# STA355 Lecture 6

Midterm: 10/17, 5:10pm – 7:40pm

(2 hrs and 30 min)

in the classroom.

- 4 pages (double-sided) cheat sheet
- no electronics.
- Bring your student ID.
- covers first 6 lectures (sufficient statistics)
- Format. (4 questions in total)

  1. True or false

  2.3.4 calculations/proofs/descriptions.

  (show your reasoning).

  (multiple parts).

From last lecture: consistency of M-estimators (incl. MLE).

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

$$F(\theta) := \mathbb{E}[f(\theta; X)], \quad \theta^* := \underset{\theta \in \Theta}{\operatorname{argmin}} \; F(\theta).$$

(for MLE, $f(\theta; x) = -\log p_\theta(x)$).

$$\hat{\theta}_n \xrightarrow{p} \theta^*.$$

Holds true under the following conditions.

(i) (ULLN). $\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{p} 0.$

(ii) $\forall \varepsilon > 0, \; \exists \delta > 0$

s.t. $F(\theta) \geq F(\theta^*) + \varepsilon$

when $\|\theta - \theta^*\| \geq \delta$.

(Equivalently. $\inf_{\theta : \|\theta - \theta^*\| \geq \delta} F(\theta) > F(\theta^*) \quad \forall \delta > 0$).

(ii) Holds true when $F$ is a continuous function, $\Theta$ is compact, and $\theta^*$ is unique minimizer

(for finite-dim models. compactness $\Leftrightarrow$ bounded and closed)

Proof: $F$ is cts function on the domain

$$\Theta \setminus \{\theta : \|\theta - \theta^*\| < \delta\}.$$

So $\inf_{\substack{\|\theta-\theta^*\|\geq\delta \\ \theta\in\Theta}} F(\theta)$ is attained by some $\theta_0$

since $\theta^*$ is unique min, $F(\theta_0) > F(\theta^*)$.

When specialized to MLE:

$$F(\theta) = -\mathbb{E}_{\theta^*}[\log p_\theta(X)]$$

$$= \underbrace{\mathbb{E}_{\theta^*}[\log p_{\theta^*}(X)]}_{\text{const, indp of } \theta} + \underbrace{\mathbb{E}_{\theta^*}\left[\log \frac{p_{\theta^*}(X)}{p_\theta(X)}\right]}_{\text{Kullback-Leibler divergence.}}$$

$$D_{KL}(P\|Q) = \mathbb{E}_P\left[\log \frac{p(X)}{q(X)}\right]$$

where $p, q$ are densities of $P, Q$.

- $D_{KL}(P\|Q) \geq 0 \quad (\forall P, Q)$

$$D_{KL}(P\|Q) \geq \frac{1}{2}\left(\int |p(x) - q(x)|\, dx\right)^2$$

(Pinsker's inequality).

$D_{KL}(P\|Q) = 0$ if and only if $P = Q$.

For MLE

$$F(\theta) = \text{constant} + D_{KL}(\mathbb{P}_{\theta^*} \| \mathbb{P}_\theta)$$

- minimized at $\theta = \theta^*$
- When the model is "identifiable", $\theta^*$ uniquely minimizes $F$.

(If $\theta_1 \neq \theta_2$ then $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$).

On identifiability:

eg. $\mathbb{P}_\theta = \frac{1}{2}\mathcal{N}(\theta, 1) + \frac{1}{2}\mathcal{N}(-\theta, 1)$.

$\theta \in \mathbb{R}$, then it's not identifiable

$$\mathbb{P}_\theta = \mathbb{P}_{-\theta}$$

$\Theta = \{\theta : \theta \geq 0\}$, the model is identifiable.

Still need to show (i)

Thm (Wald) Suppose $\Theta$ is compact (closed and bounded in $\mathbb{R}^d$)

Assuming:

- $\mathbb{E}\left[\sup_{\theta\in\Theta} |f(\theta;X)|\right] < +\infty$ (from DCT)
- $f(\theta;x)$ is a cts function of $\theta$, $\forall x$.

then $\sup_{\theta\in\Theta} |F_n(\theta) - F(\theta)| \xrightarrow{P} 0$.

---

Rmk: . Holds true under very general setting.

$\Theta$ doesn't even need to be finite-dim.

- Intuition:

If $\Theta$ is finite,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |F_n(\theta)-F(\theta)| > \varepsilon\right)$$

$$\leq \sum_{\theta\in\Theta} \mathbb{P}\left(|F_n(\theta)-F(\theta)|>\varepsilon\right) \longrightarrow 0$$

Compactness + continuity allows us to use discrete approximations to $\Theta$.

