(\theta)$

If $\epsilon_n \to 0$ *almost surely then* $\widehat{\theta}_n \to \theta_0$ *almost surely; if* $\epsilon_n \to 0$ *in probability then* $\widehat{\theta}_n \to \theta_0$ *in probability.*

> **Remark.** When $P = P_{\theta_0}$, a member of a family of probability measures $\{P_\theta : \theta \in \Theta\}$ with densities f_θ with respect to some dominating measure, and when $g(x,\theta) = -\log f_\theta(x)$, the $\widehat{\theta}_n$ from Theorem 1 can be thought of as approximate maximum likelihood estimator. As noted in Chapter 1, the minimization condition (ii) then follows from Jensen's inequality, provided we make sure that $P_\theta \neq P$ for $\theta \neq \theta_0$. More informatively,
>
> $$\begin{aligned} G(\theta) - G(\theta_0) &= P^x\left(-\log f_\theta(x) + \log f_{\theta_0}(x)\right) \\ &= \int f_{\theta_0}(x) \log\left(f_{\theta_0}(x)/f_\theta(x)\right), \end{aligned}$$
>
> the Kullback-Leibler distance between P_{θ_0} and P_θ.

Proof For each subset A of Θ define $h(x,A) := \inf_{\theta\in A} g(x,\theta)$. If we replace $g(x,\theta)$ by $g(x,\theta) - h(x,\Theta)$ neither the lower semicontinuity nor the defining properties for θ_0 and $\widehat{\theta}_n$ would be affected. We may, therefore, assume without loss of generality that $g(x,\theta) \geq 0 = h(x,\Theta)$.

> **Remark.** Even though nonmeasurability needs to be taken seriously when one manipulates uncountable families of random variables—for example, an infimum of an uncountable family of measurable functions need not itself be measurable—I will continue to sidestep such issues. A completely rigorous proof would need to establish measurability of $x \mapsto h(x,A)$ for each open ball A. See the Notes at the end of the Chapter for further discussion of measurability.

Fix an open neighborhood U of θ_0. Fatou's lemma and (iii) ensure that the function $\theta \mapsto G(\theta) := P^x g(x,\theta)$ is everywhere lower semi-continuous. It must achieve its infimum on the compact set $\Theta\backslash U$. By (ii), there must therefore be some $\epsilon > 0$ for which

$$G(\theta) \geq 3\epsilon + G(\theta_0) \qquad \text{for each } \theta \text{ in } \Theta\backslash U.$$

Consider a fixed t in $\Theta\backslash U$. There exists a sequence of open balls $N_i(t)$ that shrinks to $\{t\}$ as $i \to \infty$. By (iii), $0 \leq h(x, N_i(t)) \uparrow g(x,t)$ a.e. $[P]$ as $i \to \infty$. Monotone Convergence then ensures existence of at least one open neighborhood, call it $N(t)$, of t for which

$$P^x h(x, N(t)) > G(t) - \epsilon \geq G(\theta_0) + 2\epsilon$$

By compactness of the set $\Theta\backslash U$, there exists some finite subset F for which $\Theta\backslash U \subseteq \cup_{t\in F} N(t)$, which implies

$$\inf_{\theta\in\Theta\backslash U} G_n(\theta) \geq \min_{t\in F} \frac{1}{n} \sum\nolimits_{i\leq n} h(X_i, N(t)).$$

By finitely many appeals to the SLLN, as n tends to infinity the right-hand side of the last inequality converges almost surely to

$$\min_{t\in F} P^x h(x, N(t)) \geq G(\theta_0) + 2\epsilon$$

and

$$G_n(\theta_0) \to G(\theta_0) \qquad \text{almost surely.}$$

If $\epsilon_n \to 0$ almost surely then, for almost all ω,

$$G_n(\widehat{\theta}_n) \leq \epsilon_n + G_n(\theta_0) < G(\theta_0) + \epsilon < \inf_{\theta\in\Theta\backslash U} G_n(\theta) \qquad \text{eventually.}$$

It follows that $\widehat{\theta}_n(\omega) \in U$ eventually. Complete the proof of almost sure convergence by casting out a sequence of negligible sets, one for each U in a sequence of neighborhoods that shrinks to $\{\theta_0\}$.

If $\epsilon_n \to 0$ in probability, we have instead that

$$\mathbb{P}\{G_n(\widehat{\theta}_n) < \inf_{\theta\in\Theta\backslash U} G_n(\theta)\} \to 1$$

and so $\mathbb{P}\{\widehat{\theta}_n \in U\} \to 1$ for each neighborhood U of θ_0.

□

The natural parameter space for many problems will not be compact. To apply Theorem <1> we must either provide some ad hoc, preliminary argument to force $\widehat{\theta}_n$ into some compact subset, or we must compactify Θ.

Cauchy <2>

Example. Consider the maximum likelihood estimator $\widehat{\theta}_n$ for a sample of size n from the Cauchy(θ) location family. That is, the observations $\{X_i\}$ are assumed to come from a density $f_{\theta_0}(\cdot)$ (with respect to Lebesgue measure) for an unknown θ_0, where

$$f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

The estimator minimizes the $G_n(\theta)$ corresponding to the nonnegative function

$$g(x,\theta) = \log\left(1+(x-\theta)^2\right),$$

with θ ranging over the whole real line.

Assumptions (iii) and (iv) of the Theorem are trivially satisfied. Assumption (ii) follows via the Jensen inequality or the by the argument in the Remark that followed the statement of the Theorem.

