Sufficiency

Outline

1) Review
2) Sufficiency
3) Factorization Theorem

Sufficiency

Motivation: Coin flipping

Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$

$$\Rightarrow X \sim \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \quad \text{on} \quad \{0,1\}^n$$

Then $T(X) = \sum X_i \sim \text{Binom}(n, \theta)$

$$= \theta^t (1-\theta)^{n-t} \binom{n}{t} \quad \text{on} \quad \{0, \dots, n\}$$

$(X_1, \dots, X_n) \rightsquigarrow T(X)$ is throwing away data. How do we justify this?

In exp. fam. lingo, $T(X)$ is the "sufficient statistic" for X. Today we'll see why we call it that.

Definition Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for data X. $T(X)$ is __sufficient__ for $\mathcal{P}$ if $P_\theta(X \mid T)$ does not depend on θ

Example (Cont'd)

$$\mathbb{P}_\theta(X = x \mid T = t) = \frac{\mathbb{P}_\theta(X = x, T = t)}{\mathbb{P}_\theta(T = t)}$$

$$= \frac{\cancel{\theta^{\sum x_i}} \cancel{(1-\theta)^{n-\sum x_i}} \; 1\{\sum x_i = t\}}{\cancel{\theta^t} \cancel{(1-\theta)^{n-t}} \binom{n}{t}}$$

$$= 1\{\sum x_i = t\} / \binom{n}{t}$$

So given $T(X) = t$, X is uniform on all seq.s with $\sum x_i = t$

Interpretations of Sufficiency

Recall we only care about X in the first place because it is (indirectly) informative about θ

Sufficiency means only $T(x)$ is informative

We can think of the data as being generated in two stages:

"Pick a slice of $\mathcal{X}$" 1) Generate T : distribution dep. on θ

"Generate within the slice" 2) Generate $X \mid T$: does <u>not</u> dep on θ

Dist. over slices depends on θ, not dist. within each slice.

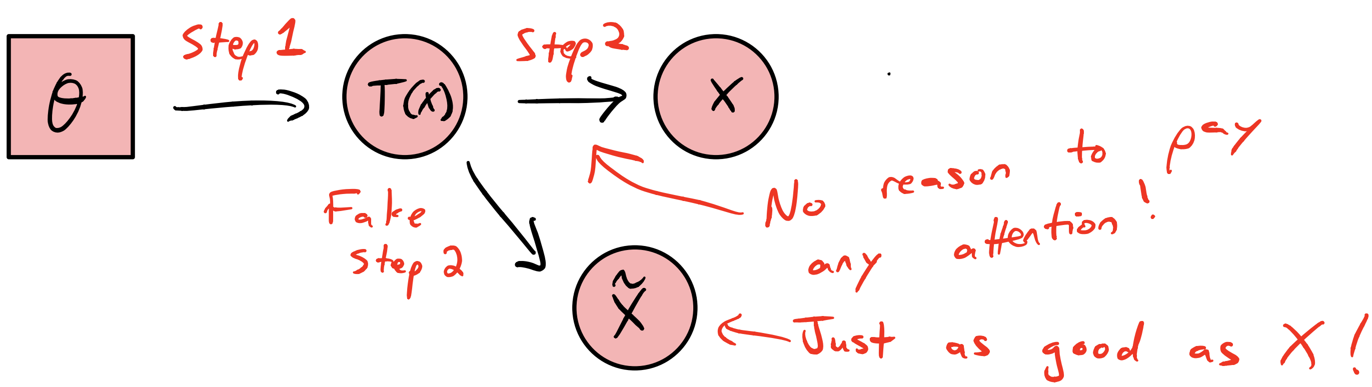
Sufficiency Principle

If $T(X)$ is sufficient for $\mathcal{P}$ then any statistical procedure should depend on X only through $T(X)$

In fact, we could throw away X and generate a new $\tilde{X} \sim P(X \mid T)$ and it would be just as good as X

no θ

In graphical model form:



Factorization Theorem

There is a very convenient way to verify sufficiency of a statistic based only on the density:

Theorem (Factorization Theorem)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of distributions dominated by μ $(P_\theta \ll \mu \;\forall \theta)$, densities p_θ.

