L05 Informant (score) and information

1. Informant (score) for $f(x;\theta)$

   (1) Logarithm of $f(x;\,\theta)$ and its derivatives
   Let $f(x;\,\theta)$ be pdf or pmf for X with parameter vector $\theta \in R^k$. Then $\ln f(x;\,\theta)$ is a 1-1 increasing function of $f(x;\,\theta)$. The gradient vector of $\ln f(x;\,\theta)$ with respect to θ accommodates all first order partial derivatives

$$V(x;\theta) = \nabla \ln f(x;\,\theta) = \frac{\partial \ln f(x;\,\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \ln f(x;\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln f(x;\theta)}{\partial \theta_k} \end{pmatrix}.$$

   All second order derivatives are stored in the Hessian matrix of $\ln f(x;\,\theta)$,

$$H_{\ln f(x;\theta)} = \begin{pmatrix} \frac{\partial^2 \ln f(x;\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \ln f(x;\theta)}{\partial \theta_1 \partial \theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln f(x;\theta)}{\partial \theta_k \partial \theta_1} & \cdots & \frac{\partial^2 \ln f(x;\theta)}{\partial \theta_k^2} \end{pmatrix}$$

   (2) Informant for $f(x;\,\theta)$
   When x is replaced by X, $V(X;\,\theta) = \nabla \ln f(x;\,\theta)$ is random vector and $H_{\ln f(X;\theta)}$ is a random matrix, both measure the sensitivity of $f(x;\,\theta)$ to the change in θ. Specifically we call $V(X;\,\theta)$ the informant or the score function.

   (3) $E[V(X;\,\theta)] = 0$
   Under regularity conditions $\left|\frac{\partial f(x;\theta)}{\partial \theta_i}\right| < g(x)$, $\left|\frac{\partial^2 f(x;\theta)}{\partial \theta_i \partial \theta_j}\right| < g(x)$ and $\int\int_{R^p} g(x)dx < \infty$,
   $\int\int_{R^p} \frac{\partial f(x;\theta)}{\partial \theta_i}\,dx = \frac{\partial}{\partial \theta_i}\int\int_{R^p} f(x;\,\theta)\,dx = \frac{\partial}{\partial \theta_i}1 = 0$ and
   $\int\int_{R^p} \frac{\partial^2 f(x;\theta)}{\partial \theta_i \partial \theta_j}\,dx = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\int\int_{R^p} f(x;\,\theta)\,dx = \frac{\partial^2}{\partial \theta_i \partial \theta_j}1 = 0$.
   Consequently, $E[V(X;\,\theta)] = 0$ since in $E[V(X;\,\theta)] = \begin{pmatrix} E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta_1}\right] \\ \vdots \\ E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta_k}\right] \end{pmatrix}$
   $E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta_i}\right] = \int\int_{R^p} \frac{\partial \ln f(x;\theta)}{\partial \theta_i} f(x;\,\theta)\,dx = \int\int_{R^p} \frac{\partial f(x;\theta)}{\partial \theta_i}\,dx = 0$.

2. Fisher's Information

   (1) Information about θ from $f(x;\,\theta)$
   The variance-covariance matrix of score $V(X;\,\theta)$ is called the Fisher's information about θ from $f(x;\,\theta)$ and is denoted as $I(\theta)$.

   (2) $I(\theta) = E(VV^T) = E\left(\frac{f'_{\theta_i} f'_{\theta_j}}{f^2}\right)_{k\times k}$

   Proof: $I(\theta) = \mathrm{Cov}\,(V) = E\left(VV^T\right) - E(V)[E(V)]^T = E\left(VV^T\right) = E(V_iV_j)_{k\times k}$
   $= E\left(\frac{\partial \ln f(X;\theta)}{\partial \theta_i} \cdot \frac{\partial \ln f(X;\theta)}{\partial \theta_j}\right)_{k\times k} = E\left(\frac{f'_{\theta_i} f'_{\theta_j}}{f^2}\right)_{k\times k}$ □

   Comment: $'$ in f'_{θ_i} is for derivative and T in V^T is for transposing.