Of course $\mathbb{R}$ is not compact. However, if we enlarge the parameter space to $\Theta = [-\infty, \infty]$ and define $g(x, \pm\infty) \equiv \infty$, the argument from the Theorem carries over. At the two new compactification points, $\pm\infty$, the $g(x,\cdot)$ functions are lower semi-continuous; and $G(\pm\infty) = \infty > G(\theta_0)$; and $\widehat{\theta}_n$ also minimizes $G_n(\theta)$ over the compact Θ.

Theorem <1> applies to the compactified problem, establishing almost sure convergence of $\widehat{\theta}_n$ to θ_0.

The assumption that P corresponds to a density $f_{\theta_0}(\cdot)$ in the model family is not essential. It was used only to identify the θ that minimizes $Pg(\cdot,\theta)$. More generally, if $f_{\theta_0}(\cdot)$ were merely the Cauchy location density that gave the closest approximation to P, in the sense that

$$Pg(\cdot,\theta) > Pg(\cdot,\theta_0) > -\infty \qquad \text{for } \theta \neq \theta_0,$$

then the Theorem would again prove almost sure convergence of $\widehat{\theta}_n$ to θ_0, without θ_0 having any interpretation as a *true* value of the parameter.

□

2.3 Global difficulties

Consistency::global

It is not always possible to justify an application of Theorem <1> by means of a compactification of Θ, as in Example <2>. There are very simple cases where the conclusion of the Theorem holds, even though one of its assumptions is violated. The usual obstacle is the global integrability Assumption (iv). Wald's theorem is not the last word in consistency proofs.

normal.mle <3>

Example. Let $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$ be the maximum likelihood estimator for the $N(\mu, \sigma^2)$ family, with the natural parameter space

$$\Theta = \{(\mu, \sigma) : -\infty < \mu < \infty,\ \sigma > 0\}.$$

Under sampling from $P = N(\mu_0, \sigma_0^2)$, it is easy to prove directly that $\widehat{\mu}_n \to \mu_0$ and $\widehat{\sigma}_n \to \sigma_0$, both with probability one.

The estimator is defined by the minimization problem treated in Theorem <1> for the special case

$$g(x, \mu, \sigma) = \log \sigma + (x - \mu)^2/2\sigma^2.$$

The infimum over μ and σ equals $-\infty$, for every x: put μ equal to x then let σ tend to zero. Assumption (iv) of Theorem <1> fails.

If we restrict the parameters to the subset

$$\Theta_\epsilon = \{(\mu, \sigma) \in \Theta : \sigma \geq \epsilon\},$$

for any fixed $\epsilon > 0$, the difficulty disappears; the same sort of compactification argument as in Example <2> would allow us to invoke Theorem <1> when the observations come from some $N(\mu_0, \sigma_0^2)$ distribution.

A more subtle argument is needed to overcome the embarrassment for σ near zero. The traditional solution to the problem involves a pairing trick, which takes advantage of some special features of the normal distribution. For simplicity, suppose the sample size is even, $n = 2m$. Treat the observations $(X_1, X_2), \ldots, (X_{n-1}, X_n)$ as a sample of size m from the bivariate normal density. The maximum likelihood problem then corresponds to minimization with

$$g(x_1, x_2, \mu, \sigma) = 2 \log \sigma + (x_1 - \mu)^2/2\sigma^2 + (x_2 - \mu)^2/2\sigma^2.$$

If $x_1 \neq x_2$ the infimum is attained when μ equals $\bar{x} = (x_1 + x_2)/2$ and σ equals T, where $2T^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 = (x_1 - x_2)^2/2$:

$$\inf_{\mu,\sigma} g(x_1, x_2, \mu, \sigma) = 1 + 2 \log T.$$

When the $\{X_i\}$ are sampled from a $N(\mu_0, \sigma_0^2)$ distribution, the infimum is integrable: the logarithm of a χ_1^2 random variable has a finite expectation. A compactification of Θ, and an argument analogous to that for the Cauchy location example, would lead us to a successful application of Theorem <1> for the bivariate samples.

□

I find the pairing remedy for the failure of Theorem <1> in the $N(\mu, \sigma^2)$ problem unsatisfying; it gives little insight into why small σ cause so much trouble. You might be puzzled about why I should be at all concerned with the failure of the Theorem in this case. After all, we have closed-form expressions for both $\widehat{\mu}_n$ and $\widehat{\sigma}_n$; a direct proof of almost sure convergence is not difficult. The real reason for concern is the shortcoming the Example exposes in the Theorem. What would happen in other problems where closed-form solutions are not available? It would be preferable to have general theorems that cover even the difficult cases, and not have to rely on ad hoc arguments to force parameters into compact regions where local arguments have global consequences. Unfortunately, general theorems tend to be conglomerates of special devices, aimed at eliminating difficulties specific to well known bad cases.

There have been many attempts at formulating general conditions that can handle the global problem. One of the most elegant is due to Huber (1967). See Problem [1]. Unfortunately, I have not been able to use Huber's method to eliminate the problem with small σ for the $N(\mu, \sigma^2)$ model. Fortunately, the problem can be remedied by using one of the general theorems from the next Section.

2.4 Comparison arguments

Consistency::comparison

As noted in Section 1, most proofs for M-estimators involve comparisons between individual sample paths of stochastic processes. Proofs for random processes become simpler if we first isolate and study the comparison arguments for deterministic functions.

