16. Cramer Rao Bounds

ECE 830, Spring 2014

The Cramer-Rao Lower Bound

The Cramer-Rao Lower Bound (CRLB) sets a lower bound on the variance of any unbiased estimator. This can be extremely useful in several ways:

1. If we find an estimator that achieves the CRLB, then we know that we have found an MVUE estimator!
2. The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (We know we're doing very well if our estimator is "close" to the CRLB)
3. The CRLB enables us to rule-out impossible estimators. That is, we know that it is physically impossible to find an unbiased estimator that beats the CRLB. This is useful in feasibility studies.
4. The theory behind the CRLB can tell us if an estimator exists which achieves the bound.

The CRLB in a Nutshell

We wish to estimate a parameter

$$\theta^* = [\theta_1^*, \theta_2^*, \cdots, \theta_p^*]^\top.$$

What can we saw about the (co)variance of any unbiased estimator $\widehat{\theta}$, where $\mathbb{E}\left[\widehat{\theta}_i\right] = \theta_i^*, i = 1, \cdots, p$?
The CRLB tells us that

$$\mathsf{Var}(\widehat{\theta}_i) \geq \left[I^{-1}(\theta^*)\right]_{ii}$$

where $I(\theta^*)$ is the Fisher Information Matrix:

$$[I(\theta^*)]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\theta^*}\right], i, j = 1, \ldots, p$$

Fisher Information Matrix

Definition: Fisher Information Matrix

For $\theta^* \in \mathbb{R}^p$, the *Fisher Information Matrix* is

$$I(\theta^*) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right) \left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^\top \Big|_{\theta=\theta^*} \right]$$
$$= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \log p(x|\theta) \Big|_{\theta=\theta^*} \right] \in \mathbb{R}^{p \times p}.$$

so that

$$[I(\theta^*)]_{j,k} = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta_j \partial \theta_k} \Big|_{\theta=\theta^*} \right]$$

Note that if $\theta^* \in \mathbb{R}$ (i.e. $p = 1$), then

$$I(\theta^*) := \mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta^*} \right] = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \Big|_{\theta=\theta^*} \right].$$

Note about vector calculus

Recall that if $\phi : \mathbb{R}^p \longrightarrow \mathbb{R}$, then

$$\begin{aligned}
\frac{\partial \phi}{\partial \theta} &:= \left[\frac{\partial \phi}{\partial \theta_1} \cdots \frac{\partial \phi}{\partial \theta_p} \right]^\top, \\
\frac{\partial \phi}{\partial \theta^\top} &:= \left[\frac{\partial \phi}{\partial \theta_1} \cdots \frac{\partial \phi}{\partial \theta_p} \right] \equiv \left(\frac{\partial \phi}{\partial \theta} \right)^\top, \\
\frac{\partial^2 \phi}{\partial \theta \partial \theta^\top} &:= \begin{bmatrix} \frac{\partial^2 \phi}{\partial \theta_1^2} & \frac{\partial^2 \phi}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \phi}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \phi}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \phi}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \phi}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \phi}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \phi}{\partial \theta_p \partial \theta_p} & \cdots & \frac{\partial^2 \phi}{\partial \theta_p^2} \end{bmatrix}
\end{aligned}$$

Theorem: Cramer-Rao Lower Bound

Assume that the pdf $p(x|\theta)$ satisfies the *"regularity" condition*

$$\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right] = 0 \forall \theta.$$

Then the covariance matrix of any unbiased estimator $\widehat{\theta}$ satisfies

$$C_{\widehat{\theta}} \succeq I^{-1}(\theta^*).$$

Moreover, an unbiased estimator $\widehat{\theta}(x)$ may be found that attains the bound $\forall \theta^*$ if and only if

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = I(\theta)(\widehat{\theta}(x) - \theta).$$

