Cramér-Rao Bound

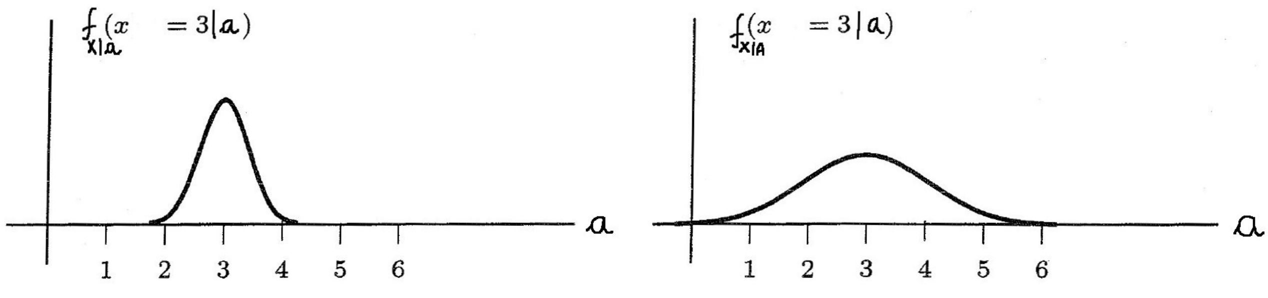
Aleksandar Dogandžić

2017-05-14

Contents

Reading: §3 in the textbook, (Hero 2015, §4.4.4), and (Sadler and Moore 2014).

How accurately we can estimate a parameter θ depends on the probability density function (pdf) or probability mass function (pmf) of the observation X given θ (i.e., on the likelihood function of θ). Intuitively, sharpness of the pdf (pmf) $f_{X\,|\,\Theta}(x \,|\, \theta)$ $[p_{X\,|\,\Theta}(x \,|\, \theta)]$ determines how accurately we can estimate θ.



$\sigma_1 = 1/3$ $\sigma_2 = 1$

Figure 1: Pdf dependence on unknown parameter for two different values of noise standard deviation σ.

Curvature of Log Likelihood

CONSIDER

Example 3.1 in §3 of the textbook

$$X = a + W$$

where a is the unknown parameter and

$$W \sim \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 > 0$ is known.

Suppose we choose $\hat{a} = \hat{a}(x) = x$. If σ^2 is large, then $\hat{a}$ is poor; if σ^2 is small, then $\hat{a}$ is good.

See Fig. 1 and recall

$$f_{X|A}(x \,|\, a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x-a)^2\Big].$$

Therefore,

$$f_{X|A}(x = 3 \,|\, a)\delta \approx \Pr\{3 - 0.5\delta < X < 3 + 0.5\delta \,|\, a\}.$$

$f_{X\,|\,A}(x = 3\,|\,a)\delta$ is the probability of observing X in a stripe (of width δ) around 3 when a is the true value.

Concentration of $f_{X|A}(x = 3 \,|\, a)$ restricts possible values of a.

pdf concentration $\nearrow$ parameter estimation accuracy $\nearrow$.

To measure the sharpness of the pdf, use curvature:

$$-\frac{\partial^2}{\partial a^2} \ln f_{X|A}(X \,|\, a).$$

But,

$$\begin{aligned}
\ln f_{X|A}(x \,|\, a) &= -0.5 \ln\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}(x-a)^2 \\
\frac{\partial \ln f_{X|A}(x \,|\, a)}{\partial a} &= \frac{x[0] - a}{\sigma^2} \\
-\frac{\partial^2 \ln f_{X|A}(x \,|\, a)}{\partial a^2} &= \frac{1}{\sigma^2}.
\end{aligned}$$

As $\sigma^2 \searrow 0$, curvature increases.

The curvature of the log likelihood function $f_{X|\Theta}(x \mid \theta)$ is depicted in Fig. 2. In general, $\partial^2 \ln f_{X|\Theta}(X \mid \theta)/\partial\theta^2$ is a random variable, so we use

$$-\mathrm{E}_{X|\Theta}\left[\frac{\partial^2}{\partial\theta^2} \ln f_{X|\Theta}(X \mid \theta) \,\middle|\, \theta\right] \tag{1}$$

to measure the average curvature. Indeed, we will show that (1) is a measure of achievable estimation performance.

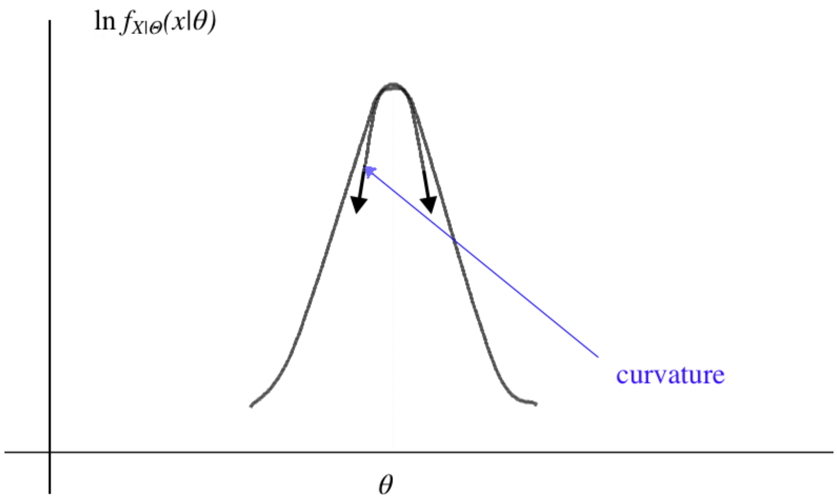


Figure 2: The curvature of the log likelihood function $f_{X|\Theta}(x \mid \theta)$.