Suppose f is a real-valued function defined on some set T. If T_0 is a subset of T and t^* is a point of T for which $f(t^*) < \inf_{t \in T \setminus T_0} f(t)$ then t^* must lie in T_0. This trivial idea lies at the heart of most analyses of minimization estimators.

Often the argument appears in a slightly disguised form involving two real-valued functions, f_1 and f_2 on the set T with finite infima $M_i := \inf_{t \in T} f_i(t)$. The set T_0 might equal $\{t \in T : f_2(t) \leq M_2 + \gamma\}$ for some (small) nonnegative γ and t^* might be a point at which f_1 is (almost) minimized, $f_1(t^*) \leq M_1 + \epsilon$ for some (small) nonnegative ϵ. If f_1 and f_2 are close enough in some uniform sense then we can hope that $f_2(t^*)$ also lies close enough to M_2 to force t^* into the region T_0.

The easiest case occurs when we have a uniform bound on $|f_1(t) - f_2(t)|$.

compare1 <4> **Lemma.** *Let f_1 and f_2 be two real-valued function defined on a set T for which $M_i := \inf_{t\in T} f_i(t)$ is finite for $i = 1, 2$. Let t^* be a point of T such that, for fixed $\epsilon \geq 0$ and $\delta > 0$,*

C1.1 *(i)* $f_1(t^*) \leq M_1 + \epsilon$

C1.2 *(ii)* $\sup_{t\in T} |f_1(t) - f_2(t)| \leq \delta$

Then $f_2(t^*) \leq M_2 + \epsilon + 2\delta$.

PROOF From (ii), $f_1(t) \leq f_2(t)+\delta$ for all t, which implies that $M_1 \leq M_2+\delta$. Thus

$$\begin{aligned} f_2(t^*) &\leq f_1(t^*) + \delta \qquad \text{by (ii)} \\ &\leq M_1 + \delta + \epsilon \\ &\leq M_2 + 2\delta + \epsilon. \end{aligned}$$

□

ULLN <5> **Theorem.** *Suppose $\{G_n(\theta) : \theta \in \Theta\}$ is a random criterion function indexed by a metric space (Θ, d). Let G be a deterministic function on Θ. Suppose $\widehat{\theta}_n$ is a random element of Θ. Suppose Θ_n is a subset of Θ for which*

(i) $\sup_{\theta\in\Theta_n} |G_n(\theta) - G(\theta)| \leq \Delta_n$

(ii) $G_n(\widehat{\theta}_n) \leq \epsilon_n + \inf_{\theta\in\Theta} G_n(\theta)$

(iii) there exists a $\theta_0 \in \Theta_n$ for which $\inf\{G(\theta) - G(\theta_0) : d(\theta, \theta_0) \geq \delta\} > 0$ for each $\delta > 0$.

Then

(a) If $\max(\epsilon_n, \Delta_n) = o_p(1)$ and $\mathbb{P}\{\widehat{\theta}_n \in \Theta_n\} \to 1$ then $d(\widehat{\theta}_n, \theta_0) = o_p(1)$.

(b) If $\max(\epsilon_n, \Delta_n) \to 0$ almost surely and $\mathbb{P}\{\omega : \widehat{\theta}_n(\omega) \in \Theta_n \text{ eventually}\} = 1$ then $d(\widehat{\theta}_n, \theta_0) \to 0$ almost surely.

PROOF Homework exercise for 618.

□

min.dist <6> **Example.** Needs editing Minimum distance with $\|P - Q\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |Ph - Qh|$. Maybe discuss classical case. cf. Problem [6]. Maybe discuss role as preliminary $\sqrt{n}$-consistent estimator. Adapt argument from the OLD comparison chapter.

□

An attempt at invoking Theorem <5> to prove consistency sometimes fails because the variability of $G_n(\theta)$ about $G(\theta)$ increases with $G(\theta)$. The failure can sometimes be averted if $|G_n(\theta)-G(\theta)|$ is small relative to $G(\theta)-G(\theta_0)$, uniformly in θ. The next Lemma gives one suitable meaning to the concept of relative closeness for deterministic functions.

compare2 <7> **Lemma.** *Let f_1 and f_2 be real-valued functions defined on a set T for which $M_i := \inf_{t\in T} f_i(t)$ is finite for $i = 1, 2$. Let t^* be a point of T such that, for a fixed $\epsilon \geq 0$,*

C2.1 *(i)* $f_1(t^*) \leq M_1 + \epsilon$

C2.2 *(ii) for fixed $\delta \geq 0$ and $\eta_i \in [0, 1)$,*

$$|f_1(t) - f_2(t)| \leq \delta + \eta_1|f_1(t)| + \eta_2|f_2(t)| \qquad \text{for every } t \text{ in } T$$

Then $f_2(t^*) \leq M_2 + \gamma$ where

$$\gamma = \frac{4\delta}{(1-\eta_2)(1-\eta_1)} + \frac{2\epsilon}{1-\eta_2} + \frac{2(\eta_1+\eta_2+\eta_1\eta_2)}{(1-\eta_1)(1-\eta_2)}|M_2|.$$

Remark. The coefficients of δ and ϵ are not important provided $\max(\eta_1, \eta_2)$ is bounded away from 1. For the whole inequality to be of much value we will need $|M_2|\max(\eta_1, \eta_2) \approx 0$.

