Lecture 32: Information inequality

Suppose that we have a lower bound for the variances of all unbiased estimators of ϑ.
There is an unbiased estimator T of ϑ whose variance is always the same as the lower bound.
Then T is a UMVUE of ϑ.
Although this is not an effective way to find UMVUE's, it provides a way of assessing the performance of UMVUE's.

Theorem 3.3 (Cramér-Rao lower bound). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where Θ is an open set in $\mathcal{R}^k$. Suppose that $T(X)$ is an estimator with $E[T(X)] = g(\theta)$ being a differentiable function of θ; P_θ has a p.d.f. f_θ w.r.t. a measure ν for all $\theta \in \Theta$; and f_θ is differentiable as a function of θ and satisfies

$$\frac{\partial}{\partial\theta}\int h(x) f_\theta(x) d\nu = \int h(x) \frac{\partial}{\partial\theta} f_\theta(x) d\nu, \qquad \theta \in \Theta, \tag{1}$$

for $h(x) \equiv 1$ and $h(x) = T(x)$. Then

$$\mathrm{Var}(T(X)) \geq \left[\tfrac{\partial}{\partial\theta} g(\theta)\right]^\tau [I(\theta)]^{-1} \tfrac{\partial}{\partial\theta} g(\theta), \tag{2}$$

where

$$I(\theta) = E\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^\tau\right\} \tag{3}$$

is assumed to be positive definite for any $\theta \in \Theta$.
Proof. We prove the univariate case ($k = 1$) only.
When $k = 1$, (2) reduces to

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2}. \tag{4}$$

From the Cauchy-Schwartz inequality, we only need to show that

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)$$

and

$$g'(\theta) = \mathrm{Cov}\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right).$$

From condition (1) with $h(x) = 1$,

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int \frac{\partial}{\partial\theta} f_\theta(X) d\nu = \frac{\partial}{\partial\theta}\int f_\theta(X) d\nu = 0.$$

From condition (1) with $h(x) = T(x)$,

$$E\left[T(X)\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int T(x)\frac{\partial}{\partial\theta} f_\theta(X) d\nu = \frac{\partial}{\partial\theta}\int T(x) f_\theta(X) d\nu = g'(\theta).$$

The $k \times k$ matrix $I(\theta)$ in (3) is called the *Fisher information matrix.*
The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values and, therefore, the more accurately θ can be estimated. Thus, $I(\theta)$ is a measure of the information that X contains about the unknown θ.
The inequalities in (2) and (4) are called *information inequalities.*
The following result is helpful in finding the Fisher information matrix.

Proposition 3.1. (i) Let X and Y be independent with the Fisher information matrices $I_X(\theta)$ and $I_Y(\theta)$, respectively. Then, the Fisher information about θ contained in (X, Y) is $I_X(\theta) + I_Y(\theta)$. In particular, if $X_1, ..., X_n$ are i.i.d. and $I_1(\theta)$ is the Fisher information about θ contained in a single X_i, then the Fisher information about θ contained in $X_1, ..., X_n$ is $nI_1(\theta)$.
(ii) Suppose that X has the p.d.f. f_θ that is twice differentiable in θ and that (1) holds with $h(x) \equiv 1$ and f_θ replaced by $\partial f_\theta / \partial \theta$. Then

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X)\right]. \tag{5}$$

Proof. Result (i) follows from the independence of X and Y and the definition of the Fisher information. Result (ii) follows from the equality

$$\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial\theta\partial\theta^\tau} f_\theta(X)}{f_\theta(X)} - \frac{\partial}{\partial\theta} \log f_\theta(X) \left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]^\tau.$$

Example 3.9. Let $X_1, ..., X_n$ be i.i.d. with the Lebesgue p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $f(x) > 0$ and $f'(x)$ exists for all $x \in \mathcal{R}$, $\mu \in \mathcal{R}$, and $\sigma > 0$ (a location-scale family). Let $\theta = (\mu, \sigma)$. Then, the Fisher information about θ contained in $X_1, ..., X_n$ is (exercise)

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \int \frac{[f'(x)]^2}{f(x)} dx & \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx \\ \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx & \int \frac{[xf'(x)+f(x)]^2}{f(x)} dx \end{pmatrix}.$$

Note that $I(\theta)$ depends on the particular parameterization.
If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information that X contains about η is

$$\tfrac{\partial}{\partial\eta}\psi(\eta) I(\psi(\eta)) \left[\tfrac{\partial}{\partial\eta}\psi(\eta)\right]^\tau.$$

However, the Cramér-Rao lower bound in (2) or (4) is not affected by any one-to-one reparameterization.

If we use inequality (2) or (4) to find a UMVUE $T(X)$, then we obtain a formula for $\text{Var}(T(X))$ at the same time.
On the other hand, the Cramér-Rao lower bound in (2) or (4) is typically not sharp.

