STA3000F Lecture 3

Sufficiency (ct'd).

How to find sufficient stats?

Thm (factorization).

$\{P_\theta : \theta \in \Theta\}$, assuming that $P_\theta = \frac{dP_\theta}{d\mu}$ exists $(\forall \theta)$ (for some base msr μ).

then T is sufficient for the class

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

$g_\theta(T(x))$: generating $T(x)$ based on θ

$h(x)$: Conditionally on T, generate X indep of θ

(Proof assumes $\exists$ joint density).

Proof. "$\Leftarrow$":

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

$$= \frac{h(x)\, 1\{T(x)=t\}}{\int_{T(z)=t} h(z)\, d\mu(z)}$$

(indep of θ).

"$\Rightarrow$" Construction:

marginal density

$$g_\theta(t) = P_\theta(T(X)=t)$$

conditional density.

$$h(x) = P_{\theta_0}(X=x \mid T(X)=t)$$

(for arbitrary $\theta_0 \in \Theta$).

$$P_\theta(x) = P_\theta(T(X)=t,\ X=x) = P_\theta(T(X)=t) \cdot P_\theta(X=x \mid T(X)=t)$$

indep of θ.

(for $t=T(x)$.) $g_\theta = h(x)$.

(See Keener textbook for a rigorous proof).

eg. Exponential family.

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

By factorization, T is sufficient

- $N(\mu, \Sigma)$ in $\mathbb{R}^d$

$$P_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$\propto \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}x\right).$$

$$T(x) = \begin{bmatrix} x \\ \mathrm{vec}(xx^T) \end{bmatrix} \in \mathbb{R}^{d+d^2} \qquad \eta(\mu,\Sigma) = \begin{bmatrix} \Sigma^{-1}\mu \\ \mathrm{vec}(\Sigma^{-1}) \end{bmatrix}.$$

(actually, we only need $d + \frac{d(d+1)}{2}$ dimensions).

- $X \sim N(\mu, I_d)$. $T(X) = X \in \mathbb{R}^d$ gives an exp family
- $X \sim N(0, \Sigma)$ $T(x) = \mathrm{vec}(XX^T)$ gives an exp family
- $X \sim \mathrm{Binom}(n, p)$ with known n.

$$P_p(x) = \binom{n}{x} \cdot p^x (1-p)^{n-x}$$
$$= \binom{n}{x} \cdot \exp\left(x \cdot \log \frac{p}{1-p} + n \log(1-p) \right).$$

$T(X) = X$. $\eta(p) = \log \frac{p}{1-p}$.

$\eta(\theta)$: natural parametrization

Fact. $X_1, X_2, \dots, X_n \overset{iid}{\sim} P_\theta$ where $\{P_\theta : \theta \in \Theta\}$ is an exp family.

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

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

is also an exp family w/ $\sum_{i=1}^n T(x_i)$ sufficient.

Corollary. $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \Sigma)$

then $(X_1, \dots, X_n)$ follows an exp family.

with $$T(X) = \begin{bmatrix} \sum_1^n X_i \\ \sum_{i=1}^n \mathrm{vec}(X_i X_i^T) \end{bmatrix}$$

e.g. $\mathrm{Unif}([\theta-1, \theta+1])$ $\quad \theta \in \mathbb{R}$.

$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} \mathrm{Unif}([\theta-1, \theta+1])$.

$T(X) = (X_{(1)}, X_{(n)})$

where $X_{(1)}, X_{(2)}, \cdots X_{(n)}$

are sorted version of $X_1, \cdots X_n$

$p_\theta(x_1, \cdots x_n) = 2^{-n} \cdot \mathbb{1}\{x_{(1)} \geq \theta-1, x_{(n)} \leq \theta+1\}$.

Use of sufficient stats: unbiased estimation.

Estimate $g(\theta)$ from $X \sim P_\theta$.