Regularity Conditions and Fisher Information

We make two assumptions on $f_{X|\Theta}(x \mid \theta)$:

Assumption 1. The support set of $f_{X|\Theta}(x|\theta)$:

$$\operatorname{supp}\{f_{X|\Theta}(x \mid \theta)\} \triangleq \{x \mid f_{X|\Theta}(x \mid \theta) > 0\} \tag{2}$$

does not depend on θ. For all $x \in \operatorname{supp}\{f_{X|\Theta}(x \mid \theta)\}$ and θ in the parameter space sp_Θ,

$$\frac{\partial \ln f_{X|\Theta}(x \mid \theta)}{\partial\theta}$$

exists and is finite.

Assumption 2. If $T(x)$ is any statistic satisfying $\mathrm{E}_{X|\Theta}(|T(X)| \mid \theta) < +\infty$ for all $\theta \in \mathrm{sp}_\Theta$, then integration over x and differentiation by θ can be interchanged when applied to $\int T(x) f_{X|\Theta}(x|\theta)\, \mathrm{d}x$, i.e.,

$$\frac{\partial}{\partial\theta}\left[\int T(x) f_{X|\Theta}(x \mid \theta)\, \mathrm{d}x\right] = \int T(x) \frac{\partial}{\partial\theta} f_{X|\Theta}(x \mid \theta)\, \mathrm{d}x \tag{3}$$

whenever the right-hand side is finite.[1]

[1] In particular, (3) should hold for $T(x) = 1$: We will use this special case in Lemma 1.

☞ NOTE: Checking Assumption 2 is not practical. We need simple sufficient conditions on $f_{X|\Theta}(x \,|\, \theta)$ so that Assumption 2 holds. The above assumptions are coupled: If Assumption 1 does not hold, it does not make sense to talk about changing the order of integration and differentiation in Assumption 2. The following proposition describes a family of distributions that satisfy Assumptions 1 and 2.

✻ ONE-PARAMETER exponential family of distributions. If

$$f_{X|\Theta}(x \,|\, \theta) = h(x) \exp\big[\eta(\theta) T(x) - B(\theta)\big] \tag{4}$$

and $\eta(\theta)$ has a nonvanishing continuous derivative on the parameter space Θ, then Assumptions 1 and 2 hold.

If Assumption 1 holds, it is possible to define an important characteristic of $f_{X|\Theta}(x \,|\, \theta)$, the *Fisher information $\mathcal{I}(\theta)$*:

$$\begin{aligned} \mathcal{I}(\theta) &= \mathrm{E}_{X|\Theta}\bigg\{\Big[\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(X \,|\, \theta)\Big]^2 \,\Big|\, \theta\bigg\} \\ &= \int \Big[\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(x|\theta)\Big]^2 f_{X|\Theta}(x \,|\, \theta)\, \mathrm{d}x. \end{aligned} \tag{5}$$

Note that $0 \le \mathcal{I}(\theta) \le \infty$.

✻ TERMINOLOGY:

$$\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(x \,|\, \theta)$$

is the *score function* for the parameter θ and data x.

Lemma 1. *Suppose that Assumptions 1 and 2 hold and that*

$$\mathrm{E}_{X|\Theta}\Big|\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(X \,|\, \theta) \,\Big|\, \theta\Big| < \infty. \tag{6a}$$

Then

$$\mathrm{E}_{X|\Theta}\Big[\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(X \,|\, \theta) \,\Big|\, \theta\Big] = 0 \tag{6b}$$

and, thus,

$$\mathcal{I}(\theta) = \mathrm{var}_{X|\Theta}\Big[\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(X \,|\, \theta) \,\Big|\, \theta\Big]. \tag{6c}$$

Proof:

$$\begin{aligned} \mathrm{E}_{X|\Theta}\Big[\frac{\partial}{\partial\theta} \ln f_{X|\Theta}(X|\theta) \,\Big|\, \theta\Big] &= \int \frac{\frac{\partial}{\partial\theta} f_{X|\Theta}(x \,|\, \theta)}{f_{X|\Theta}(x \,|\, \theta)} f_{X|\Theta}(x \,|\, \theta)\, \mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta} f_{X|\Theta}(x \,|\, \theta)\, \mathrm{d}x \\ &= \frac{\partial}{\partial\theta} \int f_{X|\Theta}(x \,|\, \theta)\, \mathrm{d}x = 0. \end{aligned}$$

Here, we have used the chain rule for differentiation:

$$\frac{\mathrm{d}g(p(z))}{\mathrm{d}z} = \frac{\mathrm{d}g(w)}{\mathrm{d}w}\Big|_{w=p(z)} \frac{\mathrm{d}p(z)}{\mathrm{d}z}$$

with $g(\cdot) = \ln(\cdot)$. □

✻ COMMENTS:

- We have just shown in Lemma 1 that the score function has mean zero and variance equal to the Fisher information $\mathcal{I}(\theta)$.
- The score function is zero when we equate θ in it to the maximum-likelihood (ML) estimator of θ.