(3) $I(\theta) = E\left[-H_{\ln f(X;\theta)}\right]$

Pf: Note that $E\left(\frac{f''_{\theta_i\theta_j}}{f}\right) = \int_{R^p} \frac{f''_{\theta_i\theta_j}}{f} f\, dx = \int_{R^p} f''_{\theta_i\theta_j} dx = \left(\int_{R^p} f\, dx\right)''_{\theta_i\theta_j} = 0$. So

$$\begin{aligned} E\left[-H_{\ln f(X;\theta)}\right] &= E\left[(-\ln f(X;\theta))''_{\theta_i\theta_j}\right]_{k\times k} = E\left[\left(-\frac{f'_{\theta_i}}{f}\right)'_{\theta_j}\right]_{k\times k} \\ &= E\left(-\frac{f f''_{\theta_i\,\theta_j} - f'_{\theta_i} f'_{\theta_j}}{f^2}\right)_{k\times k} = 0 + I(\theta) = I(\theta) \end{aligned}$$

Ex1: Find $I(\mu,\, \sigma^2)$ from $X \sim N(\mu,\, \sigma^2)$

$f(X;\, \mu,\, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$; $\ln f(X;\, \mu,\, \sigma^2) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(X-\mu)^2}{2\sigma^2}$

$$V(X;\, \mu,\, \sigma^2) = \begin{pmatrix} \frac{\partial \ln f(X;\mu,\sigma^2)}{\partial \mu} \\ \frac{\partial \ln f(X;\mu,\sigma^2)}{\partial\, \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma}\frac{X-\mu}{\sigma} \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^2}\left(\frac{X-\mu}{\sigma}\right)^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma} Z \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^2} Z^2 \end{pmatrix}$$

where $Z = \frac{X-\mu}{\sigma} \sim N(0,\, 1^2)$. Thus $V(X;\, \mu,\, \sigma^2) \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}\right)$

So $I(\mu,\, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$.

Ex2: From $V(X;\mu,\sigma^2)$ in Ex1, $H_{\ln f(X;\,\mu,\sigma^2)} = \begin{pmatrix} \frac{-1}{\sigma^2} & \frac{-(X-\mu)}{\sigma^4} \\ \frac{-(X-\mu)}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \end{pmatrix} = \begin{pmatrix} \frac{-1}{\sigma^2} & \frac{-Z}{\sigma^3} \\ \frac{-Z}{\sigma^3} & \frac{1}{2\sigma^4} - \frac{Z^2}{\sigma^4} \end{pmatrix}$

So $E\left[H_{\ln f(X;\,\mu,\sigma^2)}\right] = \begin{pmatrix} \frac{-1}{\sigma^2} & 0 \\ 0 & \frac{-1}{2\sigma^4} \end{pmatrix}$. Thus $I(\mu,\, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$

3. Information from a sample

(1) Information from independent $X_1, ..., X_n$
Suppose $X_1, ..., X_n$ are independent, $f(x_i;\, \theta)$ is pdf or pmf for X_i with score $V(X_i;\theta)$ and information $I_{X_i}(\theta)$. Then the joint pdf or pmf for $X_1, .., X_n$ is $\prod_i f(x_i;\, \theta)$ with score $\sum_i \nabla \ln f(X_i;\, \theta)$ and information $I(\theta) = \sum_i I_{X_i}(\theta)$.

(2) Information from iid $X_1, ..., X_n$
Suppose $X_1, .., X_n$ are iid with pdf or pmf $f(x;\, \theta)$, score $V(X;\, \theta)$ and information $I(\theta)$. Then the score from $X_1, .., X_n$ is $\sum_i \nabla \ln f(X_i;\, \theta)$ and information is $\sum_i I(\theta) = nI(\theta)$.

Ex3: $X \sim N(\mu,\, \sigma^2) \Rightarrow I(\mu,\, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$. Then the information based on the distribution of a random sample $X_1, ..., X_n$ from X is $nI(\mu,\, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$

L06 Cramer-Rao lower bound

1. Cramer-Rao lower bound

   (1) Derivative matrix of $y(x) \in R^n$ to $x \in R^p$.