Def. δ is unbiased for $g(\theta)$ if

$$\mathbb{E}_\theta[\delta(X)] = g(\theta) \quad (\forall \theta \in \Theta)$$

Def. (UMVU) For any other δ' unbiased.

we have $\mathrm{var}_\theta(\delta(X)) \leq \mathrm{var}_\theta(\delta'(X))$ $(\forall \theta \in \Theta)$

Among all the possible unbiased estimators, δ is the optimal.

Def (complete stats)

We call $T(X)$ is complete for $\{P_\theta : \theta \in \Theta\}$ if

$\mathbb{E}_\theta[f(T(X))] = 0, \ \forall \theta \in \Theta$ implies $f(T) = 0$ a.e.

A system of integral eq. on the function f.

only have trivial solutions.

Idea: no redundancy of info in T.

eg. If $X_1, X_2, \cdots X_n \overset{iid}{\sim} \text{Unif}\left((\theta-1, \theta+1)\right)$.

$T(X_1 \cdots X_n) = (X_1, \cdots, X_n)$.

$f(T) = X_1 - X_2$

But for $T(X_1 \cdots X_n) = (X_{(1)}, X_{(n)})$.

you cannot find such functions.

---

eg. $X_1 \cdots X_n \overset{iid}{\sim} \text{Unif}([0, \theta])$.

$T(X) = \max_{1 \le i \le n} X_i$. $\quad P_\theta(T \le t) = \left(\frac{t}{\theta}\right)^n$ for $t \in [0, \theta]$.

$$\mathbb{E}_\theta\left[f(T(X))\right] = \frac{n}{\theta^n} \int_0^\theta f(t) t^{n-1} dt = 0 \quad (\forall \theta).$$

$$\Rightarrow \int_0^\theta f(t) t^{n-1} dt = 0 \quad \forall \theta$$

So. $f(t) = 0$ almost everywhere.

Def. Exp family $p_\theta(x) = \exp\left(\eta(\theta)^T T(x) - B(\theta)\right) h(x)$

is called full-rank if $\eta(\Theta)$ has an interior point.

$(T_1(X), T_2(X), \cdots, T_d(X))$ linearly indp.

Fact. Full-rank exp family $\Rightarrow T$ is sufficient & complete.

Proof. Let $\eta(\theta_0)$ be an interior point of $\eta(\Theta)$.

Let ν be density of $T(X)$ under θ_0.

(Idea: Local perturbation around θ_0).

$$0 = \int_{\mathbb{R}^d} f(t) \cdot \frac{P_\theta(T(X)=t)}{\nu(t)} \cdot \nu(t)\, dt$$

$$\Rightarrow 0 = \int_{\mathbb{R}^d} f(t) \cdot \exp\left((\eta(\theta) - \eta(\theta_0))^T t\right) \nu(t)\, dt \quad (\forall \theta \in \Theta).$$

So. $\int_{\mathbb{R}^d} f(t)\, e^{\eta^T t} \nu(t)\, dt = 0$ for $\eta \in B(0, r_0)$

multivariate Laplace transform of $f \cdot \nu$.

By uniqueness of Laplace transform.

$f \cdot \nu = 0$ (a.e.).

Thm (Lehmann – Scheffé)

If T is sufficient & complete for $(P_\theta : \theta \in \Theta)$

Suppose that $\exists$ unbiased estimator δ for $g(\theta)$.

then $\delta^* = \mathbb{E}[\delta(X) \mid T]$ is UMVU.

Proof: δ^* is unbiased.

Consider another unbiased estimator δ'.

$$\tilde{\delta}(T) = \mathbb{E}[\delta'(X) \mid T]$$

$\forall \theta$, $\mathrm{var}_\theta(\tilde{\delta}) \le \mathrm{var}(\delta')$.

On the other hand, by completeness.

$$\mathbb{E}_\theta[\tilde{\delta}(T) - \delta^*(T)] = 0 \quad (\forall \theta \in \Theta).$$

So $\tilde{\delta} = \delta^*$ (a.s.).

eg. $X_1, \dots X_n \overset{iid}{\sim} N(\theta, I_d)$. $g(\theta) = \theta$.