Positive semi-definite matrices

Recall that $A \succeq B$ means that $A - B \succeq 0$, or $A - B$ is a positive semi-definite (PSD) matrix, so that $x^\top (A - B)x \geq 0$ for *any* x. Let's say $A - B$ is a PSD matrix, and write its eigendecomposition as

$$A = V^\top \Lambda V = V^\top (\Lambda_A - \Lambda_B) V$$

where V is an orthogonal matrix and Λ, Λ_A and Λ_B are diagonal. Then

$$0 \leq x^\top (A - B)x = x^\top V^\top (\Lambda_A - \Lambda_B) V x = y^\top (\Lambda_A - \Lambda_B) y$$

where $y := Vx$ is x transformed or rotated into the coordinate system corresponding to V. Since the above must hold for all x, we have that $A - B$ is PSD if $y^\top (\Lambda_A - \Lambda_B) y \geq 0$ for all y, which occurs if $(\Lambda_A)_{ii} \geq (\Lambda_B)_{ii}$ for all i.

In the context of the CRLB, this suggests that we can compute the eigendecomposition of $C_{\widehat{\theta}} - I^{-1}(\theta^*) = V^\top(\Lambda_C - \Lambda_I)V$, rotate any θ into the coordinate system corresponding to V to get $\psi = V\theta$. First note that $C_{\widehat{\psi}}$ is diagonal (i.e. we have diagonalized the covariance matrix), and then

$$
\begin{aligned}
C_{\widehat{\theta}} &\succeq I^{-1}(\theta^*) \\
C_{\widehat{\psi}} = C_{V\widehat{\theta}} = VC_{\widehat{\theta}}V^\top &\succeq VI^{-1}(\theta^*)V = I^{-1}(\psi^*) \\
\mathsf{Var}(\widehat{\psi_i}) &\geq [I^{-1}(\psi^*)]_{ii} \; \forall i
\end{aligned}
$$

Example: DC Level in White Gaussian Noise

$x_n = A + w_n, w_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Assume σ^2 is known, and we wish to find the CRLB for $\theta = A$. First check the regularity condition:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right] \\
&= \mathbb{E}\left[\frac{\partial \log(1/(2\pi\sigma^2)^{N/2} \exp\left\{(1/2\sigma^2)\sum_{n=1}^{N}(x_n - A)^2\right\})}{\partial A}\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial A}\right] \\
&= \mathbb{E}\left[\qquad\right] \\
&=
\end{aligned}
$$

Example: (cont.)

Now we can compute the Fisher Information:

$$
\begin{aligned}
I(\theta) &= -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\right] \\
&= \\
&=
\end{aligned}
$$

so the CRLB is

$$\text{CRLB} =$$

$\therefore$ Any unbiased estimator $\widehat{A}$ has $\text{Var}(\widehat{A}) \geq$. But we know that $\widehat{A} = (1/N)\sum_{n=1}^{N} x_n$ has $\text{Var}(\widehat{A}) =$

Proof of CRLB Theorem (Scalar case, $p = 1$)

Now let's derive the CRLB for a scalar parameter θ where the pdf is $p(x|\theta)$. Consider any unbiased estimator of θ:

$$\widehat{\theta}(x): \;\; \mathbb{E}\left[\widehat{\theta}(x)\right] = \int \widehat{\theta}(x) p(x|\theta) dx = \theta.$$

Now differentiate both sides

$$\int \widehat{\theta}(x) \frac{\partial p(x|\theta)}{\partial \theta} dx = \frac{\partial \theta}{\partial \theta}$$

or

$$\int \widehat{\theta}(x) \qquad\qquad p(x|\theta) dx = 1.$$

Now exploiting the regularity condition, since

$$\int \theta \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx =$$

we have

$$\int \qquad\qquad \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx = 1.$$

Proof (cont.)

Now apply the Cauchy-Schwarz inequality to the integral above

$$
\begin{aligned}
1 &= \left(\int (\widehat{\theta}(x) - \theta) \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx \right)^2 \\
&\leq
\end{aligned}
$$

Proof (cont.)

Now note that

$$\mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$$

Why? Regularity condition

$$0 = \mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right] = \int \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx$$

$$\Rightarrow 0 = \frac{\partial}{\partial \theta} \int \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx$$

$$\Rightarrow 0 =$$

Proof (cont.)