Proof Recall that $\operatorname{sgn}(x) := \{x > 0\} - \{x < 0\}$. To avoid the need for separate examination of many combinations of signs, define $\sigma_i(t) = \operatorname{sgn}(f_i(t))$ and $\mu_i = \operatorname{sgn}(M_i)$. Abbreviate $\sigma_i(t^*)$ to σ_i^*.

Inequality (ii) splits into two one-sided inequalities. Write the first as

split.f1f2<8> $$f_1(t)\,[1 - \eta_1\sigma_1(t)] \leq \delta + f_2(t)\,[1 + \eta_2\sigma_2(t)] \qquad \text{for all } t,$$

Bound the left-hand side from below by

$$M_1\,(1 - \eta_1\sigma_1(t)) \geq M_1\,(1 - \eta_1\mu_1)\,.$$

Then, on the right-hand side, take a limit along a sequence of t values for which $f_2(t) \to M_2$, to deduce that

M1.M2<9> $$M_1(1 - \eta_1\mu_1) \leq \delta + M_2(1 + \eta_2\mu_2).$$

The analogous inequality, with the roles of f_1 and f_2 reversed, is needed only at t^*.

$$\begin{aligned} f_2(t^*)\,(1 - \eta_2\sigma_2^*) &\leq \delta + f_1(t^*)\,(1 + \eta_1\sigma_1^*) \\ &\leq \delta + (M_1 + \epsilon)\,(1 + \eta_1\sigma_1^*) \qquad \text{by assumption (i)} \end{aligned}$$

Multiply through by the nonnegative value $1-\eta_1\mu_1$ then substitute in the upper bound for $M_1(1-\eta_1\mu_1)$ from inequality <9>.

$$\begin{aligned}
f_2(t^*)&(1-\eta_2\sigma_2^*)(1-\eta_1\mu_1) \\
&\le [\delta+\epsilon(1+\eta_1\sigma_1^*)](1-\eta_1\mu_1)+[\delta+M_2(1+\eta_2\mu_2)](1+\eta_1\sigma_1^*) \\
&\le \delta(2-\eta_1\mu_1+\eta_1\sigma_1^*) \\
&\quad +\epsilon(1+\eta_1\sigma_1^*)(1-\eta_1\mu_1)+ \\
&\quad +M_2(1-\eta_2\sigma_2^*)(1-\eta_1\mu_1) \\
&\quad +|M_2|\times|\eta_2\sigma_2^*+\eta_1\mu_1+\eta_2\sigma_2^*\eta_1\mu_1+\eta_2\mu_2+\eta_1\sigma_1^*+\eta_2\mu_2\eta_1\sigma_1^*|
\end{aligned}$$

Divide through by $(1-\sigma_2^*\eta_2)(1-\eta_1\mu_1)$ then (crudely) argue worst cases for the signs to derive the asserted bound.

□

The stochastic analog of the Lemma gives a useful method for establishing consistency. The Theorem is an example of a "high-level result". It makes no assumption that G_n be an average of independent $g(x_i,\theta)$, and no assumption that $H_n=\mathbb{P}G_n$; they could be any two (possibly random) processes satisfying the uniform approximation requirement of the Lemma. The result could be applied to all manner of processes constructed from dependent variables. In any particular application it might take some effort to verify the high-level assumption, but the rest of the comparison argument need not be repeated.

Remark. I am very much in favour of high-level theorems. They eliminate a lot of the repetitive details that often consume journal pages; they focus attention on the important approximation requirements; and they do not build in too many low-level assumptions about the random processes under consideration. The difference between high-level and low-level approaches is like the difference between a program written in Perl (or whatever your favourite programming language might be) and a program written in assembly language.

To be replaced by a similar appeal.

transfer.min<10>

Theorem. *Suppose $\{G_n(\theta):\theta\in\Theta\}$ and $\{H_n(\theta):\theta\in\Theta\}$ are (possibly) random criterion functions indexed by a metric space (Θ,d). Suppose $\widehat{\theta}_n$ is a random element of Θ and Θ_n is a subset of Θ for which*

(i) $\mathbb{P}\{\widehat{\theta}_n\in\Theta_n\}\to 1$;

(ii)

$$\sup_{\theta\in\Theta_n}\frac{|G_n(\theta)-H_n(\theta)|}{1+|G_n(\theta)|+|H_n(\theta)|}=o_p(1);$$

Then $H_n(\widehat{\theta}_n) \le o_p(1) + (1 + o_p(1)) \inf_{\theta\in\Theta_n} H_n(\theta)$.

Suppose also that there exists a (possibly random) $\theta_n^* \in \Theta_n$ *and a function* $\kappa_n : \theta \to \mathbb{R}^+$ *for which*

(iii) $H_n(\theta) \ge H_n(\theta_n^*) + \kappa_n(\theta)^2$ *for all* $\theta \in \Theta_n$.

(iv) $H_n(\theta_n^*) = O_p(1)$

Then $\kappa_n(\widehat{\theta}_n) = o_p(1)$.

Proof Exercise for Stat 618. □

I leave the formulation and proof of an almost sure analogue of Theorem <10> to the motivated reader.

In applications of the Theorem, one often discards some of the weighting factors in the denominator from Assumption (ii), most typically the $|G_n(\theta)|$. The symmetry between G_n and H_n simplify the proof, but it is often easier to prove the stronger uniform convergence result with the smaller denominator.

more.normal.mle <11>

Example. You have seen in Example <3> how Theorem <1> fails for the $N(\mu, \sigma^2)$ model. The last Lemma solves the difficulty with small σ by means of a weighting factor that becomes infinite as σ tends to zero.

