CRLB ≠ UMVU

Example: $x_i \sim \mathcal{N}_\vartheta(\vartheta, \sigma^2)$ σ^2: known, fixed

$i = 1, \dots, n$

$g(\vartheta) = \vartheta$

Joint probability distribution:

$$P_\vartheta(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \vartheta)^2}{2\sigma^2}\right)$$

log likelihood

$$\log P_\vartheta(x_1, \dots, x_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \vartheta)^2$$

→ Calculate CRLB.

$$\frac{\partial}{\partial \vartheta} \log P_\vartheta(x) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \vartheta) = \frac{n(\bar{x} - \vartheta)}{\sigma^2}$$

(how much the log likelihood changes when ϑ changes)

Fisher Information Matrix:

$$I(\vartheta) = \mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial \vartheta} \log P_\vartheta(x)\right)^2\right]$$

$$= \frac{n^2}{\sigma^4} \operatorname{Var}(\bar{x}) = \frac{n^2}{\sigma^4} \cdot \frac{\sigma^2}{n} = \frac{n}{\sigma^2}$$

$$I_n(\vartheta) = n/\sigma^2$$

The UMVU estimator for ϑ

$T = \frac{1}{n}\sum_{i=1}^{n} x_i$ sufficient, complete statistic

$\mathrm{Var}_\vartheta(T) = \frac{\sigma^2}{n}$

T attains the CRLB $\Rightarrow$ UMVU for ϑ.

An alternative estimator for ϑ that attains CRLB but it is <u>not</u> UMVU

$T(x_1, \ldots, x_n) = x_1$ (first sample) for $g(\vartheta) = \vartheta$

$\mathbb{E}[T] = \mathbb{E}[x_1] = \vartheta$ (unbiased)

$\mathrm{Var}_\vartheta[T] = \mathrm{Var}_\vartheta(x_1) = \sigma^2$

$CRLB = \frac{1}{n}(I(\vartheta))^{-1} = \sigma^2/n$

$\mathrm{Var}_\vartheta(x_1) = \sigma^2 > \frac{\sigma^2}{n}$

Estimator unbiased but does not attain CRLB.

Alternative estimator for $g(\theta)=\theta$

randomized.

$$T(x_1,\dots,x_n,\underbrace{u}_{\text{randomness}}) = \begin{cases} x_1 & u < 1/n \\ x_2 & 1/n \le u < 2/n \\ \vdots & \vdots \\ x_n & \frac{n-1}{n} \le u < 1 \end{cases}$$

We pick one sample at random (uniformly) from $\{x_1,\dots,x_n\}$ this estimator is not deterministic in data.

X_i are i.i.d. and we randomly pick one:

$$\mathbb{E}[T] = \mathbb{E}_u[\mathbb{E}[T|u]] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) = \theta.$$

(unbiased)

$$Var_\theta(T) = Var_\theta(\text{randomly pick from } \{x_i\}) =$$

$$= \frac{1}{n}\sum Var(x_i) = \sigma^2$$

$$CRLB := \frac{\sigma^2}{n} \Longrightarrow Var_\theta(T) > \frac{\sigma^2}{n} \text{ CRLB}$$

Estimator T does not attain the CRLB

NOT UMVU because T is not a function of the complete sufficient statistic

CRLB $\Rightarrow$ UMVU estimator

Exponential families & CRLB

We consider 1-dimensional exponential family:

$$P_\theta(x) = \exp\left(C(\theta)T(x) - d(\theta)\right)h(x)$$

where: $\theta \in \mathbb{R}$: scalar parameter

$T(x)$: sufficient statistic

$C(\theta)$: natural parameter

$d(\theta)$: function of parameter

$h(x)$: function of data

Theorem: Let $X_i \overset{i.i.d}{\sim} P_\theta(x)$ from the exponential family. Then the sufficient statistic T is an efficient estimator for $g(\theta) = E_\theta(T)$

[This means that $T(x)$ is unbiased and its variance achieves CRLB.]