Rearranging terms we find

$$-\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\right] \quad =$$

$$=$$

Thus,

$$\text{Var}(\widehat{\theta}(x)) \geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\right]}$$

The CRLB is not always attained.

Example: Phase Estimation (Kay p 33)

$$x_n = A\cos(2\pi f_0 n + \phi) + w_n, n = 1, \cdots, N$$

The amplitude and frequency are assumed known, and we want to estimate the phase ϕ. We assume

$$w_n \sim \mathcal{N}(0, \sigma^2)\text{iid.}$$

$$p(x|\phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(x_n - A\cos(2\pi f_0 n + \phi)\right)^2 \right\}$$

Example: (cont.)

$$\frac{\partial \log p(x|\phi)}{\partial \phi} =$$

$$\frac{\partial^2 \log p(x|\phi)}{\partial \phi^2} =$$

$$-\mathbb{E}\left[\frac{\partial^2 \log p(x|\phi)}{\partial \phi^2}\right] =$$

$$=$$

Example: (cont.)

Since $\frac{1}{N}\sum \cos(4\pi f_0 n) \approx 0$ for f_0 not near 0 or $1/2$,

$$I(\phi) \approx$$

$$\text{Var}(\widehat{\phi}) \geq$$

In this case, it can be shown that there does not exist a g such that

$$\frac{\partial \log p(x|\phi)}{\partial \phi} =$$

Therefore an unbiased phase estimator that attains the CRLB does not exist. However, a MVUE estimator may still exist – only its variance will be larger than the CRLB. Sufficient statistics can help us determine whether a MVUE still exists.

CRLB For Signals in White Gaussian Noise

(Kay p 35)

$$
\begin{aligned}
x_n &= s_n(\theta) + w_n, \qquad n = 1, \cdots, N \\
p(x|\theta) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - s_n(\theta))^2 \right\} \\
\frac{\partial \log p(x|\theta)}{\partial \theta} &= \frac{1}{\sigma^2} \sum_{n=1}^{N} \\
\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= \frac{1}{\sigma^2} \sum_{n=1}^{N} \\
\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \right] &= -\frac{1}{\sigma^2} \sum_{n=1}^{N}
\end{aligned}
$$

CRLB For Signals in WGN, cont.

$$\mathsf{Var}(\widehat{\theta}) \geq \frac{\sigma^2}{\sum_{n=1}^{N}\left(\frac{\partial s_n(\theta)}{\partial \theta}\right)^2}$$

Signals that change rapidly as θ changes result in more accurate estimators.

Definition: Efficient estimator

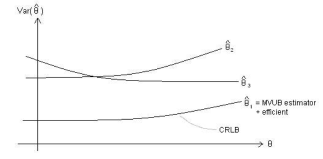
An estimator which is unbiased and attains the CRLB is said to be efficient.

Example:

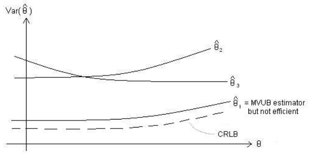
Sample-mean estimator is efficient.

Example: Efficient estimators are MVUE, but MVUE may not be efficient.

Suppose three unbiased estimators exist for a parameter θ.



Efficient and MVUE



MVUE but not efficient

Example: Sinusoidal Frequency Estimation (Kay p 36)

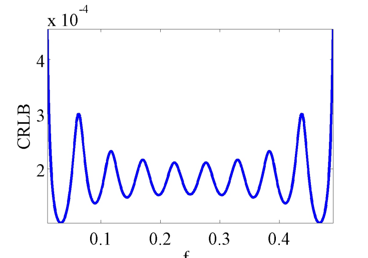
$$s_n(f_0) = A\cos(2\pi f_0 n + \phi), 0 < f_0 < 1/2$$

$$x_n = s_n(f_0) + w_n, n = 1, \cdots, N$$

A, ϕ known, f_0 unknown.

$$\text{Var}(\widehat{f_0}) \geq$$

(from CRLB for signals in AWGN earlier.) Suppose: $A^2/\sigma^2 = 1$ (SNR), $N = 10$, $\phi = 0$.



