Chapter 6. Principles of Data Reduction

6.1 Introduction

추론의 대상.

$X_1, \cdots, X_n \overset{iid}{\sim} f(x|\theta)$

$(x_1, \cdots, x_n)$: obs., just a long list of numbers that is hard to interpret.

$\underset{\sim}{X} = (X_1, \cdots, X_n) \xrightarrow{T} T(\underset{\sim}{X})$, statistic

"data reduction", "data summary".

Three principles

(i) sufficiency : data reduction that does not discard information about θ.

(ii) Likelihood : a function of θ that contains all the information about θ that is available from the sample

(iii) equivariance : data reduction that preserves some important features of the model.

6.2 The sufficiency principle.

Sufficiency: If $T(\underset{\sim}{X})$ is a suff. statistic for θ, then any inference about θ should depend on the sample $\underset{\sim}{X}$ only through the value $T(\underset{\sim}{X})$.

That is, if $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are two sample points such that $T(\underset{\sim}{x}) = T(\underset{\sim}{y})$, then the inference about θ should be the same whether $\underset{\sim}{X} = \underset{\sim}{x}$ or $\underset{\sim}{X} = \underset{\sim}{y}$ is observed.

Defn: A statistic $T(\underset{\sim}{X})$ is a SS for θ if the conditional distribution of the sample $\underset{\sim}{X}$ given the value of $T(\underset{\sim}{X})$ does not depend on θ.

Thm: $\underset{\sim}{X} \sim p(\underset{\sim}{x}|\theta)$, $T(\underset{\sim}{X}) \sim f(t|\theta)$ is a SS for θ, if $p(\underset{\sim}{x}|\theta) / f(T(\underset{\sim}{x})|\theta)$ is constant as a function of θ.

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $0<\theta<1$.

$T(\underset{\sim}{X}) = X_1 + \cdots + X_n$: SS for θ?

sol. To check if $p(\underset{\sim}{x}|\theta) / q(T(\underset{\sim}{x})|\theta)$ does not depend on θ.

$$p(\underset{\sim}{x}|\theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$q(T(\underset{\sim}{x})|\theta) = \binom{n}{\sum x_i} \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$p(\underset{\sim}{x}|\theta) / q(T(\underset{\sim}{x})|\theta) = \binom{n}{\sum x_i}$$: does not depend on θ.

$\therefore T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$ is a SS for θ. //

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$

$\bar{X}$ is a SS for θ.

Sol. $$\frac{f(\underline{x}|\theta)}{g(T(\underline{x})|\theta)} = \frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)}{(2\pi\cdot\frac{1}{n})^{-1/2} \exp\left(-\frac{n}{2}(\bar{x}-\theta)^2\right)}$$

$$= n^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left\{\sum x_i^2 - 2n\theta\bar{x} + n\theta^2\right\} + \frac{n}{2}\left\{(\bar{x})^2 - 2\bar{x}\cdot\theta + \theta^2\right\}\right)$$

$$= n^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum x_i^2 + \frac{n}{2}(\bar{x})^2\right) \quad : \text{does not depend on } \theta. \quad //$$

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} f$. Then $X_{(1)}, \cdots, X_{(n)}$ is a sufficient statistic. (not much of a reduction)

Outside of the exp'l family, it is rare to have a SS of smaller dimension than the sample size, so in many cases it will turn out that the o.s. is the best way that we can do.

Th^m : (factorization theorem) $(X_1, \cdots, X_n) \sim f(\underline{x}|\theta)$

$T(\underline{X})$; a SS for θ $\Leftrightarrow$ $f(\underline{x}|\theta) = g(T(\underline{x})|\theta)\, h(\underline{x})$

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

$$f(\underline{x}|\theta) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_1^n (x_i - \theta)^2\right)$$

$$= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\Sigma x_i^2 + n\theta \cdot \bar{x} - \frac{n}{2}\theta^2\right)$$

Put $g(t|\theta) = \exp\left(n\theta \cdot t - \frac{n}{2}\theta^2\right)$ and $h(\underline{x}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\Sigma x_i^2\right)$. //

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} Unif(1, 2, \cdots, \theta)$

$$f(x|\theta) = \frac{1}{\theta}\, I(x = 1, 2, \cdots, \theta)$$

$$f(\underline{x}|\theta) = \theta^{-n}\, I(x_i = 1, 2, \cdots, \theta \text{ for } i = 1, 2, \cdots, n) = \theta^{-n}\, I(\underbrace{\max_{1 \le i \le n} x_i}_{T(\underline{x})} \le \theta)$$

Put $g(t|\theta) = \theta^{-n} I(t \le \theta)$ and $h(\underline{x}) = 1$. //

(e.g.) $X_1, \cdots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, $\underline{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$

$$f(\underline{x} \mid \underline{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \Sigma x_i^2 + \frac{\mu}{2\sigma^2} \Sigma x_i - \frac{n\mu^2}{2\sigma^2} \right)$$

$\therefore (\Sigma X_i, \Sigma X_i^2)$ is a SS for $\underline{\theta} = (\mu, \sigma^2)$.

