STAT6018 Research Frontiers in Data Science

Topic II: Introduction to empirical process theory

Yu Gu, PhD
Assistant Professor

Department of Statistics & Actuarial Science
The University of Hong Kong

Table of Contents

M-estimators

- *M-estimators* are (approximate) maximizers (or minimizers) $\hat{\theta}_n$ of criterion functions $\mathbb{M}_n(\theta)$, i.e., $\hat{\theta}_n = \arg\max \mathbb{M}_n(\theta)$.
- For i.i.d. observations, a common empirical criterion function is of the form $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$.
- Examples:
  - maximum likelihood estimators
  - least squares estimators
- Asymptotic properties of $\hat{\theta}_n$:
  - consistency for the true parameter θ_0
  - rate of convergence r_n
  - weak convergence of $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$ to some random point $\hat{h}$

Table of Contents

Preliminary arguments

- If the argmax function were continuous w.r.t. some metric on the space of criterion functions, then weak convergence of $\mathbb{M}_n(\theta)$ would imply weak convergence of $\hat{\theta}_n$ by the continuous mapping theorem.
- Let $\{\mathbb{M}(\theta) : \theta \in \Theta\}$ be the limiting process of $\mathbb{M}_n(\theta)$.
- The argmax function is continuous at $\mathbb{M}$ if $\mathbb{M}$ has a unique, well-separated maximizer $\hat{h}$. That is, $\mathbb{M}(\hat{h}) > \sup_{h \notin G} \mathbb{M}(h)$ almost surely for any neighborhood G of $\hat{h}$.

Preliminary result

Lemma 1

Let $\mathbb{M}_n$, $\mathbb{M}$ be stochastic processes indexed by a metric space H. Let A and B be arbitrary subsets of H. Suppose that

(i) $\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in A} \mathbb{M}(h)$ *almost surely, for every open set G that contains $\hat{h}$.*

(ii) $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)$.

(iii) $\mathbb{M}_n \xrightarrow{d} \mathbb{M}$ *in* $\ell^\infty(A \cup B)$.

Then, for every closed set F,

$$\limsup_{n \to \infty} P^*(\hat{h}_n \in F \cap A) \leq P(\hat{h} \in F \cup B^c).$$

- $A = B = H \Rightarrow \hat{h}_n \xrightarrow{d} \hat{h}$ (by portmanteau theorem[1]).
- See Lemma 3.2.1 of VW for the proof.

[1] $X_n \xrightarrow{d} X$ *if and only if* $\limsup_{n \to \infty} P^*(X_n \in F) \leq P(X \in F)$ *for every closed F.*

Remarks

- The assumption that $\mathbb{M}_n \xrightarrow{d} \mathbb{M}$ uniformly in the whole parameter space is too strong.
- If dropping this assumption, additional properties of $\hat{h}_n$ need to be established in order to obtain $\hat{h}_n \xrightarrow{d} \hat{h}$.
- The Argmax theorem requires uniform tightness[2] of $\hat{h}_n$ and uniform convergence of $\mathbb{M}_n$ on compact subspace.

[2] $\forall \epsilon > 0$, $\exists$ *a compact set* $V_\epsilon \in H$ *s.t.* $P(\hat{h}_n \in V_\epsilon) > 1 - \epsilon$.

Argmax theorem

Theorem 2 (Argmax theorem)

Let $\mathbb{M}_n$, $\mathbb{M}$ be stochastic processes indexed by a metric space H. Suppose that

(i) *Almost all sample paths $h \mapsto \mathbb{M}(h)$ are upper semicontinuous*[a] *and possess a unique maximum at a (random) point $\hat{h}$, which as a random map in H is tight.*

(ii) *The sequence $\hat{h}_n$ is uniformly tight and satisfies $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_p(1)$.*

(iii) *$\mathbb{M}_n \xrightarrow{d} \mathbb{M}$ in $\ell^\infty(K)$ for every compact $K \subset H$.*

Then $\hat{h}_n \xrightarrow{d} \hat{h}$ in H.

[a] *A function $f : \mathbb{D} \mapsto \mathbb{R}$ is upper semicontinuous if for all $x_0 \in \mathbb{D}$, $\limsup_{x \to x_0} f(x) \leq f(x_0)$.*

See Theorem 3.2.2 of VW for the proof.

Remarks

- The preceding lemma and the Argmax theorem are typically applied to a local parameter h, but they can also be applied to the original parameter θ.

- Since the limiting criterion function $\mathbb{M}(\theta)$ is typically nonrandom, the approach turns into a consistency proof.