Some frequencies are easier to estimate (lower CRLB) than others!

Example: DC Level in White Gaussian Noise (Kay p 40).

$$
\begin{aligned}
x_n &= A + w_n, & w_n &\sim \mathcal{N}(0, \sigma^2) & n &= 1, \cdots, N \\
x &= A\mathbf{1} + w & w &\sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{N \times N}) \\
\theta &= [A, \sigma^2]^\top & & A \text{ and } \sigma^2 \text{ both unknown}
\end{aligned}
$$

$$
\begin{aligned}
\log p(x|\theta) &= \\
\frac{\partial \log p(x|\theta)}{\partial A} &= \\
\frac{\partial \log p(x|\theta)}{\partial \sigma^2} &= \\
\frac{\partial^2 \log p(x|\theta)}{\partial A^2} &=
\end{aligned}
$$

Example: (cont.)

$$\frac{\partial^2 \log p(x|\theta)}{\partial A \partial \sigma^2} =$$

$$\frac{\partial^2 \log p(x|\theta)}{\partial (\sigma^2)^2} =$$

$$\Rightarrow I(\theta) =$$

$$\mathsf{Var}(\widehat{A}) \geq$$

$$\mathsf{Var}(\widehat{\sigma^2}) \geq$$

Remarks

Note that CRLB for $\widehat{A}$ is the same whether or not σ^2 is known. This happens in this case due to the diagonal nature of the Fisher Information Matrix.

In general the Fisher Information Matrix is not diagonal and consequently the CRLBs will depend on other unknown parameters.

CRLBs for Subspace Models

We saw before that we could compute a minimum variance unbiased estimator (MVUE) for θ in the subspace model

$$x = H\theta + w$$

by using sufficient statistics and the Rao-Blackwell Theorem.

Does this estimator achieve the Cramer-Rao Lower Bound?

Linear models

General Form of Linear Model (LM)

$$x = H\theta + w$$

x = observation vector
H = known matrix ("observation" or "system" matrix)
θ = unknown parameter vector
w = vector of white Gaussian noise $w \sim \mathcal{N}(0, \sigma^2 I)$

Probability Model for LM

$$\begin{aligned} x &= H\theta + w \\ x &\sim p(x|\theta) = \mathcal{N}(H\theta, \sigma^2 I) \end{aligned}$$

The CRLB and MVUE Estimator

Recall

$$\widehat{\theta} = g(x)$$

achieves the CRLB if and only if

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = I(\theta)(g(x) - \theta)$$

In the case of the linear model,

$$\frac{\partial \log p(x|\theta)}{\partial \theta} =$$

Now using identities

$$\frac{\partial \boldsymbol{b}^\top \theta}{\partial \theta} = \boldsymbol{b} \text{ and } \frac{\partial \theta^\top \boldsymbol{A} \theta}{\partial \theta} = 2\boldsymbol{A}\theta \text{ for } \boldsymbol{A} \text{ symmetric}$$

we have

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = \qquad .$$

Now

$$
\begin{aligned}
I(\theta) =& \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^\top\right] \\
=& \\
=& \\
=&
\end{aligned}
$$

Assuming $H^\top H$ is invertible, we can write

$$
\frac{\partial \log p(x|\theta)}{\partial\theta} = \underbrace{\frac{H^\top H}{\sigma^2}}_{I(\theta)}
$$

$\widehat{\theta} = (H^\top H)^{-1}H^\top x$ is MVUE Estimator for $x = H\theta + w$

Theorem: MVUE Estimator for the LM

If the observed data can be modeled as

$$x = H\theta + w$$

where $w \sim \mathcal{N}(0, \sigma^2 I)$ and $H^\top H$ is invertible, then the MVUE estimator is

$$\widehat{\theta} = (H^\top H)^{-1} H^\top x = H^\# x,$$

the covariance of $\widehat{\theta}$ is

$$C_{\widehat{\theta}} = \sigma^2 (H^\top H)^{-1}$$

and $\widehat{\theta}$ attains the CRLB. *Note:*

$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2 (H^\top H)^{-1})$$

Linear Model Examples

Example: Curve fitting.

$$
\begin{aligned}
x(t_n) &= \theta_1 + \theta_2 t_n + \cdots + \theta_p t_n^{p-1} + w(t_n), n = 1, \cdots, N \\
w(t_n) &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \\
x &= [x(t), \cdots, x(t_n)]^\top \\
\theta &= [\theta_1, \theta_2, \cdots, \theta_p]^\top
\end{aligned}
$$

$$
H =
$$

The MVUE estimator for θ is

$$
\widehat{\theta} = (H^\top H)^{-1} H^\top x
$$

Example: System Identification.

