Sufficiency

Outline

1) Review
2) Sufficiency
3) Factorization Theorem

Review

Exp. Fam : \( \rho_\theta(x) = e^{\gamma(\theta)'x} - B(\theta) h(x) \)

Canonical form : \( \rho_\eta(x) = e^{\gamma(x)'x} - A(\gamma) h(x) \)

If \( X_1, \ldots, X_n \sim \text{iid exp. fam.} , (X_1, \ldots, X_n) \sim \text{exp. fam} \)

some nat. param. , suff. stat. \( \Xi = T(x) \)

If \( X \sim e^{\gamma(x)'x} - A(\eta) h(x) \) then \( T(x) \sim e^{\gamma(x)'x} - A(\eta) h_T(x) \)

where \( \int_B h^T(t) \, d\mu^T(t) = \int h(x) \, d\mu(x) \)

Differentiating \( e^{A(\eta)} = \int e^{\gamma(x)'x} \, d\mu(x) \) \( T(x) \)

Once \( \Rightarrow \nabla A(\eta) = E_x T(x) \)

Twice \( \Rightarrow \nabla^2 A(\eta) = \text{Var}_x T(x) \) on \( \Xi^0 \)
Sufficiency

Motivation: Coin flipping

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

\[
\Rightarrow X \sim \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \quad \text{on } \{0,1\}^n
\]

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

\[
= \theta^t(1-\theta)^{n-t}(n^t) \quad \text{on } \{0,\ldots,n\}
\]

\((X_1, \ldots, X_n) \rightarrow T(X)\) 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 \( \theta \)

Example (Cont'd)

\[
P_\theta(X = x \mid T = t) = \frac{P_\theta(X = x, T = t)}{P_\theta(T = t)}
\]

\[
= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} 1\{\sum x_i = t\}}{\theta^t(1-\theta)^{n-t}(n^t)}
\]

\[
= \frac{1\{\sum x_i = t\}}{\binom{n}{t}}
\]

So given \( T(X) = t \), \( X \) is uniform on all seqs with \( \sum x_i = t \)
Interpretations of Sufficiency

Recall we only care about $X$ in the first place because it is (indirectly) informative about $\Theta$. Sufficiency means only $T(X)$ is informative.

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

1) Generate $T$ : distribution dep on $\Theta$
2) Generate $X \mid T$ : does not dep on $\Theta$

Sufficiency Principle

If $T(X)$ is sufficient for $P$ then any statistical procedure should depend on $X$ only through $T(X)$.

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

In graphical model form:

\[
\Theta \rightarrow T(X) \rightarrow X
\]

Fake step 2

No reason to pay any attention!

\[
\sim \hat{X}
\]

Just as good as $X$!
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 \( \mu \) (\( P_\theta \ll \mu \Theta \)) densities \( P_\theta \).

\( T \) is sufficient for \( \mathcal{P} \) iff there exist non-neg. functions \( g_\theta \), \( h \) such that

\[
\rho_\theta(x) = g_\theta(T(x)) h(x) \quad \text{for a.e.} \ x \ \text{under} \ \mu
\]

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

Rigorous proof in Keener 6.4

"Physics proof": (rigorous for discrete \( X \))

\[
(\Leftarrow) \quad \rho_\theta(x \mid T = t) = 1\{ T(x) = t \} \cdot \frac{g_\theta(x) \cdot h(x)}{\int_{T(z) = t} g_\theta(z) h(z) \, d\mu(z)}
\]

\[
(\Rightarrow) \quad \text{Take} \ g_\theta(t) = \int_{T(x) = t} \rho_\theta(x) \, d\mu(x) = P_\theta(T(X) = t)
\]

\[
h(x) = \frac{\rho_\theta(x)}{\int_{T(z) = t} \rho_\theta(z) \, d\mu(z)} = \frac{P_\theta(X = x \mid T(X) = T(x))}{\int_{T(x) = t} P_\theta(T = T(x)) \, d\mu(X = x \mid T = T(x))}
\]
Examples

Ex. Exponential Families

\[ p_\theta(x) = e^{\gamma(\theta', \tau(x)) - B(\theta)} \frac{g_\theta(T(x))}{h(x)} \]

Ex. \( X_1, \ldots, X_n \overset{iid}{\sim} p_\theta^{(1)} \) for any model

\[ P^{(1)} = \{ p_\theta^{(1)} : \theta \in \Theta \} \] on \( X \subseteq \mathbb{R} \)

\( p_\theta \) is invariant to perms of \( X = (X_1, \ldots, X_n) \)

\[ \Rightarrow \text{ order statistics } (X_{(i)})_{i=1}^n \quad (X_k = k^{th} \text{ smallest}) \]

are sufficient. \[ \text{[Note } (X_i)_{i=1}^n \sim (X_{(i)})_{i=1}^n \text{ loses information, specifically the orig. ordering]} \]

For more general \( X \) we can say the empirical distribution

\[ \hat{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot) \]

is sufficient, where \( \delta_{X_i}(A) = \mathbb{1}\{x_i \in A\} \)

