Applications of Uniform Central Limit Theorems and Laws of Large Numbers

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Goodness of fit tests
- Convergence of M-estimators
  - rates of convergence
  - non-smooth losses and CLT-type expansions

Reading: van der Vaart, *Asymptotic Statistics*, Chapter 19.3–19.6, Chapter 5.8

Refined continuous mapping theorem

- Metric spaces $\mathbb{D}_n \subset \mathbb{D}$ and $\mathbb{E}$
- Sequence of functions $g_n : \mathbb{D}_n \to \mathbb{E}$,
- Continuous-ish: for some $g : \mathbb{D}_0 \to \mathbb{E}$, if $x_n \in \mathbb{D}_n$ has subsequence $x_{n(m)} \to x \in \mathbb{D}_0 \subset \mathbb{D}$, then

$$g_{n(m)}(x_{n(m)}) \to g(x)$$

Theorem (18.11 in van der Vaart)

If $X_n \in \mathbb{D}_n$ and $X \in \mathbb{D}$ are random elements and $X \in \mathbb{D}_0$ with probability 1,

(i) *If $X_n \stackrel{d}{\to} X$, then $g_n(X_n) \stackrel{d}{\to} g(X)$*

(ii) *If $X_n \stackrel{p}{\to} X$, then $g_n(X_n) \stackrel{p}{\to} g(X)$*

(iii) *If $X_n \stackrel{a.s.}{\to} X$, then $g_n(X_n) \stackrel{a.s.}{\to} g(X)$*

Basic approach

- have empirical process $\mathbb{G}_n = \sqrt{n}(P_n - P)$ in $L^\infty(T)$
- if it's Donsker, i.e. $\mathbb{G}_n \xrightarrow{d} \mathbb{G}$ in $L^\infty(T)$, then

$$\phi(\mathbb{G}_n) \xrightarrow{d} \phi(\mathbb{G})$$

whenever ϕ is continuous for $L^\infty(T)$

Goodness of fit tests

- null $H_0 : X \sim P$ with cdf F
- would like a test of H_0

Two statistics:

$$\begin{aligned} \sqrt{n}\,\|F_n - F\|_\infty & \quad \text{Kolmogorov-Smirnov} \\ n\int (F_n - F)^2 dF & \quad \text{Cramér-von Mises} \end{aligned}$$

- both F_n and F belong to *càdlàg* functions (continuous from right, limits from the left)
- space $D[a, b] =$ càdlàg on $[a, b]$

Kolmogorov-Smirnov Statistics

Corollary

For $K_n = \sqrt{n}\,\|F_n - F\|_\infty$,

$$K_n \xrightarrow{d} \|\mathbb{G}_F\|_\infty \quad \textit{under } H_0 : X_i \overset{\text{iid}}{\sim} F,$$

where $\mathrm{Cov}(\mathbb{G}_F(t), \mathbb{G}_F(s)) = F(s \wedge t) - F(s)F(t)$, *and* $\|\mathbb{G}_F\|_\infty$ *has identical distribution for all continuous* F

Cramér-von Mises Statistics

Corollary

For $C_n := n \int (F_n - F)^2 dF$,

$$C_n \xrightarrow{d} \int \mathbb{G}_F^2 dF \quad \textit{under } H_0 : X_i \overset{\text{iid}}{\sim} F,$$

where $\text{Cov}(\mathbb{G}_F(t), \mathbb{G}_F(s)) = F(s \wedge t) - F(s)F(t)$, *and* $\int \mathbb{G}_F^2 dF$ *has identical distribution for all continuous* F

An approach to multiple hypothesis testing

- nulls $H_{0,n} : X_i \sim P_i$ where P_i has continuous cdf G_i
- statistic $U_i = G_i(X_i) \overset{\text{iid}}{\sim} \text{Uni}[0,1]$ (p-value) under $H_{0,n}$
- $F_n = \frac{1}{n} \sum_{i=1}^n 1\{U_i \le t\}$

$$A_n := \sup_t \sqrt{n}(F_n(t) - t)w(t) \quad \text{Anderson-Darling}$$

Corollary

Under $H_{0,n}$, weighted process $\mathbb{G}_n^w = [\sqrt{n}(F_n - t)w(t)]_{t \in [0,1]}$ has

$$\mathbb{G}_n^w \xrightarrow{d} \mathbb{G}^w \quad \text{in } L^\infty([0,1])$$

whenever $\int_0^1 w^2(t)dt < \infty$.

Proof of convergence for Anderson-Darling statistics

- weighted class $\mathcal{F} \cdot w = \{fw : f \in \mathcal{F}\}$ is VC if $\mathcal{F}$ is VC-subgraph (generally true)

- envelope function $F(t) = w(t)$ for entire class $\mathcal{F}_{\text{indicators}} = \{f(x) = 1\{x \le t\}\}_{t \in [0,1]}$

Convergence of M-estimators

recall M-estimators:

- loss function $\ell_\theta(x)$ in θ
- sample and population losses $L(\theta) = P\ell_\theta(X)$ and $L_n(\theta) = P_n\ell_\theta(X)$
- M-estimator
$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} L_n(\theta)$$
- global minimizer $\theta_0 = \operatorname{argmin}_{\theta \in \Theta} L(\theta)$

idea: to get rate of convergence, argue that growth $L(\theta) - L(\theta_0)$ dominates noise $L_n(\theta) - L_n(\theta_0)$

