Advanced Signal Processing

Minimum Variance Unbiased Estimation

Danilo Mandic

room 813, ext: 46271

Department of Electrical and Electronic Engineering

Imperial College London, UK

d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

Objectives

- Learn the concept of a **minumum variance unbiased (MVU)** estimator
- Investigate how the accuracy of an estimator depends upon the relationship between the unknown parameter(s) and the PDF of noise
- Study the requirements for the design of an efficient estimator
- Analyse the Cramer–Rao Lower Bound (CRLB) for the scalar case
- Extension to the Cramer–Rao Lower Bound (CRLB) for the vector case
- Dependence on data length (motivation for 'sufficient statistics')
- Examples:
  - DC level in WGN (frequency estimation in power, bioengineering)
  - Parameters of a sinusoid (scalar case, vector case)
  - A new view of Fourier analysis
  - System identification

Example 1: Consider a single observation $x[0] = A + w[0]$, where $w[0] \sim \mathcal{N}(0, \sigma^2)$

The simplest estimator of the DC level A in white noise $w[0] \sim \mathcal{N}(0, \sigma^2)$ is

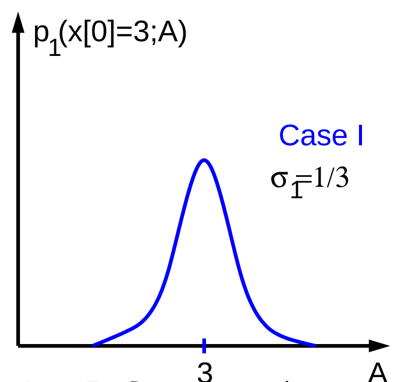
$$\hat{A} = x[0] \quad \Rightarrow \quad \hat{A} \quad \textbf{is unbiased, with variance } \sigma^2$$

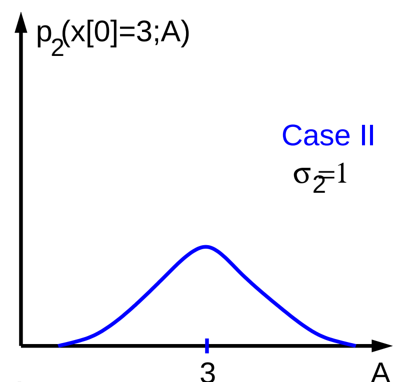
To show that the estimator accuracy improves as σ^2 decreases:

∘ Consider

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

∘ For $\underbrace{x[0] = 3}_{fundamental\ step,\ we\ are\ fixing\ the\ data\ value}$ and $i = 1, 2$ with $\sigma_1 = \frac{1}{3}$ and $\sigma_2 = 1$





Clearly, as $\sigma_1 < \sigma_2$, the DC level A is estimated more accurately with $p_1(x[0]; A)$

Likely candidates for values of $A \in 3 \pm 3\sigma \quad \Rightarrow \quad$ therefore [2, 4] for I and [0, 6] for II.

Likelihood function

When the PDF is viewed as a function of an unknown parameter **(with the dataset $\{x\} = x[0], x[1], \ldots$ fixed)** it is termed the **"likelihood function"**.

- The **"sharpness"** of the likelihood function determines the accuracy with which the unknown parameter may be estimated.
- Sharpness is measured by the **"curvature"** - a negative of the second derivative of the logarithm of the likelihood function **at its peak**.

Example 2: One sample of a DC level in WGN

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

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

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

Therefore, as expected, the curvature increases as σ^2 decreases

Likelihood function: Curvature

Since we know the variance of the estimator equals σ^2, then

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

and **the variance decreases as the curvature increases.**

Generally, the second derivative does depend upon $x[0]$, and hence a more appropriate measure of curvature is the statistical measure

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

which measures the average curvature of the log-likelihood function

Recall: The likelihood function is a random variable due to $x[0]$

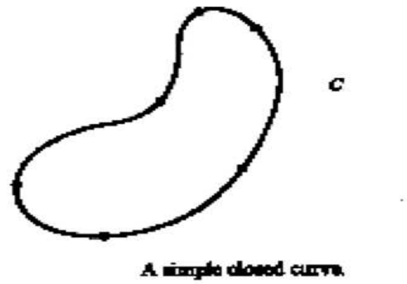
Recall: The Mean Square Error $\looparrowright$ MSE $=$ Bias^2 $+$ variance

