Sufficiency in an exponential family I

Random Sample $X_1, \ldots, X_n$

Likelihood

$$
\begin{aligned}
L(\theta; x) &= \prod_{i=1}^{n} f(x_i; \theta) \\
&= \prod_{i=1}^{n} \exp\left\{ \sum_{j=1}^{k} A_j(\theta) B_j(x_i) + C(x_i) + D(\theta) \right\} \\
&= \exp\left\{ \sum_{j=1}^{k} A_j(\theta) \left(\sum_{i=1}^{n} B_j(x_i) \right) + nD(\theta) + \sum_{i=1}^{n} C(x_i) \right\}.
\end{aligned}
$$

Exponential family form again.

Sufficiency in an exponential family II

Suppose the family is in canonical form, and let $t_j = \sum_{i=1}^n B_j(x_i)$, $C(x) = \sum_{i=1}^n C(x_i)$.

$$L(\theta; x) = \exp\left\{\sum_{j=1}^k \theta_j t_j + nD(\theta) + C(x)\right\}.$$

By the factorization criterion $t_1, \ldots, t_k$ are sufficient statistics for $\theta_1, \ldots, \theta_k$. In fact, we do not need canonical form. If

$$L(\theta; x) = \exp\left\{\sum_{j=1}^k A_j(\theta) t_j + nD(\theta) + C(x)\right\}$$

is a minimal k-dimensional linear exponential family then (by the regularity conditions above) $t_1, \ldots, t_k$ are minimal sufficient for $\theta_1, \ldots, \theta_k$.

Estimators

Classical estimation of parameters.

A point estimate for θ is a statistic of the data.

$$\widehat{\theta} = \widehat{\theta}(x) = t(x_1, \ldots, x_n).$$

An interval estimate is a set valued function $C(X) \subseteq \boldsymbol{\Theta}$ such that $\theta \in C(X)$ with a specified probability.

Maximum likelihood estimation

If $L(\theta)$ is differentiable and there is a unique maximum in the interior of $\theta \in \Theta$, then $\widehat{\theta}$ is the solution of

$$\frac{\partial}{\partial \theta} L(\theta; x) = 0 \quad \text{or} \quad \frac{\partial}{\partial \theta} \ell(\theta) = 0,$$

where $\ell(\theta) = \log L(\theta; x)$.

$T = t(\mathbf{x})$ is unbiassed for a function $g(\theta)$ if

$$\mathbb{E}_\theta(T) = \int_\chi t(x) f(\mathbf{x};\theta) d\mathbf{x} = g(\theta), \quad \text{for all } \theta \in \Theta.$$

The Bias of an estimator T is

$$\text{bias}_\theta(T) = \mathbb{E}_\theta\left[T - g(\theta)\right]$$

and the Mean square error (MSE) of T is

$$\text{mse}_\theta(T) = \mathbb{E}_\theta\left[T - g(\theta)\right]^2 = V_\theta(T) + \left[\text{bias}_\theta(T)\right]^2$$

Example: $N(\mu, \sigma^2)$. $\widehat{\mu} = \bar{X}$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ are unbiassed estimates of μ and σ^2.

Maximum likelihood estimation and exponential families

Consider a k-dimensional exponential family in canonical form

$$L(\theta; x) = \exp\left\{\sum_{j=1}^{k} \theta_j \left(\sum_{i=1}^{n} B_j(x_i)\right) + nD(\theta) + \sum_{i=1}^{n} C(x_i)\right\}.$$

Let $T_j(X) = \sum_{i=1}^{n} B_j(X_i)$, $j = 1, \ldots, k$. If the realized data are $X = x$, then the statistics evaluated on the data are $T_j(x) = t_j$.

The MLE of $\theta_1, \ldots, \theta_k$ are the solution of

$$t_j = \mathbb{E}_\theta(T_j),\ j = 1, \ldots, k.$$

[If the family is not in canonical form, there is a similar slightly more complicated matrix equation]

Proof

$$\ell = \log L = \text{const} + \sum_{j=1}^{k} \theta_j t_j + nD(\theta)$$

so

$$\frac{\partial}{\partial \theta_j}\ell = t_j + n\frac{\partial}{\partial \theta_j}D(\theta)$$

However we know that

$$\mathbb{E}_\theta[T_j] = -n\frac{\partial}{\partial \theta_j}D(\theta), \text{ so}$$

$$\frac{\partial}{\partial \theta_j}\ell = t_j - \mathbb{E}_\theta(T_j) = 0$$

is equivalent to $t_j = \mathbb{E}(T_j)$.