✻ EXAMPLE. Independent, identically distributed (i.i.d.) Poisson measurements. Suppose $(X[n])_{n=0}^{N-1}$ are i.i.d. measurements from Poisson(λ) distribution:

$$p_{X\mid\Lambda}(x[n]\mid\lambda) = \frac{\lambda^{x[n]}}{x[n]!}\mathrm{e}^{-\lambda}.$$

Then

$$\begin{aligned} p_{\boldsymbol{X}\mid\Lambda}(\boldsymbol{x}\mid\lambda) &= \prod_{n=0}^{N-1}\frac{\lambda^{x[n]}}{x[n]!}\mathrm{e}^{-\lambda} \\ &= \frac{\lambda^{\sum_n x[n]}}{\prod_{n=0}^{N-1} x[n]!}\mathrm{e}^{-N\lambda} \end{aligned} \tag{7a}$$

where $\boldsymbol{x} = (x[n])_{n=0}^{N-1}$ and

$$\frac{\partial}{\partial\lambda}\ln p_{\boldsymbol{X}\mid\Lambda}(\boldsymbol{x}\mid\lambda) = \frac{\sum_{n=0}^{N-1} x[n]}{\lambda} - N \tag{7b}$$

$$\begin{aligned} \mathcal{I}(\lambda) &\overset{\text{see (6c)}}{=} \operatorname{var}_{\boldsymbol{X}\mid\Lambda}\left(\frac{\sum_{n=0}^{N-1} X[n]}{\lambda}\,\middle|\,\lambda\right) \\ &= \frac{1}{\lambda^2}\operatorname{var}_{\boldsymbol{X}\mid\Lambda}\left(\sum_{n=0}^{N-1} X[n]\,\middle|\,\lambda\right) \qquad X[n] \text{ i.i.d.} \\ &= \frac{1}{\lambda^2}N\lambda = \frac{N}{\lambda}. \end{aligned} \tag{7c}$$

Here, we have used the fact that, for $\{X[n]\mid\lambda\} \sim$ Poisson(λ),

$$\operatorname{var}_{X\mid\Lambda}\big(X[n]\mid\lambda\big) = \lambda \tag{7d}$$

see table of probability distributions.

Fisher Information as Expected Curvature of the Log Likelihood

AN alternative expression for Fisher information.

Assumption 3. The model distribution $f_{X\mid\Theta}(x\mid\theta)$ is twice differentiable and it is permitted to interchange integration with respect to x and differentiation with respect to θ.

If, in addition to Assumptions 1 and 2, $f_{X\mid\Theta}(x\mid\theta)$ satisfies Assumption 3, then

$$\mathcal{I}(\theta) = -\mathrm{E}_{X\mid\Theta}\left[\frac{\partial^2}{\partial\theta^2}\ln f_{X\mid\Theta}(X\mid\theta)\,\middle|\,\theta\right]. \tag{8}$$

Proof.

$$\frac{\partial^2}{\partial\theta^2}\ln f_{X\,|\,\Theta}(x\,|\,\theta) = \frac{1}{f_{X\,|\,\Theta}(x\,|\,\theta)}\frac{\partial^2}{\partial\theta^2} f_{X\,|\,\Theta}(x\,|\,\theta) - \Big[\frac{\partial}{\partial\theta}\ln f_{X\,|\,\Theta}(x\,|\,\theta)\Big]^2$$

and apply expectation with respect to X given θ to both sides, i.e., multiply by $f_{X\,|\,\Theta}(x\,|\,\theta)$ and integrate. □

The result in (8) provides another way to compute the Fisher information, which may be more convenient than taking the expectation of the squared score function. Probability distributions from the exponential family satisfy Assumption 3.

✻ Example. I.I.D. Poisson measurements. Back to the Poisson example in (7a):

$$p_{\boldsymbol{X}\,|\,\Lambda}(\boldsymbol{x}\,|\,\lambda) = \frac{\lambda^{\sum_n x[n]}}{\prod_{n=0}^{N-1} x[n]!}\mathrm{e}^{-N\lambda}.$$

