Chapter 6
Principle of Data Deduction
Minimal Sufficiency & Ancillarity

Outline

Bayesian Sufficiency

Definition. T is Bayesian sufficient for every prior density π, there exist posterior densities $f_{\Psi|\mathbf{X}}$ and $f_{\Psi|T(\mathbf{X})}$ so that

$$f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) = f_{\Psi|T(\mathbf{X})}(\theta|T(\mathbf{x})).$$

The posterior density is a function of the sufficient statistic.

Theorem. If T is sufficient in the classical sense, then T is sufficient in the Bayesian sense.

Proof. Bayes formula states the posterior density

$$f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f_{\mathbf{X}}(\mathbf{x}|\psi)\pi(\psi)\nu(d\psi)}$$

By the Neyman-Fisher factorization theorem, $\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))$

$$\begin{aligned} f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) &= \frac{\mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))\pi(\theta)}{\int_\Theta \mathbf{h}(\mathbf{x})g(\psi, T(\mathbf{x}))\pi(\psi)\nu(d\psi)} \\ &= \frac{g(\theta, T(\mathbf{x}))\pi(\theta)}{\int_\Theta g(\psi, T(\mathbf{x}))\pi(\psi)\nu(d\psi)} = f_{\Psi|T(\mathbf{X})}(\theta|T(\mathbf{x})) \end{aligned}$$

Bayesian Sufficiency

We have shown that classical sufficiency implies Bayesian sufficiency. Now we show the converse. Thus, assuming Bayesian sufficiency, we apply Bayes formula twice

$$\begin{aligned}\frac{f_X(\mathbf{x}|\theta)\pi(\theta)}{f_X(\mathbf{x})} &= f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) = f_{\Psi|T(\mathbf{X})}(\theta|T(\mathbf{x})) \\ &= \frac{f_{T(X)}(T(\mathbf{x}))|\theta)\pi(\theta)}{f_{T(X)}(T(\mathbf{x}))}\end{aligned}$$

Thus,

$$f_X(\mathbf{x}|\theta) = f_X(\mathbf{x})\frac{f_{T(X)}(T(\mathbf{x}))|\theta)}{f_{T(X)}(T(\mathbf{x}))}.$$

which can be written in the form $\mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))$ and T is classically sufficient

Introduction

While entire set of observations $X_1 \dots, X_n$ is sufficient, this choice does not result in any reduction in the data used for formal statistical inference.

Recall that any statistic U induces a partition $\mathcal{A}_U$ on the sample space $\mathcal{X}$.

Exercise. The partition $\mathcal{A}_T$ induced by $T = c(U)$ is coarser than $\mathcal{A}$.

Let $A_{\mathbf{x}} = \{\tilde{\mathbf{x}}; U(\tilde{\mathbf{x}}) = U(\mathbf{x})\} \in \mathcal{A}$, then $A_{\mathbf{x}} \subset \tilde{A}_{\mathbf{x}} = \{\tilde{\mathbf{x}}; c(U(\tilde{\mathbf{x}})) = c(u(\mathbf{x}))\}$

Moreover $\mathcal{A}_c = \mathcal{A}$ if and only if c is one-to-one.

Thus, if T is sufficient, then so is U and we can proceed using T to perform inference with a further reduction in the data.

Is there a sufficient statistic that provides *maximal* reduction of the data?

Minimal Sufficiency

Definition. A sufficient statistic T is called a minimal sufficient statistic provided that any sufficient statistic U, T is a function $c(U)$ of U.

- T is a function of U if and only if $U(\mathbf{x}_1) = U(\mathbf{x}_2)$ implies that $T(\mathbf{x}_1) = T(\mathbf{x}_2)$
- In terms of partitions, if T is a function of U, then

  $$\{\tilde{\mathbf{x}}; U(\tilde{\mathbf{x}}) = U(\mathbf{x})\} \subset \{\tilde{\mathbf{x}}; T(\tilde{\mathbf{x}}) = T(\mathbf{x})\}$$

  In other words, the minimal sufficient statistic has the coarsest partition and thus achieves the greatest possible data reduction among sufficient statistics.
- If both U and T are minimal sufficient statistics then

  $$\{\tilde{\mathbf{x}}; U(\tilde{\mathbf{x}}) = U(\mathbf{x}) = \{\tilde{\mathbf{x}}; T(\tilde{\mathbf{x}}) = T(\mathbf{x})\}$$

  and c is one-to-one,

Minimal Sufficiency

The following thereom will be used to find minimal sufficient statistics.

Theorem. Let $f_X(\mathbf{x}|\theta); \theta \in \Theta$ be a parametric family of densities and suppose that T is a sufficient statistic for θ. Assume that for every pair $\mathbf{x}_1, \mathbf{x}_2$ chosen so that at least one of the points has non-zero density. If the ratio

$$\frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)}$$

does not depend on θ implies that $T(\mathbf{x}_1) = T(\mathbf{x}_2)$, then T is a minimal sufficient statistic.