   For $y(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{pmatrix} \in R^n$ where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in R^p$, define

   $$\frac{\partial y}{\partial x^T} = \begin{pmatrix} (y_1)'_{x_1} & \cdots & (y_1)'_{x_p} \\ \vdots & \ddots & \vdots \\ (y_n)'_{x_1} & \cdots & (y_n)'_{x_p} \end{pmatrix} = \left(\frac{\partial y_i}{\partial x_j}\right)_{n\times p} \text{ and } \frac{\partial y^T}{\partial x} = \left(\frac{\partial y}{\partial x^T}\right)^T.$$

   (2) A covariance matrix

   Let $V = V(X;\, \theta)$ be the score from $f(x;\, \theta)$ where $\theta \in R^k$ and $Y = Y(X) \in R^s$ be a vector valued function of X. Then $\text{Cov}_\theta\,[Y,\, V(X;\, \theta)] = \frac{\partial E_\theta(Y)}{\partial \theta^T} \in R^{s\times k}$

   Proof: Note that $E_\theta(V) = 0$. So $\text{Cov}_\theta\,(Y,\, V) = E_\theta\left(YV^T\right) = [E_\theta(Y_iV_j)]_{s\times k}$. But

   $$\begin{aligned} E_\theta(Y_iV_j) &= E_\theta\left[Y_i(X)\,\frac{\partial \ln f(X;\theta)}{\partial \theta_j}\right] = \int\int Y_i(x)\frac{\partial \ln f(x;\theta)}{\partial \theta_j}\, f(x;\, \theta)\, dx \\ &= \int\int Y_i(x)\frac{\partial f(x;\theta)}{\partial \theta_j}dx = \frac{\partial}{\partial \theta_j}\int\int Y_i(x) f(x;\, \theta)dx = \frac{\partial\, E_\theta(Y_i)}{\partial\, \theta_j} \end{aligned}$$

   Comment: $\text{Cov}_\theta(V,\, Y) = [\text{Cov}_\theta(Y,\, V)]^T = \frac{\partial\,[E_\theta(Y)]^T}{\partial\,\theta}$

   (3) An inequality from a variance-covariance matrix

   Suppose $\begin{pmatrix} X \\ Y \end{pmatrix} \sim (\mu,\, \Sigma)$ where $X \in R^p$, $Y \in R^q$ and $\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$.

   Then $\Sigma_{YY} \geq \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$

   Proof: For all $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in R^{p+q}$, $0 \leq \text{var}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^T \Sigma \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. So $\Sigma \geq 0$.

   For all γ, $0 \leq (A\gamma)^T\Sigma(A\gamma) = \gamma^T(A^T\Sigma A)\gamma$. So $A^T\Sigma A \geq 0$.

   With $A = \begin{pmatrix} -\Sigma_{XX}^{-1}\Sigma_{XY} \\ I \end{pmatrix}$, $A^T = (-\Sigma_{YX}\Sigma_{XX}^{-1},\, I)$ and $A^T\,\Sigma\,A = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$.

   So $\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \geq 0$, i.e., $\Sigma_{YY} \geq \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$.

   (4) Cramer-Rao lower bound

   $I(\theta) \in R^{k\times k}$ is the Fisher's information from the distribution of X with parameter $\theta \in R^k$. $Y(X) \in R^s$ is a vector-valued function of X. Then

   $$\left[\frac{\partial\, E_\theta(Y)}{\partial\, \theta^T}\right][I(\theta)]^{-1}\left[\frac{\partial\, E_\theta(Y)}{\partial\, \theta^T}\right]^T \leq \text{Cov}_\theta(Y).$$

   Proof: Let $V = V(X;\, \theta) \in R^k$ be the score. Then

   $$\text{Cov}_\theta\left(\begin{pmatrix} V \\ Y \end{pmatrix}\right) = \begin{pmatrix} \text{Cov}_\theta(V) & \text{Cov}_\theta(V,\, Y) \\ \text{Cov}_\theta(Y,\, V) & \text{Cov}_\theta(Y) \end{pmatrix} = \begin{pmatrix} I(\theta) & \frac{\partial\, E_\theta(Y^T)}{\partial\,\theta} \\ \frac{\partial\, E_\theta(Y)}{\partial\, \theta^T} & \text{Cov}_\theta(Y) \end{pmatrix}.$$

   By (3), the inequality holds.

