STA 355 Lecture 5

Parametric inference

- Z-estimators

$h(\cdot,\cdot) \in \mathbb{R}^d$
param data.

Suppose $H(\theta) := \mathbb{E}[h(\theta; X)]$

satisfies $H(\theta^*) = 0$ $(\theta \in \mathbb{R}^d)$.

Goal: estimate θ^* (point estimation, CI).

$$\hat{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(\theta; X_i).$$

Try to find θ^* by solving $\hat{H}_n(\theta) = 0$.

e.g. "method of moments".

Given $(\mathbb{P}_\theta : \theta \in \Theta)$, for each $\theta \in \Theta$, $j \in \mathbb{N}_+$

we have moment

$$\alpha_j(\theta) = \mathbb{E}_\theta[X^j].$$

Suppose we have data $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_{\theta^*}$

$$\forall j, \quad \mathbb{E}_{\theta^*}[X^j - \alpha_j(\theta^*)] = 0$$

Let $h(\theta;X) = \begin{bmatrix} X - \alpha_1(\theta) \\ X^2 - \alpha_2(\theta) \\ \vdots \\ X^d - \alpha_d(\theta) \end{bmatrix} \in \mathbb{R}^d$.

$$H(\theta^*) = \mathbb{E}[h(\theta^*;X)] = 0$$

We can estimate θ^* by $\hat{H}_n(\hat{\theta}_n) = 0$

$$\forall j \in \{1, 2, \cdots, d\}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^j = \alpha_j(\hat{\theta}_n).$$

Rmk: does not have to be first d moments.

any set of positive integers could possibly work.

e.g. "M-estimation",

$$\hat{\theta}_n := \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} f(\theta; X_i)$$

where we also define $F(\theta) := \mathbb{E}[f(\theta;X)]$

and $\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} F(\theta)$.

Generally, M-estimators and Z-estimators are not comparable.

When $f(\cdot;X)$ is differentiable $\forall X$.

and $\Theta = \mathbb{R}^d$, then M-estimator

is also a Z-estimator,

$$h(\theta;X) = \nabla_\theta f(\theta;X).$$

(in that case, M-estimator is a Z-estimator,

while Z-estimators may not be M-estimators

eg. first-order saddle point, local minima)

In general, Z-estimators may not exist / satisfy good properties.

But it works "locally".

Thm: Assume $\exists \theta^*$ s.t. $H(\theta^*)=0$.

Assume $h(\theta;X)$ is cts differentiable

at a local neighborhood of θ^*, $\forall X$.

Assume $\nabla_\theta H(\theta^*) \in \mathbb{R}^{d\times d}$ is non-degenerate.

then $\mathbb{P}\left(\exists \hat{\theta}_n, \hat{H}_n(\hat{\theta}_n)=0\right) \to 1$, and

$\hat{\theta}_n \xrightarrow{P} \theta^*$, (Assuming that we search $\hat{\theta}_n$ at sufficiently small neighborhood of θ^*)

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla H(\theta^*)^{-1} \Sigma^* \nabla H(\theta^*)^{-T}\right)$$

where $\Sigma^* = \operatorname{cov}_{\theta^*}\left(h(\theta^*; X)\right)$

$$= \mathbb{E}\left[h(\theta^*; X) \cdot h(\theta^*; X)^T\right].$$

$A^{-T} := (A^{-1})^T$.

(Assuming all the expectations in the formula exist).

Remark:

- Σ^*, $\nabla H(\theta^*)$ can be estimated through empirical data

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} h(\hat{\theta}_n; X_i)\, h(\hat{\theta}_n; X_i)^T$$

$$\hat{J}_n = \frac{1}{n} \sum_{i=1}^{n} \nabla h(\hat{\theta}_n; X_i)$$

Easy to show $\hat{\Sigma}_n \xrightarrow{P} \Sigma^*$

$\hat{J}_n \xrightarrow{P} \nabla H(\theta^*)$.

So we also have

$$\sqrt{n}\left(\hat{J}_n^{-1} \hat{\Sigma}_n \hat{J}_n^{-T}\right)^{-1/2} \left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, I_d)$$

(in 1-d, $\dfrac{\sqrt{n}\,(\hat{\theta}_n - \theta^*) \cdot \hat{J}_n}{\sqrt{\hat{\Sigma}_n}} \xrightarrow{d} \mathcal{N}(0,1)$)

· Alternatively, CI can be built using bootstrap.