Suppose the observations are sampled from $P = N(\mu_0, \sigma_0^2)$. The maximum likelihood estimators minimize the G_n corresponding to

$$g(x, \mu, \sigma) = \log(\sigma/\sigma_0) + (x - \mu)^2/2\sigma^2.$$

The extra σ_0 has no effect on the location of the maximizing value, but it does ensure that the expected value

$$G(\mu, \sigma) = Pg(x, \mu, \sigma) = \log(\sigma/\sigma_0) + \frac{(\mu - \mu_0)^2 + \sigma_0^2}{2\sigma^2}$$

has a strictly positive minimum: it achieves its minimum value of $1/2$ cleanly at $\mu = \mu_0$ and $\sigma = \sigma_0$. If we can show that $G(\widehat{\mu}, \widehat{\sigma}) \to 1/2$, then it will follow that $\widehat{\mu} \to \mu_0$ and $\widehat{\sigma} \to \sigma_0$. Lemma <10> with $H_n = G$ will establish the almost sure version of this convergence if we check that

$$\sup_{\mu,\sigma} \frac{|G_n(\mu, \sigma) - G(\mu, \sigma)|}{G(\mu, \sigma)} \to 0 \qquad \text{almost surely.}$$

Such a weighted USLLN shows that the bad behaviour for σ near zero is unimportant compared to the divergent behaviour of $G(\mu, \sigma)$ as $\sigma \to 0$.

To simplify the notation, reparametrize by putting $\mu = \mu_0 + \sigma_0 t$ and $\sigma^2 = \sigma_0^2 s$, with $t \in \mathbb{R}$ and $s > 0$. Thus Reparametrization too

$$G(\mu, \sigma) = \frac{1}{2}\log s + \frac{1+t^2}{2s}.$$

Also, write η_i for the standardized, $N(0,1)$ random variable $(x_i - \mu_0)/\sigma_0$. Then we need to show

$$\sup_{s,t} \left| \frac{1}{n} \sum_{1 \le n} \frac{(\eta_i - t)^2/2s - (1+t^2)/2s}{(\log s)/2 + (1+t^2)/2s} \right| \to 0 \qquad \text{almost surely}$$

or

$$\sup_{s,t} \left| \frac{1}{n} \sum_{1 \le n} (\eta_i^2 - 1 - 2t\eta_i)/(s \log s + 1 + t^2) \right| \to 0 \qquad \text{almost surely.}$$

The $s \log s$ factor achieves its minimum value $-e^{-1}$ at e^{-1}. With $c = 1 - e^{-1}$, the last supremum is less than

$$\left| \frac{1}{n} \sum_{i \le n} (\eta^2 - 1) \right| + \sup_t \frac{|t|}{c + t^2} \left| \frac{1}{n} \sum_{1 \le n} \eta_i \right| ,$$

which tends to zero almost surely by virtue of two applications of the SLLN.

□

The last Example might seem like an awful lot of work for a modest application, but it does illustrate the effect of allowing weights into the uniform bounds.

chi.square <12>

Example. Method of minimum chi-square as in Pakes and Pollard (1989) perhaps? Maybe give some other example where the uniform convergence breaks down near the edge of the parameter space.

□

2.5 Z-estimators

Consistency::Zest

Give theorem and examples via first order conditions.

2.6 A nonparametric example

Consistency::monotone

Unedited This section treats the consistency problem for the maximum likelihood esimator (MLE) defined by a model whose parameter space Θ is much more complex than a subset of (finite dimensional) Euclidean space. Specifically, Θ will be the set of all densities with respect to Lebesgue measure on $\mathbb{R}^+$ that are monotone decreasing.

The model posits $X_1, X_2, \ldots$ to be independent observations from a probability distribution P defined by some unknown θ_0 in Θ. The MLE, $\widehat{\theta}_n$, is defined as

$$\widehat{\theta}_n(\omega) = \operatorname*{argmax}_{\theta\in\Theta} \prod\nolimits_{i\le n} \theta(X_i) = \operatorname*{argmax}_{\theta\in\Theta} \sum\nolimits_{i\le n} \log\theta(X_i).$$

The function $\theta \mapsto \sum_{i\le n} \log\theta(X_i)$ is usually called the ***log-likelihood***.

For each θ in Θ, the left and right limits, $\theta(t-)$ and $\theta(t+)$, must exist at each $t > 0$; the function θ must be continuous except for possibly countably many points of discontinuity. Those discontinuities have no effect on the corresponding probability measure on $\mathbb{R}^+$. We may therefore assume that each θ in Θ is left continuous everywhere.

> **Remark.** Left continuity actually ensures that the supremum in the definition of the maximum likelihood estimator is actually achieved (van der Vaart 1998, Section 24). This fact will have little effect on the explanations that follow, which are written to emphasize how much we can learn about an estimator from the mere fact that it maximizes some random criterion function.

Ideally, a consistency result should show, at least under each θ_0 in Θ, that

monotone.consistent<13>

$$\mathbb{P}_{n,\theta_0}\{d(\widehat{\theta}_n, \theta_0) \ge \epsilon\} \to 0 \qquad \text{as } n \to \infty$$

for some metric d on Θ. Unfortunately, I cannot establish the result for all θ_0, but only those that satisfy some extra assumptions. Such a defect is quite common in proofs with large parameter spaces. Ideally, we want at least to describe behavior of an estimator under every possible distribution prescribed by a model; in practice, only some subset of the possible distributions is tractable.

