Computer Science Year 2

Algorithms & Data

Estimation, Regression, Classification

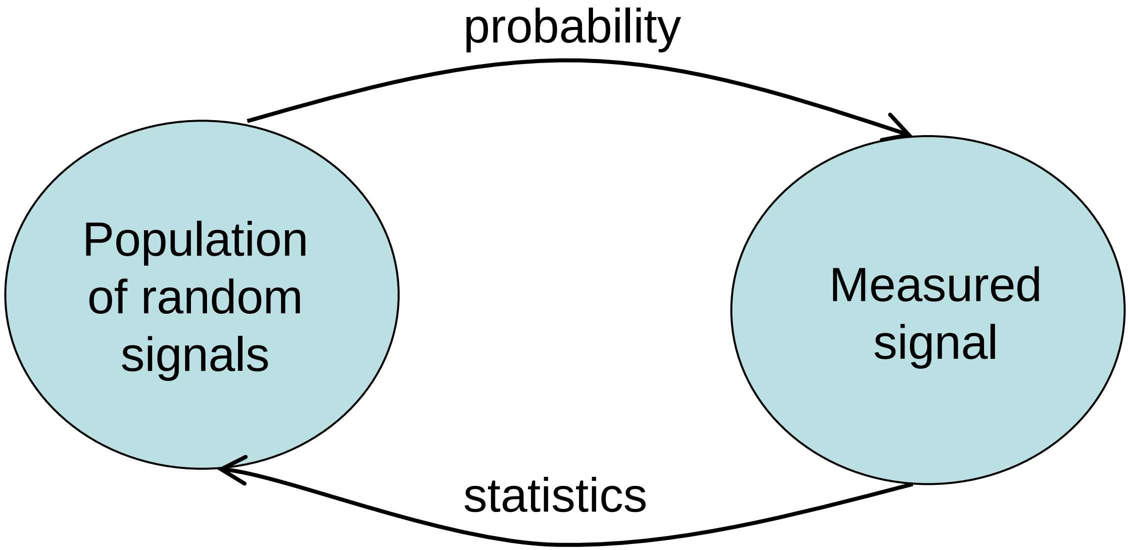
Prof Alin Achim

Statistical Inference

- A statistic is a function of the observed data.
- Example: Suppose we observe N scalar values $x_1, x_2, \ldots x_N$. The following are statistics:
  - Sample mean: $$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$
  - The data itself: $x_1, x_2, \ldots x_N$
  - Order statistic: *min*$\{x_1, x_2, \ldots x_N\}$
- A statistic CANNOT depend on unknown parameters!!

Statistical Inference

- Probability is used to model uncertainty
- Statistics are used to draw conclusions about probability models



- **Probability models our uncertainty about signals we may observe.**
- **Statistics reason from the measured signal to the population of possible signals.**

Statistical Signal Processing (SSP)

- A Three-step approach:
  - Step 1 - Postulate a probability model (or models) that reasonably capture the uncertainties at hand
  - Step 2 - Collect data
  - Step 3 - Formulate statistics that allow us to interpret or understand our probability model(s)
- There are two major kinds of problems that are studied: detection and estimation. Most SSP problems fall under one of these two headings.

Stochastic Processes:

- A stochastic process is a ***random*** process.

- A discrete random signal is defined as a sequence of indexed random variables assuming values:

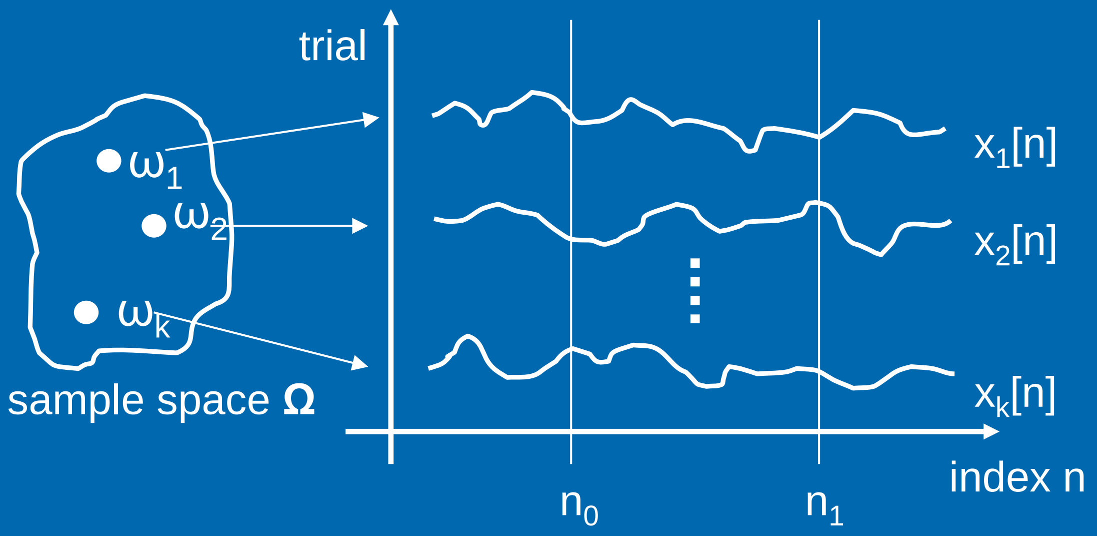
x[0], **x**[1], **x**[2],…., **x**[i], ….

where it is assumed that:

- samples are evenly spaced in time
- samples are continuous in amplitude (infinite precision representation)
- samples are taken at a rate greater than twice the highest frequency component present (i.e. Nyquist satisfied)
- sample period is normalised to unity

Stochastic Processes:

Consider an ensemble of sample functions:



- A ***stochastic process*** is an ensemble of time (or spatial) variables together with a probability rule which assigns a probability to any event observed

- The figure shows a set of sample functions, or ***realisations***, $x_k[n]$, corresponding to a sample point ω_k in the sample space $\mathbf{\Omega}$.

Observation of sample waveforms at some point n_0. Each sample has a value $x_k[n_0]$ and a probability $P(\omega_k)$. The set of numbers $\{x_k[n_0]\}$ k = 1..K form a ***random variable***. Observation at n_1 results in a second random variable $\{x_k[n_1]\}$ k = 1..K.

Objectives

- Minimum Variance Unbiased Estimator
- Cramer-Rao Lower Bound (CRLB)

The Minimum Variance Unbiased Estimator (MVUE)

- In parameter estimation we observe a vector x consisting of N measurements. The distribution of x is governed by a probability density function p(x,θ) parameterized by an unknown parameter θ.
- Our goal is to establish a useful optimality criterion for guiding the design and assessing the quality of an estimator $\hat{\theta}(x)$
- One possibility – the mean square error

$$MSE(\theta) = E[(\hat{\theta} - \theta)^2]$$

- The MSE is a perfectly reasonable way of assessing estimator quality but does NOT lead to a useful design criterion since $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}$, that is the estimator depends on the value of the unknown parameter!!

MVUE: The Bias-variance decomposition of the MSE

- By rewriting the MSE as below, a useful optimality criterion emerges:

$$
\begin{aligned}
MSE(\hat{\theta}) &= E\left\{\left[\left(\hat{\theta} - E(\hat{\theta})\right) + \left(E(\hat{\theta}) - \theta\right)\right]^2\right\} \\
&= \mathrm{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\
&= \mathrm{var}(\hat{\theta}) + b^2(\theta)
\end{aligned}
$$

Where:

- Definition 1: variance $\mathrm{var}(\theta) = E[(\theta - E(\theta))^2]$

- Definition 2: bias $b(\theta) = E(\hat{\theta}) - \theta$

MVUE example: DC level in WGN

- Consider the observations $x[n] = A + w[n]$

  Where A is the parameter to be estimated, which can take on any real value. A possible estimator for the average of x[n] is then

  $$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

  Taking the expectation

  $$E(\hat{A}) = E\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\sum_{n-0}^{N-1} E(x[n])$$
  $$= \frac{1}{N}\sum_{n=0}^{N-1} A = A$$