Then,

$$\frac{\partial}{\partial\lambda}\ln p_{\boldsymbol{X}\,|\,\Lambda}(\boldsymbol{x}\,|\,\lambda) = \frac{\sum_{n=0}^{N-1} x[n]}{\lambda} - N \tag{9a}$$

$$\frac{\partial^2}{\partial\lambda^2}\ln p_{\boldsymbol{X}\,|\,\Lambda}(\boldsymbol{x}\,|\,\lambda) = -\frac{\sum_{n=0}^{N-1} x[n]}{\lambda^2} \tag{9b}$$

obtained by differentiating (7b) with respect to λ

and apply (8):

$$\begin{aligned}\mathcal{I}(\lambda) &= \mathrm{E}_{\boldsymbol{X}\,|\,\Lambda}\Big[-\frac{\partial^2}{\partial\lambda^2}\ln f_{\boldsymbol{X}\,|\,\Lambda}(\boldsymbol{X}\,|\,\lambda)\,\Big|\,\lambda\Big] \\ &= \frac{1}{\lambda^2}\,\mathrm{E}_{\boldsymbol{X}\,|\,\Lambda}\Big(\sum_{n=0}^{N-1} X[n]\,\Big|\,\lambda\Big) \\ &= \frac{N}{\lambda}\end{aligned}$$

which is easier than before. Here, we have used the fact that, for $\{X[n]\,|\,\lambda\} \sim \text{Poisson}(\lambda)$,

$$\mathrm{E}_{X\,|\,\Lambda}(X[n]\,|\,\lambda) = \lambda$$

see the table of probability distributions.

I.I.D. Measurements

Assume that $(X[n])_{n=0}^{N-1}$ are i.i.d. with density $f_{X\,|\,\Theta}(x\,|\,\theta)$ and that Assumptions 1 and 2 hold. Define the contribution of a single measurement ($X[0]$, say) to the Fisher information:

$$\mathcal{I}_1(\theta) = \mathrm{E}_{X\,|\,\Theta}\Big\{\Big[\frac{\partial}{\partial\theta}\ln f_{X\,|\,\Theta}(X[0]\,|\,\theta)\Big]^2\,\Big|\,\theta\Big\}. \tag{10a}$$

Here, we pick $\boldsymbol{X}[0]$ arbitrarily as our single measurement, its contribution to the Fisher information is equal to that of $\boldsymbol{X}[1]$ etc.

Then, the Fisher information $\mathcal{I}(\theta)$ for θ based on all observations $\boldsymbol{x}$ is easy to compute using $\mathcal{I}_1(\theta)$:

$$\mathcal{I}(\theta) = N\mathcal{I}_1(\theta). \tag{10b}$$

Proof:

$$\begin{aligned}
\mathcal{I}(\theta) &= \mathrm{var}_{\boldsymbol{X}\,|\,\Theta}\Big[\frac{\partial}{\partial\theta}\ln f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\theta)\,\Big|\,\theta\Big] \\
&= \mathrm{var}_{\boldsymbol{X}\,|\,\Theta}\Big[\sum_{n=0}^{N-1}\frac{\partial}{\partial\theta}\ln f_{X\,|\,\Theta}(X[n]\,|\,\theta)\,\Big|\,\theta\Big] \\
&= \sum_{n=0}^{N-1}\mathrm{var}_{X\,|\,\Theta}\Big[\frac{\partial}{\partial\theta}\ln f_{X\,|\,\Theta}(X[n]\,|\,\theta)\,\Big|\,\theta\Big] \\
&= N\mathcal{I}_1(\theta). \qquad \square
\end{aligned}$$

$X[n]$ i.i.d.

✻ EXAMPLE. Suppose that, conditional on μ, $(X[n])_{n=0}^{N-1}$ are i.i.d. observations from $\mathcal{N}(\mu,\sigma^2)$. Here, μ is the unknown parameter and σ^2 is a known constant. Note that the Assumptions 1 and 2 hold because $\mathcal{N}(\mu,\sigma^2)$ with parameter μ is a member of the exponential family of distributions. Then

$$\ln f_{X\,|\,\mu}(x[0]\,|\,\mu) = -0.5\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x[0]-\mu)^2 \tag{10a}$$

$$\frac{\partial \ln f_{X\,|\,\mu}(x[0]\,|\,\mu)}{\partial\mu} = \frac{x[0]-\mu}{\sigma^2} \tag{10b}$$

and

$$\begin{aligned}
\mathcal{I}_1(\mu) &= \mathrm{E}_{X\,|\,\mu}\Big[\Big(\frac{\partial \ln f_{X\,|\,\mu}(X[0]\,|\,\mu)}{\partial\mu}\Big)^2\,\Big|\,\mu\Big] \\
&= \mathrm{E}_{X\,|\,\mu}\Big[\Big(\frac{X[0]-\mu}{\sigma^2}\Big)^2\,\Big|\,\mu\Big] \\
&= \frac{1}{\sigma^2}.
\end{aligned} \tag{11}$$

We could differentiate the score function with respect to μ:

$$\frac{\partial^2 \ln f_{X\,|\,\mu}(x[0]\,|\,\mu)}{\partial\mu^2} = -\frac{1}{\sigma^2}$$

and apply (8):

$$\begin{aligned}
\mathcal{I}_1(\mu) &= -\,\mathrm{E}_{X\,|\,\mu}\Big[\frac{\partial^2}{\partial\mu^2}\ln f_{X\,|\,\mu}(X[0]\,|\,\mu)\,\Big|\,\mu\Big] \\
&= \frac{1}{\sigma^2}.
\end{aligned}$$

