Chapter 3. M-estimator based on Le Cam

Setup: $X_1, \dots, X_n \overset{iid}{\sim} P_{\theta_0}$ $(=: P_0)$

θ_0: the true DG Parameter.

$\in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$

$\Phi : \Theta \longmapsto \mathbb{R}^k$

Goal of interest: $\phi_0 = \Phi(\theta_0)$ ← e.g.: population mean

M-estimator Framework

① $\mathrm{Im}(\Phi) \subset S$

↑ image

② Restriction on Φ:

$$\forall \theta \in \Theta,\ \Phi(\theta) \in \underset{\phi \in S}{\arg\max}\ \mathbb{E}_\theta[m_\phi(X)] \qquad X \sim P_\theta$$

where m_ϕ can be understood as

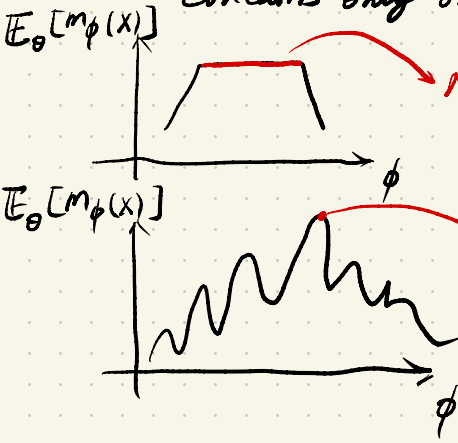
(a) negative of a loss function;

(b) pseudo likelihood function;

M-estimation inference

① Identification

"m_ϕ" identifies $\Phi(\theta)$ iff the argmax contains only one element for each $\theta \in \Theta$.



(partial identification) NOT IN S&I.

② Regarding estimation:

Recall $\Phi(\theta_0) = \underset{\phi \in S}{\operatorname{argmax}}\, \mathbb{E}_{\theta_0}[m_\phi(X)]$

$\hat{\phi}_n \in \underset{\phi}{\operatorname{argmax}}\, \hat{\mathbb{E}}[m_\phi(X)]$

↑ m-estimator of $\Phi(\theta_0)$

$= \underset{\phi}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} m_\phi(X_i)$

Notation simplification

$$Pf := \int f(x) P(dx) \quad (\mathbb{E}_{X \sim P})$$

P_n : the empirical probability measure

e.g. $Pm_\phi = \int m_\phi(x) P(dx)$

$P_\theta m_\phi = \int m_\phi(x) P_{\theta_0}(dx)$

$P_n m_\phi = \frac{1}{n} \sum_{i=1}^{n} m_\phi(X_i)$

$\phi_0 \in \arg\max_\phi P_\theta m_\phi$

$\phi_n \in \arg\max_\phi P_n m_\phi$

CLT: $\sqrt{n}(P_n - P_\theta) f \Rightarrow N(0, \Sigma_f)$

Example (MLE as an M-est.)

KL-divergence: Given any P, Q, assuming

$P \ll \mu$ and $Q \ll \mu$ → Lesbeague / Counting measure usually.

($\Leftrightarrow p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$)

$D_{KL}(P \| Q) := -P\left[\log \frac{q}{p}\right]$ → m_θ

$\theta_0 \in \arg\max_{\theta \in \Theta} \left[-D_{KL}(P_{\theta_0} \| P_\theta)\right]$

$\theta_n \in \operatorname{argmax} P_n m_\theta$ → $\log \frac{Q}{P}$

We can show, by Jensen ≤ (Gibbs ineq) that

(1) $D_{KL}(P\|Q) \geq 0$

(2) $D_{KL}(P\|Q) = 0$ iff $P = Q$.

$\Downarrow$

① $\theta_0 \in \underset{\theta\in\Theta}{\arg\max} \left[-D_{KL}(P_{\theta_0}\|P_\theta)\right]$

$= \underset{\theta\in\Theta}{\operatorname{argmax}} P_{\theta_0}\left[\log \frac{P_\theta}{P_{\theta_0}}\right]$

$= \operatorname{argmax} P_{\theta_0}[\log P_\theta]$

if $P_0|\log P_{\theta_0}| < \infty$; $\log P_\theta = m_\phi$ with $\Phi: \theta \mapsto \theta$.

$= \underset{\theta\in\Theta}{\arg\max}\ P_{\theta_0} m_\phi$

② Identifiability: $m_\phi(m_\theta)$ identifies the M-est if

$P_{\theta_1} = P_{\theta_2}$ iff $\theta_1 = \theta_2$.

③ $\theta_n \in \operatorname{argmax} P_n m_\theta$

$= \underset{\theta\in\Theta}{\operatorname{argmax}} P_n[\log P_\theta]$

$= \underset{\theta\in\Theta}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(X_i)$

$= \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{i=1}^{n} P_\theta(X_i)$ classic def. of MLE

Z-est (a cousin of M-est)

Setup: the same as M-est

framework: $Z_\phi : x \longmapsto \mathbb{R}^b$

$\Phi(\theta)$ is the root of

$$\mathbb{E}_\theta[Z_\phi(X)]$$

i.e. $\Phi(\theta) \in \{\phi : \mathbb{E}_\theta[Z_\phi(X)] = 0\}$

Inference:

① Z_ϕ identifies $\Phi(\theta)$ if it has a unique root.

② $\phi_n \in \{\phi : \underbrace{P_n Z_\phi}_{\frac{1}{n}\sum_{i=1}^n Z_\phi(X_i)} = 0\}$

Ex1 [Sample median]

if $\Phi(\theta) := \text{median}_\theta(X)$ is unique

then $0 \overset{\text{by def. of median}}{=} P_\theta[\mathbb{1}\{X > \Phi(\theta)\}] - P_\theta[\mathbb{1}\{X < \Phi(\theta)\}]$

$= P_\theta[\underbrace{\text{sign}(X - \Phi(\theta))}_{Z_\phi := \text{sign}(x-\phi)}]$

$\Downarrow$

$$\phi_n \in \{\phi : P_n Z_\phi \left(= \frac{1}{n}\sum_{i=1}^n \operatorname{sign}(X_i - \phi)\right) = 0\}$$

matches the sample median

Ex. [sample mean] m-est: $m_\phi = -|X-\phi|^2$

Ex. [sample median]
also an M-est: $m_\phi := -|X-\phi|$.

Ex. [Pearson method of moments]
Suppose $\Theta = \mathbb{R}^k$ and $\Phi : \Theta \to \Theta$, i.e.
we are interested in estimating θ_0.
Suppose $X \sim P_{\theta_0}$ and the first k moments

$$\begin{pmatrix} \mathbb{E}_\theta[X^1] \\ \vdots \\ \mathbb{E}_\theta[X^k] \end{pmatrix}$$

are well-defined and finite

Then $P_0 X \left(= \mathbb{E}_{\theta_0}[X]\right) \approx \frac{1}{n}\sum_{i=1}^n X_i$

$P_0 X^2 \approx \frac{1}{n}\sum_{i=1}^n X_i^2$

$\vdots$

$P_0 X^k \approx \frac{1}{n}\sum_{i=1}^n X_i^k$

Then the MoM est. is the root of

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_\theta[X] \\ \frac{1}{n}\sum_{i=1}^n X_i^2 - \mathbb{E}_\theta[X^2] \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^k - \mathbb{E}_\theta[X^k] \end{pmatrix}$$

the blue parts are known functions of θ.