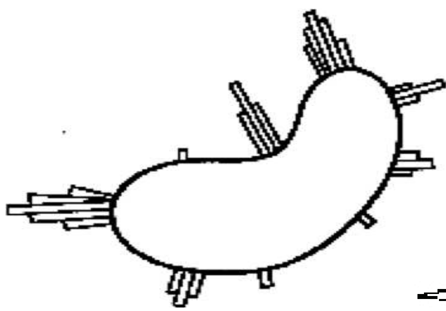
☞ it makes perfect sense to look for a minimum variance unbiased (MVU) solution

Link with human perception

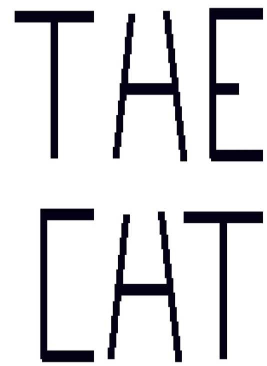
In the 50's phychologist Fred Attneave recorded eye dwellings on objects

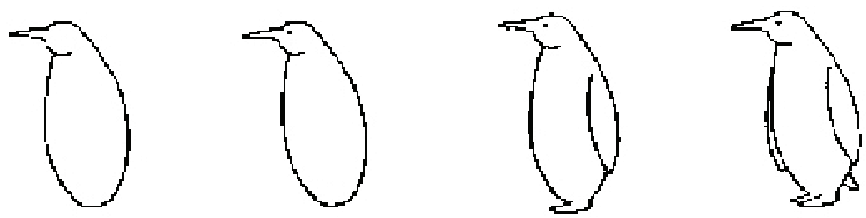
Example 3: The drawing of a bean (top) and the histogram of eye dwellings (bottom)





Example 4: Read the words below ... now read letter by letter ... are you still sure?





Example 3: Is the drawing on the left still a penguin?

So, what is the sufficient information to 'estimate' an object?

Cramer–Rao Lower Bound (CRLB)

Theorem: [CRLB] **Assumption:** The PDF $p(\mathbf{x};\theta)$ satisfies the "**regularity**" condition

$$E\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right] = 0, \qquad \forall \theta$$

where the expectation is taken with respect to $p(\mathbf{x};\theta)$.

Then, the variance of any **unbiased** estimator $\hat{\theta}$ must satisfy

$$Var(\hat{\theta}) \geq \frac{1}{-E\left\{\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right\}}$$

where the derivative is evaluated at the true value of θ.

CRLB for scalar parameter, continued

Moreover, an unbiased estimator may be found that attains the bound for all θ, if and only if for some functions g and $\mathcal{I}$

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

That estimator is the **minimum variance unbiased (MVU) estimator**, for which

$$\hat{\theta} = g(\mathbf{x})$$

and its minimum variance

$$\frac{1}{\mathcal{I}(\theta)}$$

—— end of CRLB theorem ——

Remark: Since the variance $Var(\hat{\theta}) \geq \frac{1}{-E\left\{\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right\}}$, the evaluation of the "curvature term" gives

$$E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} p(\mathbf{x};\theta) d\mathbf{x}$$

Obviously, in general the bound depends on the parameter θ and the data length

Example 5: Estimation of a DC level in WGN

Consider the estimation of a DC level in WGN, assume N observations

$$x[n] = \underbrace{A}_{unknown\ DC\ level} + \underbrace{w[n]}_{noise\ with\ known\ pdf} \qquad n = 0, 1, 2, \ldots, N-1$$

where $w[n] \sim \mathcal{N}(0, \sigma^2)$.

Determine the CRLB for the unknown DC level A, starting from $(\theta = A)$

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

☞ Estimation of a DC level is very useful, e.g. in the time-frequency plane a sinusoid of frequency f is represented by a straight line ☺

Example 5: DC level in WNG $\looparrowright$ continued

Taking the first derivative

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

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

where $\bar{x}$ denotes the sample mean.