MVUE example continued

- Consider now the *modified* (biased) estimator:

$$\breve{A} = a\frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

Since $E(\breve{A}) = aA$ and $\mathrm{var}(\breve{A}) = \frac{a^2\sigma^2}{N}$, using the bias-variance decomposition we have

$$MSE(\breve{A}) = \frac{a^2\sigma^2}{N} + (a-1)^2A^2$$

Differentiating with respect to a:

$$\frac{dMSE(\breve{A})}{da} = \frac{2a\sigma^2}{N} + 2(a-1)A^2$$

Setting to zero and solving yields: $a_{opt} = \frac{A^2}{A^2 + \frac{\sigma^2}{N}}$

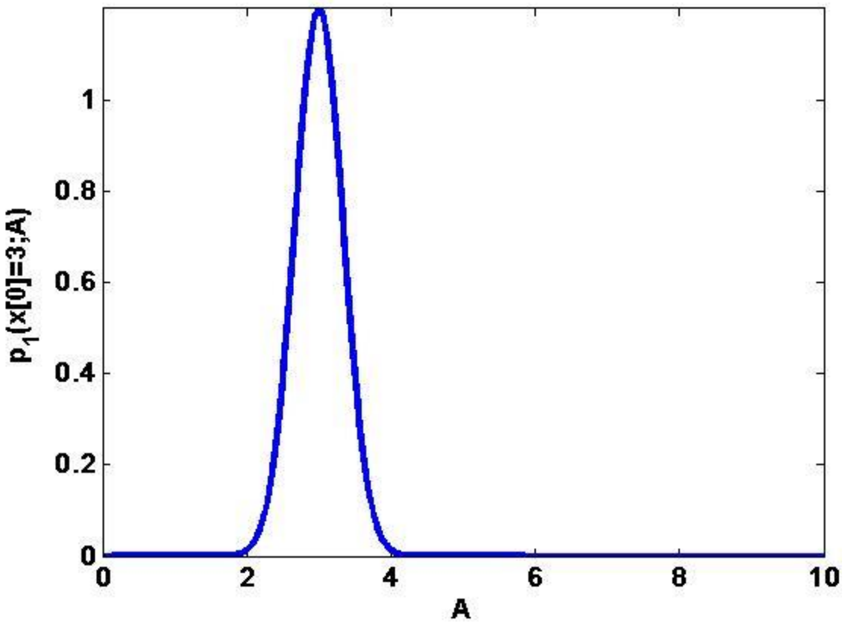
Minimum Variance Unbiased Estimation (MVUE)

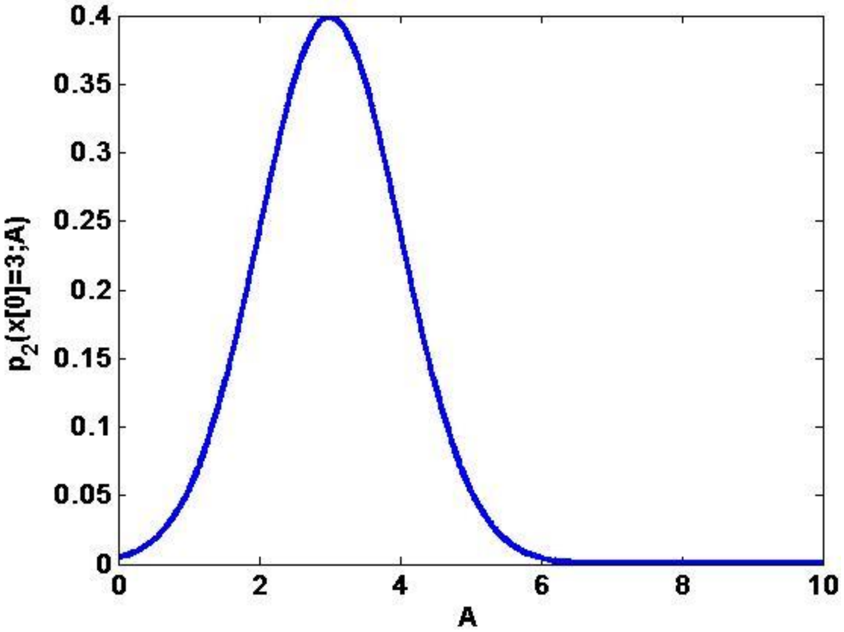
- MSE criterion sometimes leads to unrealistic estimators
- An alternative approach is to constrain the bias to be zero and minimise the variance which is known as the **Minimum Variance Unbiased Estimate**
- Since the bias is fixed to be zero, the MVU estimate will be equivalent to the MSE criterion.
- There is no known standard procedure to find a MVU estimator. However, there are few possible approaches for finding the MVU estimator:
  - Determine the Cramer-Rao Lower Bound (CRLB) and check to see whether some estimator satisfies it.
  - Restrict the class of estimators to be not only unbiased but also linear.