Then, MoM ests are Z ests:

$$Z_\theta : x \longmapsto \begin{pmatrix} x - \mathbb{E}_\theta[X] \\ x^2 - \mathbb{E}_\theta[X^2] \\ \vdots \\ x^k - \mathbb{E}_\theta[X^k] \end{pmatrix}$$

← plz verify it

Ex. [M-ests, but more general (U-stat)] ← 583

PLM: $Y_i = X_i^\top \beta_0 + g(U_i) + \varepsilon_i$

- Y_i: outcome
- $X_i^\top \beta_0$: linear
- $g(U_i)$: nonlinear
- ε_i: noise
- X_i, U_i: covariates
- β_0, g: unknown params.

a benchmark "semi-parametric" model since

- $\beta_0 \in \mathbb{R}^d$ is finite-dim
- g is infinite-dim

Goal of interest: $\Phi(\theta) = \beta_0$, where $\theta = (\beta_0, g)$

Difference-based estimator:

$$Y_i = X_i^\top \beta_0 + g(U_i) + \varepsilon_i$$
$$Y_j = X_j^\top \beta_0 + g(U_j) + \varepsilon_j$$

Pairwise diff

$$Y_i - Y_j = (X_i - X_j)^\top \beta_0 + [g(U_i) - g(U_j)] + \varepsilon_i - \varepsilon_j$$

Then, if $U_i - U_j \approx 0$ and g is smooth, then

$$[g(U_i) - g(U_j)] \approx 0$$

$$\| $$

$$Y_i - Y_j \approx (X_i - X_j)^\top \beta_0 + \varepsilon_i - \varepsilon_j$$

$$\| $$

$$\underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \sum_{i<j} \left[K_h(U_i - U_j) \left(Y_i - Y_j - (X_i - X_j)^\top \beta \right)^2 \right] \right\}$$

$K_h(U_i - U_j)$: encouraging closeness of U_i and U_j

$\left(Y_i - Y_j - (X_i - X_j)^\top \beta \right)^2$: LSE loss

$\sum_{i<j} \left[K_h(U_i - U_j) \left(Y_i - Y_j - (X_i - X_j)^\top \beta \right)^2 \right]$: U-statistics for any β

$\left\{ \beta : \sum_{i<j} K_h(\cdot) \cdots \right\}$ is a U-process indexed by β.

Full generalization of M- and Z-

M-est: $\Phi(\theta) \in \underset{\phi \in S}{\operatorname{argmax}} M_\theta(\phi)$ ← population

special case: $P_\theta m_\phi$

$\phi_n \in \underset{\phi \in S}{\operatorname{argmax}} M_n(\phi)$ ← data-based

special case: $P_n m_\phi$

We believe as long as $M_n(\cdot)$ is close to $M_{\theta_0}(\cdot)$, then their maximizer ϕ_n should be close to ϕ_0 $(:= \Phi(\theta_0))$

Z-est ① $\phi_0 \in \{\text{roots of } Z_{\theta_0}(\phi).\}$

② $\phi_n \in \{\text{roots of } Z_n(\phi)\}$

Similarly, as long as $Z_{\theta_0}(\cdot)$ is close to $Z_n(\cdot)$, ϕ_n should be close to ϕ_0.

Remark:

(i) Usually, M-est. can be also formulated as an Z-est.

$$Z_\theta : \phi \longmapsto \nabla M_\theta(\phi)$$

$$Z_n : \phi \longmapsto \nabla M_n(\phi)$$

But not always true.

Manski's estimator:

Binary choice model — goal of interest

$$Y_i = \mathbb{1}(X_i^\top \beta_0 + \varepsilon_i > 0)$$

Y_i: outcome; X_i: covariates; ε_i: noise

$$\hat{\beta}_n := \underset{\beta \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ Y_i \mathbb{1}\{X_i^\top \beta > 0\} \right\}$$

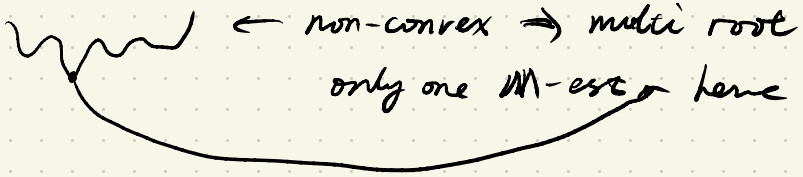
It doesn't have a natural Z-estimator.

(ii) Z-est. can always be formulated as an M-est:

$$M_\theta : \phi \longmapsto \|Z_\theta(\phi)\|^2$$

$$M_n : \phi \longmapsto \|Z_n(\phi)\|^2$$

Intuition: the global max is easier to be unique than the root.

Ex. 

Q: 1. consistency? 2. Rate of convergence?
3. Distr.?

1st Q: Is the ϕ_n above a consistent est. of ϕ_0??
i.e., $\phi_n \xrightarrow{P} \phi_0$?

Example [1-dim Z-est.]

$\Phi : \Theta \mapsto \mathbb{R}$ s.t. $\phi_0 \in \mathbb{R}^1$.

ϕ_0 = root of $Z_0(\phi)$ ← population score function

ϕ_n = root of $Z_n(\phi)$ empirical score func.

Conditions:

(i) $\phi_0, \phi_n \in \mathbb{R}^1$

(ii) [Point-wise consistency]: $\forall$ given ϕ,

$$Z_n(\phi) \xrightarrow{P} Z_0(\phi)$$

(iii) One of the following two is true

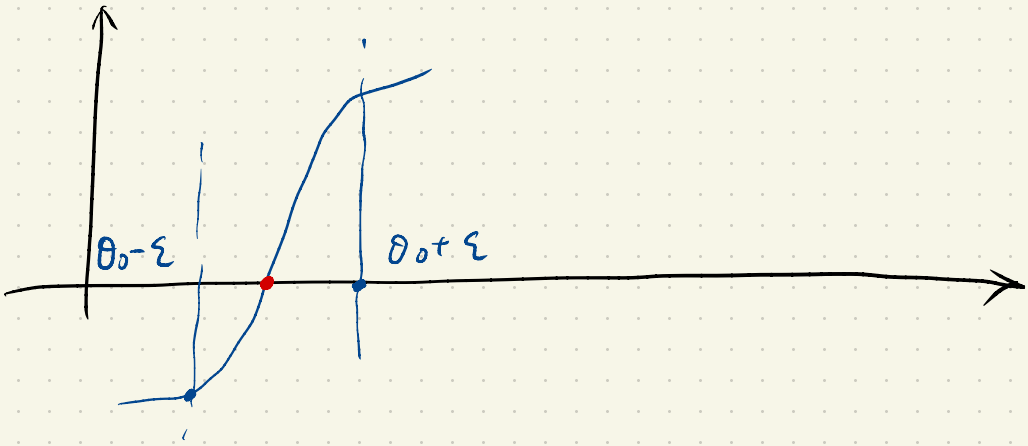
(a) Each $\phi \mapsto Z_n(\phi)$ is cont. and has exactly one root

or (b) $\phi \mapsto Z_n(\phi)$ is non-decreasing.

(iv) $\forall \varepsilon > 0$, $Z_0(\phi_0 - \varepsilon) < 0 < Z_0(\phi_0 + \varepsilon)$

Proof. We only show

(i) + (ii) + (iii) a + (iv)



Step 1: $P(Z_n(\phi_0-\varepsilon)<0,\ Z_n(\phi_0+\varepsilon)>0)$

$\leq P(\exists$ a root of Z_n between $\phi_0-\varepsilon$ and $\phi_0+\varepsilon)$

$= P(\phi_n \in (\phi_0-\varepsilon,\ \phi_n+\varepsilon))$

On the other hand, we have known

(iv): $Z_0(\phi_0-\varepsilon)<0$, $Z_0(\phi_0+\varepsilon)>0$

(ii): $Z_n(\phi_0-\varepsilon) \xrightarrow{P} Z_0(\phi_0-\varepsilon)$

$Z_n(\phi_0+\varepsilon) \xrightarrow{P} Z_0(\phi_0+\varepsilon)$

$\Downarrow$

$P(Z_n(\phi_0-\varepsilon)<0,\ Z_n(\phi_0+\varepsilon)>0) \xrightarrow{P} 1$

$\Downarrow$

$\phi_n \xrightarrow{P} \phi_0$

Remark: This proof doesn't work for general $\mathbb{R}^k$ param.