Connection with CRLB: $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x(n) = g(\mathbf{x})$, $\quad \mathcal{I}(A) = N/\sigma^2$

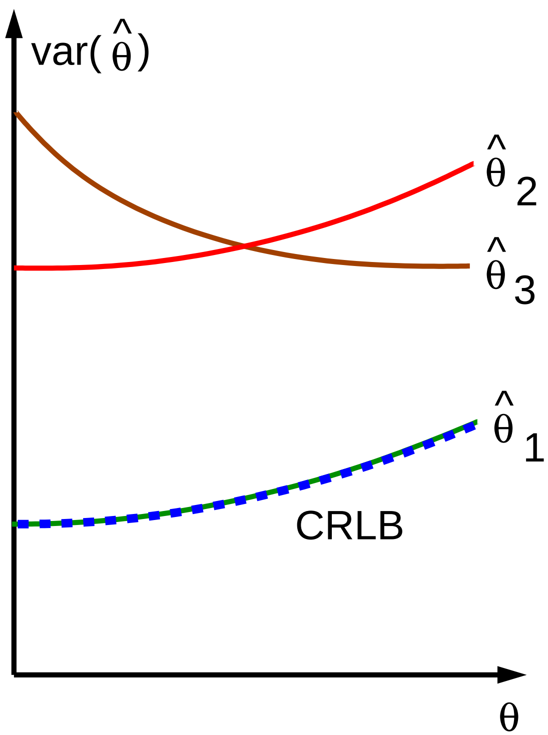
Differentiating again

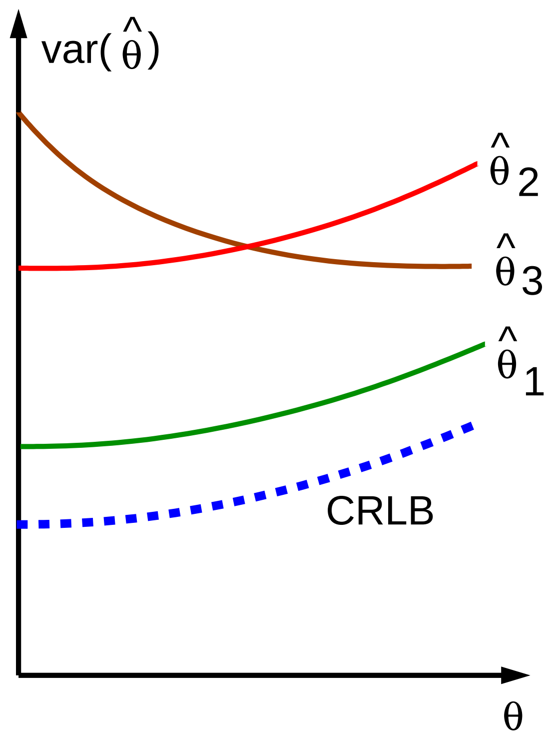
$$\frac{\partial^2 \ln p(\mathbf{x};A)}{\partial A^2} = -\frac{N}{\sigma^2}$$

Therefore $Var(\hat{A}) \geq \frac{\sigma^2}{N}$ is the CRLB, which implies that **the sample mean estimator attains the Cramer-Rao lower bound and must, therefore, be an MVU estimator in WGN.**

Efficient estimator $\looparrowright$ concept

An estimator which is unbiased and attains the CRLB is said to be **efficient** in that it is an Minimum Variance Unbiased (MVU) etimator and that it efficiently uses the data.





$\hat{\theta}_1$ is efficient and MVU, $\hat{\theta}_2, \hat{\theta}_3$ are not

$\hat{\theta}_1$ may be MVU but is not efficient

Fisher information

The denominator in the CRLB

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]$$

is referred to as the **Fisher information.**

Intuitively:

the more information available $\rightsquigarrow$ the lower the bound $\rightsquigarrow$ less variance

Essential properties of an information measure:

- ☞ Non–negative
- ☞ Additive for independent observations

General CRLB for **arbitrary signals** in WGN (see the next slide)

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

Accurate estimators: signals change rapidly with the parameter changes.

General case: Arbitrary signal in noise

Consider a deterministic signal $s[n;\theta]$ observed in WGN, $w \sim \mathcal{N}(0, \sigma^2)$