$\Big\updownarrow$ 1-1

$(\bar{X}, S^2)$

$\therefore (\bar{X}, S^2)$ is a SS for (μ, σ^2). //

Thm: $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\underline{\theta}) = h(x)\, c(\underline{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\underline{\theta})\, t_j(x) \right)$, $\underline{\theta} = (\theta_1, \cdots, \theta_d)$, $d \le k$.

Then, $T(\underline{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i) \right)$ is a SS for $\underline{\theta}$.

Pf: Trivial by the factorization theorem. //

Defn: A suff. stat. $T(\underline{X})$ is called a MSS if, for any other SS $T'(\underline{X})$, $T(\underline{x})$ is a func. of $T'(\underline{x})$.

"If $T'(\underline{x}) = T'(\underline{y})$, then $T(\underline{x}) = T(\underline{y})$." //

- the greatest possible data reduction for a sufficient statistic.
- any 1-1 func. of a MSS is also a MSS.

Thm: $\underset{\sim}{X} \sim f(\underset{\sim}{x}|\theta)$.

$\exists\, T(\cdot)$ s.t. $f(\underset{\sim}{x}|\theta)/f(\underset{\sim}{y}|\theta)$ is constant in θ iff $T(\underset{\sim}{x}) = T(\underset{\sim}{y})$.

Then, $T(\underset{\sim}{X})$ is a MSS for θ.

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

$$\frac{f(\underset{\sim}{x}|\theta)}{f(\underset{\sim}{y}|\theta)} = \exp\left(-\frac{1}{2}\left\{\sum_{i=1}^{n}(x_i-\theta)^2 - \sum_{i=1}^{n}(y_i-\theta)^2\right\}\right) = \exp\left(-\frac{1}{2}\left\{\Sigma x_i^2 - \Sigma y_i^2 - \theta\cdot(\Sigma x_i - \Sigma y_i)\right\}\right)$$

is const. in θ iff $\Sigma x_i = \Sigma y_i$.

$\therefore T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$ is a MSS for θ. //

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} \text{Unif}(\theta, \theta+1)$.

$$f(\underline{x}|\theta) = I\left(\max_i x_i - 1 < \theta < \min_i x_i\right)$$

$f(\underline{x}|\theta) \,/\, f(\underline{y}|\theta)$ is const. in θ iff $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$.

$\therefore (X_{(1)}, X_{(n)})$ is a MSS for θ.

Defn: A statistic $S(\underline{X})$ whose distribution does not depend on the paramter θ is called an ancillary statistic.

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} \text{Unif}(\theta, \theta+1)$

$R = X_{(n)} - X_{(1)}$ is an ancillary statistic for θ. //

(e.g.) (Location family ancillary) $X_1, \cdots, X_n \overset{iid}{\sim} f(x-\theta)$, $\theta \in \mathbb{R}$.

$R = X_{(n)} - X_{(1)}$ is an ancillary statistic for θ.

sol. Consider $Z_1, \cdots, Z_n \overset{iid}{\sim} f(x)$. "$\theta = 0$"

$X_1 \overset{d}{=} Z_1 + \theta$, $X_2 \overset{d}{=} Z_2 + \theta$, $\cdots$, $X_n \overset{d}{=} Z_n + \theta$.

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \le r) \\ &= P_\theta(X_{(n)} - X_{(1)} \le r) \\ &= \underbrace{P_\theta(Z_{(n)} - Z_{(1)} \le r)}_{\text{does not depend on } \theta.} \end{aligned}$$

"

(e.g.) (scale family ancillary) $X_1, \cdots, X_n \overset{iid}{\sim} \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$, $\sigma > 0$.

$\left(X_1/X_n, X_2/X_n, \cdots, X_{n-1}/X_n\right)$ is an ancillary stat. for σ.

- A MSS achieves the maximal data reduction possible while retaining "all" the information about θ.

$\Rightarrow$ MSS $\perp$ ancillary stat. ?

NOT necessarily!

BUT in many important cases, yes!

Defn: $T(\underline{X})$ is called a complete statistic if

$E_\theta\, g(T) = 0 \ \forall\theta$ implies $P_\theta(g(T)=0) = 1 \ \forall\theta$.

$\left(\int g(t) f_T(t|\theta)\, dt \overset{\forall\theta}{=} 0 \implies \int_{g(t)=0} f_T(t|\theta)\, dt \overset{\forall\theta}{=} 1 \right)$

"Completeness is a property of a family of prob. dist^n's."

(e.g.) $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, $0<p<1$

$T = \sum_{i=1}^{n} X_i \sim B(n,p)$ is a complete statistic.

Sol. Let g be a func. s.t. $E_p\, g(T) = 0 \quad \forall p \in (0,1)$.

Then $$0 = E_p\, g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t}$$

$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \quad \forall p \in (0,1).$$

$$\sum_{t=0}^{n} g(t) \underbrace{\binom{n}{t}}_{>0 \ \forall t=0,1,\cdots,n} r^t = 0 \quad \forall r > 0.$$

$\therefore g(t) = 0$ for all $t = 0, 1, 2, \cdots, n$.