Real meat: $\phi_0 \in \operatorname{argmax} M_{\infty}(\phi)$

$\phi_n \in \operatorname{argmax} M_n(\phi)$ (*)

Goal: $\phi_n \xrightarrow{P} \phi_0$

Remark: (i) Normally, we can relax (*) to allow for computation error.

$$M_n(\phi_n) \geq \sup_{\phi} M_n(\phi) - o_p(1)$$

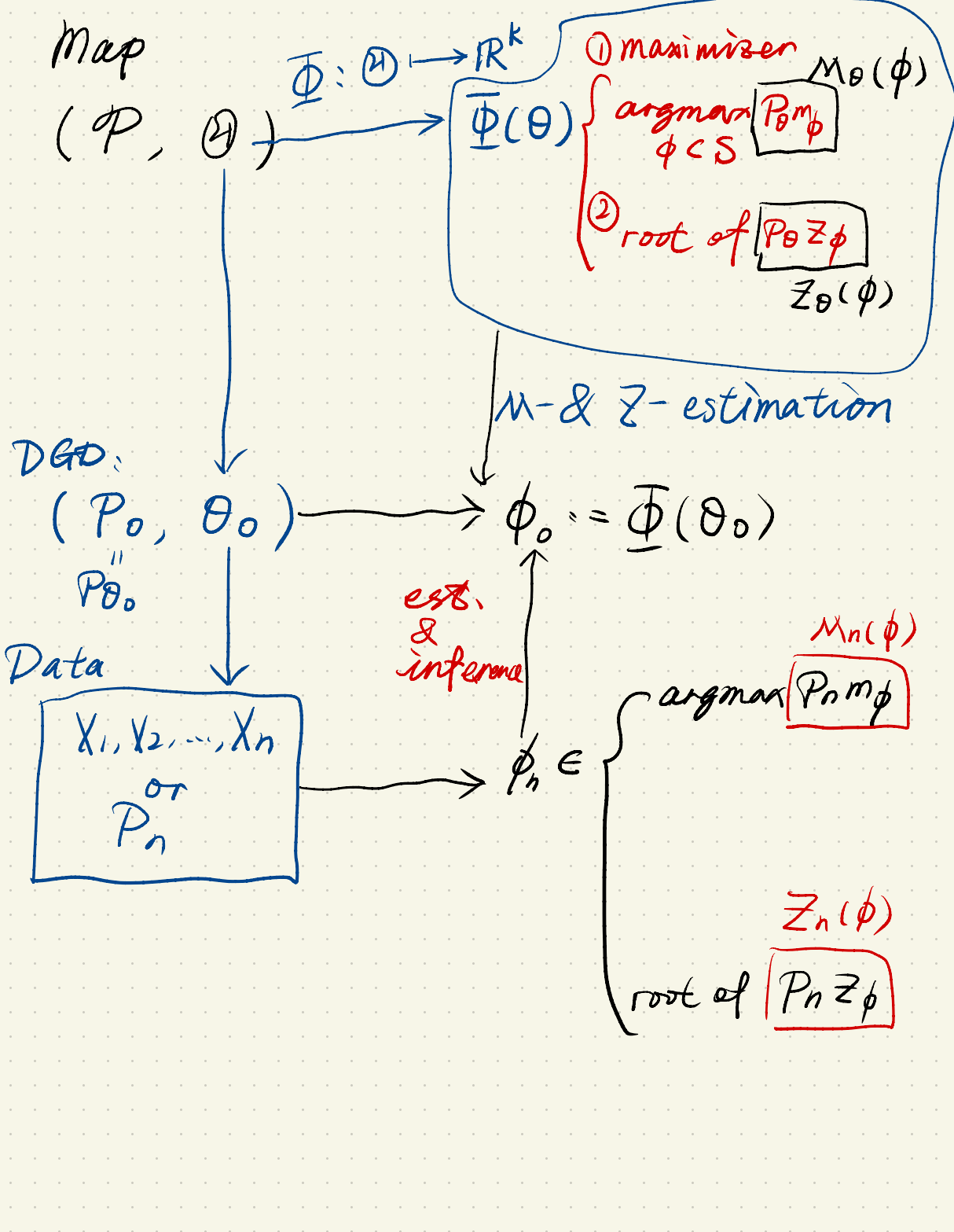
(ii) Identification: $\forall \varepsilon > 0$,

$$M_0(\phi_0) > \sup_{\phi : \|\phi - \phi_0\| > \varepsilon} M_0(\phi)$$

(iii) [Uniform consistency M_n to M_0]

$$\sup_{\phi \in S} |M_n(\phi) - M_0(\phi)| \xrightarrow{P} 0$$

Claim: (i) + (ii) + (iii) $\Longrightarrow \phi_n \xrightarrow{P} \phi_0$


Map
$(\mathcal{P}, \Theta)$
$\Phi : \Theta \mapsto \mathbb{R}^k$
$\Phi(\theta)$
① maximizer
$M_\theta(\phi)$
$\underset{\phi \in S}{\operatorname{argmax}}$ $P_\theta m_\phi$
② root of $P_\theta z_\phi$
$Z_\theta(\phi)$
M- & Z- estimation
DGD:
(P_0, θ_0)
P_{θ_0}
$\phi_0 := \Phi(\theta_0)$
est. & inference
Data
$X_1, X_2, \dots, X_n$
or
P_n
$\phi_n \in$
$M_n(\phi)$
$\underset{\phi}{\operatorname{argmax}}$ $P_n m_\phi$
$Z_n(\phi)$
root of $P_n z_\phi$

Consistency & rate of convergence

(i) $M_n(\phi_n) \geq \sup_{\phi} M_n(\phi) - o_p(1)$

minor computation error

(ii) $\forall \varepsilon > 0$,

$$M_0(\phi_0) > \sup_{\phi : \|\phi - \phi_0\| > \varepsilon} M_0(\phi)$$

(iii) Uniform convergence

$$\sup_{\phi \in S} |M_n(\phi) - M_0(\phi)| \xrightarrow{P} 0$$

Goal: $\phi_n \xrightarrow{P} \phi_0$.

Proof:

$$0 \leq M_0(\phi_0) - M_0(\phi_n)$$

$$\leq [M_0(\phi_0) - M_0(\phi_n)] - [M_n(\phi_0) - \sup_{\phi \in S} M_n(\phi)]$$

$[M_n(\phi_0) - \sup_{\phi \in S} M_n(\phi)] \leq 0$

$$\leq [M_0(\phi_0) - M_0(\phi_n)] - [M_n(\phi_0) - M_n(\phi_n)] + o_p(1)$$

(i)

$$= \underbrace{[M_0(\phi_0) - M_n(\phi_0)]}_{\le \sup_\phi |(M_n - M_0)(\phi)|} + \underbrace{[M_n(\phi_n) - M_0(\phi_n)]}_{\le \sup_\phi |(M_n - M_0)(\phi)|} + o_p(1)$$

$$\le 2 \sup_\phi |(M_n - M_0)(\phi)| + o_p(1) \overset{(iii)}{=} o_p(1)$$

$$\Rightarrow M_0(\phi_0) - M_0(\phi_n) = o_p(1) \qquad (*)$$

It remains to show

$$(*) \overset{(ii)}{\Longrightarrow} \phi_n \xrightarrow{P} \phi_0$$

Fix $\varepsilon > 0$, and let

$$\delta = M_0(\phi_0) - \sup_{\|\phi - \phi_0\| > \varepsilon} M_0(\phi) \overset{(ii)}{>} 0$$

Notice

$$\{\|\phi_n - \phi_0\| > \varepsilon\} \subset \{M_0(\phi_0) - M_0(\phi_n) \ge \delta\}$$

$$\| \qquad\qquad \|$$

$$P_0(\|\phi_n - \phi_0\| > \varepsilon) \le P_0(M_0(\phi_0) - M_0(\phi_n) \ge \delta)$$

$$\xrightarrow{n \to \infty} = 0 \qquad \| \qquad\qquad \|$$

$$\phi_n \xrightarrow{P} \phi_0$$

Uniform consistency

Setup: $M_0(\phi) = P_0 m_\phi$

$$M_n(\phi) = P_n m_\phi = \frac{1}{n}\sum_{i=1}^{n} m_\phi(X_i)$$

Prop. As long as

(i) $\forall \phi \in S,\ P_0|m_\phi| < \infty$

(ii) $|S| < \infty$

We have

$$\sup_{\phi \in S} |(P_n - P_0) m_\phi| \xrightarrow{P} 0$$

Proof. SLLN $\overset{(i)}{\leftarrow}$ + (ii)