T is sufficient for $\mathcal{P}$ iff there exist non-neg. functions g_θ, h such that

$$p_\theta(x) = g_\theta(T(x))\, h(x) \quad \text{for a.e. } x \text{ under } \mu$$

$$[\mu(\{x : p_\theta \neq g_\theta(T(x)) \cdot h(x)\}) = 0]$$

Rigorous proof in Keener 6.4

"Physics proof": (rigorous for discrete $\mathcal{X}$)

$(\Leftarrow)$ $p_\theta(x \mid T=t) = 1\{T(x)=t\} \cdot \dfrac{\cancel{g_\theta(t)}\, h(x)}{\int_{T(z)=t} \cancel{g_\theta(t)}\, h(z)\, d\mu(z)}$

$(\Rightarrow)$ Take $g_\theta(t) = \int_{T(x)=t} p_\theta(x)\, d\mu(x) = \mathbb{P}_\theta(T(X)=t)$

$$h(x) = p_{\theta_0}(x) \Big/ \int_{T(z)=t} p_{\theta_0}(z)\, d\mu(z)$$

$$= \mathbb{P}_{\cancel{\theta_0}}(X = x \mid T(X) = T(x))$$

$$g_\theta(T(x))\, h(x) = \mathbb{P}_\theta(T = T(x))\, \mathbb{P}(X = x \mid T = T(x)) \quad \boxtimes$$

Examples

Ex. Exponential Families

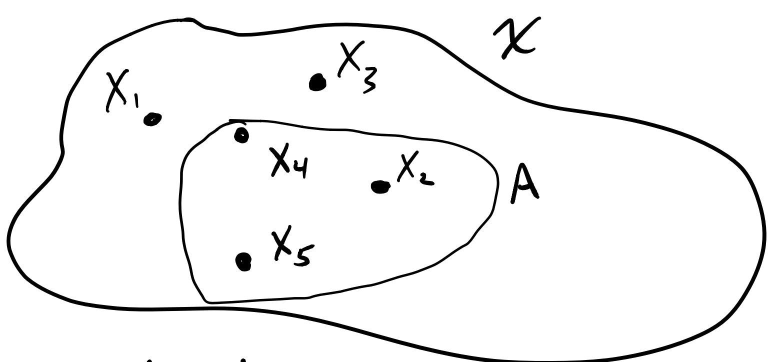
$$p_\theta(x) = \underbrace{e^{\eta(\theta)' T(x) - B(\theta)}}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)}$$

Ex. $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta^{(1)}$ for <u>any</u> model $\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ on $\mathcal{X} \subseteq \mathbb{R}$

P_θ is invariant to perm.s of $X = (X_1, \ldots, X_n)$

$\Rightarrow$ <u>order statistics</u> $(X_{(i)})_{i=1}^n$ $(X_{(k)} = k^{th}$ smallest$)$ are sufficient. [Note $(X_i)_{i=1}^n \rightsquigarrow (X_{(i)})_{i=1}^n$ loses information, specifically the orig. ordering]

For more general $\mathcal{X}$ we can say the <u>empirical distribution</u> $\hat{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot)$ is sufficient, where $\delta_{X_i}(A) = 1\{X_i \in A\}$



$$\hat{P}_n(A) = \frac{3}{5}$$

Not important that it's a measure in this context; just keeps track of which values came up how many times

Ex. $X_1, \dots, X_n \overset{iid}{\sim} U[\theta, \theta+1]$

$$= 1\{\theta \le x \le \theta + 1\}$$

$$p_\theta(x) = \prod_{i=1}^{n} 1\{\theta \le x_i \le \theta + 1\}$$

$$= 1\{\theta \le X_{(1)}\}\, 1\{X_{(n)} \le \theta + 1\}$$

$\Rightarrow (X_{(1)}, X_{(n)})$ is sufficient.