$\delta(X) = T(X) = \frac{1}{n}\sum_1^n X_i$ is UMVU.

eg. $X_1, \dots X_n \overset{iid}{\sim} \text{Ber}(\theta)$ $\quad g(\theta) = \theta^k$

$T(X) = \sum_{i=1}^n X_i$.

Unbiased estimator $\quad \delta(X) = X_1 X_2 \cdots X_k$.

UMVU: $\mathbb{E}[\delta(X) \mid T(X)] = \dfrac{T(T-1)\cdots(T-k+1)}{n(n-1)\cdots(n-k+1)}$.

eg. $X_1, \dots X_n \overset{iid}{\sim} N(0, \sigma^2)$ $\quad g(\sigma^2) = \sigma^2$.

$T(X) = \sum_{i=1}^n X_i^2$ $\quad \dfrac{T(X)}{n}$ is UMVU.

Indeed, UMVU is not the "optimal" estimator

$\delta_c(T) = cT$.

$\underset{c}{\text{argmin}}\ MSE(c) = \dfrac{1}{n+2}$.

Additional remarks. (sufficiency, unbiased estimation)

Least useful $\qquad$ most useful

UMVU. $\qquad$ sufficiency $\qquad$ debiasing $\qquad$ Cramer-Rao.

Completeness $\qquad$ Rao-Blackwell. $\qquad$ (approx unbias).

Thm. (Cramér-Rao)

If $\delta(X)$ is unbiased estimator for $g(\theta) \in \mathbb{R}$.

Assuming $(P_\theta : \theta \in \Theta)$ has shared support $(\forall \theta)$.

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

twice cts differentiable

then we have

$$\mathrm{var}_\theta(\delta(X)) \geq \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$$

where $I(\theta) := \mathbb{E}_\theta[\nabla \log p_\theta(X) \cdot \nabla \log p_\theta(X)^T]$.

(usually not achieved by unbiased estimator.

achievable (asymptotically) by MLE plug-in).

Proof. two simple facts. $\ell(\theta; X) = \log p_\theta(X)$.

- $\mathbb{E}_\theta[\nabla_\theta \ell(\theta; X)] = \int \nabla_\theta \ell(\theta; x) \cdot e^{\ell(\theta;x)} \, d\mu(x)$.

$\nabla_\theta e^{\ell(\theta;x)}$

$$= \nabla_\theta \left(\int e^{\ell(\theta;x)} d\mu(x) \right) = 0.$$

$= 1$.

- (Fisher's identity).

Note that $\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_j} \ell(\theta; X)\right] = 0.$

take more derivative.

$$0 = \frac{\partial}{\partial \theta_i} \mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_j} \ell(\theta; X)\right].$$

$$= \frac{\partial}{\partial \theta_i} \int \frac{\partial}{\partial \theta_j} \ell(\theta; x) \cdot e^{\ell(\theta; x)} \, d\mu(x).$$

$$= \int \left(\frac{\partial^2 \ell(\theta; x)}{\partial \theta_j \partial \theta_i} + \frac{\partial \ell(\theta; x)}{\partial \theta_i} \cdot \frac{\partial \ell}{\partial \theta_j}(\theta, x) \right) e^{\ell(\theta; x)} \, d\mu(x)$$

So we have

$$\mathbb{E}_\theta\left[\nabla \ell(\theta; x) \nabla \ell(\theta; x)^T\right] = -\mathbb{E}_\theta\left[\nabla^2 \ell(\theta; x)\right].$$

(holding true only for well specified model at true θ).

Proof of CRLB.

$$\nabla_\theta g(\theta) \overset{\text{(unbias)}}{=} \nabla_\theta \left(\int \delta(x) \cdot e^{\ell(\theta; x)} \, d\mu(x) \right)$$

$$= \int \delta(x) \cdot \nabla_\theta \ell(\theta; x) \cdot e^{\ell(\theta; x)} \, d\mu(x).$$

How to insert $g(\theta)$?

$$\int \nabla_\theta \ell(\theta; x) e^{\ell(\theta; x)} \, d\mu(x) = 0.$$

$g(\theta)$ indp of x.