Proof of Z-estimator result:

$$0 = \hat{H}_n(\hat{\theta}_n)$$

$$= \underbrace{\hat{H}_n(\theta^*)}_{\frac{1}{n}\sum_{i=1}^n h(\theta^*; X_i)} + \underbrace{\nabla \hat{H}_n(\theta^*)}_{= \frac{1}{n}\sum_{i=1}^n \nabla h(\theta^*; X_i)} \cdot (\hat{\theta}_n - \theta^*) + \underbrace{o(|\hat{\theta}_n - \theta^*|)}_{\text{ignore}}$$

$$\frac{1}{n}\sum_{i=1}^n h(\theta^*; X_i) \begin{cases} \xrightarrow{P} 0 \\ \sqrt{n}\, \hat{H}_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*) \end{cases}$$

$$\xrightarrow{P} \nabla H(\theta^*)$$

invertible matrix.

Putting them together

$$\nabla H(\theta^*) \cdot (\hat{\theta}_n - \theta^*) = -\frac{1}{n}\sum_{i=1}^n h(\theta^*; X_i) + o_p(|\hat{\theta}_n - \theta^*|)$$

So $\sqrt{n} \cdot \nabla H(\theta^*)(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$

Applying $\nabla H(\theta^*)^{-1}$, we get the result.

(Existence of $\hat{\theta}_n$ can be shown e.g. by limit of an iterative procedure).

Maximal Likelihood.

General principle:

$$X \sim \mathbb{P}_{\theta^*}$$ (θ^* does not have to be parametric)

Suppose $\forall \theta \in \Theta$, $\mathbb{P}_\theta$ has a density $p_\theta(X)$

- cts r.v. pdf
- discrete r.v. pmf.

suppose that $p_\theta(x)$ has common support $\forall \theta \in \Theta$.

then we can estimate θ^* by solving

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} p_\theta(X)$$

For this lecture: $\Theta \subseteq \mathbb{R}^d$

$$X_1, X_2, \dots, X_n \overset{iid}{\sim} \mathbb{P}_{\theta^*}.$$

In such a case

$$p_\theta^{(n)}(X_1, X_2, \dots, X_n) = \prod_{i=1}^n p_\theta(X_i).$$

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i)$$

e.g. $X_1, X_2, \dots X_n \overset{iid}{\sim} Ber(p)$.

$$p_p(x) = \begin{cases} p & x=1 \\ 1-p & x=0. \end{cases}$$

$$\hat{p}_n = \underset{p\in[0,1]}{\arg\max} \ \frac{1}{n}\sum_{i=1}^{n}\left(1_{X_i=1}\cdot \log p + 1_{X_i=0}\cdot \log(1-p)\right)$$

$$= \underset{p\in[0,1]}{\arg\max} \left\{ \bar{X}_n \log p + (1-\bar{X}_n)\cdot \log(1-p) \right\}$$

Solve F.O.C. for p, get $\hat{p}_n = \hat{X}_n$.

e.g. $X_1, X_2, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$p_{\mu,\sigma^2}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

$$L(X_1^n) = \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \log(2\pi\sigma^2)\right).$$

$\hat{\mu}_n = \frac{1}{n}\sum_1^n X_i$ (easily from quadratic function)

$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ (by F.O.C. after plugging in $\hat{\mu}_n$)

Applying Z-estimator results. we get consistency & asymptotic normality e.g. $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathrm{Unif}([0,\theta])$.

$$p_\theta(x_i) = \begin{cases} 1/\theta & x_i \in [0,\theta] \\ 0 & x_i \notin [0,\theta]. \end{cases}$$

$$p_\theta^{(n)}(x_1, x_2, \ldots, x_n) = \begin{cases} 1/\theta^n & \theta \geq \max_{1\leq i\leq n} X_i \\ 0 & \theta < \max_{1\leq i\leq n} X_i. \end{cases}$$

MLE exists. $\hat{\theta}_n = \max_{1\leq i\leq n} X_i$.

$$\mathbb{P}(\hat{\theta}_n \leq t) = \mathbb{P}(X_1 \leq t)^n = (t/\theta)^n.$$

$$\mathbb{E}_\theta\left[|\hat{\theta}_n - \theta|^2\right] = O(1/n^2)$$

faster than MLE convergence rate in the smooth case.

Properties of MLE.