For nonparametric problems the choice of metric can also be more subtle than for parametric problems, where Euclidean distance usually recommends itself. Ideally, if consistency is just the first step in a more detailed analysis, closeness in d distance should simplify subsequent calculations based on local

approximations. Unfortunately, we sometimes have to settle for a weaker metric. For example, I would like to be able to establish some sort of uniform convergence of $\widehat{\theta}_n$ to θ_0, but instead have to settle for the $\mathcal{L}^1$ distance,

$$\|p - q\|_1 = \int_0^\infty |p(x) - q(x)|\, dx.$$

monotone.consistency<14>

Theorem. *For each bounded θ_0 in Θ with bounded support, if*

(i) the X_i's are independent observations on the probability distribution defined by θ_0,

(ii) $\int_0^\infty \theta_0(x) \log \theta_0(x)\, dx > -\infty$,

then $\|\widehat{\theta}_n - \theta_0\|_1 \to 0$ in probability.

Remark. Assumption (iii) is equivalent to integrability of $\theta_0 \log \theta_0$, because θ_0 is bounded. I write the condition as a one-sided bound to emphasize the similarity to Theorem <1>.

In fact, a much stronger result—convergence at an $O_p(n^{-2/3})$ rate in $\mathcal{L}^2$ norm—is possible (van der Vaart 1998, Theorem 24.6) under the same assumptions on θ_0 if we make explicit use of details about the form of $\widehat{\theta}_n$. However, as my current purpose is to illustrate what might be possible in cases where we do not know the explicit form of an estimator defined by an optimization, I feel the weaker result does have merit.

PROOF It will be notationally cleaner to express the argument in terms of the empirical measure P_n, which puts mass $1/n$ at each X_i, for $i = 1, \ldots, n$. With this notation, the log-likelihood equals $P_n \log \theta$.

The fact that $\widehat{\theta}_n$ maximizes the log-likelihood suggests that we try to apply one of the limit theorems from this Chapter to the process $\{-P_n \log \theta : \theta \in \Theta\}$. Unfortunately, I am unable to obtain uniform bounds for this process. Instead, consider the process

Whose idea?

$$G_n(\theta) = P_n^x g(x, \theta) \qquad \text{where } g(x, \theta) := -\log\left(\frac{\theta(x) + \theta_0(x)}{2}\right).$$

The MLE need not minimize G_n but, by concavity of the log function and the fact that $P_n \log \widehat{\theta}_n \geq P_n \log \theta_0$, it does satisfy the inequality

Gn<15>

$$G_n(\widehat{\theta}_n) \leq -P_n\left(\tfrac{1}{2}\log \widehat{\theta}_n + \tfrac{1}{2}\log \theta_0\right) \leq G_n(\theta_0).$$

Remark. To argue for consistency I will need some preliminary information about $\widehat{\theta}_n$, which might seem to contradict my stated purpose of arguing directly from the fact of optimization without specific knowledge of the form of the MLE. I prefer to think of it as analogous to needing a preliminary argument to force a parametric estimator into a compact subset. Some authors might interpret the preliminaries as a replacement of the MLE by a more tractable ***sieve estimator***.

Note that the MLE must be a step function with jumps at the order statistics $X_{(1)} < \cdots < X_{(n)}$ with $\widehat{\theta}_n(x) = 0$ for $x > X_{(n)}$. Otherwise the step function

$$\theta^*(x) = \{0 \le x \le X_{(1)}\}\widehat{\theta}_n(X_{(1)}) + \sum\nolimits_{2\le i\le n}\{X_{(i-1)} < x \le X_{(i)}\}\widehat{\theta}_n(X_{(i)})$$

would have the same log-likelihood as $\widehat{\theta}_n$ and $1 > \int_0^\infty \theta^*(x)\,dx$. For some $\delta > 0$, the function $(1+\delta)\theta^*$ would belong to Θ and have a larger log-likelihood than $\widehat{\theta}_n$. Consequently,

$$1 = \int_0^\infty \widehat{\theta}_n(x)\,dx \ge X_{(1)}\widehat{\theta}_n(0).$$

Under P, the random variable $1/X_{(1)}$ is of order $O_p(n)$ because

$$\mathbb{P}\{X_{(1)} > \epsilon/n\} = (1 - P[0,\epsilon/n])^n \ge (1-\theta_0(0)\epsilon/n)^n \to \exp(-\theta_0(0)\epsilon).$$

With high probability we still capture the MLE if we restrict the optimization to those θ for which $\theta(0)$ is bounded by a quantity, such as n^2, that grows more rapidly than n.

Similarly, if $L_n = \inf\{x : \theta_0(x) < n^{-2}\}$ then $P(L_n,\infty) = O(n^{-2})$ because θ_0 has bounded support, which implies

$$\mathbb{P}\{X_{(n)} > L_n\} \le \sum\nolimits_{i\le n}\mathbb{P}\{X_i > L_n\} = O(n^{-1}).$$

With high probability, $P_n(L_n,\infty) = 0$ and $\widehat{\theta}_n$ must therefore concentrate on $[0, L_n]$.