Fisher Information (scalar parameter θ)

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right]$$

where $l = \log L(\theta; x)$. Under regularity conditions

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right],$$

and, as the sample size $n \longrightarrow \infty$, the MLE $\widehat{\theta} \approx N(\theta, I_\theta^{-1})$. The score function $s(x;\theta)$ is defined as

$$s(x;\theta) = \frac{\partial}{\partial\theta} l(\theta) = \frac{f'(x;\theta)}{f(x;\theta)}$$

so that

$$I_\theta = \text{Var}[S(X;\theta)]$$

Identity $-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right] = \mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right]$.

$$\begin{aligned}\frac{\partial^2 l}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left\{\frac{1}{L}\frac{\partial L}{\partial\theta}\right\} \\ &= -\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 + \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2} \\ &= -\left(\frac{\partial l}{\partial\theta}\right)^2 + \frac{1}{L}\left(\frac{\partial^2 L}{\partial\theta^2}\right)\end{aligned}$$

The second term has expectation zero because

$$\mathbb{E}\left[\frac{1}{L}\left(\frac{\partial^2 L}{\partial\theta^2}\right)\right] = \int \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2}L dx = \int \frac{\partial^2 L}{\partial\theta^2}dx = 0$$

The alternative form $I_\theta = \mathrm{Var}[S(X;\theta)]$ follows from $\mathbb{E}\left[\frac{\partial\ell}{\partial\theta}\right] = 0$.

Sample of size n.

$$f(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$i_n(\theta) = -\int \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta) f(x;\theta) dx = n i_1(\theta).$$

That is, $i_1(\theta)$ is calculated from the density as

$$i_1(\theta) = -\int \frac{\partial^2}{\partial\theta^2} \log f(x;\theta) f(x;\theta) dx$$

Minimum variance estimator T.

If T and T' are unbiassed estimators of θ, then T is a MVE if

$$\text{var}_\theta(T) \leq \text{var}_\theta(T'), \text{ for all } \theta \in \Theta$$

Variance-Covariance inequality

Let U and V be scalar rv. We will shortly make use of the inequality

$$\text{cov}(U,V)^2 \leq \text{var}(U)\text{var}(V)$$

with equality if and only if $U = aV + b$ for constants and $a \neq 0$.

Cramér-Rao inequality.

If $\widehat{\theta}$ is an unbiassed estimator of θ, then

$$\mathsf{Var}(\widehat{\theta}) \geq I_\theta^{-1}.$$

Proof of the inequality

$$\mathbb{E}(\widehat{\theta}) = \int_\chi \widehat{\theta}(x) L(\theta; x) dx = \theta$$

Differentiate both sides w.r.t. θ

$$\int_\chi \widehat{\theta} \frac{\partial L}{\partial \theta} dx = 1$$

Now

$$\frac{\partial L}{\partial \theta} = L \frac{\partial l}{\partial \theta}$$

so

$$1 = \int_{\chi} \widehat{\theta} \frac{\partial l}{\partial \theta} L dx = \mathbb{E}\left[\widehat{\theta} \frac{\partial l}{\partial \theta}\right]$$

Now we use the inequality that for two random variables U, V

$$\mathsf{Cov}[U, V]^2 \leq \mathsf{Var}[U]\mathsf{Var}[V]$$

with $U = \widehat{\theta}$, $V = \frac{\partial l}{\partial \theta}$.

$$\begin{aligned} \mathbb{E}[V] &= \int_{\chi} \frac{\partial l}{\partial \theta} L dx = \int_{\chi} \frac{\partial L}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta}\left[\int_{\chi} L dx\right] \\ &= \frac{\partial}{\partial \theta}[1] = 0 \end{aligned}$$

Thus Cov$[U,V] = \mathbb{E}[UV] = 1$, and

$$\text{Var}[U] = \text{Var}[\widehat{\theta}] \geq \frac{\text{Cov}[U,V]^2}{\text{Var}[V]} = \frac{1^2}{I_\theta} = I_\theta^{-1}$$

(Corollary 1) There exists an unbiased estimator $\widehat{\theta}$ which attains the CR lower bound (under regularity conditions) if and only if

$$\frac{\partial l}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

Proof In the CR proof

$$\text{Cov}[U,V]^2 \leq \text{Var}[U]\text{Var}[V]$$

and the lower bound is attained if and only equality is achieved. $U = \widehat{\theta}, V = \frac{\partial l}{\partial \theta}$, so $\frac{\partial l}{\partial \theta} = c + d\widehat{\theta}$, where c, d are constants.

$\mathbb{E}[V] = 0$ so $c = -d\theta$ and $\frac{\partial l}{\partial \theta} = d(\widehat{\theta} - \theta)$. Multiply by $\partial l / \partial \theta$ and take expectations. The LHS is I_θ and the RHS is

$$d\mathbb{E}\left[\frac{\partial l}{\partial \theta}\widehat{\theta}\right] - d\theta\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = d \times 1 - 0 = d.$$

That is $d = I_\theta$ and

$$\frac{\partial l}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

(Corollary 2) If there exists an unbiased estimator $\widehat{\theta}(X)$ which attains the CR lower bound (under regularity conditions) it follows that X must be in an exponential family since (taking $n = 1$)

$$\frac{\partial \log f(x;\theta)}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

and

$$\log f(x;\theta) = \widehat{\theta} A(\theta) + D(\theta) + C(x)$$

which is an exponential family form. The constant of integration $C(x)$ is a function of x.