$$\nabla g(\theta) = \int \big(\delta(x) - g(\theta)\big) \cdot \nabla_\theta \ell(\theta; x) \cdot e^{\ell(\theta;x)} \, d\mu(x)$$

$$= \mathbb{E}_\theta\Big[\big(\delta(x) - g(\theta)\big) \cdot \nabla_\theta \ell(\theta; x) \Big].$$

in 1D. Cauchy–Schwarz gives the result.

in general, consider quadratic form

$$u \in \mathbb{R}^d \mapsto \mathbb{E}_\theta\Big[\big(u^T \nabla_\theta \ell(\theta; x) - (\delta(x) - g(\theta))\big)^2 \Big] \geq 0.$$

Expanding the quadratic form.

$$0 \leq u^T I(\theta) u - 2 u^T \underbrace{\mathbb{E}_\theta\Big[\big(\delta(x) - g(\theta)\big) \cdot \nabla_\theta \ell(\theta; x) \Big]}_{= \nabla g(\theta)} + \mathrm{var}_\theta(\delta(x))$$

This implies $\mathrm{var}_\theta(\delta(x)) \geq \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$

Corollary. $g: \mathbb{R}^d \to \mathbb{R}^k$. δ unbiased.

$$\mathbb{E}\Big[\big(\delta(x) - g(\theta)\big)^{\otimes 2} \Big] \succeq \big(\nabla g(\theta)\big)^T I(\theta)^{-1} \cdot \big(\nabla g(\theta)\big).$$

(PSD)

In particular, $g(\theta) = \theta$

$$\mathbb{E}\Big[\|\delta(x) - \theta\|_2^2 \Big] \geq \mathrm{Tr}\big(I(\theta)^{-1}\big).$$

e.g. $X_1, \cdots X_n \overset{i.i.d}{\sim} P_\theta$

$$\ell_n(\theta; X_1, \cdots X_n) = \sum_{i=1}^{n} \ell(\theta; X_i).$$

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

and CRLB $\dfrac{\left(\nabla g(\theta)\right)^T I(\theta)^{-1} \nabla g(\theta)}{n}$.

(asymptotically achieved by MLE, $n \to +\infty$)

When $P_\theta(x)$ is singular / has discontinuities

CRLB is false.

e.g. $X_1, X_2, \cdots X_n \overset{i.i.d}{\sim} \mathrm{Unif}([0, \theta])$.

UMVU. $\delta(X) = \max_i X_i \cdot \frac{n+1}{n}$.

$\mathrm{Var}_\theta(\delta(X)) = \dfrac{\theta^2}{n(n+2)}$ faster rate.

Bayesian CRLB.

Idea. key step in CRLB

$$\nabla g(\theta) = \int \left(\delta(X) - g(\theta)\right) \nabla \ell(\theta; X) \cdot e^{\ell(\theta; X)} \, d\mu(x).$$

relying on unbiased of δ.

look like "integration-by-parts formula".

use a prior distribution to integrate.

$$r_\pi(\delta) := \int_\Theta \mathbb{E}_\theta\left[\left(\delta(X) - g(\theta)\right)^2\right] \pi(d\theta).$$

Thm (van Trees inequality). for any estimator

$$r_\pi(\delta) \geq \left(\int_\Theta \nabla g(\theta)\, \pi(d\theta)\right)^T \cdot \left(\int_\Theta I(\theta)\pi(d\theta) + J(\pi)\right)^{-1} \cdot \left(\int_\Theta \nabla g(\theta)\, \pi(d\theta)\right)$$

where $J(\pi) := \int_\Theta \nabla_\theta \log \pi(\theta) \nabla_\theta \log \pi(\theta)^T \pi(d\theta)$

"Information theorist's Fisher info".

Compared to CRLB:

- Pay a $J(\pi)$ term
- No need for unbiasedness.

Proof: see next lecture.

eg. $v_0(x) = \cos^2\left(\frac{\pi x}{2}\right) 1_{|x|\leq 1}$ supported on $[-1,1]$.

$$J(v_0) = \pi^2.$$