PDF Dependence on Unknown Parameter

- Suppose we observe a single sample

$$x[0] = A + w[0], \text{ where } w[0] \sim N(0, \sigma^2)$$





$$p_i(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}(x[0] - A)^2\right]$$

with $i = 1{,}2$ and $\sigma_1 = 1/3; \sigma_2 = 1$

Estimator accuracy

- The more peaky or spiky a likelihood function is, the easier it is to determine the unknown parameter.
- The peakyness is effectively measured by the negative of the second derivative of the log-likelihood at its peak (the curvature of the function).
- Consider the natural logarithm of the PDF

$$\ln p\,(x[0];A) = -\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x[0]-A)^2$$

- Taking the first derivative w.r.t. A

$$\frac{\partial \ln p\,(x[0];A)}{\partial A} = \frac{1}{\sigma^2}(x[0]-A)$$

Estimator accuracy

- The negative of the second derivative becomes

$$-\frac{\partial^2 \ln p\,(x[0];A)}{\partial A^2} = \frac{1}{\sigma^2}$$

- The curvature increases as the variance decreases

$$\mathrm{var}(\hat{A}) = \frac{1}{-\frac{\partial^2 \ln p\,(x[0];A)}{\partial A^2}}$$

- In general, the second derivative will also depend on x[0] so a more appropriate measure of curvature is

$$-E\left[\frac{\partial^2 \ln p\,(x[0];A)}{\partial A^2}\right]$$

Cramer-Rao Lower Bound (CRLB)

The variance of any unbiased estimator, $\hat{\theta}$, must satisfy

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \ln p(x;\theta)}{\partial\theta}\right]}$$

where the derivative is evaluated at the true value of θ and the expectation is taken with respect to PDF. Furthermore, an unbiased estimator may be found that attains the bound for all θ iff

$$\frac{\partial \ln p(x;\theta)}{\partial\theta} = I(\theta)(g(x)-\theta)$$

For some function of g and I (the Fisher information matrix). The MVU estimator is $\hat{\theta} = g(x)$ and the minimum variance is $1/I(\theta)$.

CRLB for x[0] = A+w[0]

$$-\frac{\partial^2 \ln p\,(x[0];A)}{\partial A^2} = \frac{1}{\sigma^2}$$

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \ln p\,(x;\theta)}{\partial \theta}\right]}$$

$$\mathrm{var}(\hat{A}) \geq \sigma^2$$

- No unbiased estimator can exist whose variance is lower than σ^2

$$\frac{\partial \ln p\,(x[0];A)}{\partial A} = \frac{1}{\sigma^2}(x[0]-A)$$

$$\frac{\partial \ln p\,(x;\theta)}{\partial \theta} = I(\theta)(g(x)-\theta)$$

$$\theta = A$$

$$I(\theta) = \frac{1}{\sigma^2}$$

$$g(x[0]) = x[0]$$

Example: DC level in WGN

- Consider now the multiple observations

$$x[n] = A + w[n], \; n = 0,1,\ldots,N-1$$

- To determine the CRLB for A we start again by writing the likelihood function

$$p(x;A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right]$$
$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$