Proof: $S_\theta(x) = \frac{\partial}{\partial \theta} \log p_\theta(x) = -d'(\theta) + c'(\theta) T(x)$

The score is linear in the sufficient statistic T.

$$\Rightarrow \mathbb{E}_\theta\left(S_\theta(x) \left(T(x) - \mathbb{E}_\theta(T(x))\right)\right) = \frac{d}{d\theta} \mathbb{E}_\theta[T(x)] = g'(\theta)$$

↑

By C-R proof: Covariance between the score and the estimators gives the derivative of the expectation

By C-S inequality $\left(\mathbb{E}[X \cdot Y]\right)^2 \leq \mathbb{E}(X^2) \cdot \mathbb{E}(Y^2)$

$X = S(\theta)$

$Y = T(x) - \mathbb{E}_\theta(T(x))$

$$\left(g'(\theta)\right)^2 \leq \mathbb{E}_\theta\left(S_\theta(x)^2\right) \cdot \mathrm{Var}_\theta(T) \Rightarrow$$

$$\mathrm{Var}_\theta(T) \geq \frac{\left(g'(\theta)\right)^2}{\mathbb{E}_\theta\left(S_\theta(x)\right)^2} = \frac{\left(g'(\theta)\right)^2}{I(\theta)}$$

$$\Rightarrow \mathrm{Var}_\theta(T) = \frac{\left(g'(\theta)\right)^2}{I(\theta)}$$

Because the sufficient statistic $T(x)$ is linearly related to the score function.

$$T(x) - E_\theta(T(x)) = \lambda \cdot S_\theta(x) \qquad \lambda: \text{constant}$$

$$T(x) = E_\theta(T(x)) + \lambda \cdot S_\theta(x)$$

$$S_\theta(x) = c'(\theta) T(x) - d'(\theta)$$

$$T(x) = \frac{1}{c'(\theta)} \cdot S_\theta(x) + \frac{d'(\theta)}{c'(\theta)}.$$

Lemma: Assume A1 - A3 hold, and that $T: \mathcal{X} \to \mathbb{R}$ is unbiased estimator for $g(\theta) \in \mathbb{R}$ reaching the CRLB. Then there exists differentiable $c(\theta)$, $d(\theta)$ and $h(x)$ s.t.

$$p_\theta(x) = \exp(c(\theta) T(x) - d(\theta) h(x)) \quad \forall \theta \in \Theta.$$

By Cramér - Rao lower bound: Let T be an unbiased estimator of $g(\theta) \in \mathbb{R}^k$ with $E_\theta(T^2) < \infty$, $\forall \theta \in \Theta$. Then assuming I

is invertible:

$$\mathrm{Cov}(T) \geq \nabla_\theta g(\theta) \left(I(\theta)\right)^{-1} \nabla_\theta g(\theta)^T.$$

Proof: T_j unbiased $\Rightarrow$ $\mathrm{bias}_\theta(T_j) = 0 \quad \forall j \in \{1,\dots,k\}$

$$\Rightarrow \mathrm{bias}_\theta\left(\sum_j a_j T_j\right) = 0 \qquad \sum a_j \frac{\partial}{\partial \theta_j} g(\theta)$$

$$\Rightarrow a^T \mathrm{Cov}(T)\, a = \mathrm{Cov}(a^T T) \geq a^T \nabla_\theta g(\theta) \left(I(\theta)\right)^{-1} \nabla_\theta g(\theta)^T a$$

∎

Fisher Information Matrix:

Definition: We define the fisher Information Matrix as:

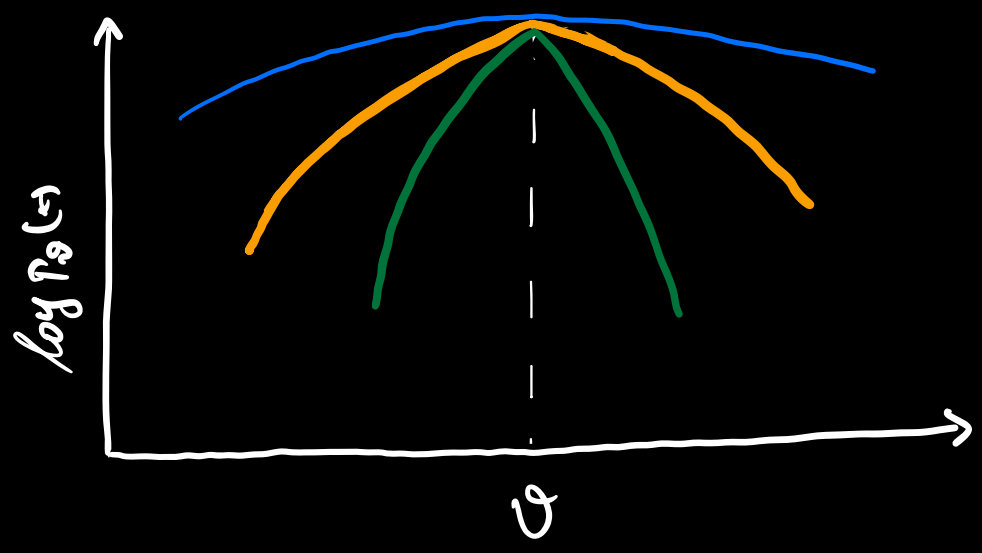
$$I(\theta)_{ij} = -\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_i} \log P_\theta(x) \frac{\partial}{\partial \theta_j} \log P_\theta(x)\right]$$

for $i,j \in \{1,\dots,k\}$

Note: · Fisher Information matrix is a covariance matrix.

· FIM shows us how sensitive the model is to changes of θ.

Curvature of log likelihood function



- $\sigma^2 = 2.0$ sharp peak
- $\sigma^2 = 1.0$ moderate curvature.
- $\sigma^2 = 0.2$ flat curvature

$$\text{Curvature} = -\frac{d^2}{d\theta} \log p_\theta(x) = \frac{1}{\sigma^2}$$

more curvature => sharper likelihood

=> smaller estimator variance

high fisher Information matrix
=> low uncertainty

low fisher Information Matrix
$\Rightarrow$ higher uncertainty.

Kullback-Leibler divergence:

Q: If I observe some data, can I tell whether it came from $P_{\vartheta_1}(x)$ or $P_{\vartheta_2}(x)$? Is it easy to distinguish between two parameter choises ϑ_1, ϑ_2 in the same statistical model?

Example 1: $P_{\vartheta_1}(x) = \mathbb{1}(x \in \{1,2\})$

$P_{\vartheta_2}(x) = \mathbb{1}(x \in \{2,3\})$

(a) if $x = 1$ $\quad P_{\vartheta_1}(x) = 1 \quad P_{\vartheta_2}(x) = 0$

(b) if $x = 2$ $\quad P_{\vartheta_1}(x) = 1 \quad P_{\vartheta_2}(x) = 1$

(c) if $x = 3$ $\quad P_{\vartheta_1}(x) = 0 \quad P_{\vartheta_2}(x) = 1$

(a), (c) are distinguishable.

(b) distributions are identical. I cannot tell them apart.

Definition: The discriminating power between ϑ_1, ϑ_2 of result x is

$$h(\vartheta_1 | \vartheta_2)(x) = \log \frac{P_{\vartheta_1}(x)}{P_{\vartheta_2}(x)}$$

defined on $\{P_{\vartheta_1}(x) > 0\} \cup \{P_{\vartheta_2}(x) > 0\}$

[This power tells you how much evidence the data x gives in favor of ϑ_1 over ϑ_2.]

To get an overall measure of distinguishability, we average the log likelihood ratio under P_{ϑ_1}:

$$D_{KL}(P_{\vartheta_1} \| P_{\vartheta_2}) = \mathbb{E}_{\vartheta_1}\left[\log \frac{P_{\vartheta_1}(x)}{P_{\vartheta_2}(x)}\right]$$

$$= \int \log \frac{P_{\vartheta_1}(x)}{P_{\vartheta_2}(x)} P_{\vartheta_1}(x) \mu(dx)$$

Properties of KL divergence:

1. $D_{KL}(P_{\vartheta_1} \| P_{\vartheta_2}) \geq 0$
2. $D_{KL}(P_{\vartheta_1} \| P_{\vartheta_2}) = 0 \iff P_{\vartheta_1}(x) = P_{\vartheta_2}(x)$ μ-a.s.