In short, with probability tending to one, $\widehat{\theta}_n$ must lie in the set

$$\Theta_n = \{\theta \in \Theta : \theta(0) \le n^2 \text{ and } \theta(x) = 0 \text{ for } x \le L_n\}.$$

Remark. By left continuity, $\theta_0(L_n) \ge n^{-2}$. We might also have $\widehat{\theta}_n(x) > 0$ for some $x > L_n$, which would imply $\theta_0 \notin \Theta_n$.

We can hope that $G_n(\theta)$ is close to $G(\theta) = P^x g(x,\theta)$, which will let us exploit a clean minimization at θ_0:

$$G(\theta) - G(\theta_0) = \int_0^\infty \theta_0(x) \log\left(\frac{2\theta_0(x)}{\theta(x)+\theta_0(x)}\right) = D(\theta_0 \| (\theta+\theta_0)/2),$$

the Kullback-Leibler distance between θ_0 and $(\theta+\theta_0)/2$. For general probability densities recall (Pollard 2001, Section 3.2) that the KL distance dominates both the squared Hellinger distance and the squared $\mathcal{L}^1$ distance,

$$D(p\|q) \geq \max\left(H^2(p,q), \tfrac{1}{2}\|p-q\|_1^2\right)$$

For not particularly compelling reasons, I prefer to work with the $\mathcal{L}^1$ metric via the lower bound

G.lower<16>
$$G(\theta) \geq \tfrac{1}{2}\left(\int |(\theta+\theta_0)/2 - \theta_0|\right)^2 = \tfrac{1}{8}\|\theta_0 - \theta\|_1^2 .$$

The insertion of the θ_0 into the definition of $g(x,\theta)$ and the restriction to Θ_n have some pleasant boundedness consequences, namely,

$$g(x,\theta) \leq -\log(\theta_0(x)/2) \leq -\log(1/2n^2)) = O(\log n) \qquad \text{for } x \leq L_n$$

and

$$-g(x,\theta) \leq \log(n^2 + \theta_0(0)) = O(\log n) \qquad \text{for each } \theta \text{ in } \Theta_n.$$

Thus there exists a sequence of constants M_n for which $|g(x,\theta)| \leq M_n = O(\log n)$ whenever $x \leq L_n$ and $\theta \in \Theta_n$.

When $|g(x,\theta)| \leq M_n$, the representation

$$M_n + g(x,\theta) = \int_0^{2M_n} \{M_n + g(x,\theta) \geq t\}\, dt$$

and the monotonicity of the map $x \mapsto (x,\theta)$ leads to a bound

$$\begin{aligned}
&\sup_{\theta\in\Theta_n} |(P_n - P)\{x \leq L_n\}(M_n + g(x,\theta))| \\
&\qquad \leq \int_0^{2M_n} \sup_{\theta\in\Theta_n, t\geq 0} |(P_n - P)\{x \leq L_n, M_n + g(x,\theta) \geq t\}|\, dt \\
&\qquad \leq 2M_n \sup_{I\in\mathcal{I}} |P_n I - PI|
\end{aligned}$$

where $\mathcal{I}$ denotes the set of all subintervals of the real line. Remember (SCME ?) that $\sup_{I\in\mathcal{I}} |P_n I - PI| = O_p(n^{-1/2})$.

SCME = first part of Stat 618 notes

For the contributions from $\{x > L_n\}$, use the definition of Θ_n and the fact that $\mathbb{P}\{P_n(L_n, \infty) = 0\} \to 1$ to get a bound that holds with probability tending to one,

$$\sup_{\theta\in\Theta_n} |(P_n - P)\{x > L_n\}(M_n + g(x,\theta))| \leq \int_{L_n}^{\infty} \theta_0(x)\,(M_n + |\log(\theta_0(x)/2)|)\,dx = O(M_n/n^2) + o(1),$$

the $o(1)$ term coming from the integrability of $\theta_0 \log(\theta_0)$

Remark. If we were prepared to make further assumptions about θ_0 the $o(1)$ term could be improved.

Combine the contributions from $x \leq L_n$ and $x > L_n$ to get

UWLLN<17> $$\Delta_n := \sup_{\theta\in\Theta_n} |G_n(\theta) - G(\theta)| = o_p(1).$$

The rest of the proof of consistency fits the pattern of Lemma <7>. With probability tending to one,

$$\begin{aligned} G(\widehat{\theta}_n) &\leq G_n(\widehat{\theta}_n) + \Delta_n \qquad \text{by } <17> \\ &\leq G_n(\theta_0) + \Delta_n \qquad \text{by } <15> \\ &\leq G(\theta_0) + o_p(1) + \Delta_n. \end{aligned}$$

Notice that we cannot bound $|G_n(\theta_0) - G(\theta_0)|$ by Δ_n, because θ_0 needn't belong to Θ_n; instead we can appeal to a SLLN for $P_n g(x, \theta_0)$. An appeal to the lower bound <16> completes the proof.

□

Problems

Unedited

[1] Suppose $\beta(\cdot)$ is a nonnegative function defined on a subset Θ_0^c of Θ such that for some integrable $h(\cdot)$ with $Ph > 0$, *Huber.caseA*

$$g(x,\theta) \geq \beta(\theta)h(x) \qquad \text{for all } \theta \in \Theta_0^c \text{ and all } x.$$

Define G_n as in Theorem <1>. Suppose θ_0 is a point in Θ_0 for which $\inf_{\theta\in\Theta_0^c} \beta(\theta)Ph > Pg(\cdot,\theta_0)$. If $G_n(\widehat{\theta}_n) \leq o(1) + \inf_\theta G_n(\theta)$ almost surely, show by the following steps that, with probability one, $\widehat{\theta}_n$ must eventually lie in Θ_0.

