Slide 4: More About Sufficient Statistics

STATS 511: Statistical Inference

Kean Ming Tan

Curved Exponential Family

Suppose that $X_1, \ldots, X_n \sim N(\mu, \mu^2)$. We have shown that the distribution belongs to an exponential family with

$$T(\mathbf{x}) = \left(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i \right) \qquad \text{and} \qquad w(\boldsymbol{\theta}) = \left(-\frac{1}{2\mu^2}, \frac{1}{\mu} \right).$$

Here, $w(\boldsymbol{\theta})$ forms a curve in the 2-dimensional space. That is, as μ varies, we get a curve in the "xy" plane.

An Example on Non Exponential Family

Suppose that $X_1, \ldots, X_n$ is a random sample from a "right-half" normal distribution with $\sigma^2 = 1$. Find a sufficient statistic for μ.

Minimal Sufficient Statistic

As in previous example, there are many choices of sufficient statistics for the parameter of interest, and of course the "smaller" ones are more useful for data reduction. This motivates the concept on minimal sufficient statistic.

Minimal Sufficient Statistic: $T(\mathbf{X})$ is a minimal sufficient statistic for θ if $T(\mathbf{X})$ is sufficient, and is a function of any other sufficient statistic.

How to Find Minimal Sufficient Statistic (Theorem 6.2.13)

Theorem: Let $f(\mathbf{x} \mid \boldsymbol{\theta})$ be the pdf and pmf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x} \mid \boldsymbol{\theta})/f(\mathbf{y} \mid \boldsymbol{\theta})$ is constant as a function of $\boldsymbol{\theta}$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.

$$f(x|\theta) = h(x)\, c(\theta) \exp\left\{ \sum_{j=1}^{k} w_j(\theta)\, T_j(x) \right\}$$

let $k = 3$

Proof idea: if $T(x)$ is not linearly independent

then $T_1(x) = a_2 T_2(x) + a_3 T_3(x)$

Ratio:
$$\frac{h(x)}{h(y)} \exp\Big\{ w_1(\theta)\big(T_1(x) - T_1(y)\big) + w_2(\theta)\big(T_2(x) - T_2(y)\big) + w_3(\theta)\big(T_3(x) - T_3(y)\big) \Big\}$$

$$\Rightarrow \frac{h(x)}{h(y)} \exp\Big\{ w_1(\theta)\Big(a_2\big(T_2(x) - T_2(y)\big) + a_3\big(T_2(x) - T_2(y)\big)\Big) + w_2(\theta)\big(T_2(x) - T_2(y)\big) + w_3(\theta)\big(T_3(x) - T_3(y)\big) \Big\}$$

$$\Rightarrow \frac{h(x)}{h(y)} \exp\Big\{ \big(w_2(\theta) + a_2 w_1(\theta)\big)\big(T_2(x) - T_2(y)\big) + \big(w_3(\theta) + a_3 w_1(\theta)\big)\big(T_3(x) - T_3(y)\big) \Big\}$$

so, if $w(\theta)$ is not linearly independent, it does not imply $T(x) = T(y)$

An Example: Multinomial Distribution

We have three boxes labelled Box 1, Box 2, and Box 3. We toss n balls into Box 1, Box 2, or Box 3 (n is given). Suppose that Box 1, Box 2, and Box 3 each has probability p_1, p_2, and p_3 of a ball landing in their respective box. Let X_1, X_2, and X_3 be the number of balls that land in Box 1, Box 2, and Box 3 respectively. Then,

$$(X_1, X_2, X_3) \sim \mathit{Multinomial}(\overset{\text{known}}{n}, \underbrace{p_1, p_2, p_3}_{\text{unknown}}).$$

Ignore

Claim: (X_1, X_2) is the minimal sufficient statistic for $\theta = (p_1, p_2, p_3)$.