Union bound: as long as

$A_1, A_2, \ldots \in \mathcal{A}$

then $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} \mu(A_i)$

Step 1:

By LLN, $\forall \varepsilon > 0$,

$$P_0(|P_n - P_0| m_\phi > \varepsilon)$$

$$= P_0\left[\left|\frac{1}{n}\sum_{i=1}^{n}(m_\phi(X_i) - P_0 m_\phi)\right| > \varepsilon\right]$$

$\overset{\text{LLN + (i)}}{\longrightarrow} 0$, as $n \to \infty$

Step 2. $P_0\left(\sup_{\phi \in S} |P_n - P_0| m_\phi > \varepsilon\right)$

$= P_0\left(\bigcup_{\phi \in S} \{|P_n - P_0| m_\phi > \varepsilon\}\right)$

$\leq \sum_{\phi \in S} P_0(|P_n - P_0| m_\phi > \varepsilon)$

↑ union bound

$\leq |S| \cdot \sup_{\phi \in S} P_0(|P_n - P_0| m_\phi > \varepsilon)$

$\downarrow n \to \infty$

0

$n \to \infty$ 0

For studying

$$\sup_{\phi \in S} |P_n - P_0| m_\phi$$

it suffices to study

$$\{m_\phi : \phi \in S\}$$

Def. [Po-Glivenko-Cantelli]

The class of functions

$$\{m_\phi : \phi \in S\}$$

is called a P_0-GC class/set

if it satisfies

$$\sup_{\phi \in S} |P_n - P_0| m_\phi = o_p(1)$$

Remark G&C independently studied

$$\sup_{t \in \mathbb{R}} |F_n(t) - F_0(t)| \xrightarrow{P} 0$$

$F_n(t)$: empirical CDF $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le t\}$ $=$ $P_n \mathbb{1}(X \le \cdot)$

$F_0(t)$: population CDF $P(X \le t)$ $=$ $P_0 \mathbb{1}(X \le \cdot)$

$$\Leftrightarrow \sup_{\phi \in \mathbb{R}} |P_0 - P_0| m_\phi \xrightarrow{P} 0$$

w.r.t. $\{m_\phi := \mathbb{1}\{X \le \phi\}, \phi \in \mathbb{R}\}$

Remark. Methods to prove a certain function class is P_0-GC:

(i) Symmetrization + VC-type bound

(ii) Martingale theory [520s]

(iii) Entropy argument [581]

metric-entropy, bracketing-entropy ...

Def: [bracketing entropy]

Setup: $\mathcal{F} = \{f \in \mathcal{F}\}$

$f : \mathcal{X} \longmapsto \mathbb{R}^1$ (real line)

(i) [Bracket] Given any two functions l and u [$l < u$] in $L^1(P_0)$ [i.e., $\int |l(x)| P_0(dx) < \infty$, $\int |u(x)| P_0(dx) < \infty$]

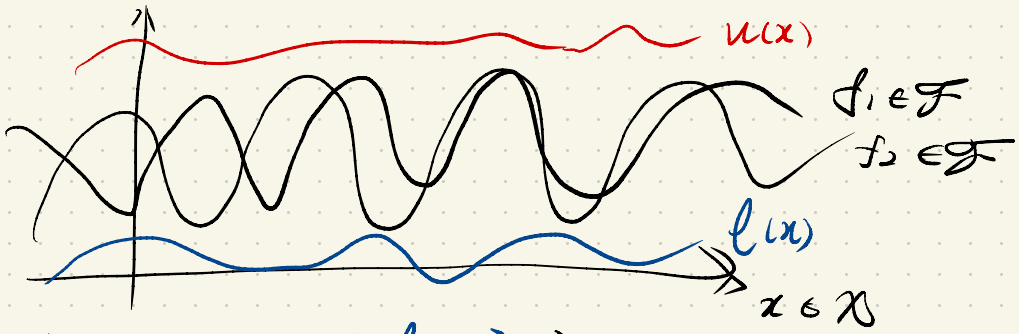
the bracket of l and u,

$[l, u]$

containing all functions

$f \in \mathcal{F}$

s.t. $l(x) \le f(x) \le u(x)$, $\forall x \in \mathcal{X}$.



(ii) A bracket $[l, u]$ is said to be an ε-bracket if

$\| u - l \|_{L_1(P_0)} \le \varepsilon$.

$\left(\|f\|_{L_1(P_0)} := \int |f(x)| P_0(dx) \right)$

(iii) Bracketing entropy of $\mathcal{F}$.

$$N_{[]}(\varepsilon, L^1(P_0), \mathcal{F})$$

represents the smallest number of ε-brackets needed to cover $\mathcal{F}$.

Remark. Notice the l's and u's do not have to be in $\mathcal{F}$, but they should $\in L^1(P_0)$.

Thm. [P_0-GC using bracketing entropy]

Suppose, $\forall \varepsilon > 0$,

$$N_{[]}(\varepsilon, L^1(P_0), \mathcal{F}) < \infty$$

Then, $\mathcal{F}$ is P_0-GC, i.e.,

$$\|P_n - P_0\|_{\mathcal{F}} =: \sup_{f \in \mathcal{F}} |P_n - P_0| f$$

$$= o_p(1)$$

Proof. (i) with each ε-bracket, f's are nearly the same (identical)

(ii) $\exists$ only finitely many ε-brackets

(iii) Union bound proof.

Fix $\varepsilon > 0$. By cond.,

$\exists\, [\ell_j, u_j]$, $j = 1, 2, \cdots, \boxed{N_{[\varepsilon]}(\cdots)}$ ← N

s.t. it covers $\mathcal{F}$.

Define

$$A_{n,\varepsilon} := \{ \sup_{f \in \mathcal{F}} (P_n - P) f > 2\varepsilon \}$$

$$B_{n,\varepsilon} := \{ \inf_{f \in \mathcal{F}} (P_n - P) f < -2\varepsilon \}$$

By symmetry, only study $A_{n,\varepsilon}$.

Fix $f \in \mathcal{F}$, By cond.,

$\exists$ one $j \in \{1, 2, \cdots, N\}$ s.t.

$$\ell_j \le f \le u_j$$

implying

$$P_n f \le P_n u_j$$

$$P_0 f \ge P_0 u_j - P_0(u_j - \ell_j)$$

$$\Downarrow$$

$$(P_n - P_0) f \le P_n u_j - P_0 u_j + P_0(u_j - \ell_j)$$

$$= (P_n - P_0) u_j + P_0(u_j - \ell_j)$$

$$= (P_n - P_0) u_j + \underbrace{P_0 |u_j - \ell_j|}_{= \|u_j - \ell_j\|_{L^2(P_0)} \le \varepsilon}$$

$$\le (P_n - P_0) u_j + \varepsilon$$

$$
\begin{aligned}
& P_0(A_{n,\varepsilon}) \\
&= P_0\Big(\sup_{f\in\mathcal{F}} (P_n - P_0) f > 2\varepsilon\Big) \\
&\le P_0\Big(\sup_{j=\{1,2,\dots,N\}} (P_n - P_0) u_j > \varepsilon\Big) \\
&\le \sum_{j=1}^{N} P_0\big((P_n - P_0) u_j > \varepsilon\big) = o(1)
\end{aligned}
$$

□

Real Q: How to show a class of functions have finite bracketing entropy

$N_{[]}(\cdot)$???

Thm. [Example 19.8 Van der Vaart]

① Suppose

$$\mathcal{F} := \{ f_\phi , \phi \in K \subset \mathbb{R}^d \}$$

where K is compact.

② $\forall$ given x,

$\phi \mapsto f_\phi(x)$ is continuous w.r.t. ϕ.