(i) With no loss of generality suppose $Ph = 1$. Write β for the infimum appearing in (ii) and ϵ_n for the $o(1)$ quantity identified by (iii). Choose $\epsilon > 0$ and γ such that

$$\beta(1-\epsilon) \geq \gamma > Pg(\cdot, \theta_0).$$

Use the SLLN to show that, with probability one, we eventually have

$$G_n(\theta) \geq \beta(\theta) \frac{1}{n} \sum_{i \leq n} h(\xi_i) \geq \beta(1-\epsilon) \qquad \text{for all } \theta \text{ in } \Theta_0^c.$$

(ii) Appeal to the SLLN again to show that, with probability one, we eventually have $G_n(\widehat{\theta}_n) \leq \epsilon_n + G_n(\theta_0) < \gamma$, which forces $\widehat{\theta}_n$ into Θ_0.

xxx [2] Suppose $\xi_1, \ldots, \xi_n$ are independent observations from the $\text{Bin}(1, \theta_0)$ distribution, for a θ_0 with $0 < \theta_0 < 1$. The maximum likelihood estimator $\widehat{\theta}_n$ for the $\text{Bin}(1, \theta)$ model minimizes

$$G_n(\theta) = \frac{X_n}{n} \log\left(\frac{\theta_0}{\theta}\right) + \left(1 - \frac{X_n}{n}\right) \log\left(\frac{1-\theta_0}{1-\theta}\right)$$

over $0 < \theta < 1$, where $X_n = \sum_{i \leq n} \xi_i$. Without using the explicit form for $\widehat{\theta}_n$, but appealing only to the general theorems in Chapter C, prove that $\widehat{\theta}_n = \theta_0 + O_p(1/\sqrt{n})$. [I am interested in seeing which of the general assumptions hold in the present special case, rather than in verifying the elementary result about $\widehat{\theta}_n$.]

wald2 [3] Let $\{g(\cdot, \theta) : \theta \in \Theta\}$ be a family of functions indexed by a compact metric space Θ. Suppose $\xi_1, \ldots, \xi_n$ are independent observations from a distribution P for which $P \sup_\theta |g(\cdot, \theta)| < \infty$. Suppose also that at each θ the function $g(x, \cdot)$ is continuous, for P almost all x. Adapt Wald's method to prove that

$$\sup_\theta \left| \frac{1}{n} \sum_{i \leq n} \big(g(\xi_i, \theta) - Pg(\cdot, \theta)\big) \right| \to 0 \qquad \text{almost surely.}$$

one.sided.ULLN [4] State and prove a useful one-sided analogue of Theorem <5>.

one.sided.weighted.ULLN [5] State and prove a useful one-sided analogue of Theorem <10>.

dist.fn [6] Let $\{F(x, \theta) : \theta \in \Theta\}$ be a family of distribution functions indexed by

a subset Θ of $\mathbb{R}$. Suppose $F(x,\theta)$ corresponds to the measure having density $f(x,\theta)$ with respect to a measure μ. Suppose there exists a neighborhood of a fixed θ_0 within which $\frac{\partial f}{\partial \theta}(x,\theta)$ exists, is continuous in θ, and is bounded in absolute value by some μ-integrable function $M(x)$. Suppose

$$2C_0 = \sup_x \left| \frac{\partial f}{\partial \theta}(t,\theta_0)\{t \le x\}\mu(dx) \right| > 0.$$

Show that

$$\sup_x |F(x,\theta_1) - F(x,\theta_2)| \ge C_0|\theta_1 - \theta_2|$$

for all θ_1, θ_2 in some neighborhood of θ_0. [Hint: Show that

$$F(x,\theta_1) - F(x,\theta_2) = (\theta_1 - \theta_2) \int \frac{\partial f}{\partial \theta}(t,\theta_0)\{t \le x\}\mu(dx) + o(|\theta_1 - \theta_2|),$$

by means of a dominated convergence argument.]

[7] Generalize the previous problem to Θ a subset of $\mathbb{R}^k$. *dist.fn2*

Notes

Discuss measurability issues.

Wald (1949) did not require a compact parameter space. Instead he made assumptions that, essentially, allowed an argument like the one in Theorem <1> to be applied to a compactification of the parameter space, in much the same way as for Example <2>.

Consistency: Zaman (1989) Perlman (1972) Jennrich (1969), Wu (1981)

Monotone density: van der Geer (2000), van der Vaart and Wellner (1996, Sections 3.2 and 3.4), (van der Vaart 1998, Section 24), Groeneboom (1985), Prakasa Rao (1969), Kim and Pollard (1990).

References

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume II*, pp. 539–555. Belmont, CA: Wadsworth Publishing Co.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In L. Le Cam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, pp. 221–233. University of California Press.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics 40*, 633–643.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1058.

Perlman, M. (1972). On the strong consistency of maximum likelihood estimators. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 1*, 236–281.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Prakasa Rao, B. (1969). Estimation of a unimodal density. *Sankhyā Ser. A 31*, 23–36.

van der Geer, S. A. (2000). *Applications of Empirical Process Theory*. Cambridge Unversity Press.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics 20*, 595–601.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics 9*, 501–513.

Zaman, A. (1989). Consistency via type 2 inequalities: A generalization of Wu's theorem. *Econometric Theory 5*, 272–286.