Minimal Sufficiency

Consider $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$:

$T(x) = \sum X_i$ sufficient

$\bar{X} = \frac{1}{n} \sum X_i$ also

$S(x) = (X_{(1)}, \dots, X_{(n)})$ too

$X = (X_1, \dots, X_n)$ too

Which can be recovered from which others?

$$\begin{tikzcd}
 & X \arrow[d] \arrow[ddl, bend right] \arrow[ddr, bend left] & \\
 & S(x) \arrow[dl] \arrow[dr] & \\
\sum X_i \arrow[rr, leftrightarrow] & & \bar{X}
\end{tikzcd}$$

these can be compressed further

These are the most compressed. Are they as compressed as possible?

Prop If $T(X)$ is sufficient and $T(x) = f(S(x))$ then $S(X)$ is sufficient

Proof:
$$\begin{aligned} p_\theta(x) &= g_\theta(T(x))\, h(x) \\ &= (g_\theta \circ f)(S(x))\, h(x) \end{aligned}$$
$\boxtimes$

Definition: $T(X)$ is minimal sufficient if

1) $T(X)$ is sufficient
2) For any other sufficient $S(X)$,
$$T(X) = f(S(X)) \text{ for some } f$$
(a.s. in $\mathcal{P}$)

So, no matter how many more suff. stats we add to our diagram, they will all have arrows pointing to ΣX_i

How to check minimal sufficiency? Basically, "equivalent to knowing likelihood ratios"

Definition Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ has densities $p_\theta(x)$ wrt common μ, data X. The (log) likelihood function is the (log) density, reframed as a random function of θ:

$$\mathrm{Lik}(\theta; X) = p_\theta(x), \qquad \ell(\theta; X) = \log \mathrm{Lik}(\theta; X)$$

Note if $T(X)$ is sufficient then

$$\text{Lik}(\theta; x) = \underbrace{g_\theta(T(x))}_{\substack{T \text{ determines the} \\ \text{"shape"}}} \underbrace{h(x)}_{\text{scaling}}$$

Thm 3.11 Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, densities p_θ

$T(X)$ sufficient for X.

If $\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y) \Rightarrow T(x) = T(y)$

then $T(X)$ is minimal sufficient

"T determines the likelihood shape in a one-to-one fashion"

Proof Suppose S is sufficient and $\nexists f$

s.t. $f(S(X)) = T(X)$

Then $\exists x, y$ with $S(x) = S(y)$, $T(x) \neq T(y)$

$$\begin{aligned} \text{Lik}(\theta; x) &= g_\theta(S(x)) h(x) \\ &\propto_\theta g_\theta(S(y)) h(y) \\ &= \text{Lik}(\theta; y) \end{aligned}$$

But that implies $T(x) = T(y)$ by assumption.

Ex. $$p_\theta(x) = e^{\eta(\theta)'T(x) - \underbrace{B(\theta)}_{\text{doesn't change with } x}} \underbrace{h(x)}_{\text{const. in } \theta}$$

Is $T(x)$ minimal?

Assume $\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y)$. WTS $T(x) = T(y)$

$$\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y)$$

$$\iff e^{\eta(\theta)'T(x)} = e^{\eta(\theta)'T(y)} c(x,y) \quad \forall \theta$$

$$\iff \eta(\theta)'T(x) = \eta(\theta)'T(y) + a(x,y) \quad \forall \theta$$

$$\iff \big(\eta(\theta_1) - \eta(\theta_2)\big)'T(x) = \big(\eta(\theta_1) - \eta(\theta_2)\big)'T(y) \quad \forall \theta_1, \theta_2$$

$$\iff \eta(\theta_1) - \eta(\theta_2) \perp T(x) - T(y) \quad \forall \theta_1, \theta_2$$

$$\iff T(x) - T(y) \perp \text{Span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\}$$

We were trying to show $T(x) - T(y) = 0$, not quite there yet.