Implication for MLE:

identifiability + cts log-likelihood + $\mathbb{E}[\sup |\cdot|] < \infty$

$$\Rightarrow \hat{\theta}_n \xrightarrow{p} \theta^*.$$

furthermore, (from last lecture)

under additional smoothness assumptions & $I(\theta^*) > 0$.

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

(since we can show MLE consistency, we don't need to restrict in a local neighborhood of $\theta^*$).

---

Jackknife.

Motivation: in many cases, the estimator satisfies asymptotic expansion

$$\hat{\theta}_n - \theta^* = \frac{1}{n}\sum_{i=1}^{n} \xi_i + \frac{a}{n} + \frac{b}{n^2} + O_p\left(\frac{1}{n^3}\right)$$

$(\xi_i \overset{iid}{\sim}, \quad \mathbb{E}[\xi]=0, \ \mathbb{E}|\xi|^2 < +\infty)$.

$$\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| \simeq O\left(\frac{1}{\sqrt{n}}\right)$$

Jackknife bias estimator

Let $\hat{\theta}_{(-i)}$ be the estimator computed from $(X_j)_{j\neq i}$. $(i=1,2,\cdots n)$

$$\hat{\theta}_{(-i)} - \theta^* = \frac{1}{n-1}\sum_{j\neq i} \xi_j + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O_p\left(\frac{1}{n^3}\right)$$

"Leave-one-out" estimators.

$$\hat{b}_{jack} := \frac{n-1}{n}\sum_{i=1}^{n} \hat{\theta}_{(-i)} - (n-1)\hat{\theta}_n$$

$$= \left(\frac{n-1}{n}\sum_{i=1}^{n} \xi_i + \frac{(n-1)}{n}\cdot n \cdot \frac{a}{n-1} + O_p\left(\frac{1}{n^2}\right)\right) \quad \text{ignore}$$

$$- \left(\frac{n-1}{n}\sum_{i=1}^{n} \xi_i + \frac{n-1}{n}\cdot a + O_p\left(\frac{1}{n^2}\right)\right)$$

$\frac{a}{n}$ difference

$\hat{b}_{jack}$ as estimator for $\frac{a}{n}$.

Jackknife debiasing.

$$\hat{\theta}_{debias} = \hat{\theta}_n - \hat{b}_{jack}.$$

- useful in some high-dim applications.
- expansion holds true only w/ $n \to +\infty$

  for finite $n$, $\hat{\theta}_{debias}$ may still be biased.

---

Computational algorithms.

- Newton's method.

$$\min_{\theta \in \mathbb{R}^d} \{F(\theta)\}.$$

Start from $\theta^{(0)}$, iterate

$$\theta^{(t+1)} = \theta^{(t)} - \nabla^2 F(\theta^{(t)})^{-1} \nabla F(\theta^{(t)})$$

$(t = 0, 1, 2, \cdots)$.

. Local super-exponential convergence

One-step Newton estimator:

Start from $\tilde{\theta}_n$

$$\hat{\theta}_n := \tilde{\theta}_n - \nabla^2 F_n(\tilde{\theta}_n)^{-1} \cdot \nabla F_n(\tilde{\theta}_n).$$

Fact. If $\|\tilde{\theta}_n - \theta^*\| = O_p\left(\frac{1}{\sqrt{n}}\right)$

(this holds true when $\sqrt{n}(\tilde{\theta}_n - \theta^*) \xrightarrow{d} N(0, \Sigma_0)$ for some potentially large $\Sigma_0$).

then $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, H_*^{-1} \Sigma_* H_*^{-1})$

where $H_* = \nabla^2 F(\theta^*)$

$\Sigma_* = \mathrm{cov}(\nabla f(\theta^*, X))$

(the same limiting distribution as M-estimator).

Proof. $\tilde{\theta}_n - \nabla^2 F_n(\tilde{\theta}_n)^{-1} \nabla F_n(\tilde{\theta}_n)$

$\approx \tilde{\theta}_n - \nabla F_n(\theta^*)^{-1} \cdot \left(\nabla F_n(\theta^*) + \nabla^2 F_n(\theta^*) \cdot (\tilde{\theta}_n - \theta^*)\right)$

Cancellation

$\approx \theta^* - \nabla^2 F(\theta^*)^{-1} \cdot \nabla F_n(\theta^*)$

$\sqrt{n}(\cdots) \xrightarrow{d} N(0, H_*^{-1} \Sigma_* H_*^{-1})$.