(Corollary 3) Suppose $\widetilde{\theta}(X)$ is an MVUE which attains the CRLB. Suppose that the MLE $\widehat{\theta}$ is a solution to $\partial l/\partial\theta = 0$ (so, not on boundary). Then $\widetilde{\theta} = \widehat{\theta}$, by evaluating $\partial l/\partial\theta = I_\theta(\widehat{\theta} - \theta)$ at $\theta = \widehat{\theta}$.

Extensions to the Cramér-Rao inequality

1. If $\widehat{\theta}$ is an estimator with bias $b(\theta) = \text{bias}(\widehat{\theta})$, then

$$\text{Var}[\widehat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 I_\theta^{-1}$$

2. If $\widehat{g}(x)$ is an unbiased estimator for $g(\theta)$, then

$$\text{Var}[\widehat{g}(X)] \geq \left(\frac{\partial g}{\partial \theta}\right)^2 I_\theta^{-1}.$$

Proof of the above extensions begins with $\mathbb{E}_\theta(\widehat{\theta}(X)) = \theta + b(\theta)$ (in 1.) and $\mathbb{E}_\theta(\widehat{g}(X)) = g(\theta)$ (in 2.). Differentiate both sides and proceed as above to find $\text{Cov}[U, V] = (1 + \partial b/\partial\theta)$ (in 1.) and $\text{Cov}[U, V] = (1 + \partial b/\partial\theta)$ (in 2., with $U = \widehat{g}$). The bound is against $\text{Cov}[U, V]^2$ which leads to the results above.

Fisher Information for a d-dimensional parameter

Information matrix:

$$I_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right]$$

under regularity conditions. The CR inequality is

$$\text{Var}(\widehat{\theta}_i) \geq [I^{-1}]_{ii},\ i = 1, \ldots, d.$$

For an Exponential family in canonical form,

$$I_{ij} = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} nD(\theta).$$

Exercise: verify that we have already proved $\text{Var}(\widehat{\theta}_i) \geq [I_{ii}]^{-1}$.
Note that $[I^{-1}]_{ii} \geq [I_{ii}]^{-1}$ (GJJ) so bound above is stronger.

The (Bahadur) efficiency of an estimator $\widetilde{\theta}$ is defined as a comparison of the variance of $\widetilde{\theta}$ with the CR bound I_θ^{-1}. That is

$$\text{Efficiency of } \widetilde{\theta} = \frac{I_\theta^{-1}}{\mathrm{Var}[\widetilde{\theta}]} = \frac{1}{I_\theta \mathrm{Var}[\widetilde{\theta}]}$$

The asymptotic efficiency is the limit as $n \to \infty$.

There are similar definitions for the relative efficiency of two estimators.

Rao-Blackwell Theorem (GJJ 2.5.2) Let $X_1, \ldots, X_n$ be a random sample of observations from $f(x; \theta)$. Suppose that T is a sufficient statistic for θ and that $\widehat{\theta}$ is any unbiased estimator for θ. Define $\widehat{\theta}_T = \mathbb{E}[\hat{\theta} \mid T]$. Then

1. $\widehat{\theta}_T$ is a function of T alone;
2. $\mathbb{E}[\widehat{\theta}_T] = \theta$; (partition theorem for expectation)
3. $\text{Var}(\widehat{\theta}_T) \leq \text{Var}(\widehat{\theta})$.

Corollary If an MVUE $\widehat{\theta}$ for θ exists, then there is a function $\widehat{\theta}_T$ of the minimal sufficient statistic T for θ which is an MVUE.

Proof: T is sufficient so (2.) $\widehat{\theta}_T$ is an unbiased estimator which is (1.) a function of T alone. By (3.) $\text{Var}(\widehat{\theta}_T) \leq \text{Var}(\widehat{\theta})$, but $\widehat{\theta}$ is already minimum variance, so $\text{Var}(\widehat{\theta}_T)$ is also.

Complete Sufficient Statistic

Let $T(X_1, \ldots, X_n)$ be a sufficient statistic for θ. The statistic T is complete if, whenever $h(T)$ is a function of T for which $\mathbb{E}[h(T)] = 0$ for all θ, then $h(T) \equiv 0$ almost everywhere.

Suppose $h = h(T)$ with T complete and sufficient for θ, and $h(T)$ unbiased for θ. Then $h(T)$ is the unique function of T which is an unbiased estimator of θ.

Proof If there were two such unbiased estimators $h_1(T), h_2(T)$, then $\mathbb{E}[h_1(T) - h_2(T)] = \theta - \theta = 0$ for all θ, so $h_1(T) = h_2(T)$ almost everywhere.

Sufficient condition for estimator to be MVUE

An unbiased estimator with efficiency 1 (ie variance at the CRB) is clearly MVUE (subject to regularity conditions). What if we have an unbiased estimator with efficiency less than one. Could it be MVUE?

Suppose $h = h(T)$ with T complete and minimal sufficient for θ, and $h(T)$ unbiased for θ. If an MVUE for θ exists then $h(T)$ is a MVUE.

Proof: if an MVUE exists then there is a function of T which is an MVUE, by the RB corollary. But $h(T)$ is the only function of T which is unbiased for θ. So h must be the function of T which an MVUE.