Consistency

Corollary 3 (Consistency)

Let $\mathbb{M}_n$ *be stochastic processes indexed by a metric space* Θ*, and let* $\mathbb{M} : \Theta \mapsto \mathbb{R}$ *be a deterministic function.*

(A) *Suppose that*

- (i) $\mathbb{M}(\theta_0) > \sup_{\theta \notin G} \mathbb{M}(\theta)$ *for every open set* G *that contains* θ_0*.*
- (ii) $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_p(1)$*.*
- (iii) $\|\mathbb{M}_n - \mathbb{M}\|_\Theta \to 0$ *in outer probability.*

Then $\hat{\theta}_n \to \theta_0$ *in outer probability.*

(B) *Suppose that*

- (i) *The map* $\theta \mapsto \mathbb{M}(\theta)$ *is upper semicontinuous with a unique maximum at* θ_0*.*
- (ii) *The sequence* $\hat{\theta}_n$ *is uniformly tight and satisfies* $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_p(1)$*.*
- (iii) $\|\mathbb{M}_n - \mathbb{M}\|_K \to 0$ *in outer probability for every compact* $K \subset \Theta$*.*

Then $\hat{\theta}_n \to \theta_0$ *in outer probability.*

Under i.i.d. setting

In the case of i.i.d. data, $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M} = \mathbb{P} m_\theta$, the uniform convergence in (iii) is valid if and only if the class of functions $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli.

Table of Contents

Preliminary arguments

- If $\mathbb{M}(\theta)$ is twice differentiable at a point of maximum θ_0, then $\mathbb{M}'(\theta_0) = 0$ and $\mathbb{M}''(\theta_0)$ is negative definite.

- It is natural to assume that $\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0)$ for every θ in a neighborhood of θ_0.

- The *modulus of continuity* of a stochastic process $\{X(t) : t \in T\}$ is defined by

$$m_X(\delta) := \sup_{s,t \in T : d(s,t) \leq \delta} |X(s) - X(t)|.$$

  An upper bound for the rate of convergence of $\hat{\theta}_n$ can be obtained from the modulus of continuity of $\mathbb{M}_n - \mathbb{M}$ at θ_0.

Rate of convergence

Theorem 4 (Rate of convergence)

Let $\mathbb{M}_n$ be stochastic processes indexed by a semimetric space Θ and $\mathbb{M} : \Theta \to \mathbb{R}$ a deterministic function. Suppose that

(i) *For every θ in a neighborhood of θ_0,*

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0).$$

(ii) *For every n and sufficiently small δ, the centered process $\mathbb{M}_n - \mathbb{M}$ satisfies*

$$E^* \sup_{d(\theta,\theta_0)<\delta} \left| (\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0) \right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on n.

(iii) *The sequence $\hat{\theta}_n$ converges in outer probability to θ_0 and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(r_n^{-2})$ for some sequence r_n such that*

$$r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n} \quad \text{for every } n.$$

Then $r_n d(\hat{\theta}_n, \theta_0) = O_p^*(1)$. If the displayed conditions are valid for every θ and δ, then the condition that $\hat{\theta}_n$ is consistent is unnecessary.

See Theorem 3.2.5 of VW for the proof.

Remarks

- The theorem remains true if replacing the metric function d by an arbitrary function $\tilde{d} : \Theta \times \Theta \mapsto [0, \infty)$ that satisfies $\tilde{d}(\theta_n, \theta_0) \to 0$ whenever $d(\theta_n, \theta_0) \to 0$.
- When $\phi(\delta) = \delta^\alpha$, the rate r_n is at least $n^{1/(4-2\alpha)}$.
- In particular, the "usual" rate $\sqrt{n}$ corresponds to $\phi(\delta) = \delta$.

Under i.i.d. setting

- Recall Condition (ii) in the preceding theorem:

$$E^* \sup_{d(\theta,\theta_0)<\delta} \left| (\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0) \right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}$$

- For i.i.d. data and empirical criterion functions $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M}(\theta) = P m_\theta$, Condition (ii) involves the suprema of the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ indexed by classes of functions

$$\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}.$$

- It is reasonable to assume that these suprema are bounded uniformly in n.

Rate of convergence under i.i.d. setting

Corollary 5

In the i.i.d. case, assume that

(i) *For every* θ *in a neighborhood of* θ_0,

$$P\left(m_\theta - m_{\theta_0}\right) \lesssim -d^2(\theta, \theta_0).$$

(ii) *There exists a function* ϕ *such that* $\delta \mapsto \phi(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$ *and, for every* n,

$$E^* \left\|\mathbb{G}_n\right\|_{\mathcal{M}_\delta} \lesssim \phi(\delta).$$

(iii) *The sequence* $\hat{\theta}_n$ *converges in outer probability to* θ_0 *and satisfies* $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \mathbb{P}_n m_\theta - O_p(r_n^{-2})$ *for some sequence* r_n *such that*

$$r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n} \quad \text{for every } n.$$

Then $r_n d(\hat{\theta}_n, \theta_0) = O_p^*(1)$.

Bounds on continuity modulus

- It is important to derive a sharp bound on the modulus of continuity of $\mathbb{G}_n$ before applying the corollary.
- A simple but not necessarily efficient approach is to apply the maximal inequalities to the class $\mathcal{M}_\delta$, which yield

$$E_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J(1, \mathcal{M}_\delta)(P^* M_\delta^2)^{1/2},$$
$$E_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J_{[]}\big(1, \mathcal{M}_\delta, L_2(P)\big)(P^* M_\delta^2)^{1/2}.$$

- These bounds depend mostly on the envelope function M_δ.
- Assuming that the entropy integrals are bounded as $\delta \downarrow 0$, we obtain an upper bound $\phi(\delta) = (P^* M_\delta^2)^{1/2}$ on the modulus.
- By the preceding corollary, r_n is at least the solution of

$$r_n^4 P^* M_{1/r_n}^2 \sim n.$$