If $\text{Span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s$, we are done.

Otherwise, $T(x)$ really might not be minimal!

e.g., $\eta(\theta) = \begin{pmatrix} \theta \\ 0 \end{pmatrix}$: then $T_1(x)$ sufficient

[could $T(x)$ still be minimal?]

Ex. $X \sim N_2(\mu(\theta), I_2) \qquad \theta \in \mathbb{R}$

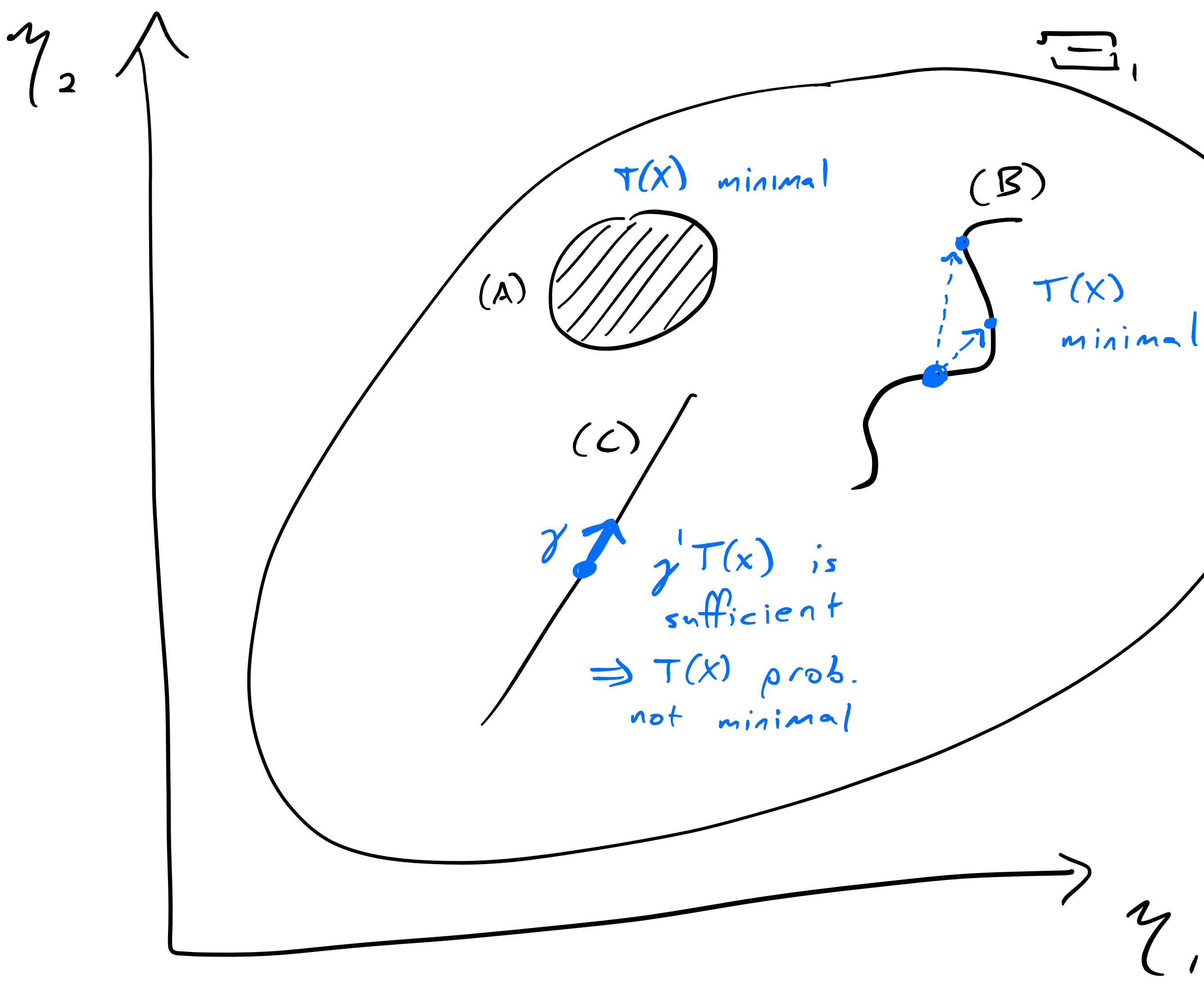
$$= e^{\mu(\theta)'x - B(\theta)} \, e^{-\frac{1}{2}x'x}$$