③ [Envelope] Suppose

$\exists$ a function F satisfying

(a) $\sup_{\phi\in K} |f_\phi(x)| \le F(x)$, $\forall x$

(b) $P_0|F| = P_0 F < \infty$

Then $\forall \varepsilon > 0$, $N_{[]}(\varepsilon, L^1(P_0), \mathcal{F})$ is finite.

Proof: HW!

Ex. [logistic reg.]

Setup: $X = (Z, Y)$, $Z \in \mathbb{R}^d$, $Y \in \{0,1\}$

Data — covariates — outcome

GLM:
$$\begin{cases} Y|Z \sim \mathrm{Ber}(g_{\Phi(\theta)}(Z)) \\ \quad \text{w. } g_\phi(z) = \dfrac{1}{1+\exp(-\phi^T z)} \\ Z \sim P_\theta \quad \{ \text{Gaussian, others} \end{cases}$$

n IID obs. of (Z, Y)

$X_1, X_2, \cdots, X_n$

Goal: Estimate ϕ_0

Assumption: $\Phi(\theta_0) = \phi_0 \in K \subseteq \mathbb{R}^d$ (compact)

Procedure: MLE / M-est.

$$\hat{\phi}_n = \arg\max_{\phi \in K} P_n m_\phi$$

with $m_\phi : x = (z, y) \longmapsto y \log g_\phi(z) + (1-y)\log(1-g_\phi(z))$

$$= y \,\mathrm{logit}(g_\phi(z)) + \log(1 - g_\phi(z))$$

$$= y \cdot \phi^T z - \log(1 + \exp(\phi^T z))$$

① $\phi \in K$ compact ✓

② $\forall x = (y, z)$, it is cont. w.r.t. ϕ. ✓

③ It remains to find the envelope F:

$$\sup_{\phi \in K} |m_\phi(x)| \le |y \cdot \phi^T z| + |\log(1+\exp(\phi^T z))| \quad (\log(\cdot) > 0)$$

$$\le |\phi^T z| + \log(1+\exp(\phi^T z))$$

$$\left[\begin{array}{l} 0 \le \log(1+\exp(t)) \\ \quad \le \max(t, 0) + 1 \\ \forall t \end{array}\right]$$

$$\le |\phi^T z| + \max(\phi^T z, 0) + 1$$

$$= 2|\phi^T z| + 1$$

[Cauchy-schwarz] $\le 2\|\phi\|\|z\| + 1$

$$\le 2\|z\| \sup_{\phi \in K} \|\phi\| + 1$$

since K compact, it is finite, $< \infty$.

$:= F_0 := x \mapsto 2\|z\| \sup_{\phi \in K} \|\phi\| + 1$

If $\|F_0\|_{L^1(P_0)} < \infty$, then done.

$\Leftrightarrow P_0\|z\| < \infty$

$\Leftrightarrow P_0 z_j^2 < \infty$, $\forall j = 1, \dots, d$.

By VdV

$N_{[]}(\varepsilon, L'(P_0), \{m_\phi : \phi \in K\}) < \infty$

UCLLN is true

$P_0 m_\phi$ is convex ← verify

$\phi_n \xrightarrow{P} \phi_0$

3.4 ASN of M- & Z- est.

$\sqrt{n}(\phi_n - \phi_0) \Rightarrow N(0, \sigma^2)$

asymptotic normality, a.k.a. ASN

To establish ASN of ϕ_n, usually we aim to show

Asymptotic linear estimator (ALE)

$$\sqrt{n}(\phi_n - \phi_n) = \boxed{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i} + \boxed{R_n} + o_p(1)$$

linear expansion / Hajek projection

Donsker property

$\sqrt{n}(P_n - P)(\hat{f}_n - \hat{g}_n)$

with $d(\hat{f}_n, \hat{g}_n) \xrightarrow{n \to \infty} 0$

The most challenging part is to show

$R_n = o_p(1)$ (Take supremum)

stochastic equicontinuity

$$\lim_{\varepsilon \to 0} \sup_{f,g : d(f,g) < \varepsilon} \sqrt{n}(P_n - P)(f - g) = 0$$

e.g.

$\mathcal{F}$ s.t. $\square$ is true is called P_0-Donsker

Example (mean absolute deviation, David Pollard)

$$\nu_n := \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}_n|$$

Robust statistics

[Setup: $X_1, \dots, X_n \overset{iid}{\sim} P_0$,

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$]

1st [Consistency]

$$\mu_n \xrightarrow{P} \mu_0 := \mathbb{E}[|X - \mathbb{E}X|]$$

2nd Mean absolute deviation is ASN

Goal: $\sqrt{n}(\mu_n - \mu_0) \Rightarrow N(0, \sigma^2)$

Proof:

Def. $\mathcal{F} := \{ \underbrace{|x - t|}_{f_t(x)} : t \in \mathbb{R} \}$

$$\mu_n = P_n f_{\bar{X}_n}$$

$$\mu_0 = P_0 f_{\mu (:= \mathbb{E}X)}$$

$$\sqrt{n}(\mu_n - \mu_0) = \sqrt{n}(P_n f_{\bar{X}_n} - P_0 f_\mu)$$

$$= \sqrt{n}(P_n - P_0) f_\mu \quad \leftarrow \text{CLT}$$

$$+ \sqrt{n}(P_n f_{\bar{X}_n} - P_n f_\mu)$$

$$= \sqrt{n}(P_n - P_0) f_\mu \quad \leftarrow \text{CLT}$$

ψ: smooth

$$+ \sqrt{n}(\psi(\bar{X}_n) - \psi(\mu)) \quad \leftarrow \text{Delta method}, \quad \psi(t) = P_0 f_t$$

$$+ \sqrt{n}(P_n - P_0)(f_{\bar{X}_n} - f_\mu)$$

↑ Donsker

$$\Rightarrow N(0, \sigma^2)$$

$$= \sqrt{n}(P_n - P_0)(f_\mu(x) + \psi'(\mu) \cdot x) + o_p(1) + o_p(1)$$

Delta method main term by Taylor expansion ($\psi'(\mu) \cdot x$)

STAT 582

Delta method (first $o_p(1)$) — Donsker's $o_p(1)$ (second $o_p(1)$)

$$\Rightarrow N(0, \sigma^2)$$

please calculate it (σ^2)

ASN of Z-est.

Cramer 1930

Cond. too strong and eliminate very important application like quantile regression.

Setup [simple]

$\phi_0 \in \mathbb{R}$

ϕ_0 : $Z_0(\phi_0) = 0$

ϕ_n : $Z_n(\phi_n) = 0$

Goal: $\sqrt{n}(\phi_n - \phi_0) \Rightarrow N(0, \theta^2)$

513-level :

① Taylor expansion on Z_n

② Cramer-type cond.

③ smoothness on Z_n's function

smoothness is required naguthy might not be 0

They are strong conditions and not necessary.

Claim: ①-③ cannot handle many estimators, such as quantile reg:

$$\operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} |Y_i - X_i^T \beta|$$

Insight: too much assumption on Z_n which can be very non-smooth.

Instead of Z_n, we can look at Z_0.
Because Z_0 we can add expectation, which will smoothen the original function.

Use Donsker to connect Z_n and Z_0.

Details (heuristic)

$$0 = -Z_0(\phi_0)$$

$$= [Z_n(\phi_0) - Z_0(\phi_0)] - Z_n(\phi_0)$$

$$= [Z_n(\phi_0) - Z_0(\phi_0)] + [Z_n(\phi_n) - Z_n(\phi_0)]$$

($Z_n(\phi_n) = 0$)

$$\text{I} = [Z_n(\phi_0) - Z_n(\phi_0)] \to \text{CLT}$$

$$\text{II} + Z_0(\phi_n) - Z_0(\phi_0) \to \text{Taylor expansion}$$

(Z_0 → usually smooth)

$$\text{III} + Z_n(\phi_n) - Z_n(\phi_0) - Z_0(\phi_n) + Z_0(\phi_0) \to \text{a Donsker term}$$

When $Z_n(\phi) = P_n Z_\phi$

$Z_0(\phi) = P_0 Z_\phi$

the last term $= (P_n - P_0)(Z\phi_n - Z\phi_0)$

Step I: $\text{Term I} = (P_n - P_0) Z\phi_0$

$\Rightarrow \sqrt{n}\,\text{Term I} \Rightarrow N(0, \text{var}(Z\phi_0))$

Step II: $\text{Term II} = Z_0(\phi_n) - Z_0(\phi_0)$

$\Rightarrow \text{Term II} = (\phi_n - \phi_0)\dot{Z}_0(\phi_0) + \frac{1}{2}(\phi_n - \phi_0)^2 \ddot{Z}_0(\tilde{\phi}_n)$

for some $\tilde{\phi}_n$ between ϕ_0 and ϕ_n.