The Fisher information for μ based on $(X[n])_{n=0}^{N-1}$ is

$$\mathcal{I}(\mu) = N\mathcal{I}_1(\mu) = \frac{N}{\sigma^2}. \tag{12}$$

Information Inequality

CONSIDER statistics $T(X)$ that satisfy

$$\mathrm{var}_{X\,|\,\Theta}[T(X)\,|\,\theta] < +\infty$$

for all θ. Suppose that Assumptions 1 and 2 hold and $0 < \mathcal{I}(\theta) < +\infty$, where $\mathcal{I}(\theta)$ is the Fisher information for θ, defined in (5). Define

$$\psi(\theta) \triangleq \mathrm{E}_{X\,|\,\Theta}[T(X)\,|\,\theta].$$

Then, for all θ,

$$\mathrm{var}_{X\,|\,\Theta}[T(X)\,|\,\theta] \geq \frac{|\psi'(\theta)|^2}{\mathcal{I}(\theta)} \tag{13a}$$

where

$$\psi'(\theta) \triangleq \frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\theta}. \tag{13b}$$

Proof: Use Assumptions 1 and 2:

$$\begin{aligned}
\psi'(\theta) &= \frac{\partial}{\partial\theta}\int T(x)\, f_{X\,|\,\Theta}(x\,|\,\theta)\,\mathrm{d}x \\
&= \int T(x)\frac{\partial f_{X\,|\,\Theta}(x\,|\,\theta)}{\partial\theta}\,\mathrm{d}x \\
&= \int T(x)\frac{\partial \ln f_{X\,|\,\Theta}(x\,|\,\theta)}{\partial\theta}\, f_{X\,|\,\Theta}(x\,|\,\theta)\,\mathrm{d}x \\
&= \mathrm{E}_{X\,|\,\Theta}\Big[T(X)\frac{\partial \ln f_{X\,|\,\Theta}(X\,|\,\theta)}{\partial\theta}\,\Big|\,\theta\Big]
\end{aligned}$$

and, therefore,[2]

$$\psi'(\theta) = \mathrm{cov}_{X\,|\,\Theta}\Big[\frac{\partial \ln f_{X\,|\,\Theta}(X\,|\,\theta)}{\partial\theta}, T(X)\Big].$$

Apply the Cauchy-Schwartz inequality[3]

$$[\mathrm{cov}_{P,Q}(P,Q)]^2 \leq \mathrm{var}_P(P)\,\mathrm{var}_Q(Q)$$

to the random variables $\overbrace{\frac{\partial \ln f_{X\,|\,\Theta}(X\,|\,\theta)}{\partial\theta}}^{P}$ and $\overbrace{T(X)}^{Q}$:

$$|\psi'(\theta)|^2 \leq \mathrm{var}_{X\,|\,\Theta}[T(X)\,|\,\theta]\,\mathrm{var}_{X\,|\,\Theta}\Big[\frac{\partial \ln f_{X\,|\,\Theta}(X\,|\,\theta)}{\partial\theta}\,\Big|\,\theta\Big].$$

Apply Lemma 1. □

✻ COMMENTS:

- The information inequality holds for vector measurements $\boldsymbol{X}$ as well;

[2] If $\mathrm{E}_P[P] = 0$ or $\mathrm{E}_Q[Q] = 0$, then $\mathrm{cov}_{P,Q}(P,Q) = \mathrm{E}_{P,Q}[PQ]$, which follows from the fact that $\mathrm{cov}_{P,Q}(P,Q) = \mathrm{E}_{P,Q}[PQ] - \mathrm{E}_P[P]\,\mathrm{E}_Q[Q]$.

[3] proved in the Appendix

- If we view $T(X)$ as a (generally biased) estimator of θ, then

$$\begin{aligned} \mathrm{E}_{X\,|\,\Theta}[T(X)\,|\,\theta] &= \psi(\theta) \\ &= \theta + \underbrace{b(\theta)}_{\text{bias}} \end{aligned} \tag{14}$$

  and (13a) is a lower bound on the variance of $T(X)$:

$$\mathrm{var}_{X\,|\,\Theta}[T(X)\,|\,\theta] \geq \frac{|1+b'(\theta)|^2}{\mathcal{I}(\theta)}. \tag{15}$$

✻ Example. Continue with the Poisson example and consider statistics $S(\boldsymbol{x})$ with expectation $\mathrm{E}_{\boldsymbol{X}\,|\,\Lambda}[S(\boldsymbol{x})\,|\,\lambda] = \psi(\lambda) = \lambda^2$. Here,

$$\psi(\lambda) = \lambda^2 \quad \text{and, therefore,} \quad \psi'(\lambda) = 2\lambda$$

and

$$\begin{aligned} \mathrm{var}_{\boldsymbol{X}\,|\,\Lambda}[S(\boldsymbol{x})\,|\,\lambda] &\overset{\text{see (13a)}}{\geq} \frac{|\psi'(\lambda)|^2}{\mathcal{I}(\lambda)} \\ &= \frac{4\lambda^2}{N/\lambda} \\ &= \frac{4\lambda^3}{N} \end{aligned}$$