There is an FIR filter h, and we want to know its impulse response. To estimate this, we send a probe signal u through the filter and observe

$$x[n] = \sum_{k=0}^{m-1} h[k]u[n-k] + w[n], \quad n = 0, \cdots, N-1.$$

Goal: given x and u estimate h. In matrix form

$$x = \underbrace{\qquad\qquad\qquad\qquad}_{H}\underbrace{\qquad\qquad}_{\theta} + w$$

$$\widehat{\theta} = (H^\top H)^{-1}H^\top x \leftarrow \text{MVUE estimator}$$

$$\text{Cov}(\widehat{\theta}) = \sigma^2(H^\top H)^{-1} = C_{\widehat{\theta}}$$

Example: (cont.)

An important question in system id is how to chose the input $u[n]$ to "probe" the system most efficiently. First note that

$$\mathsf{Var}(\widehat{\theta}_i) = e_i^\top C_{\widehat{\theta}} e_i$$

where $e_i = [0, \cdots, 0, 1, 0, \cdots, 0]^\top$. Also, since $C_{\widehat{\theta}}^{-1}$ is symmetric and positive definite, we can factor it using Cholesky factorization:

$$C_{\widehat{\theta}}^{-1} = D^\top D,$$

where D is invertible. Note that

$$(e_i^\top D^\top (D^\top)^{-1} e_i)^2 = 1$$

Example: (cont.)

The Schwarz inequality shows

$$
\begin{aligned}
1 &= (e_i^\top D^\top (D^\top)^{-1} e_i)^2 \\
&= \langle De_i, (D^\top)^{-1} e_i \rangle \\
&\leq \qquad\qquad (1) \\
&= \\
&= \\
\Rightarrow \text{Var}(\widehat{\theta}_i) &\geq
\end{aligned}
$$

The minimum variance is achieved when equality is attained above in (1).

Example: (cont.)

This happens only if $\xi_1 = De_i$ is proportional to $\xi_2 = (D^\top)^{-1} e_i$. That is $\xi_1 = c\xi_2$for some constant c. Equivalently

$$D^\top D e_i = c_i e_i \text{ for } i = 1, 2, \cdots, m$$

$$D^\top D = C_{\widehat{\theta}}^{-1} = \frac{H^\top H}{\sigma^2}$$

$$D^\top D e_i = \frac{H^\top H}{\sigma^2} e_i = c_i e_i$$

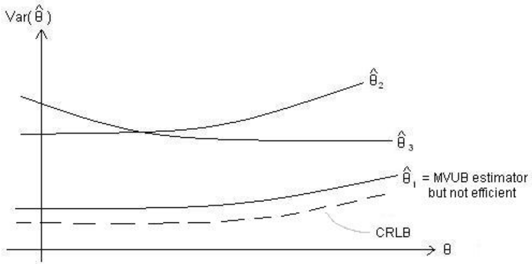
Combining these equations in matrix form

$$H^\top H = \sigma^2 \begin{bmatrix} c_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & c_m \end{bmatrix}.$$

$\therefore$ In order to minimize the variance of the MVUE estimator, $u[n]$ should be chosen to make $H^\top H$ diagonal.

When the CRLB Doesn't Help

The Cramer-Rao lower bound gives a necessary and sufficient conditionfor the existence of an efficient estimator. However, MVUEs are not necessarily efficient. What can we do in such cases?



The Rao-Blackwell theorem, when applied in conjunction with a complete sufficient statistic, gives another way to find MVUEs that applies even when the CRLB is not defined.