$= (\phi_n - \phi_0)\dot{Z}_0(\phi_0) + o_p(\phi_n - \phi_0)$

If ① $\phi_n \xrightarrow{P} \phi_0$

② $\sup_{\phi \in U(\phi_0)} |\ddot{Z}_0(\phi)| < \infty$

small neighbourhood

Step 3: We need to show

$\sqrt{n} \cdot \text{Term II} = o_p(1)$

An incorrect heuristic to step 3.

LeCam's thinking: Instead of having a random seq. of $\{\phi_n\}$. Let's think about "$\{\phi_n\}$ to be deterministic"

INCORRECT but nearly the sample splitting.

Under the Incorrect assumption:

$$\text{Term III} = (P_n - P_0)(Z_{\phi_n} - Z_{\phi_0})$$

Fixing any $t > 0$,

$$P_0\left[\sqrt{n}\,\text{Term III} > t\right]$$

Chebyshev ↘

$$\leq \frac{\text{Var}\left(\sqrt{n}\,(P_n - P_0)(Z_{\phi_n} - Z_{\phi_0})\right)}{t^2}$$

$$= \frac{\text{Var}\left(Z_{\phi_n}(X_1) - Z_{\phi_0}(X_1)\right)}{t^2}$$

↑ please verify

Then, under certain conditions,

$$\text{Var}\left(Z_{\phi_n}(\cdot) - Z_{\phi_0}(\cdot)\right) \to 0$$

as $\phi_n \to 0$

e.g. Lipschitz cond:

$$|Z_\phi(x) - Z_{\phi_0}(x)| \leq \|\phi - \phi_0\| \cdot G(x)$$

and $P_0 G^2 < \infty$.

Step 4: $0 = \text{Term I} + \text{Term II} + \text{Term III}$

$$\Rightarrow \phi_n - \phi_0 = -(P_n - P_0) Z_{\phi_0} \leftarrow \text{Term I}$$

$$\dot{Z}_0(\phi_0) + o_p(1) \leftarrow \text{Term II}$$

$$+ o_p(n^{-1/2})$$

$$\Rightarrow \sqrt{n}(\phi_n - \phi_0) = \underbrace{-\sqrt{n}(P_n - P_0) Z_{\phi_0}}_{\dot{Z}_0(\phi_0)} + o_p(1) \quad \leftarrow \text{ASL}$$

$$\Rightarrow N\left(0, \frac{\text{Var}_0(Z_{\phi_0})}{(\dot{Z}_0(\phi_0))^2}\right)$$

$$\Uparrow$$

$$\ell_n \text{ MLE}_n = \frac{1}{I(\theta_0)}$$

The following 2 Thms will not be proved.

Thm[General Z-est. ASN]

① $Z_0 = P_0 Z_\phi$, $Z_n = P_n Z_\phi$, and $\forall \phi$ in an open subset of $\mathbb{R}^d$

$$Z_\phi : \mathcal{X} \longmapsto \mathbb{R}^d$$

② $\mathbb{E}_0 \| Z_{\phi_0}(X) \|^2 < \infty$

③ $\phi \longmapsto P_0 Z_\phi$ is differentiable (Taylor expansion) at ϕ_0 with nonsingular Jacobian V_{ϕ_0}. ⟶ In MLE, Fisher info matrix

④ Assume $\exists\, G : \mathcal{X} \longmapsto \mathbb{R}$ (smoothness for Donsker)

s.t. ① $P_0 G^2 < \infty$

② $\forall x \in \mathcal{X}$, $\forall \phi, \tilde{\phi} \in U(\phi_0)$,

$$\| Z_\phi(x) - Z_{\tilde{\phi}}(x) \| \le \| \phi - \tilde{\phi} \| \cdot G(x)$$

⑤ $\{\phi_n\}$ is a sequence of est. ϕ_0 s.t.

$$P_n Z_{\phi_n} = o_p(n^{-1/2})$$

and $\phi_n \xrightarrow{P} \phi_0$

Claim: Under ①-⑤,

$$\sqrt{n}(\phi_n - \phi_0) \Rightarrow N\left(0, V_{\phi_0}^{-1} \cdot P_0[Z_{\phi_0} Z_{\phi_0}^T][V_{\phi_0}^{-1}]\right)$$

Thm [ASN of M-est.] Assume

① $\phi_0 = \operatorname{argmax} P_0 m_\phi$

$[\phi_n \approx \operatorname{argmax} P_n m_\phi]$

$m_\phi : \mathcal{X} \mapsto \mathbb{R}$

② $\forall \phi \in$ open subset $\subset \mathbb{R}^d$,

$\phi \mapsto m_\phi(x)$ is differentiable at ϕ_0 for P_0-almost everywhere x.

Note: think about $|X|$  has measure 0

∴ still satisfies cond. ②

We take $\dot{m}_{\phi_0}$ to be the derivative.

Note: even weaker than Z because M-est. acts like integral of Z-est.

③ $\forall \phi, \tilde{\phi} \in U(\phi_0)$, assume

$\exists G : \mathcal{X} \mapsto \mathbb{R}$ s.t.

① $P_0 G^2 < \infty$

② $|m_\phi(x) - m_{\tilde{\phi}}(x)| \le \|\phi - \tilde{\phi}\| G(x)$, $\forall x \in \mathcal{X}$.

④ Assume $\exists$ a non-singular symmetric matrix V_{ϕ_0} s.t.

$$\lim_{\varepsilon \to 0} \frac{\sup_{\|h\|=1} \left| P_0 m_{\phi_0 + \varepsilon h} - P_0 m_{\phi_0} - \frac{1}{2} \varepsilon^2 h^\top V_{\phi_0} h \right|}{\varepsilon^2} = 0$$

[We will use QMD to verify]

⑤ $P_n m_{\phi_n} \geq \sup_{\phi} P_n m_\phi - o_p(n^{-1})$

& $\phi_n \xrightarrow{P} \phi_0$

Claim: ① – ⑤ imply

$$\sqrt{n}(\phi_n \overset{\text{ALE}}{-} \phi_n) \Rightarrow N(0, V_{\phi_0}^{-1} [P_0 \dot{m}_{\phi_0} \dot{m}_{\phi_0}^\top] [V_{\phi_0}^{-1}]^\top)$$

$$= V_{\phi_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{m}_{\phi_0}(X_i) + o_p(1)$$

Slutsky

As long as it is not ASL, Bootstrap cannot work.

Bootstrap consistency $\approx$ ASL.

BS can improve finite sample accuracy.

BS can accelerate rate.

Chapter 3.5 Parametric models

(MLE + QMD)

K-L divergence — Hellinger distance

Setup: $\mathcal{P} := \{P_\theta : \theta \in \underset{\text{open}}{H} \subset \mathbb{R}^d \to \text{finite dim}\}$

MLE, maximize log-likelihood.

is a M-est:

$$M_{\theta_0} : \theta \longmapsto \mathbb{E}_0\left[\log \frac{dP_\theta}{d\mu}(X)\right]$$

① μ is the reference measure { Lesbgue / Counting

② Choose μ to be Lebesgue measure.

$\frac{dP_\theta}{d\mu}$ is the pdf.

③ Choose μ to be counting measure,

$\frac{dP_\theta}{d\mu}$ is the pmf.