Or, consider statistics $U(\boldsymbol{x})$ with expectation $\mathrm{E}_{\boldsymbol{X}\,|\,\Lambda}[U(\boldsymbol{x})\,|\,\lambda] = \psi(\lambda) = \exp(-\lambda)$. Here,

$$\psi(\lambda) = \mathrm{e}^{-\lambda}, \qquad \psi'(\lambda) = -\mathrm{e}^{-\lambda}$$

and

$$\begin{aligned} \mathrm{var}_{\boldsymbol{X}\,|\,\Lambda}[U(\boldsymbol{X})\,|\,\lambda] &\overset{\text{see (13a)}}{\geq} \frac{|\psi'(\lambda)|^2}{\mathcal{I}(\lambda)} \\ &= \frac{\mathrm{e}^{-2\lambda}}{N/\lambda} \\ &= \frac{\lambda \mathrm{e}^{-2\lambda}}{N}. \end{aligned}$$

For example, consider the proportion of observations $X[n]$ that are equal to zero:

$$U(\boldsymbol{x}) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{1}_{\{0\}}(X[n])$$

and note that

$$\begin{aligned} \mathrm{E}_{\boldsymbol{X}\,|\,\Lambda}[U(\boldsymbol{X})\,|\,\lambda] &= \mathrm{e}^{-\lambda} \\ \mathrm{var}_{\boldsymbol{X}\,|\,\Lambda}[U(\boldsymbol{X})\,|\,\lambda] &= \frac{1}{N}\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) \\ &\geq \frac{\lambda \mathrm{e}^{-2\lambda}}{N}. \end{aligned}$$

as expected

Unbiased Estimators

THE lower bound in (13a) depends on $T(X)$ through its expectation $\psi(\theta)$. If we consider *only the class of unbiased estimators $T(X)$ of θ*, i.e.,

$$\psi(\theta) = \theta$$

we obtain a *universal lower bound* on variance of such estimators, given by the following.

Cramér-Rao bound

SUPPOSE that Assumptions 1 and 2 hold and that $T(X)$ is an unbiased estimator of θ, i.e.,

$$\mathrm{E}_{X \mid \Theta}[T(X) \mid \theta] = \theta.$$

Then

$$\mathrm{var}_{X \mid \Theta}[T(X) \mid \theta] \geq \frac{1}{\mathcal{I}(\theta)} \tag{16}$$

where equality is attained if and only if, for some deterministic scalar c_θ,

$$\frac{\partial}{\partial \theta} \ln f_{X \mid \Theta}(x \mid \theta) = c_\theta [T(x) - \theta] \tag{17}$$

see also eq. (3.7) in the textbook.

(17) follows from that fact that, for random variables P and Q, equality in the Cauchy-Schwartz inequality

$$[\mathrm{cov}_{P,Q}(P,Q)]^2 \leq \mathrm{var}_P(P)\,\mathrm{var}_Q(Q)$$

occurs if and only if

$$P - \mathrm{E}_P(P) = c\,[Q - \mathrm{E}_Q(Q)]$$

for some deterministic scalar c.

Here,

$$\mathrm{CRB}(\theta) \triangleq \frac{1}{\mathcal{I}(\theta)} \tag{18}$$

is often referred to as the Cramér-Rao bound (CRB) on the variance of an unbiased estimator of θ.

When the CRB is attainable, it is said to be a tight bound and (17) is called the *CRB tightness condition*.

One-parameter Canonical Exponential Family

IN handout `expon_family`, we introduced the one-parameter canonical exponential family:

$$f_{\boldsymbol{X} \mid \eta}(\boldsymbol{x} \mid \eta) = h(\boldsymbol{x}) \exp[\eta T(\boldsymbol{x}) - A(\eta)]. \tag{19}$$

which satisfies Assumptions 1 and 2, see also p. 4.

Now,

$$\frac{\partial \ln f_{\boldsymbol{X} \mid \eta}(\boldsymbol{x} \mid \eta)}{\partial \eta} = T(\boldsymbol{x}) - \frac{\mathrm{d}A(\eta)}{\mathrm{d}\eta} \tag{20}$$

which implies

$$\begin{aligned}\mathrm{E}_{\boldsymbol{X}\mid\eta}\Big[\frac{\partial \ln f_{\boldsymbol{X}\mid\eta}(\boldsymbol{x}\mid\eta)}{\partial\eta}\,\Big|\,\eta\Big] &= \mathrm{E}_{\boldsymbol{X}\mid\eta}[T(\boldsymbol{X})\mid\eta] - \frac{\mathrm{d}A(\eta)}{\mathrm{d}\eta}\\ &= 0 \qquad (21)\\ \mathcal{I}(\eta) &= \mathrm{var}_{\boldsymbol{X}\mid\eta}\Big[\frac{\partial}{\partial\eta}\ln f_{X\mid\eta}(X\mid\eta)\,\Big|\,\eta\Big]\\ &= \mathrm{var}_{\boldsymbol{X}\mid\eta}[T(\boldsymbol{X})\mid\eta] \qquad (22)\end{aligned}$$

by Lemma 1. We also know that, if Assumption 3 holds,

$$\begin{aligned}\mathcal{I}(\eta) &= -\mathrm{E}_{X\mid\eta}\Big[\frac{\partial^2}{\partial\eta^2}\ln f_{X\mid\eta}(X\mid\eta)\,\Big|\,\eta\Big]\\ &= \frac{\mathrm{d}^2A(\eta)}{\mathrm{d}\eta^2}\end{aligned}$$

see (8)

which implies

$$\mathrm{var}_{\boldsymbol{X}\mid\eta}[T(\boldsymbol{X})\mid\eta] = \tfrac{\mathrm{d}^2A(\eta)}{\mathrm{d}\eta^2}. \qquad (23)$$