n = known

p_1, p_2, p_3 = unknown

$p_1 + p_2 + p_3 = 1$

$x_1 + x_2 + x_3 = n$

$$f(x|\theta) = \frac{n!}{x_1!\,x_2!\,x_3!}\, p_1^{x_1}\, p_2^{x_2}\, p_3^{x_3}$$

$$= \frac{n!}{x_1!\,x_2!\,x_3!} \exp\left\{ x_1 \log p_1 + x_2 \log p_2 + x_3 \log p_3 \right\}$$

$$= \frac{n!}{x_1!\,x_2!\,x_3!} \exp\left\{ x_1 \log p_1 + x_2 \log p_2 + (n - x_1 - x_2) \log(1 - p_1 - p_2) \right\}$$

$T(x) = (x_1, x_2, x_3)$ is not minimal sufficient statistic because those are linearly dependent.

Proof: In these situation's we need to rewrite our distribution function's.

Proof

$$f(x|\theta) = \frac{n!}{x_1!\, x_2!\, (n-x_1-x_2)!} \exp\left\{ x_1 \log p_1 + x_2 \log p_2 + (n-x_1-x_2)\log p_3 \right\}$$

$$= \frac{n!}{x_1!\, x_2!\, (n-x_1-x_2)!} \exp\left\{ x_1 \log \frac{p_1}{p_3} + x_2 \log \frac{p_2}{p_3} + n \log p_3 \right\}$$

$$= \frac{n!}{x_1!\, x_2!\, (n-x_1-x_2)!}\, p_3^{\,n} \exp\left\{ x_1 \log \frac{p_1}{p_3} + x_2 \log \frac{p_2}{p_3} \right\}$$

$$T(x) = (x_1, x_2)$$

$$W(\theta) = \left(\log \frac{p_1}{p_2}, \log \frac{p_1}{p_3} \right)$$

$T_1(x) = \sum_{i=1}^{n} X_1$, $T_2(x) = \sum_{i=1}^{n} X_2$ $\Rightarrow$ is Minimal sufficient statistic of (p_1, p_2, p_3)

Ancillary Statistics (Exact opposite of sufficient statistic)

Sufficient statistics contain all the information about $\boldsymbol{\theta}$ that is available in the sample. Now, we consider a statistic $S(\mathbf{X})$ that has no information about $\boldsymbol{\theta}$.

Definition: A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter $\boldsymbol{\theta}$ is called an ancillary statistic.

Examples: $\underbrace{\text{see 6.2.17–6.2.19 in your textbook}}$

$X_1, X_2, \ldots X_n \sim \text{unif}(\theta, \theta+1)$

$R = X_{(n)} - X_{(1)}$ be the range

$f_R(r) = n(n-1) r^{n-2} (1-r)$ does not depend on θ

Complete Statistics

Definition: Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistics $T(\mathbf{X})$. The family of probability distributions is called complete if $E_\theta\{g(T)\} = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. $T(\mathbf{X})$ is also called a complete statistic.

Complete Sufficient Statistic: If $T(\mathbf{X})$ is a complete statistic and a sufficient statistic, then we call $T(\mathbf{X})$ the complete sufficient statistic.

Theorem 6.2.25: Let $X_1, \ldots, X_n$ be iid observations from an exponential family. Then the statistic

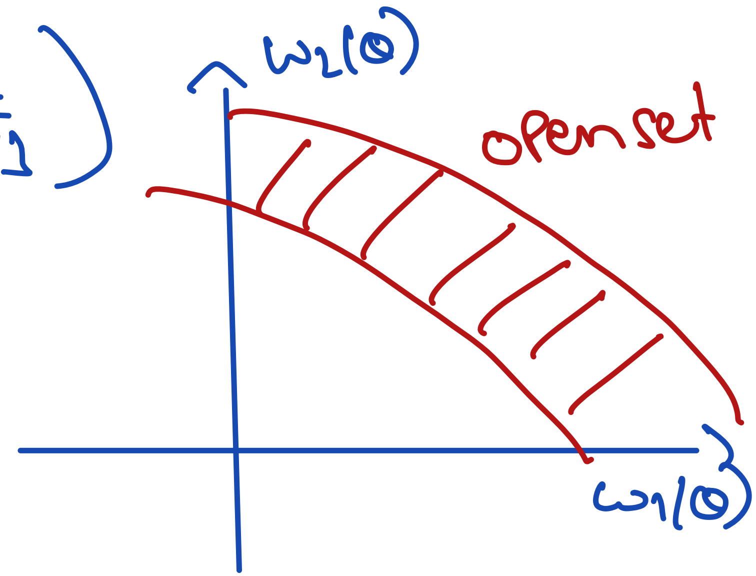
$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_1(X_k) \right)$$