Minimal Sufficiency

Proof. Choose a sufficient statistic U. The plan is to show that $U(\mathbf{x}_1) = U(\mathbf{x}_2)$ implies that $T(\mathbf{x}_1) = T(\mathbf{x}_2)$. If this holds, then T is a function of U and consequently T is a minimal sufficient statistic.

We return to the ratio and use the Neyman-Fisher factorization theorem on the sufficient statistic U to write the density as a product $\mathbf{h}(\mathbf{x})g(\theta, U(\mathbf{x}))$

$$\frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)} = \frac{\mathbf{h}(\mathbf{x}_1)g(\theta, U(\mathbf{x}_1))}{\mathbf{h}(\mathbf{x}_2)g(\theta, U(\mathbf{x}_2))}$$

If $U(\mathbf{x}_1) = U(\mathbf{x}_2)$, then the ratio

$$\frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)} = \frac{\mathbf{h}(\mathbf{x}_1)}{\mathbf{h}(\mathbf{x}_2)}$$

does not depend on θ and T is a minimal sufficient statistic.

Examples

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be Bernoulli trials. Then $T(\mathbf{x}) = x_1 + \cdots + x_n$ is sufficient.

$$\begin{aligned}\frac{f_X(\mathbf{x}_1|p)}{f_X(\mathbf{x}_2|p)} &= \frac{p^{T(\mathbf{x}_1)}(1-p)^{n-T(\mathbf{x}_1)}}{p^{T(\mathbf{x}_2)}(1-p)^{n-T(\mathbf{x}_2)}} \\ &= \left(\frac{p}{1-p}\right)^{T(\mathbf{x}_1)-T(\mathbf{x}_2)}\end{aligned}$$

This ratio does not depend on p if and only if $T(\mathbf{x}_1) = T(\mathbf{x}_2)$. Thus T is a minimal sufficient statistic.

Examples

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent $N(\mu, \sigma^2)$ random variables. Then $T(\mathbf{x}) = (\bar{x}, s^2)$ is sufficient. To check if its minimal, note that

$$\frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)} = \frac{(2\pi\sigma)^{-n/2}\exp -\left(n(\bar{x}_1-\mu)^2+(n-1)s_1^2\right)/(2\sigma^2)}{(2\pi\sigma)^{-n/2}\exp -\left(n(\bar{x}_2-\mu)^2+(n-1)s_2^2\right)/(2\sigma^2)}$$
$$= \exp -\left(n(\bar{x}_1^2-\bar{x}_2^2)-2n\mu(\bar{x}_1-\bar{x}_2)+(n-1)(s_1^2-s_2^2)\right)/(2\sigma^2)$$

This ratio does not depend on $\theta = (\mu, \sigma^2)$ if and only if $\bar{x}_1 = \bar{x}_2$ and $s_1^2 = s_2^2$, i.e., $T(\mathbf{x}_1) = T(\mathbf{x}_2)$. Thus T is a minimal sufficient statistic.

Examples

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent $Unif(\theta, \theta + 1)$ random variables.

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= I_{[\theta,\theta+1]}(x_1) \cdots I_{[\theta,\theta+1]}(x_n) \\ &= \begin{cases} 1 & \text{if all } x_i \in [\theta, \theta + 1] \\ 0 & \text{otherwise} \end{cases} \\ &= I_{A_{\mathbf{x}}}(\theta). \end{aligned}$$

For the density to be equal to 1, we must have $\theta \leq x_{(1)} \leq x_{(n)} \leq \theta + 1$, $A_{\mathbf{x}} = [x_{(n)} - 1, x_{(1)}]$. Thus,

$$\frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)} = \begin{cases} 0 & \text{if } \theta \in A^c_{\mathbf{x}_1} \cap A_{\mathbf{x}_2} \\ 1 & \text{if } \theta \in A_{\mathbf{x}_1} \cap A_{\mathbf{x}_2} \\ \infty & \text{if } \theta \in A_{\mathbf{x}_1} \cap A^c_{\mathbf{x}_2} \end{cases}$$

For this to be independent of θ, both $\mathbf{x}_1$ and $\mathbf{x}_2$ must have the same minimum and maximum values.