We'll write

$$p_\theta = \frac{dP_\theta}{d\mu} \quad (\text{w.r.t. } \mu)$$

M_{θ_0} will give us Z_{θ_0}

$$Z_{\theta_0} : \theta \longmapsto \nabla_\theta M_{\theta_0}$$
$$= \nabla_\theta \mathbb{E}_0[\log p_\theta(X)]$$

High level motivation of QMD

∇_θ and $\int$

Interchange the position

Ex) $Z_{\theta_0}(\theta) = \mathbb{E}_{\theta_0}[Z_\theta(x)]$

$Z_\theta := x \longmapsto \nabla_\theta \log p_\theta(x)$

$= \dfrac{\dot{p}_\theta(x)}{p_\theta(x)} \quad \dot{l}_\theta(x)$

$(\dot{p}_\theta(x) = \nabla_\theta p_\theta(x))$

Claim [5.13] "In general,"

$Z_{\theta_0}(\theta_0) = 0$

Proof. $Z_{\theta_0}(\theta_0) = \mathbb{E}_0[\dot{l}_{\theta_0}(x)]$

$= \int \dot{l}_{\theta_0}(x)\, p_{\theta_0}(x)\, \mu(dx)$

$= \int \dot{p}_{\theta_0}(x)\, \mu(dx)$

cond. is very naughty

$= \int \nabla_\theta p_{\theta_0}(x)\, \mu(dx)$

$\overset{?}{=} \nabla_\theta \underbrace{\int p_{\theta_0}(x)\, \mu(dx)}_{1} = 0$

Claim 2 [STAT 513 level]

ASN of MLE:

We can formulate the problem as

Z-est:

$$\sqrt{n}(\theta_n - \theta_0) \Rightarrow N\left(0, \frac{P_{\theta_0}[\dot{Z}_{\theta_0}]^2}{[\dot{Z}_{\theta_0}(\theta_0)]^2}\right)$$

[when $\theta_0 \in \mathbb{R}$]

Intuitively,

$$\frac{P_{\theta_0}\dot{Z}_{\theta_0}^2}{[\dot{Z}_{\theta_0}(\theta_0)]^2} = \frac{1}{I(\theta_0)} \leftarrow \text{FIN at } \theta_0 \qquad (*)$$

To prove (*), 513 tells us

$$\dot{Z}_{\theta_0}(\theta_0) = \frac{d}{d\theta}\int \dot{\ell}_{\theta_0}(x)\, P_{\theta_0}(dx)$$

$$= \int \frac{d}{d\theta}\, \dot{\ell}_{\theta_0}(x)\, P_{\theta_0}(dx)$$

Interchange

$$= \int \frac{\ddot{P}_{\theta_0}(x)\, P_{\theta_0}(x) - \dot{P}_{\theta_0}^2(x)}{[P_{\theta_0}(x)]^2} P_{\theta_0}(dx)$$

$P_{\theta_0}(dx) = P_{\theta_0}(x)\,\mu(dx)$

$$= \int \ddot{P}_{\theta_0}(x)\,\mu(dx)$$

$$\frac{d^2}{d\theta^2}\int P_{\theta_0}(x)\,\mu(dx) = 0$$

$$= -\int [\dot{\ell}_{\theta_0}(x)]^2 P_{\theta_0}(dx)$$

← var of score function = FIN

$$= -\text{FIN}$$

Chapter 3.6 QMD

Quadratic Mean Differentiability

Note: $\nabla_\theta\left(\sqrt{P_\theta(x)}\right) = \frac{1}{2}\cdot\frac{\dot{P}_\theta(x)}{\sqrt{P_\theta(x)}}$

$= \frac{1}{2}\cdot\frac{\dot{P}_\theta(x)}{P_\theta(x)}\cdot\sqrt{P_\theta(x)}$

$= \frac{1}{2}\cdot\dot{\ell}_\theta(x)\cdot\sqrt{P_\theta(x)}$

Def [QMD]

(i) The root density

$\theta \longmapsto \sqrt{P_\theta}$

is called QMD at θ if

$\exists$ a function $\dot{\ell}_\theta$

s.t.

$$\lim_{\varepsilon\to 0}\sup_{\|h\|=1}\int\left[\frac{\sqrt{P_{\theta+\varepsilon h}(x)}-\sqrt{P_\theta(x)}}{\varepsilon} - \frac{1}{2}h^\top\dot{\ell}_\theta(x)\sqrt{P_\theta(x)}\right]^2\mu(dx) = 0$$

- $\sqrt{P_{\theta+\varepsilon h}(x)}-\sqrt{P_\theta(x)}$: perturbation
- $\int$: mean
- $\frac{\sqrt{P_{\theta+\varepsilon h}(x)}-\sqrt{P_\theta(x)}}{\varepsilon}$: differentiability
- $\frac{1}{2}h^\top\dot{\ell}_\theta(x)\sqrt{P_\theta(x)}$: quadratic
- $\frac{\sqrt{P_{\theta+\varepsilon h}(x)}-\sqrt{P_\theta(x)}}{\varepsilon}$: related to Hellinger distance (metric)

$\forall P, Q \ll \mu,\ \int\left(\sqrt{p}-\sqrt{q}\right)^2\mu(dx)$

$p(x) = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$

(ii) A model $\{P_\theta : \theta\in\Theta\}$ is QMD at θ if $\theta \longmapsto \sqrt{P_\theta}$ is QMD at θ.

This model is QMD if it is QMD at $\forall\theta\in\Theta$.

Remark: An equivalent expression for QMD at θ is

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \cdot \sqrt{p_\theta} \right]^2 \mu(dx)$$

$= o(\|h\|^2)$ as $h \to 0$ [VdV book]

↑ uniform

[Please verify]

Thm. [1st part of vdV's Thm. 7.2] Suppose Θ is open subset of $\mathbb{R}^d$ and $\{P_\theta : \theta \in \Theta\}$ is QMD at θ. Then

($\dot{\ell}_\theta(\cdot)$ is the function in QMD)

① $P_\theta \dot{\ell}_\theta = 0$ (1st Eq.)

② $I_\theta = P_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^T]$ exists.

Proof of ① QMD at θ gives us

① $\frac{\sqrt{p_{\theta+\varepsilon h}(x)} - \sqrt{p_\theta(x)}}{\varepsilon}$ converge in $L_2(\mu)$ to $\frac{1}{2} h^T \dot{\ell}_\theta \cdot \sqrt{p_\theta}$

② $\sqrt{p_{\theta+\varepsilon h}(x)} - \sqrt{p_\theta(x)}$ converges in $L_2(\mu)$ to 0.

To prove $P_\theta \dot{\ell}_\theta = 0$, it suffices to show

$P_\theta[h^T \dot{\ell}_\theta] = 0$, $\forall h$ s.t. $\|h\| = 1$.

$$= \int h^T \dot{\ell}_\theta \, p_\theta \, d\mu$$

$$= \int [h^T \dot{\ell}_\theta \cdot \sqrt{p_\theta}] \sqrt{p_\theta} \, d\mu$$

$$= \lim_{\varepsilon \to 0} 2 \int \left[\frac{\sqrt{P_{\theta+\varepsilon h}} - \sqrt{P_\theta}}{\varepsilon} \right] \frac{\sqrt{P_{\theta+\varepsilon h}} + \sqrt{P_\theta}}{2} \, d\mu$$

HW

$$= \lim_{\varepsilon \to 0} \int \frac{P_{\theta+\varepsilon h} - P_\theta}{\varepsilon} \, d\mu$$

$$= 0$$ □

Thm [ASN of MLE under QMD]

Suppose

① $\{P_\theta : \theta \in \Theta\}$ is QMD at θ_0, which is an inner point of Θ.

② $\exists$ a function G satisfying

(a) $P_{\theta_0} G^2 < \infty$

(b) $\forall \theta_1, \theta_2 \in U(\theta_0)$, $\forall x$

$$|\log P_{\theta_1}(x) - \log P_{\theta_2}(x)| \le G(x) \|\theta_1 - \theta_2\|$$

③ I_{θ_0} is non-singular (invertible) and the MLE θ_n is consistent.