Example: (Poisson)

Compare $P_{\vartheta_1}(x) = \frac{\vartheta_1^x \cdot e^{-\vartheta_1}}{x!}$, $P_2(x) = \frac{\vartheta_2^x e^{-\vartheta_2}}{x!}$

KL divergence:

$$D_{KL}(P_{\vartheta_1} \| P_{\vartheta_2}) = \mathbb{E}_{\vartheta_1}\left[\log \frac{P_{\vartheta_1}(x)}{P_{\vartheta_2}(x)}\right]$$

$$= \mathbb{E}_{\vartheta_1}\left[x \log\left(\frac{\vartheta_1}{\vartheta_2}\right) - (\vartheta_1 - \vartheta_2)\right]$$

Since $\mathbb{E}_{\vartheta_1}(X) = \vartheta_1$

$$= \vartheta_1\left(\log\left(\frac{\vartheta_1}{\vartheta_2}\right) - (\vartheta_1 - \vartheta_2)\right)$$

Remark: KL divergence is not symmetric

$$D_{KL}(P_{\vartheta_1} \| P_{\vartheta_2}) \neq D_{KL}(P_{\vartheta_2} \| P_{\vartheta_1})$$

Lemma: Let Y r.v. with $\mathbb{E}[|Y|] = \mathbb{E}[Y^+] + \mathbb{E}[Y^-] < \infty$

For any convex function φ: $\mathcal{Y} \to \mathcal{Y}$, $\mathbb{E}[\varphi(Y^-)] < \infty$

Proof: Consider $Y = \frac{P_{\theta_2}(x)}{P_{\theta_1}(x)}$

$$E_{\theta_1}(|Y|) = \int \frac{P_{\theta_2}(x)}{P_{\theta_1}(x)} \cdot P_{\theta_1}(x)\, \mu(dx) = 1.$$

$$\Rightarrow E_{\theta_1}\left[\left(\log \frac{P_{\theta_2}(x)}{P_{\theta_1}(x)}\right)^-\right] = E_{\theta_1}\left[(-\log Y)^-\right] < \infty$$

$\Rightarrow$ The integral is well defined on $\mathbb{R} \cup \{+\infty\}$

By Jensen inequality

$$D_{KL}(P_{\theta_1} \| P_{\theta_2}) = E_{\theta_1}\left(-\log \frac{P_{\theta_2}(x)}{P_{\theta_1}(x)}\right) \geq -\log E_{\theta_1}\left[\frac{P_{\theta_2}(x)}{P_{\theta_1}(x)}\right]$$

$$= -\log 1 = 0$$

To show the reverse direction we need reverse Jensen.

$-\log Y$ Strictly Convex $\Rightarrow Y = \frac{P_{\theta_2}(x)}{P_{\theta_1}(x)} = C = 1$

P_{θ_1} - a.s.

Lemma: Let Y RV with $\mathbb{E}[Y]$ well defined on $\mathbb{R} \cup \{\infty\}$ and ϕ is a convex function Then

$$\mathbb{E}[\phi(Y)] \geq \phi(\mathbb{E}(Y))$$

with equality if - f $\varphi(y) = ay + b$ $\mathbb{P}$-a.s.

In particular if φ is strictly convex

then $Y = c$ $\mathbb{P}$-a.s.

Theorem: Let A1 - A3 hold, and assume that

$$\frac{\partial^2}{\partial \vartheta_j \partial \vartheta_i} \int \log \frac{p_{\vartheta_1}(x)}{p_{\vartheta_2}(x)} p_{\vartheta_1}(x) \mu(dx)$$

$$= \int \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \left(\log \frac{p_{\vartheta_1}(x)}{p_{\vartheta_2}(x)} p_{\vartheta_1}(x) \right) \mu(dx)$$

then $I(\vartheta)_{ij} = \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} D_{KL}(\vartheta \mid \vartheta') \Big|_{\vartheta = \vartheta'}$

Fisher Information Matrix

- At $\theta = \theta'$ $KL = 0$
- second derivative how fast the KL divergence grows as we move away from θ

that is exactly what F.I. quantifie