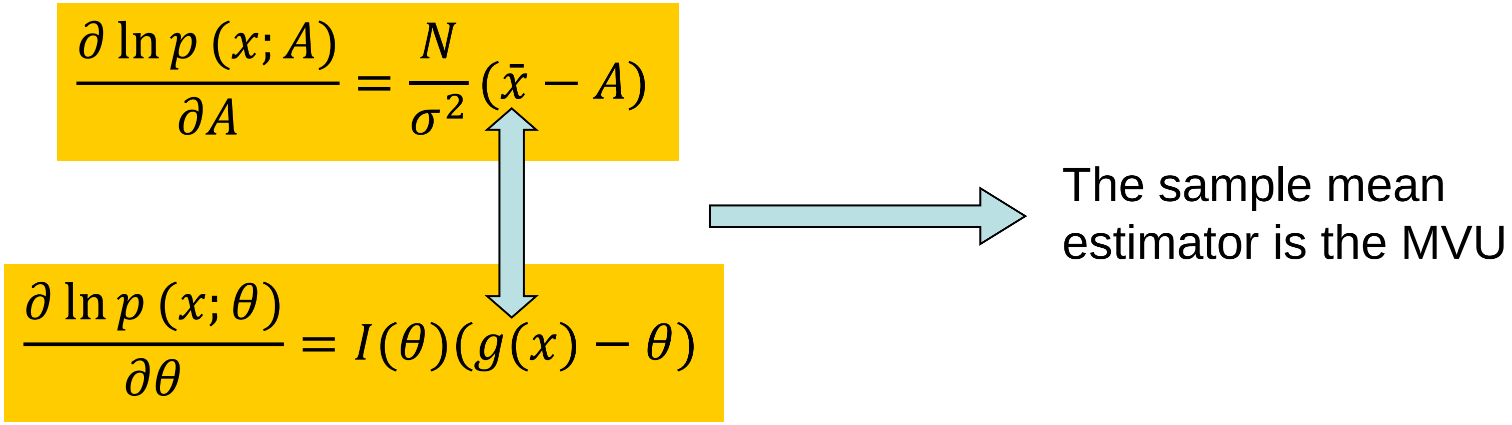
- Taking the first derivative of the log-likelihood

$$\frac{\partial \ln p\,(x;A)}{\partial A} = \frac{\partial}{\partial A}\left[-\ln\left[(2\pi\sigma^2)^{\frac{N}{2}}\right] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$
$$= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) = \frac{N}{\sigma^2}(\bar{x}-A)$$

DC level in WGN (continued)

- Differentiating again $$\frac{\partial^2 \ln p\,(x;A)}{\partial A^2} = -\frac{N}{\sigma^2}$$

- Finally, the CRLB is $$\operatorname{var}(\hat{A}) \geq \frac{\sigma^2}{N}$$



$$\frac{\partial \ln p\,(x;A)}{\partial A} = \frac{N}{\sigma^2}(\bar{x} - A)$$

$$\frac{\partial \ln p\,(x;\theta)}{\partial \theta} = I(\theta)(g(x) - \theta)$$

The sample mean estimator is the MVU

Example: CRLB for general signal in WGN

- Assume a deterministic signal depending on an unknown parameter θ is observed in WGN:

$$x[n] = s[n;\theta] + w[n],\ n = 0,1,\ldots,N-1$$

- The likelihood function:

$$p(x;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-s[n;\theta])^2\right\}$$

- Differentiate once:

$$\frac{\partial \ln p\,(x;\theta)}{\partial\theta} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-s[n;\theta])\frac{\partial s[n;\theta]}{\partial\theta}$$

- Differentiating a second time:

$$\frac{\partial^2 \ln p\,(x;\theta)}{\partial\theta^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left\{(x[n]-s[n;\theta])\frac{\partial^2 s[n;\theta]}{\partial\theta^2} - \left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2\right\}$$

Example: CRLB for general signal in WGN

- Taking the expected value yields

$$E\left(\frac{\partial^2 \ln p\,(x;\theta)}{\partial\theta^2}\right) = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2$$