Then, we have asymptotic normality.

$$\sqrt{n}(\theta_n - \theta_0) \Rightarrow N(0, I(\theta_0)^{-1})$$

Proof. We only prove ④ in M-ess. Thm,

$$\sup_{\|h\|=1} \left| P_{\theta_0} \underbrace{m_{\theta_0+\varepsilon h}}_{\ell_{\theta_0+\varepsilon h}} - P_{\theta_0} \underbrace{m_{\theta_0}}_{\ell_{\theta_0}} + \frac{1}{2} \varepsilon^2 h^T V_0 h \right| = o(\varepsilon^2)$$

as $\varepsilon \to 0$.

$$\square = P_0\left[\log \underbrace{P_{0\varepsilon h}}_{\swarrow P_\varepsilon} - \log \underbrace{P_{0_0}}_{\downarrow P_0}\right]$$

$$= P_0\left[\log P_\varepsilon - \log P_0\right]$$

$$= P_0\left[2\log\sqrt{P_\varepsilon} - 2\log\sqrt{P_0}\right]$$

$$= 2P_0\left[\log\sqrt{P_\varepsilon} - \log\sqrt{P_0}\right]$$

$$= 2P_0\left[\log\frac{\sqrt{P_\varepsilon}}{\sqrt{P_0}} - \log 1\right]$$

$$= 2\int\left[f(w_\varepsilon) - f(0)\right]P_0(dx)$$

where $f(w) = \log(1+w)$

$w_\varepsilon = \sqrt{P_\varepsilon/P_0} - 1$

Now do Taylor expansion on $f(\cdot)$: Lyapnov version of Taylor expansion

$$f(w) - f(0) = w - \frac{w^2}{2} + \underbrace{w^2 \cdot r(w)}_{\| o(w^2)}$$

with $r(w) \to 0$ as $w \to 0$

$$\Rightarrow \square = \boxed{2\int w_\varepsilon \cdot P_0(dx)} - 2\cdot\frac{1}{2}\boxed{\int w_\varepsilon^2 \cdot P_0(dx)}$$

$$= \int\left(\sqrt{\frac{P_\varepsilon}{P_0}} - 1\right)P_0\,\mu(dx)$$

$$= \int\left(\sqrt{P_\varepsilon} - \sqrt{P_0}\right)\sqrt{P_0}\,\mu(dx)$$

$$= \int\left(\frac{\sqrt{P_\varepsilon}}{\sqrt{P_0}} - 1\right)^2 P_0\,\mu(dx)$$

$$= \int\left(\frac{P_\varepsilon}{P_0} + 1 - 2\sqrt{\frac{P_\varepsilon}{P_0}}\right)P_0\,d\mu$$

$$= \int\left(P_\varepsilon + P_0 - 2\sqrt{P_\varepsilon P_0}\right)d\mu$$

$$= \int\left(\sqrt{P_\varepsilon} - \sqrt{P_0}\right)^2 d\mu$$

$$+ 2\int\left[\sqrt{P_\varepsilon} - \sqrt{P_0}\right]^2 r(w_\varepsilon)\,d\mu$$

$$= 2\int (\sqrt{P_\varepsilon} - \sqrt{P_0})\sqrt{P_0}\, d\mu \quad \text{1st}$$

$$- \int (\sqrt{P_\varepsilon} - \sqrt{P_0})^2 d\mu \quad \text{H-distance}$$

$$+ 2\int (\sqrt{P_\varepsilon} - \sqrt{P_0})^2 r(w_\varepsilon)\, d\mu \quad \text{H-distance small order term}$$

$$(1st) = 2\int (\sqrt{P_\varepsilon P_0} - P_0)\, d\mu$$

$$= -2\int (P_0 - \sqrt{P_0 P_\varepsilon})\, d\mu$$

$$= -2\int \frac{P_\varepsilon + P_0 - 2\sqrt{P_\varepsilon P_0}}{2}\, d\mu$$

$$= -\frac{2}{2}\int (\sqrt{P_\varepsilon} - \sqrt{P_0})^2 d\mu$$

$$= -H^2(P_\varepsilon, P_0)$$

$$(2nd) = H^2(P_\varepsilon, P_0)$$

$$(3rd) = o(\varepsilon^2)$$

↑ left to the students

$$P_0[m_{\theta_0 + \varepsilon h} - m_{\theta_0}]$$

$$= -2H^2(P_\varepsilon, P_0) + o(\varepsilon^2)$$

Wish to show $\frac{1}{2}\varepsilon^2 h^T V_0 h$

It remains to show

$-2H^2(P_\varepsilon, P_0)$ and $\frac{1}{2}\varepsilon^2 h^T V_0 h$ (*)

are $o(\varepsilon^2)$ close.

To show (*) holds:

① Reverse triangle inequality:

$$\left| \|a\| - \|b\| \right| \le \|a-b\|$$

$$\left(\|a\| \le \|b\| + \|a-b\| \right.$$

$$\left. \Rightarrow \|a\| - \|b\| \le \|a-b\| \right)$$

② Introduce $L^2(\mu) = \{ f : \|f\|_{L_2(\mu)} < \infty \}$

$$\|f\|_{L_2(\mu)} = \left[\int f^2(x)\, \mu(dx) \right]^{1/2}$$

is a metric space

$$\Big| \|\sqrt{P_\varepsilon} - \sqrt{P_0}\|_{L_2(\mu)} - \|\tfrac{1}{2} \varepsilon h^T \dot{\ell}_{\theta_0} \sqrt{P_0}\|_{L_2(\mu)} \Big|$$

$$\le \boxed{\| \sqrt{P_\varepsilon} - \sqrt{P_0} - \tfrac{1}{2} \varepsilon h^T \dot{\ell}_{\theta_0} \sqrt{P_0} \|_{L_2(\mu)}}$$

QMD says $\square^2 = o(\varepsilon^2)$

which means

$$\|\sqrt{P_\varepsilon} - \sqrt{P_0}\|_{L_2(\mu)} = \tfrac{1}{2} \varepsilon \|h^T \dot{\ell}_{\theta_0} \sqrt{P_0}\|_{L_2(\mu)} + o(\varepsilon)$$

$$\boxed{\|\sqrt{P_\varepsilon} - \sqrt{P_0}\|^2_{L_2(\mu)}} = \tfrac{1}{4} \varepsilon^2 \|h^T \dot{\ell}_{\theta_0} \sqrt{P_0}\|^2_{L_2(\mu)} + o(\varepsilon^2)$$

$H^2(P_\varepsilon, P_0)$

$$-2H^2(P_\varepsilon, P_0) = -\frac{1}{2}\varepsilon^T \| h^T \dot{\ell}_{\theta_0} \sqrt{p_0} \|^2_{L_2(\mu)} + o(\varepsilon^2)$$

$$\| h^T \dot{\ell}_{\theta_0} \sqrt{p_0} \|^2_{L_2(\mu)} = h^T \left[\int \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T p_0 \, d\mu \right] h$$

$$= h^T P_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T h$$

$$P_0[m_{\theta_0+\varepsilon h} - m_{\theta_0}] = \frac{1}{2}\varepsilon^2 h^T V_0 h + o(\varepsilon^2)$$

with $V_0 = P_0[\dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T]$

[Thm 7.6 vdV] Suppose

$\forall \theta \in$ open subset of $\mathbb{R}^d$,

we have

$$P_\theta = \frac{dP_\theta}{d\mu}$$

Assume

(i) $\theta \mapsto \sqrt{p_\theta}$ is cont. differentiable $\forall x$,

(ii) The elements in the FIM

$$I(\theta) = \int \frac{\dot{p}_\theta(x) \dot{p}_\theta(x)^T}{[p_\theta(x)]^2} P_\theta(dx)$$

are all well-defined and cont. w.r.t. θ

Then (a) $\theta \longmapsto \sqrt{p_\theta}$ is QMD

(b) $\dot{\ell}_\theta$ in the QMD is $\frac{\dot{p}_\theta}{p_\theta}$