$\therefore P_p(g(T) = 0) \overset{\forall p}{=} 1$ //

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, $\theta > 0$.

$T(\underline{X}) = X_{(n)}$ is a complete statistic.

sol. $f_T(t|\theta) = n t^{n-1} \theta^{-n} I(0<t<\theta)$

Let g be a func. s.t. $E_\theta\, g(T) = 0 \ \forall \theta > 0$.

Since $E_\theta\, g(T)$ is a const. in θ, $\frac{d}{d\theta} E_\theta\, g(T) = 0, \ \forall \theta > 0$.

$$0 = \frac{d}{d\theta} \int_0^\theta g(t) \cdot n t^{n-1} \theta^{-n} \, dt = \frac{d}{d\theta} \left[\theta^{-n} \int_0^\theta g(t)\, n t^{n-1} \, dt \right]$$

$$= -n\theta^{-n-1} \int_0^\theta g(t)\, n t^{n-1} \, dt + \theta^{-n} \cdot g(\theta)\, n \theta^{n-1}$$

$$= -n\theta^{-1} \int_0^\theta g(t)\, n t^{n-1} \theta^{-n} \, dt + n g(\theta) \theta^{-1}$$

$$= -n\theta^{-1} \cdot 0 + n g(\theta) \theta^{-1} = \underbrace{n\theta^{-1}}_{>0, \ \forall \theta > 0} \cdot g(\theta) \qquad \therefore g(t) = 0 \ \forall t > 0. \ //$$

Th^m: (Basu) $T(\underset{\sim}{X})$: a complete and (minimal) suff. stat. for θ.
Then, $T(\underset{\sim}{X})$ is indep. of every ancillary statistic.

pf. (only for discrete dist^n's)

$S(\underset{\sim}{X})$: an ancillary stat., i.e. $P(S(\underset{\sim}{X})=s)$ does not depend on θ.

$P(S(\underset{\sim}{X})=s \mid T(\underset{\sim}{X})=t)$ does not depend on θ.

WTS : $P(S(\underset{\sim}{X})=s \mid T(\underset{\sim}{X})=t) = P(S(\underset{\sim}{X})=s) \quad \forall s,t$

$$P(S(\underset{\sim}{X})=s) = \sum_t P(S(\underset{\sim}{X})=s,\ T(\underset{\sim}{X})=t)$$

$$= \sum_t P(S(\underset{\sim}{X})=s \mid T(\underset{\sim}{X})=t) \cdot P_\theta(T(\underset{\sim}{X})=t)$$

$$\sum_t \underbrace{\left[P(S(\underset{\sim}{X})=s) - P(S(\underset{\sim}{X})=s \mid T(\underset{\sim}{X})=t) \right]}_{g(T)} \cdot P_\theta\left(T(\underset{\sim}{X})=t\right) = 0$$

//

Thm: $X_1, \cdots, X_n \overset{\text{iid}}{\sim} f(x|\underline{\theta}) = h(x)\, c(\underline{\theta}) \exp\left(w_1(\underline{\theta})\, t_1(x) + w_2(\underline{\theta})\, t_2(x) + \cdots + w_k(\underline{\theta})\, t_k(x) \right)$

where $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$.

Then, $T(\underline{X}) = \left(\sum_{i=1}^n t_1(X_i), \sum_{i=1}^n t_2(X_i), \cdots, \sum_{i=1}^n t_k(X_i) \right)$ is complete

if $\{ (w_1(\underline{\theta}), \cdots, w_k(\underline{\theta})) : \theta \in \Theta \}$ contains an open set in $\mathbb{R}^k$.

(e.g.) $X_1, \cdots, X_n \overset{\text{iid}}{\sim} f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} I_{(x>0)}, \ \theta > 0.$

$$E\left(\frac{X_1}{X_1 + \cdots + X_n} \right) = ?$$

sol. $S(\underline{X}) = \dfrac{X_1}{X_1 + \cdots + X_n}$ is an ancillary stat. (∵ scale family)

$f(x|\theta)$ forms an exp'l family with $t_1(x) = x$ and $w_1(\theta) = \frac{1}{\theta}$. $\Rightarrow T(\underline{X}) = \sum_{i=1}^n X_i$ is complete.

$\{ \frac{1}{\theta} : \theta > 0 \} = (0, \infty)$ contains an open set in $\mathbb{R}^1$.

∴ By Basu, $S(\underline{X})$ and $T(\underline{X})$ are indep.

$$\underbrace{E(X_1)}_{\theta} = E(S(\underline{X}) \cdot T(\underline{X})) = \underbrace{E(S(\underline{X}))}_{?} \cdot \underbrace{E T(\underline{X})}_{n\theta} \quad \therefore \ \frac{1}{n} \ //$$

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

$\bar{X}$ and S^2 are indep. !

Th$^{\text{m}}$: If a MSS exists, then any complete statistic is also a MSS.