Under some regularity conditions, the Cramér-Rao lower bound is attained if and only if f_θ is in an exponential family; see Propositions 3.2 and 3.3 and the discussion in Lehmann (1983, p. 123).
Some improved information inequalities are available (see, e.g., Lehmann (1983, Sections 2.6 and 2.7)).

Proposition 3.2. Suppose that the distribution of X is from an exponential family $\{f_\theta : \theta \in \Theta\}$, i.e., the p.d.f. of X w.r.t. a σ-finite measure is

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\}c(x), \tag{6}$$

where Θ is an open subset of $\mathcal{R}^k$.
(i) The regularity condition (1) is satisfied for any h with $E|h(X)| < \infty$ and (5) holds.
(ii) If $\underline{I}(\eta)$ is the Fisher information matrix for the natural parameter η, then the variance-covariance matrix $\mathrm{Var}(T) = \underline{I}(\eta)$.
(iii) If $\overline{I}(\vartheta)$ is the Fisher information matrix for the parameter $\vartheta = E[T(X)]$, then $\mathrm{Var}(T) = [\overline{I}(\vartheta)]^{-1}$.
Proof. (i) This is a direct consequence of Theorem 2.1.
(ii) The p.d.f. under the natural parameter η is

$$f_\eta(x) = \exp\left\{\eta^\tau T(x) - \zeta(\eta)\right\} c(x).$$

From Theorem 2.1, $E[T(X)] = \frac{\partial}{\partial\eta}\zeta(\eta)$. The result follows from

$$\tfrac{\partial}{\partial\eta}\log f_\eta(x) = T(x) - \tfrac{\partial}{\partial\eta}\zeta(\eta).$$

(iii) Since $\vartheta = E[T(X)] = \frac{\partial}{\partial\eta}\zeta(\eta)$,

$$\underline{I}(\eta) = \tfrac{\partial\vartheta}{\partial\eta}\overline{I}(\vartheta)\left(\tfrac{\partial\vartheta}{\partial\eta}\right)^\tau = \tfrac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta)\overline{I}(\vartheta)\left[\tfrac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta)\right]^\tau.$$

By Theorem 2.1 and the result in (ii), $\frac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta) = \mathrm{Var}(T) = \underline{I}(\eta)$. Hence

$$\overline{I}(\vartheta) = [\underline{I}(\eta)]^{-1}\underline{I}(\eta)[\underline{I}(\eta)]^{-1} = [\underline{I}(\eta)]^{-1} = [\mathrm{Var}(T)]^{-1}.$$

A direct consequence of Proposition 3.2(ii) is that the variance of any linear function of T in (6) attains the Cramér-Rao lower bound.
The following result gives a necessary condition for $\mathrm{Var}(U(X))$ of an estimator $U(X)$ to attain the Cramér-Rao lower bound.

Proposition 3.3. Assume that the conditions in Theorem 3.3 hold with $T(X)$ replaced by $U(X)$ and that $\Theta \subset \mathcal{R}$.
(i) If $\mathrm{Var}(U(X))$ attains the Cramér-Rao lower bound in (4), then

$$a(\theta)[U(X) - g(\theta)] = g'(\theta)\frac{\partial}{\partial\theta}\log f_\theta(X) \quad \text{a.s. } P_\theta$$

for some function $a(\theta)$, $\theta \in \Theta$.
(ii) Let f_θ and T be given by (6). If $\text{Var}(U(X))$ attains the Cramér-Rao lower bound, then $U(X)$ is a linear function of $T(X)$ a.s. P_θ, $\theta \in \Theta$.

Example 3.10. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.
Let f_μ be the joint distribution of $X = (X_1, ..., X_n)$. Then

$$\tfrac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^{n} (X_i - \mu)/\sigma^2.$$

Thus, $I(\mu) = n/\sigma^2$.
It is obvious that $\text{Var}(\bar{X})$ attains the Cramér-Rao lower bound in (4).
Consider now the estimation of $\vartheta = \mu^2$.
Since $E\bar{X}^2 = \mu^2 + \sigma^2/n$, the UMVUE of ϑ is $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$.
A straightforward calculation shows that

$$\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

On the other hand, the Cramér-Rao lower bound in this case is $4\mu^2\sigma^2/n$.
Hence $\text{Var}(h(\bar{X}))$ does not attain the Cramér-Rao lower bound.
The difference is $2\sigma^4/n^2$.

Condition (1) is a key regularity condition for the results in Theorem 3.3 and Proposition 3.3.
If f_θ is not in an exponential family, then (1) has to be checked.
Typically, it does not hold if the set $\{x : f_\theta(x) > 0\}$ depends on θ (Exercise 37).
More discussions can be found in Pitman (1979).