Examples

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent random variables from an exponential family, the probability density functions can be expressed in the form

$$\mathbf{f}_X(\mathbf{x}|\eta) = \mathbf{h}(\mathbf{x}) \cdot \exp\left(\sum_{j=1}^{n} \langle \eta, \mathbf{t}(x_j) \rangle\right) e^{-nA(\eta)}, \quad x \in S.$$

We have seen that $T(\mathbf{x}) = \sum_{j=1}^{n} \mathbf{t}(x_j)$ is sufficient. To check that it is minimal sufficient.

$$\begin{aligned} \frac{f_X(\mathbf{x}_1|\theta)}{f_X(\mathbf{x}_2|\theta)} &= \frac{\mathbf{h}(\mathbf{x}) \cdot \exp(\sum_{j=1}^{n} \langle \eta, \mathbf{t}(x_{1,j}) \rangle) e^{-nA(\eta)}}{\mathbf{h}(\mathbf{x}) \cdot \exp(\sum_{j=1}^{n} \langle \eta, \mathbf{t}(x_{2,j}) \rangle) e^{-nA(\eta)}} \\ &= \exp\langle \eta, \sum_{j=1}^{n} (\mathbf{t}(x_{1,j}) - \mathbf{t}(x_{2,j})) \rangle = \exp\langle \eta, T(\mathbf{x}_1) - T(\mathbf{x}_2) \rangle \end{aligned}$$

For this to be independent of parameter η, $T(\mathbf{x}_1) - T(\mathbf{x}_2)$ must be the zero vector and T is a minimal sufficient statistic.

Ancillary Statistics

At the opposite extreme, we call a statistic V is called ancillary if its distribution does not depend on the parameter value θ

Even though an ancellary statistic V by itself fails to provide any information about the parameter, in conjunction with another statistic statistic T , e.g., the maximum likelihood estimator, it can provide valuable information, if the estimator itself is not sufficient.

Examples

Let X be a continuous (discrete) random variable with density (mass) function $f_X(x)$. Let

$$Y = \sigma X + \mu, \quad \sigma > 0, \mu \in \mathbb{R}.$$

Then Y has density (mass) function,

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma} f_X((y-\mu)/\sigma), \qquad f_Y(y|\mu,\sigma) = f_X((y-\mu)/\sigma).$$

Such a two parameter family of density (mass) functions is called a location/scale family.

- μ is the location parameter. If X has mean 0, then μ is the mean of Y. The case $\sigma = 1$ is called a location family.
- σ is the scale parameter. If X has standard deviation 1, then σ is the standard deviation of Y. The case $\mu = 0$ is called a scale family.

Location Families

Examples of (location families)

$$Unif(\mu - a_0, \mu + a_0), a_0 \text{ fixed}, \quad N(\mu, \sigma_0^2), \sigma_0^2 \text{ fixed} \quad Logistic(\mu, s_0), s_0 \text{ fixed},$$

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be independent random variables from an location family. Then,

$$P_\mu\{\mathbf{Y} \in B\} = P_0\{\mathbf{Y} - \mu \in B\} = P_0\{\mathbf{Y} \in B + \mu\}.$$

Example.

- The difference of order statistics has a distribution

$$P_\mu\{Y_{(j)} - Y_{(i)} \in A\} = P_0\{(Y_{(j)} - \mu) - (Y_{(i)} - \mu) \in A\} = P_0\{Y_{(j)} - Y_{(i)} \in A\}$$

  that does not depend on the location parameter μ and thus is ancillary.
- In particular the range,

$$R = Y_{(n)} - Y_{(1)}$$

  is ancillary

Location Families

- The variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((Y_i - \mu) - (\bar{Y} - \mu))^2$$

  is invariant under a shift by a constant μ. Thus, S^2 is an ancillary statistic.
- More generally, if T is a location invariant statistic, i.e., for any b in the state space for the Y_i,

$$T(y_1 + b, \ldots, y_n + b) = T(y_1, \ldots, y_n)$$

  then T is ancillary.

Scale Families

Examples of (scale families)

$$Unif(0,\theta) \quad Exp(\beta) \quad \Gamma(\alpha_0,\beta), \alpha_0 \text{ fixed}, \quad N(0,\sigma^2)$$

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent random variables from a scale family. Then,

$$\begin{aligned} P_\sigma\{(X_2/X_1, \ldots, X_n/X_1) \in A\} &= P_1\{((\sigma X_2)/(\sigma X_1), \ldots, (\sigma X_n)/(\sigma X_1)) \in A\} \\ &= P_1\{(X_2/X_1, \ldots, X_n/X_1) \in A\} \end{aligned}$$

  and $T(\mathbf{X}) = (X_2/X_1, \ldots, X_n/X_1)$ is ancillary.
- For $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\mu, \sigma_0)$,

$$T(\mathbf{X}) = \left(\frac{X_1 - \bar{X}}{S}, \ldots \frac{X_n - \bar{X}}{S}, \right)$$

  is ancillary.

Value of Ancillarity

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent $Unif(\theta - 1, \theta + 1)$ random variables.

Estimate θ by the mid-range $M = (X_{(1)} + X_{(n)})/2$

The range $R = X_{(n)} - X_{(1)}$ is ancillary.

Note that $0 \leq R \leq 2$. However, If R is close 2, then $X_{(1)}$ must be close to $\theta - 1$ and $X_{(n)}$ must be close to $\theta - 1$, so the M must be an accurate estimate of θ.

Thus a larger value of R increases our faith in the observed estimate.

In addition, (M, R) is a minimal sufficient statistic with R ancillary.