is complete as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

Multinomial and Curved Exponential Family Example

① $w_1(\theta) = \left(\log \frac{p_1}{p_3}, \log \frac{p_2}{p_3}\right)$

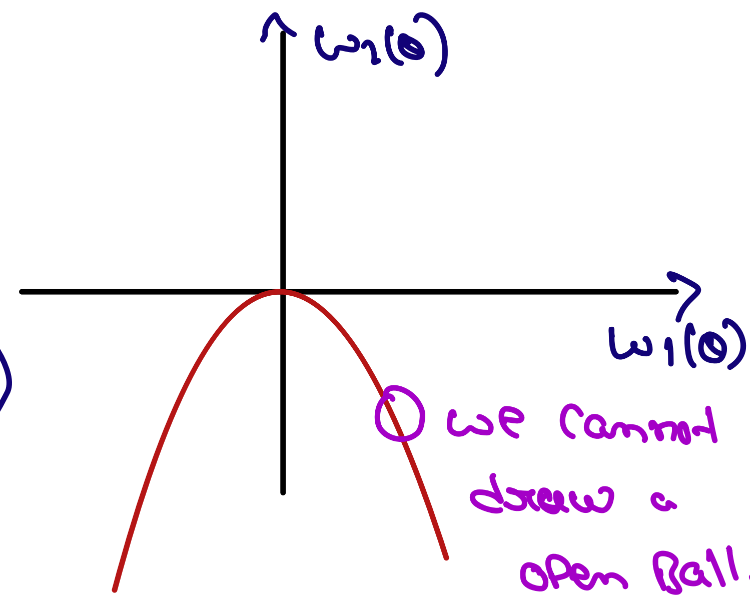
$T(x) = (x_1, x_2)$



② $X_1, X_2, \dots, X_n \sim N(\mu, \mu^2)$

$W(\theta) = \left(-\frac{1}{2\mu^2}, \frac{1}{\mu}\right)$

$T(X) = \left(\sum x_i^2, \sum x_i\right)$



$T_1(x) = \sum x_i^2 \qquad T_2(x) = \sum x_i$

$$\mathbb{E}[T_1(x)] = \mathbb{E}\left[\sum x_i^2\right]$$

$$g(T) = (n+1)T_1 - 2T_2^2$$

find a $g(t)$ s.t $E_\mu[g(T)] = 0$ but

$g(t) \neq 0$

$X \sim N(\mu, \sigma^2)$ $\quad E_\mu[T_1] = 2n\sigma^2$

$E_\mu[T_2^2] = (n^2+n)\sigma^2$

$$E_\mu[g(t)] = (n+1)E[T_1] - 2E[T_2^2]$$

$$= 0$$

But $g(t) \neq 0$ unless all x_i's are zero

$\Rightarrow$ $T(x)$ is not a C.S.S for μ

Quiz type Problem:

$(X_1, X_2, X_3) \sim$ Mutinomial (n, p_1, p_2, p_1+p_2)

Non-Exponential Family Example

Let $X_1, \ldots, X_n \sim \mathit{Unif}(0, \theta)$ be iid Uniform random variables. We know that $T(\mathbf{X}) = X_{(n)}$ is sufficient statistic for θ. Is $T(\mathbf{X})$ complete sufficient statistic?