The picture in the "standard" case

1. demonstrate population growth $L(\theta) - L(\theta_0) \geq \|\theta - \theta_0\|^2$
2. central limit behavior for localized process

$$|(L_n(\theta) - L_n(\theta_0)) - (L(\theta) - L(\theta_0))| = O_P(1)\frac{\|\theta - \theta_0\|}{\sqrt{n}}$$

3. critical radius

$$\frac{\|\theta - \theta_0\|}{\sqrt{n}} = \|\theta - \theta_0\|^2 \quad \text{i.e.} \quad \|\theta - \theta_0\| = \frac{1}{\sqrt{n}}.$$

Rates of convergence

- distance-like function $d : \Theta \times \Theta \to \mathbb{R}_+$
- population growth $L(\theta) - L(\theta_0) \geq \lambda d(\theta, \theta_0)^\beta$ near θ_0, i.e. for growth function $g(\delta) = \lambda\delta^\beta$, in a neighborhood of θ_0,

$$L(\theta) \geq L(\theta_0) + g(\delta) \quad \text{if } d(\theta, \theta_0) \geq \delta$$

- stochastic modulus $\omega(\delta) = c\delta^\alpha$, some $0 \leq \alpha < \beta$

$$\mathbb{E}\left[\sup_{\theta : d(\theta,\theta_0) \leq \delta} |\mathbb{G}_n(\ell_\theta - \ell_{\theta_0})|\right] \leq \omega(\delta)$$

Theorem

Let the rate $r_n > 0$ satisfy the critical radius condition $\frac{\omega(r_n)}{\sqrt{n}} \leq g(r_n)$. If $\widehat{\theta}_n \overset{p}{\to} \theta_0$, then $d(\widehat{\theta}_n, \theta_0) = O_P(r_n)$.

Rates of convergence: proof by *peeling*

- let $\epsilon > 0$, choose η such that $P(d(\widehat{\theta}_n, \theta_0) \geq \eta) \leq \epsilon$
- construct shells $S_{j,n} = \{\theta \in \Theta, r_n 2^{j-1} \leq d(\theta, \theta_0) \leq 2^j r_n\}$
- probability of individual shells is small:

M-estimators with non-smooth losses

some losses $\ell(\theta, x)$ we like, population loss $L(\theta) = P\ell(\theta, X)$

- $\ell(\theta, x) = |\theta - x|$ has $L(\theta)$, minimized by $\mathrm{med}(X)$
- $\ell_\alpha(\theta, x) = (1 - \alpha)(\theta - x)_+ + \alpha(x - \theta)_+$, L minimized by

$$\mathsf{Q}_P(\alpha) := \inf\{\theta \in \mathbb{R} \mid \alpha \le P(X \le \theta)\}$$

Stochastic Taylor approximations

using shorthand $\ell_\theta = \ell(\theta, \cdot)$, assume in a neighborhood of θ_0:

- Lipschitz condition

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \leq M(x) \, \|\theta_1 - \theta_2\|$$

- differentiability (in probability): $\theta \mapsto \ell_\theta(x)$ has gradient $\dot{\ell}_{\theta_0}$ at θ_0 with P-probability 1

Lemma (19.31 in van der Vaart)

If $r_n \uparrow \infty$ and $PM^2 < \infty$, then

$$\sup_{\|h\| \leq 1} \mathbb{G}_n \left(r_n \left(\ell_{\theta_0 + \frac{h}{r_n}} - \ell_{\theta_0} \right) - h^\top \dot{\ell}_{\theta_0} \right) \xrightarrow{p} 0.$$

Proof of stochastic Taylor approximation

- Finite dimensional convergence to 0

- Tightness (asymptotic stochastic equicontinuity)

Convergence of M-estimators

same conditions as lemma, and

- $L(\theta) = P\ell_\theta(X)$ is twice differentiable at $\theta_0 = \operatorname{argmin}_\theta L(\theta)$, with positive definite Hessian

$$\nabla^2 L(\theta_0) \succ 0$$

Theorem (5.23 in van der Vaart)

Assume $\widehat{\theta}_n \xrightarrow{p} \theta_0$ and $L_n(\widehat{\theta}_n) \le \inf_\theta L_n(\theta) + o_P(1/n)$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\nabla^2 L(\theta_0)^{-1} \cdot \sqrt{n} P_n \dot{\ell}_{\theta_0} + o_P(1)$$

Proof of convergence

- for any $h_n = O_P(1)$, we have

$$n(P_n \ell_{\theta_0 + h_n/\sqrt{n}} - P_n \ell_{\theta_0}) = \frac{1}{2} h_n^\top \nabla^2 L(\theta_0) h_n + h_n^\top \mathbb{G}_n \dot{\ell}_{\theta_0} + o_P(1)$$

- expand using $\widehat{h}_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ and $\widetilde{h}_n = -\nabla^2 L(\theta_0)^{-1} \mathbb{G}_n \dot{\ell}_{\theta_0}$

Example: quantile estimation

- CDF $F(t) := P(X \leq t)$ has density $f(\theta_0)$ at θ_0
- loss function $\ell_\theta(x) = (1-\alpha)(\theta - x)_+ + \alpha(x-\theta)_+$
- $P(X \leq \theta_0) = \alpha$

Corollary (Asymptotic linearity of quantile estimator)

The empirical minimizer $\widehat{\theta}_n = \operatorname{argmin} L_n(\theta)$ *satisfies*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{1}{f(\theta_0)} \cdot \sqrt{n}\left[(1-\alpha)P_n(X_i \leq \theta_0) - \alpha P_n(X_i \geq \theta_0)\right] + o_P(1)$$