Suppose $\log p_\theta(X)$ twice cts differentiable.

and assuming $\mathbb{E}\,|\nabla \log p_\theta(X)|^2 < +\infty$

$\mathbb{E}\,|\nabla^2 \log p_\theta(X)| < +\infty$.

we can then apply Z-estimator results

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} N\left(0, (J^*)^{-1} \Sigma^* (J^*)^{-T}\right)$$

where $\Sigma^* := \mathbb{E}_\theta\left[\nabla_\theta \log p_\theta(X) \nabla_\theta \log p_\theta(X)^T\right]$

$$J^* = \mathbb{E}_\theta\left[\nabla^2 \log p_\theta(X)\right]. \quad \text{(symmetric)}$$

---

score function : $\nabla \log p_\theta(X)$.

- $\mathbb{E}_{\theta^*}\left[\nabla_\theta \log p_{\theta^*}(X)\right] = 0$

(Proof: $\int \nabla_\theta \log p_{\theta^*}(x) \cdot p_{\theta^*}(x)\, dx$

$= \int \nabla_\theta \left(p_{\theta^*}(x)\right) dx = \nabla_\theta (1) = 0.$)

- (Fisher's identity) $J^* = -\Sigma^*$.

Proof: note that

$\forall \theta \in \Theta$, $\int \partial_{\theta_j} \log p_\theta(x) \cdot p_\theta(x)\, dx = 0$

Taking ∂_{θ_k} on both sides, we have

$$\int \partial_{\theta_j}\partial_{\theta_k} \log p_\theta(x) \cdot p_\theta(x)\,dx$$

$$+ \int \left(\partial_{\theta_j} \log p_\theta(x) \cdot \partial_{\theta_k} \log p_\theta(x)\right) \cdot p_\theta(x)\,dx = 0$$

holding true for any $j, k \in \{1, 2, \ldots, d\}$.

Putting together in matrix form, we have

$$-J^* = \left[-\mathbb{E}_\theta\left[\partial_{\theta_j}\partial_{\theta_k} \log p_\theta(X)\right]\right]_{j,k\in[d]} = \left[\mathbb{E}_\theta\left[\partial_{\theta_j} \log p_\theta(X) \cdot \partial_{\theta_k} \log p_\theta(X)\right]\right]_{j,k}$$

$$= \Sigma^*.$$

"Fisher information matrix"

$$I(\theta) = \mathbb{E}_\theta\left[\nabla_\theta \log p_\theta(X)\, \nabla_\theta \log p_\theta(X)^T\right]$$

$$= -\mathbb{E}_\theta\left[\nabla_\theta^2 \log p_\theta(X)\right].$$

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} N\left(0, I(\theta)^{-1}\right).$$

Remark.

- $I(\theta)$ is a measure of information about θ contained in X
- If $I(\theta)$ is not invertible, this theory

does not apply, but MLE may still exist (and it may still be consistent)

- For CI, we just need to estimate $I(\theta)$ (using either one of the representations).

Asymptotic optimality.

Under certain regularity conditions, among all the possible estimators for θ, MLE has the smallest possible MSE, asymptotically

More formally, if $\tilde{\theta}_n$ is another estimator, s.t. $\mathbb{E}_\theta\left[|\tilde{\theta}_n-\theta|^2\right]\cdot n \xrightarrow{n\to+\infty} a(\theta) \quad \forall \theta\in\Theta$.

then for almost all $\theta\in\Theta$ (up to a measure-0 set)

$$a(\theta) \geq I(\theta)^{-1}, \text{ or } tr\left(I(\theta)^{-1}\right) \text{ in the multivariate case}$$

"a partial evidence of this optimality"

Cramér-Rao lower bound.

Thm (CRLB) If $\hat{\theta}_n$ is unbiased,

then $\forall \theta \in \Theta$,

$$\mathrm{var}_\theta(\hat{\theta}_n) \geq \frac{1}{n} \mathrm{tr}(I(\theta)^{-1})$$

(only for unbiased estimators,
for biased estimator, asymptotic LB above)

Proof (1-D for simplicity)

$\forall \theta \in \Theta$ $\displaystyle\int \nabla_\theta \log p_\theta(x) \cdot p_\theta(x) = 0.$ (by previous derivation) (1)

$\displaystyle\int \hat{\theta}_n(x) \cdot p_\theta(x) dx = \theta$ (by unbiasedness) (2)

$\frac{d}{d\theta}(2)$, we get

$$\int \hat{\theta}_n(x) \cdot \nabla_\theta \log p_\theta(x) \cdot p_\theta(x)\, dx = 1 \qquad (3).$$