If $\Theta = \mathbb{R}$, $\quad \mu(\theta) = a + \theta b$ for $a, b \in \mathbb{R}^2$

$$p_\theta(x) = e^{(a+\theta b)'x - B(\theta)} \, e^{-\frac{1}{2}x'x}$$

$$= e^{\theta(b'x) - B(\theta)} \, e^{-\frac{1}{2}(x-2a)'x}$$

$b'X$ is sufficient $\Rightarrow$ X not minimal



Ex Laplace location family

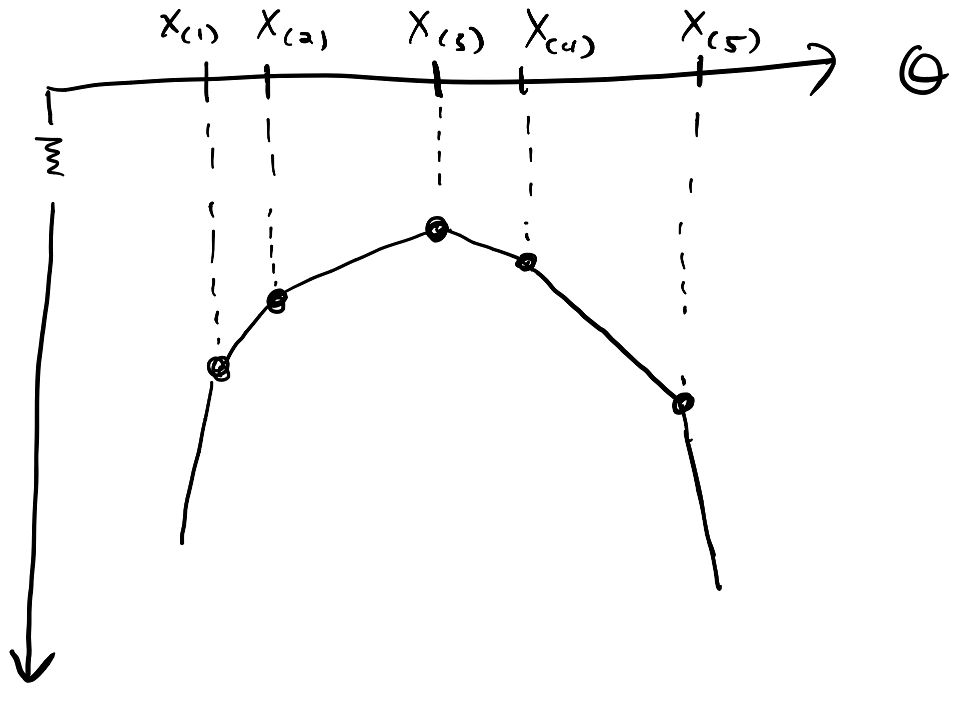
$$X_1, \ldots, X_n \overset{iid}{\sim} p_\theta^{(1)}(x) = \frac{1}{2} e^{-|x-\theta|}$$

$$p_\theta(x) = \frac{1}{2^n} \exp\left\{ -\sum_{i=1}^n |x_i - \theta| \right\}$$

$$\ell(\theta; x) = \log p_\theta(x)$$

$$= -\sum_{i=1}^n |x_i - \theta| - n \log 2$$

Piecewise linear in θ, knots at $x_{(i)}$



$$\ell(\theta; x) = \ell(\theta; y) + \text{const} \Leftrightarrow x, y \text{ same order statistics}$$

Thm 3.11 $\Rightarrow$ order stats are minimal suff.