- SGD.

$$\theta_{t+1} = \theta_t - \alpha_{t+1} \nabla f(\theta_t; X_{t+1})$$

(for $t = 0, 1, 2, \cdots, n-1$)

$(\alpha_t)_{t \geq 1}$ step sequence (tuning parameters)

(Polyak – Juditsky – Ruppert)

$$\hat{\theta}_n := \frac{1}{n} \sum_{t=1}^{n} \theta_t.$$

Fact. under certain regularity conditions (incl convexity)

with suitable stepsize sequence,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, H_*^{-1} \Sigma_* H_*^{-1})$$

(same as M-estimator).

---

Sufficient statistics.

Def. $(P_\theta : \theta \in \Theta)$ is a class of distributions

We call $T(X)$ sufficient if

$\forall \theta$ and $\forall t$, conditional distribution of $X$

under $P_\theta$ conditioned on $T(X)=t$

is indp of $\theta$.

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

$T(X_1, X_2, \cdots, X_n) = X_1 + X_2 + \cdots + X_n \sim \text{Binom}(n, p)$

Conditionally on $T(X_1 \cdots X_n) = t$

$(X_1, \cdots, X_n)$ are indicators of random subset with cardinality $t$, indp of $p$.

---

Thm (factorization)

$T$ is sufficient $\iff$ $\exists\, g_\theta, h$ s.t. $P_\theta(x) = g_\theta(T(x)) \cdot h(x)$ $(\forall \theta, x)$

e.g. $X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$

$$P_\mu(X_1, \cdots, X_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (\mu - X_i)^2\right)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \exp\left(-\mu \cdot \left(\sum_{i=1}^n X_i\right) - \frac{n\mu^2}{2}\right) \cdot \exp\left(-\frac{1}{2}\sum_1^n X_i^2\right)$$

So $T(X_1^n) = \sum_{i=1}^n X_i$ is sufficient.

Proof. "$\Leftarrow$"

$$P_\theta(x \mid T(X)=t) = \frac{g_\theta(t)\cdot h(x)\cdot 1\{T(x)=t\}}{\int_{T(z)=t} g_\theta(t)\cdot h(z)\,dz}$$

indp of $\theta$.

"$\Rightarrow$": Construction: $g_\theta(t) = P_\theta(T(X)=t)$

(density of $T(X)$ when $X \sim P_\theta$)

$$h(x) = P_{\theta_0}(X=x \mid T(X)=t) \left(= P_\theta(X=x \mid T(X)=t)\right.$$

(conditional density of $X$ conditioned on $T(X)=t$ under model $X \sim P_{\theta_0}$).

---

Fact. For any estimator $\hat{\theta}_n$ of $\theta$.

$\exists$ another estimator $\tilde{\theta}_n$ which depends on $X$ only through $T(X)$

such that $\hat{\theta}_n \overset{d}{=} \tilde{\theta}_n$ under $P_\theta$ $(\forall \theta)$.

Proof: $X \mid T(X)$ is indp of $\theta$.

So we can sample from this conditional distribution.

- Sample $\tilde{X} \sim p(x \mid T(X) = t)$
- Compute $\hat{\theta}_n$ on $\tilde{X}$, and let $\tilde{\theta}_n$ be the result.

Improving an estimator using sufficient stats.

. Rao-Blackwellization.

Given estimator $\hat{\theta}_n$ and sufficient stats $T(X)$

$$\hat{\theta}_{RB} := \mathbb{E}\left[\hat{\theta}_n \mid T(X)\right].$$

. $\mathbb{E}_\theta\left[\hat{\theta}_{RB}\right] = \mathbb{E}_\theta\left[\hat{\theta}_n\right]$

(If $\hat{\theta}_n$ is unbiased, so is $\hat{\theta}_{RB}$).

. $\mathrm{Var}_\theta\left(\hat{\theta}_{RB}\right) \leq \mathrm{Var}_\theta\left(\hat{\theta}_n\right)$.

Exponential family.

Suppose $X_1, X_2, \cdots X_n \overset{iid}{\sim} P_\theta$

$$P_\theta(x) = \exp\left( \eta(\theta)^T T(x) - B(\theta) \right) \cdot h(x).$$

For joint distribution

$$P_\theta^{(n)}(x_1, \cdots, x_n) = \exp\left( \eta(\theta)^T \cdot \sum_{i=1}^n T(x_i) - n B(\theta) \right) \cdot \prod_{i=1}^n h(x_i)$$

By factorization thm.

$$\sum_{i=1}^n T(x_i) \text{ is sufficient for } \theta.$$

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

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

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{\mu^2}{2\sigma^2} \right) \cdot \exp\left( \begin{bmatrix} \mu/\sigma^2 \\ -\frac{1}{2\sigma^2} \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} \right).$$

$$T(X_1, X_2, \cdots X_n) = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i^2 \end{bmatrix} \text{ is sufficient}$$