   Comment: $\left[\frac{\partial\, E_\theta(Y)}{\partial\, \theta^T}\right][I(\theta)]^{-1}\left[\frac{\partial\, E_\theta(Y)}{\partial\, \theta^T}\right]^T$, a lower bound of $\text{Cov}_\theta(Y)$ called the Cramer-Rao lower bound, relates to Y only through $E_\theta(Y)$ and is denoted as $\text{CRLB}(E_\theta(Y))$.

2. Efficient statistics and MVUEs

(1) CRLB for the variance-covariance matrix of a statistic
Let $X_1, ..., X_n$ be a sample from a population with Fisher information $I(\theta) \in R^{k\times k}$ and $Y \in R^s$ be a vector of statistics. Then

$$\text{CRLB}(E_\theta(Y)) = \left[\frac{\partial E_\theta(Y)}{\partial \theta^T}\right][nI(\theta)]^{-1}\left[\frac{\partial E_\theta(Y)}{\partial \theta^T}\right]^T \le \text{Cov}_\theta(Y).$$

Proof: Y is a vector-valued function of $X_1, ..., X_n$. The information on θ from the distribution of $X_1, ..., X_n$ is $nI(\theta)$. So

$$\text{CRLB}(E_\theta(Y)) = \left[\frac{\partial E_\theta(Y)}{\partial \theta^T}\right][nI(\theta)]^{-1}\left[\frac{\partial E_\theta(Y)}{\partial \theta^T}\right]^T \le \text{Cov}_\theta(Y).$$

(2) Efficient statistics and MVUEs
A statistic Y is an efficient statistic if $\text{Cov}_\theta(Y) = \text{CRLB}(E_\theta(Y))$.
If Y is an efficient statistic and $E_\theta(Y) = \eta(\theta) = \eta$, then Y is MVUE for η.

Proof: $E_\theta(Y) = \eta(\theta)$. So Y is an UE for $\eta(\theta)$. Suppose T is also an UE for $\eta(\theta)$. Then $E_\theta(T) = \eta(\theta) = E_\theta(Y)$. But Y is an efficient statistic. Thus
$\text{Cov}_\theta(Y) = \text{CRLB}(E_\theta(Y)) = \text{CRLB}(E_\theta(T)) \le \text{Cov}_\theta(T)$.

Ex: $\widehat{\theta} = \begin{pmatrix}\overline{X}\\ S^2\end{pmatrix}$ is from a random sample from $N(\mu, \sigma^2)$ with $\theta = \begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}$.

(i) Is $\widehat{\theta}$ an efficient statistic?
By L05, $I(\theta) = \begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{pmatrix}$. With $E(\widehat{\theta}) = \theta$, $\frac{\partial E(\widehat{\theta})}{\partial \theta^T} = I_2$.
So $\text{CRLB}(\theta) = [nI(\theta)]^{-1} = \begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}$. By HW02, $\text{Cov}(\widehat{\theta}) = \begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n-1}\end{pmatrix}$.
Thus $\text{Cov}(\widehat{\theta}) \ne \text{CRLB}(E(\widehat{\theta}))$. Therefore $\widehat{\theta}$ is not an efficient statistic.

(ii) Is $\widehat{\theta}$ MVUE for θ?
Yes. By HW02, $\widehat{\theta}$ is MVUE for θ.

(iii) Find $\text{CRLB}(\mu)$
$\text{CRLB}(\mu) = [\frac{\partial \mu}{\partial(\mu,\sigma^2)}][nI(\theta)]^{-1}[\frac{\partial \mu}{\partial(\mu,\sigma^2)}]^T = (1, 0)\begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}\begin{pmatrix}1\\ 0\end{pmatrix} = \frac{\sigma^2}{n}$.

(iv) Is $\overline{X}$ an efficient statistic?
$E(\overline{X}) = \mu$. $\text{Cov}(\overline{X}) = \frac{\sigma^2}{n} = \text{CRLB}(\mu) = \text{CRLB}(E(\overline{X}))$.
So $\overline{X}$ is an efficient statistic.

(v) Is $\overline{X}$ MVUE for μ?
Yes since $\overline{X}$ is efficient and $E(\overline{X}) = \mu$.