Not important that it's a measure in this context, just keeps track of which values came up how many times
\[ \begin{align*}
\Pr &. \quad X_1, \ldots, X_n \overset{iid}{=} U[\theta, \theta+1] \\
&= 1\{ \theta \leq x \leq \theta+1 \} \\
\rho_\theta(x) &= \prod_{i=1}^{n} 1\{ \theta \leq x_i \leq \theta+1 \} \\
&= 1\{ \theta \leq X_{(1)} \} \cdot 1\{ X_{(n)} \leq \theta+1 \} \\
\Rightarrow \quad (X_{(1)}, X_{(n)}) \text{ is sufficient.}
\end{align*} \]

**Minimal Sufficiency**

Consider \( X_1, \ldots, X_n \overset{iid}{=} N(\theta, 1) \):

\[ \begin{align*}
T(x) &= \Sigma X_i \text{ sufficient} \\
\bar{X} &= \frac{1}{n} \Sigma X_i \text{ also} \\
S(x) &= (X_{(1)}, \ldots, X_{(n)}) \text{ too} \\
X &= (X_1, \ldots, X_n) \text{ too}
\end{align*} \]

Which can be recovered from which others?

Which of 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: $p_{\theta}(x) = g_{\theta}(T(x)) \cdot h(x)$

$$= (g_{\theta} \circ f)(S(x)) \cdot h(x) \quad \Box$$

**Definition:** $T(X)$ is [minimal sufficient](#) if

1) $T(X)$ is sufficient

2) For any other sufficient $S(X)$, $T(X) = f(S(X))$ 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 $S(X)$.

**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 $\mu$, data $X$. The (log) likelihood function is the (log) density, rephrased as a random function of $\Theta$:

$$\text{Lik}(\theta; X) = p_{\theta}(x), \quad l(\theta; x) = \log \text{Lik}(\theta; x)$$
Note if $T(X)$ is sufficient then

$$\text{Lik}(\theta; x) = \frac{g_\theta(T(x))}{T \text{ determines the scaling}} h(x)$$

**Theorem 3.11** Assume $P = \{P_\theta : \theta \in \Theta\}$, densities $P_\theta$

$T(X)$ sufficient for $X$.

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

then $T(X)$ is minimal sufficient

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

**Proof** Suppose $S$ is sufficient and $\not\exists$

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

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

$$\text{Lik}(\theta; x) = \frac{g_\theta(S(x))}{g_\theta(S(y))} h(x)$$

$$\propto \frac{g_\theta(S(y))}{g_\theta(S(y))} h(y)$$

$$= \text{Lik}(\theta; y)$$

But that implies $T(x) = T(y)$ by assumption.
\[ \text{Ex. } \rho_\theta(x) = \mathbb{e}^{T(x)'}T(x) - B(x)h(x) \]

Is \( T(x) \) minimal? 

Assume \( \text{Lik}(\theta; x) \propto \theta \text{Lik}(\theta; y) \), WTS \( T(x) = T(y) \)

\[ \text{Lik}(\theta; x) \propto \theta \text{Lik}(\theta; y) \]

\[ \iff \mathbb{e}^{T(x)'}T(x) = \mathbb{e}^{T(y)'}T(y) \quad \forall \theta \]

\[ \iff \gamma(\theta)'T(x) = \gamma(\theta)'T(y) + a(x, y) \quad \forall \theta \]

\[ \iff (\gamma(\theta_1) - \gamma(\theta_2))'T(x) = (\gamma(\theta_1) - \gamma(\theta_2))'T(y) \quad \forall \theta_1, \theta_2 \]

\[ \iff \gamma(\theta_1) - \gamma(\theta_2) \perp T(x) - T(y) \quad \forall \theta_1, \theta_2 \]

\[ \iff T(x) - T(y) \perp \text{Span} \{ \gamma(\theta_1) - \gamma(\theta_2) : \theta_1, \theta_2 \in \Theta \} \]

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

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

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

E.g., \( \gamma(\theta) = (\theta) \) : then \( T_1(x) \) sufficient (could \( T(x) \) still be minimal?)
\[ X \sim N_2(\mu(\theta), I_2) \quad \theta \in \mathbb{R} \]

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

If \( \mathcal{H} = \mathbb{R} \), \( \mu(\theta) = a + \theta b \) 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 \( \implies X \) not minimal

\( \eta_2 \)

\( T(X)_{\text{minimal}} \)

\( (A) \)

\( \gamma' T(X) \) is sufficient

\( \implies T(X) \) prob. not minimal

\( \eta_1 \)
**Ex**  Laplace location family

\[ X_1, \ldots, X_n \overset{iid}{\sim} \rho_\theta^{(\cdot)}(x) = \frac{1}{2} e^{-|x-\theta|} \]

\[ \rho_\theta(x) = \frac{1}{a^n} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} |x_i - \theta| \right\} \]

\[ l(\theta; x) = \log \rho_\theta(x) \]

\[ = -\frac{1}{2} \sum_{i=1}^{n} |x_i - \theta| - n \log 2 \]

Piecewise linear in \( \theta \), knots at \( x_{(i)} \)

\[ l(\theta; x) = l(\theta; y) + \text{const} \iff X, Y \text{ same order statistics} \]

Thm 3.11 \( \Rightarrow \) order stats are minimal suff.