location- scale family

① let $f_\theta(x)$ be some pdf

$$f_\theta(x) = f_\theta(x-\theta)$$

location parameter

Ex: $X \sim N(\mu,1)$

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

② Scale family

$$f_\sigma(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

Ex: $X \sim EXP(\theta)$ $\quad f_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad 0 < x < \infty$

① if X is location family, difference of two X's will remove the location parameter

Ex: $X_1, X_2, \ldots, X_n \sim N(\mu,1)$

$X_1 - X_2$ is ancillary

$R := X_{(n)} - X_{(1)}$ is ancillary

② Scale family

The ratio of any two x remove the scale parameter

Ex: $x_1, x_2 \sim N(0, \sigma^2)$ scale family

$S(x) = \frac{x_1}{x_2}$ is this ancillary?

let $Y = \frac{x}{\sigma}$ the $Y \sim N(0,1)$

so $S(Y) = \frac{\sigma Y_1}{\sigma Y_2} = \frac{Y_1}{Y_2}$

another example: (Both location & scale family)

$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$

ancillary statistic $\frac{x_1 - \mu}{\sigma} \sim N(0,1)$

$$\frac{x_n - x_1}{x_2 - x_1}, \quad \frac{x_{(n)} - x_{(1)}}{x_1 - \text{median}(x)}$$

Applications of Ancillary and Complete Sufficient Statistic

Basu's Theorem: If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistics $S(\mathbf{X})$.

complete sufficient statistic

Example: $X_1, \ldots, X_n \sim Uniform(0, \theta]$. Show that $T(\mathbf{X}) = \underbrace{X_{(n)}}_{C.S.S}$ and $S(\mathbf{X}) = \underbrace{(X_{(n)} - X_{(1)})/(X_{[2n/3]} - X_{[n/3]})}_{\text{does not depend on } \theta}$.

$Y = \frac{X}{\theta} \sim \text{uniform}(0,1)$

$$S(Y) = \frac{\frac{Y_{(n)}}{\theta} - \frac{Y_{(1)}}{\theta}}{\frac{Y_{[2n/3]}}{\theta} - \frac{Y_{[n/3]}}{\theta}} = \frac{Y_{(n)} - Y_{(1)}}{Y_{(\frac{2n}{3})} - Y_{(\frac{n}{3})}}$$

(ancillary)

$T(X)$ is C.S.S

$S(X)$ is ancillary $\Longrightarrow$

$\Big\downarrow$ Basu's theorem

$T(X), S(X)$ are independent

Quiz Type Problem:

$(X_1, X_2, X_3) \sim \text{Multinomial}(n, P_1, P_2, P_1+P_2)$

(a) what's the range of P_1

(b) Find a M.S.S for (P_1, P_2)

(c) Is T is Complete

---

(1) $P_1+P_2+P_1+P_2=1 \implies P_1+P_2=\frac{1}{2}$

$\implies P_2=\frac{1}{2}-P_1 \qquad 0<P_1<\frac{1}{2}$

$f(\vec{x})= h(\vec{x})\, p_1^{x_1}\left(\frac{1}{2}-P_1\right)^{x_2}\left(\frac{1}{2}\right)^{n-x_1-x_2}$

$= \tilde{h}(x)\, P_1^{x_1}\left(\frac{1}{2}-P_1\right)^{x_2}$

$= \tilde{h}(x) \exp\left\{x_1 \log P_1 + x_2 \log\left(\frac{1}{2}-P_2\right)\right\}$

$T(\vec{x})=(X_1, X_2) \qquad W(\theta)=\left(\underset{\theta_1}{\log P_1},\ \underset{\theta_2}{\log\left(\frac{1}{2}-P_1\right)}\right)$

Both $\log P_1$ and $\log\left(\frac{1}{2}-P_1\right)$ is not linearly independent. (NONlinear dependent but we only check for linear dependence)

$\implies W(\theta) \in \mathbb{R}^2$

can't apply our exponential theorem because we only have one free parameter, (we cannot have a open subset $W(\theta) \implies T(x)$ is not C.S.S