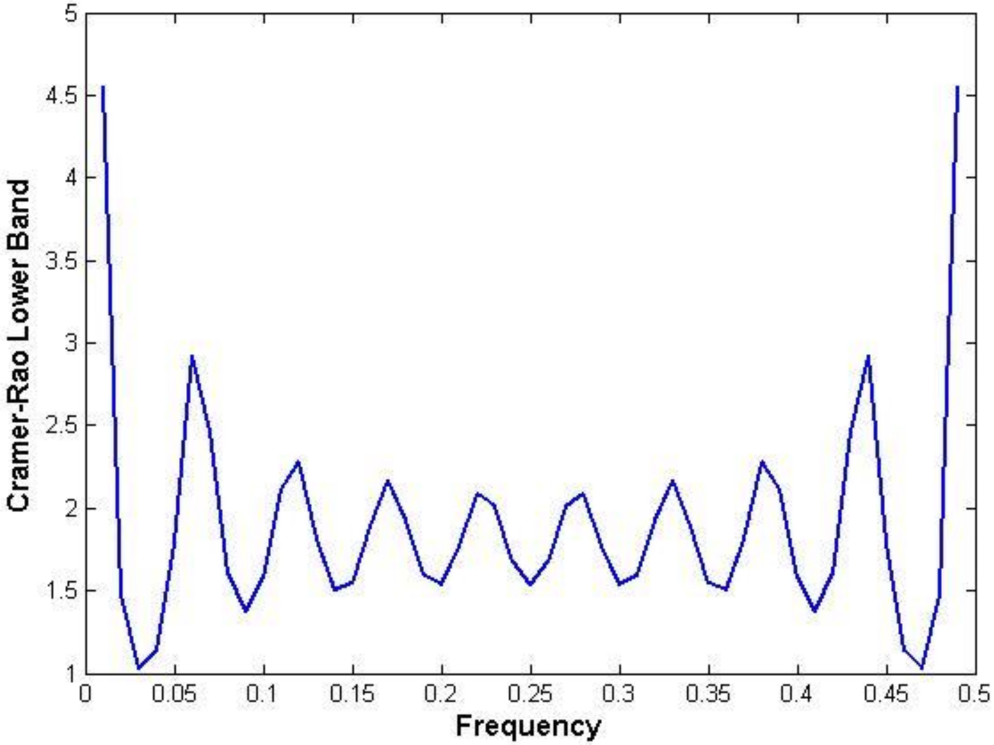
- Finally, the CRLB

$$\mathrm{var}(\hat{\theta}) \geq \frac{\sigma^2}{\sum_{n=0}^{N-1}\left(\frac{\delta s[n;\theta]}{\partial\theta}\right)^2}$$

- The form of the obtained bound demonstrates the importance of the signal dependence on θ: Signals that change rapidly as the unknown parameter changes result in accurate estimators.

Example: CRLB for sinusoidal frequency estimation

$$s[n; f_0] = A\cos(2\pi f_0 n + \Phi) \quad 0 < f_0 < \frac{1}{2}$$



$$CRLB = \mathrm{var}(f_0) \geq \frac{\sigma^2}{A^2 \sum_{n=0}^{N-1}[2\pi n \sin(2\pi f_0 n + \Phi)]^2}$$

Cramer-Rao Lower Bound

- If we find an estimator that achieves the CRLB, then we know that we have found an MVU estimator
- The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator (we know we are doing well if our estimator is close to the CRLB)
- The CRLB enables us to rule-out impossible estimators. Specifically, we know that is impossible to find an unbiased estimator that beats the CRLB (useful in feasibility studies)
- The theory behind the CRLB can tell us if an estimator exists that achieves the bound

Best Linear Unbiased Estimator (BLUE)

- In certain situations either the MVU estimate does not exist or the pdf of the data may not be known.
- A linear estimator that is unbiased and has minimum variance can be determined with the knowledge of the first and second moments of the pdf. Although this is suboptimal, it is frequently used in practice as it does not require knowledge of the data
- For linear models with Gaussian noise, BLUE is identical to the MVU estimates