ECE531 Screencast 3.4: Sufficiency and Completeness

D. Richard Brown III

Worcester Polytechnic Institute

Sufficiency

Definition

$T : \mathcal{Y} \mapsto \Delta$ is a **sufficient statistic** for the family of parameterized pdfs $\{p_Y(y\,;\,\theta)\,;\,\theta \in \Lambda\}$ if the distribution of the random observation conditioned on $T(Y)$, i.e. $p_Y\left(y \,|\, T(Y) = t\,;\,\theta\right)$, does not depend on θ for all $\theta \in \Lambda$ and all $t \in \Delta$.

Intuitively, a sufficient statistic summarizes the information contained in the observation about the unknown parameter. Knowing $T(y)$ is as good as knowing the full observation y when we wish to estimate θ.

Neyman-Fisher Factorization Theorem

Theorem (Fisher 1920, Neyman 1935)

A statistic T is sufficient for θ if and only if there exist functions g_θ and h such that the parameterized pdf of the observation can be factored as

$$p_Y(y\,;\,\theta) \quad = \quad g_\theta(T(y))h(y)$$

for all $y \in \mathcal{Y}$ and all $\theta \in \Lambda$.

The proof of this theorem is in your textbook.

Note that $h(y)$ can't be a function of θ and $g_\theta(T(y))$ must only be a function of θ and $T(y)$.

Example: Neyman-Fisher Factorization Theorem

Suppose $\theta \in \mathbb{R}$ and

$$p_Y(y\,;\,\theta) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{k=0}^{n-1} (y_k - \theta)^2 \right\}.$$

Let $T(y) = \frac{1}{n}\sum_{k=0}^{n-1} y_k$. Let's try the N-F factorization...

$$\begin{aligned} p_Y(y\,;\,\theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{-n}{2\sigma^2} \left(\frac{1}{n} \sum_{k=0}^{n-1} y_k^2 - 2\theta y_k + \theta^2 \right) \right\} \\ &= \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{-n}{2\sigma^2} \left(\theta^2 - 2\theta T(y)\right) \right\}}_{g_\theta(T(y))} \underbrace{\exp\left\{ \frac{-1}{2\sigma^2} \sum_{k=0}^{n-1} y_k^2 \right\}}_{h(y)} \end{aligned}$$

Hence $T(y)$ is a sufficient statistic.

Completeness of a Family of PDFs

Definition

The family of pdfs $\{p_Y(y\,;\,\theta)\,;\,\theta \in \Lambda\}$ is said to be complete if the condition $\mathrm{E}\,[f(Y)] = 0$ for all θ in Λ implies that $\mathrm{Prob}[f(Y) = 0] = 1$ for all θ in Λ. Note that $f : \mathcal{Y} \mapsto \mathbb{R}$ can be any function.

To get some intuition, consider the case where $\mathcal{Y} = \{y_0, \dots, y_{L-1}\}$ is a finite set. Then

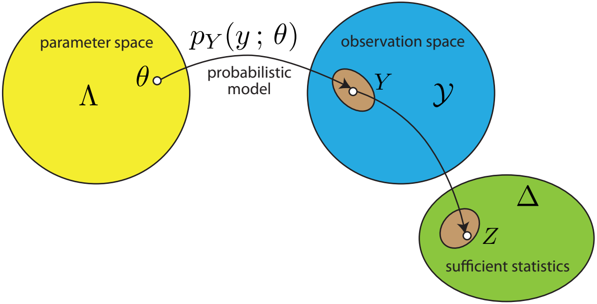
$$\begin{aligned} \mathrm{E}\,[f(Y)] &= \sum_{\ell=0}^{L-1} f(y_\ell)\mathrm{Prob}(Y = y_\ell; \theta) \\ &= f^\top(y)P(\theta) \end{aligned}$$

For a fixed θ it is certainly possible to find a non-zero f such that $\mathrm{E}\,[f(Y)] = 0$. But we have to satisfy this condition for all $\theta \in \Lambda$, i.e. we need a vector $f(y)$ that is **orthogonal to the all members of the family** of vectors $\{P(\theta)\,;\,\theta \in \Lambda\}$. If the only such vector that satisfies the condition $\mathrm{E}\,[f(Y)] = 0$ for all $\theta \in \Lambda$ is $f(y_0) = \cdots = f(y_{L-1}) = 0$, then the family $\{P\,;\,\theta \in \Lambda\}$ is complete.

Complete Sufficient Statistics

Definition

Suppose that T is a sufficient statistic for the family of pdfs $\{p_Y(y\,;\,\theta)\,;\,\theta \in \Lambda\}$. Let $p_Z(z\,;\,\theta)$ denote the distribution of $Z = T(Y)$ when the parameter is θ. If the family of pdfs $\{p_Z(z\,;\,\theta)\,;\,\theta \in \Lambda\}$ is complete, then T is said to be a complete sufficient statistic for the family $\{p_Y(y\,;\,\theta)\,;\,\theta \in \Lambda\}$.



One-Parameter Exponential Families

Theorem

Suppose $\mathcal{Y} = \mathbb{R}^n$, $\Lambda \subset \mathbb{R}$, *and*

$$p_Y(y\,;\,\theta) \quad = \quad a(\theta)\exp\left\{q(\theta)T(y)\right\}h(y)$$

where a, T, q, *and* h *are all real-valued functions. Then* $T(y)$ *is a complete sufficient statistic for the family* $\{p_Y(y\,;\,\theta)\,;\,\theta \in \Lambda\}$.

For the proof, see Poor pp. 165-166 and/or Lehmann 1986.