To prove T is not complete, we WTS

$\exists$ a $g(\cdot)$ s.t $\mathbb{E}[g(T)]$ but $g(t) \neq 0$

$T(x) = (x_1, x_2)$ $\quad (x_1, x_2, x_3) \sim \text{multinomial}(n, p_1, \frac{1}{2}p_1, \frac{1}{2})$

$E[x_1] = np_1$

$E[x_2] = \frac{n}{2} - np_1$

$\Rightarrow$ let

$g(t) = x_1 + x_2 - \frac{n}{2}$

$\Rightarrow E[g(t)] = 0$

But $g(t) \neq 0 \quad \forall t \in \mathcal{T}$

$\Rightarrow$ Therefore the $T(x)$ is not complete.

<u>Another Example:</u>

$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$

known

Show that

$\bar{X} \perp S_n^2 \qquad S_n^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

---

we show that $T(x) = \bar{X}$ is c.s.s for ie. By Basu's theorem ; $\bar{X} \perp\!\!\!\perp$ of any ancillary statistics.

Check if S_n^2 is ancillary

$\frac{1}{n}\sum(x_i-\bar{x})^2$ does not depend on

ie $\Longrightarrow$ S_n^2 is ancillary

$$\Longrightarrow \underset{(C.S.S)}{\bar{X}} \perp\!\!\!\perp \underset{(\text{ancillary})}{S_n^2}$$

Because of Basu's theorem.

<u>A non-homogenous Poission Process</u>

$\{X(t) \mid 0<t<\infty\}$ $\quad X(0)=0 \quad X(t)=\lambda t$

$\lambda>0$

$$X(t_{i+1}) - X(t_i) \sim \text{Poission}\left(\int_{t_i}^{t_{i+1}} \lambda t \, dt\right)$$

$$X(t) \perp\!\!\!\perp [X(t_2)-X(t_1)] \perp\!\!\!\perp \cdots \perp\!\!\!\perp X(t_n)-X(t_{n-1})$$

Let say we observe $X(1), X(2), \ldots, X(n)$

Find a C.S.S for λ

$Y \sim \text{Poission}$ $\qquad f(y) = \dfrac{e^{-\lambda}\lambda^{y}}{y!}$

Define $\quad Y_1 = X(1)$

$Y_2 = X(2)$

$\vdots$

$Y_n = X(n) - X(n-1)$

By * $\quad Y_1, Y_2, \ldots, Y_n$ are independent

By ** $\quad Y_i \sim \text{Poission}\left(\int_i^{i+1} \lambda t \, dt\right)$

$\sim \text{poission}\left(\lambda\left(i-\frac{1}{2}\right)\right)$

$$f(y_1, y_2, \ldots y_n \mid \lambda) = \prod_{i=1}^{n} \left[\frac{e^{-\lambda(i-\frac{1}{2})} \left[\lambda\left(i-\frac{1}{2}\right)\right]^{y_i}}{y_i!}\right]$$

$$= \frac{e^{\frac{n}{2}}}{y_1! y_2! \ldots y_n!} \cdot e^{-\lambda \frac{i(i-1)}{2}} \cdot \prod_{i=1}^{n} \left(\lambda\left(i-\frac{1}{2}\right)\right)^{y_i}$$

$$= h(y) \cdot c(\lambda) \prod_{i=1}^{n} \lambda^{y_i} \left(i-\frac{1}{2}\right)^{y_i}$$

$$= h(y) \cdot c(\lambda) \prod_{i=1}^{n} \exp\left\{y_i \log \lambda + \underbrace{y_i \log\left(i-\frac{1}{2}\right)}\right\}$$

$$= \tilde{h}(y) \cdot C(\lambda) \quad \exp\left\{ \log\lambda \sum y_i \right\}$$

$\Rightarrow \quad W(\lambda) = \log\lambda \qquad T(Y) = \sum Y_i$

$\Downarrow$

we can draw open set

$\Rightarrow \quad T(Y)$ is C.S.S

Up Next - Methods for Estimation