Hence, we have shown that

$$\mathrm{E}_{\boldsymbol{X}\mid\eta}[T(\boldsymbol{X})\mid\eta] = \frac{\mathrm{d}A(\eta)}{\mathrm{d}\eta}, \qquad \mathrm{var}_{\boldsymbol{X}\mid\eta}(T(\boldsymbol{X})\mid\eta) = \frac{\mathrm{d}^2A(\eta)}{\mathrm{d}\eta^2} \qquad (24)$$

as stated in (5) of handout `expon_family`.

Since (20) has the form in (17), we also know that CRB is attainable by $T(\boldsymbol{X})$ in this case.

Example: Sinusoidal frequency estimation

Consider a sinusoid of *unknown frequency* but known amplitude and phase:

$0 < f < 0.5$

$0 \le n \le N-1$

$$\begin{aligned}s[n;f] &= a\cos\big(2\pi f n + \phi\big)\\ X[n] &= s[n;f] + W[n].\end{aligned}$$

Assume that $W[n]$ is zero-mean additive white Gaussian noise (AWGN) with known variance σ^2.[4] Then,

[4] $W[n]$ are i.i.d. zero-mean Gaussian with constant variance σ^2.

$$f_{X[n]\mid F}(x[n]\mid f) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big\{-\frac{1}{2\sigma^2}(x[n]-s[n;f])^2\Big\}.$$

Since the observations are independent, the likelihood function of f for the data $\boldsymbol{x}$ is

$$f_{\boldsymbol{X}\mid F}(\boldsymbol{x}\mid f) = \prod_{n=0}^{N-1} f_{X[n]\mid F}(x[n]\mid f)$$

where $\boldsymbol{x} = (x[n])_{n=0}^{N-1}$. Take the logarithm:

const means not a function of f

$$
\begin{aligned}
\ln f_{\boldsymbol{X}|F}(\boldsymbol{x} \mid f) &= \sum_{n=0}^{N-1} \ln f_{X[n]|F}(x[n] \mid f) \\
&= -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n; f])^2 + \text{const}.
\end{aligned}
$$

Differentiate with respect to f:

$$
\frac{\partial \ln f_{X[n]|F}(\boldsymbol{x} \mid f)}{\partial f} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial s[n; f]}{\partial f} (x[n] - s[n; f])
$$

and once more:

$$
\begin{aligned}
\frac{\partial^2 \ln f_{X[n]|F}(\boldsymbol{x} \mid f)}{\partial f^2} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial^2 s[n; f]}{\partial f^2} (x[n] - s[n; f]) \\
&\quad - \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \Big(\frac{\partial s[n; f]}{\partial f} \Big)^2.
\end{aligned}
$$

The negative expected value of this expression is the Fisher information:

$$
\begin{aligned}
\mathcal{I}(f) &= -\mathrm{E}_{\boldsymbol{X}|F}\Big[\frac{\partial^2 \ln f_{\boldsymbol{X}|f}(\boldsymbol{x} \mid f)}{\partial f^2} \Bigm| f \Big] \\
&= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \Big(\frac{\partial s[n; f]}{\partial f} \Big)^2 \\
&= \mathrm{SNR} \sum_{n=0}^{N-1} [2\pi n \sin(2\pi f n + \phi)]^2
\end{aligned}
$$

where $\mathrm{SNR} = a^2/\sigma^2$ is the signal-to-noise ratio (SNR). The CRB is

$$
\mathrm{CRB}(f) = \frac{1}{\mathcal{I}(f)} \le \mathrm{var}_{\boldsymbol{X}|F}\big(\hat{f} \mid f \big)
$$

for unbiased frequency estimators $\hat{f}$.

Consider the case where $\mathrm{SNR} = 1$, $N = 10$, and $\phi = 0$ rad. Then

$$
s[n; f] = a \cos(2\pi f n).
$$

Recall that N, a, ϕ, and σ^2 are assumed *known*. There are preferred frequencies, see Fig. 3.

Consider now the case where $\mathrm{SNR} = 1$, $N = 10$, and $\phi = -0.5\pi$ rad. Then

$$
s[n; f] = a \sin(2\pi f n).
$$

Here, $f \searrow 0$ is good for frequency estimation because we can easily differentiate between the case of no signal at all (which happens at $f = 0$) and a sinusoid with amplitude a, see Fig. 4.