$(3) - \theta \times (1)$.

$$1 = \int (\hat{\theta}_n(x) - \theta) \cdot \nabla \log p_\theta(x) \cdot p_\theta(x)\, dx$$

$$= \mathbb{E}_\theta[(\hat{\theta}_n - \theta) \cdot (\nabla_\theta \log p_\theta(X))].$$

$$\leq \sqrt{\mathbb{E}_\theta\left[|\hat{\theta}_n - \theta|^2\right]} \cdot \sqrt{\mathbb{E}_\theta |\nabla_\theta \log p_\theta(X)|^2}$$

(Cauchy-Schwarz)

Re-arranging, we get $\mathbb{E}\left[|\hat{\theta}_n - \theta|^2\right] \geq \dfrac{1}{\mathbb{E}_\theta |\nabla \log p_\theta(X)|^2}$

Above derivations are for joint distribution

of $X = (X_1, X_2, \dots X_n)$.

$$I^{(n)}(\theta) = \mathbb{E}_\theta\left[\nabla_\theta \log p_\theta^{(n)}(X_1 \dots X_n) \cdot \nabla_\theta \log p_\theta^{(n)}(X_1 \dots X_n)^T\right]$$

$$= \sum_{i=1}^n \mathbb{E}_\theta\left[\nabla_\theta \log p_\theta(X_i) \cdot \nabla_\theta \log p_\theta(X_i)^T\right]$$

$$= n \cdot I(\theta).$$

So $\mathrm{var}_\theta(\hat{\theta}_n) \geq \dfrac{1}{n I(\theta)}$.

Asymptotic normality of MLE requires

strong assumptions (eg. differentiability, $I(\theta)$ invertible, etc).

We can still have consistency w/o strong conditions.

- Consistency arguments extend beyond parametric models.
- Consistency is "global": do not need to restrict $\hat{\theta}_n$ at a local neighborhood of θ^*.

Thm. Assume that

"ULLN"

(i) $\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) - \mathbb{E}_{\theta^*}[\log p_\theta(X_i)] \right| \xrightarrow{P} 0$ (under P_{θ^*})

(ii) $\forall \varepsilon > 0$, $\sup_{\theta : |\theta - \theta^*| \geq \varepsilon} \mathbb{E}_{\theta^*}[\log p_\theta(X)] < \mathbb{E}_{\theta^*}[\log p_{\theta^*}(X)]$

then MLE is consistent, $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Proof: Notations: $f(\theta; x) := \log p_\theta(x)$.

$$F(\theta) := \mathbb{E}_{\theta^*}[f(\theta; X)].$$

$$\hat{F}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} f(\theta; X_i).$$

Condition (ii) becomes

$$\sup_{\theta : |\theta - \theta^*| > \varepsilon} F(\theta) < F(\theta^*). \quad (*)$$

We only need to study $F(\theta^*) - F(\hat{\theta}_n)$.

$$0 \leq F(\theta^*) - F(\hat{\theta}_n)$$

$$= \underbrace{\left(F(\theta^*) - \hat{F}_n(\theta^*)\right)}_{\leq \sup_{\theta \in \Theta} |\hat{F}_n(\theta) - F(\theta)| \xrightarrow{P} 0} + \underbrace{\left(\hat{F}_n(\theta^*) - \hat{F}_n(\hat{\theta}_n)\right)}_{\leq 0} + \underbrace{\left(\hat{F}_n(\hat{\theta}_n) - F(\hat{\theta}_n)\right)}_{\leq \sup_{\theta \in \Theta} |\hat{F}_n(\theta) - F(\theta)| \xrightarrow{P} 0}$$

So we have

$$F(\hat{\theta}_n) \xrightarrow{P} F(\theta^*).$$

by $(*)$, we have $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Now about the two conditions:

Condition (ii):

If $F(\theta) := \mathbb{E}_{\theta^*}[\log p_\theta(X)]$ is cts in θ

(ii) $\iff$ θ^* uniquely maximizes F.

$$F(\theta) = \underbrace{\mathbb{E}_{\theta^*}[\log p_{\theta^*}(X)]}_{\text{indp of } \theta} - \underbrace{\mathbb{E}_{\theta^*}\left[\log \frac{p_{\theta^*}}{p_\theta}(X)\right]}_{D_{KL}(p_{\theta^*} \| p_\theta)}$$

Properties of KL divergence:

$D_{KL}(P\|Q) \geq 0$, with 0 attained by $P=Q$.