In general, CRB is used as a

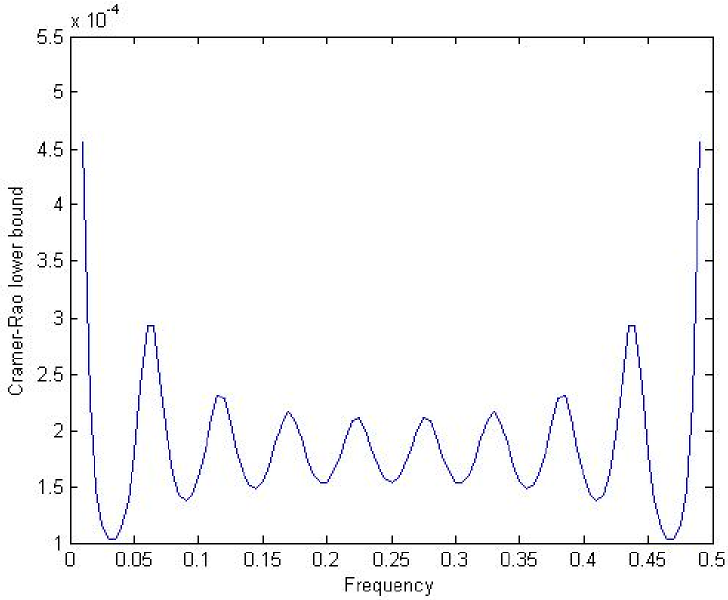


Figure 3: CRB for f as a function of f, for SNR $= 1$, $N = 10$, and $\phi = 0$ rad.

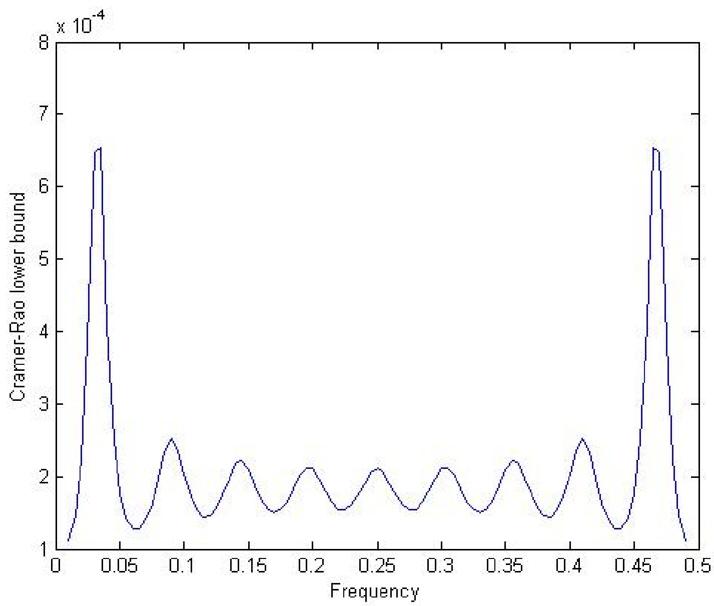


Figure 4: CRB for f as a function of f, for SNR $= 1$, $N = 10$, and $\phi = -0.5\pi$ rad.

- measure of the potential performance attainable from the system,
- benchmark for assessing algorithm performance,
- measure for system design.

Appendix

We present a proof of the Cauchy-Schwartz inequality.
Proof of the Cauchy-Schwartz inequality: First, remember that any covariance matrix needs to be positive semidefinite. Therefore,

$$\det\left(\mathrm{cov}_{P,Q}([P,Q]^T)\right) = \det\left(\begin{bmatrix} \mathrm{var}_P(P) & \mathrm{cov}_{P,Q}(P,Q) \\ \mathrm{cov}_{P,Q}(P,Q) & \mathrm{var}_Q(Q) \end{bmatrix}\right) \\ \geq 0$$

and the Cauchy-Schwartz inequality follows.

Why does a covariance matrix of $[P,Q]^T$ need to be positive semidefinite? Because the following holds for arbitrary a and b:

$$\mathrm{var}_{P,Q}(aP + bQ) \geq 0$$

which can be rewritten as

$$\begin{bmatrix} a & b \end{bmatrix} \mathrm{cov}_{P,Q}\left(\begin{bmatrix} P \\ Q \end{bmatrix}\right) \begin{bmatrix} a \\ b \end{bmatrix} \geq 0 \quad \forall a, b$$

which, by the definition of positive semidefiniteness, implies that $\mathrm{cov}_{P,Q}\left(\begin{bmatrix} P \\ Q \end{bmatrix}\right)$ is a positive semidefinite matrix. □

Acronyms

AWGN additive white Gaussian noise. 11

CRB Cramér-Rao bound. 10–13

i.i.d. independent, identically distributed. 5–7, 11

ML maximum-likelihood. 5

pdf probability density function. 1, 2

pmf probability mass function. 1

SNR signal-to-noise ratio. 12

References

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI.

Sadler, Brian M. and Terrence J. Moore (2014). "Performance Analysis and Bounds". In: *Array and Statistical Signal Processing*. Ed. by Abdelhak M. Zoubir, Mats Viberg, Rama Chellappa, and Sergios Theodoridis. Vol. 3. Academic Press Library in Signal Processing. Elsevier. Chap. 8, pp. 297–322.