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

Then the PDF for $\mathbf{x}$ parametrised by $\boldsymbol{\theta}$ has the form

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

so that

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

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

and the Fisher information

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

Example 6: Sinusoidal frequency estimation

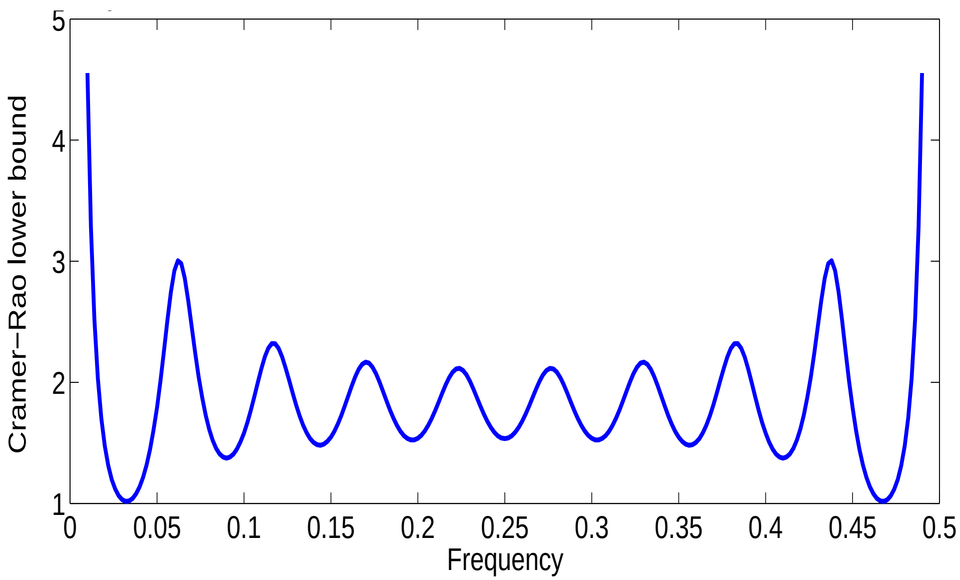
(the CRLB depends both on the unknown parameter f_0 and the data length N)

Consider a general sinewave in noise: $x[n] = A\cos\left(2\pi f_0 n + \Phi\right) + w[n]$

If only the freqency f_0 is unknown, then (for normalised frequency)

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

$$\text{and} \qquad Var(\hat{f}_0) \quad \geq \quad \frac{\sigma^2}{A^2 \sum_{n=0}^{N-1} \left[2\pi n \ \sin(2\pi f_0 n + \Phi)\right]^2}$$



Note the preferred frequencies, e.g.

$f \approx 0.03$, and that

for $f_0 \rightarrow \{0, 1/2\}$ the CRLB $\rightarrow \infty$

Paramet.: N $=$ 10, Φ $=$ 0, SNR$=A^2/\sigma^2 = 1$

Extension to a vector parameter

we now have the Fisher Information Matrix $\mathcal{I}$, s.t. $[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 ln\ p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]$

Formulation: Estimate a vector parameter $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$

- Recall that an unbiased estimator $\hat{\boldsymbol{\theta}}$ is efficient (and therefore an MVU estimator) when it satisfies the conditions of the CRLB

- It is assumed that the PDF $p(\mathbf{x};\boldsymbol{\theta})$ satisfies the **regularity conditions**

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \quad \forall \boldsymbol{\theta}$$

- Then the covariance matrix of any unbiased estimator $\hat{\boldsymbol{\theta}}$ satisfies

$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathcal{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}$ (symbol $\geq \mathbf{0}$ means that $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ is positive semidefinite)

- The Fisher Information Matrix is given by $[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]$

☞ An unbiased estimator $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ exists that satisfies the bound $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathcal{I}^{-1}(\boldsymbol{\theta})$ if and only if

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

Extension to a vector parameter: Fisher information

Some observations:

- Elements of the Information Matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ are given by

$$[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})]_{ij} = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]$$

  where the derivatives are evaluated at the true values of the parameter vector.

- The CRLB theorem provides a powerful tool for finding MVU estimators for a vector parameter.

☞ **MVU estimators for linear models are found with the Cramer–Rao Lower Bound (CRLB) theorem.**

Example 7: Sinusoid parameter estimation $\looparrowright$ vector case

Consider again a general sinewave

$$s[n] = A\cos\left(2\pi f_0 n + \Phi\right)$$

where A, f_0 and Φ are all unknown. Then, the data model becomes

$$x[n] = A\ cos(2\pi f_0 n + \Phi) + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $A > 0$, $0 < f_0 < 1/2$, and $w[n] \sim \mathcal{N}(0, \sigma^2)$.

Task: Determine CRLB for the parameter vector $\boldsymbol{\theta} = [A \quad f_0 \quad \Phi]^T$

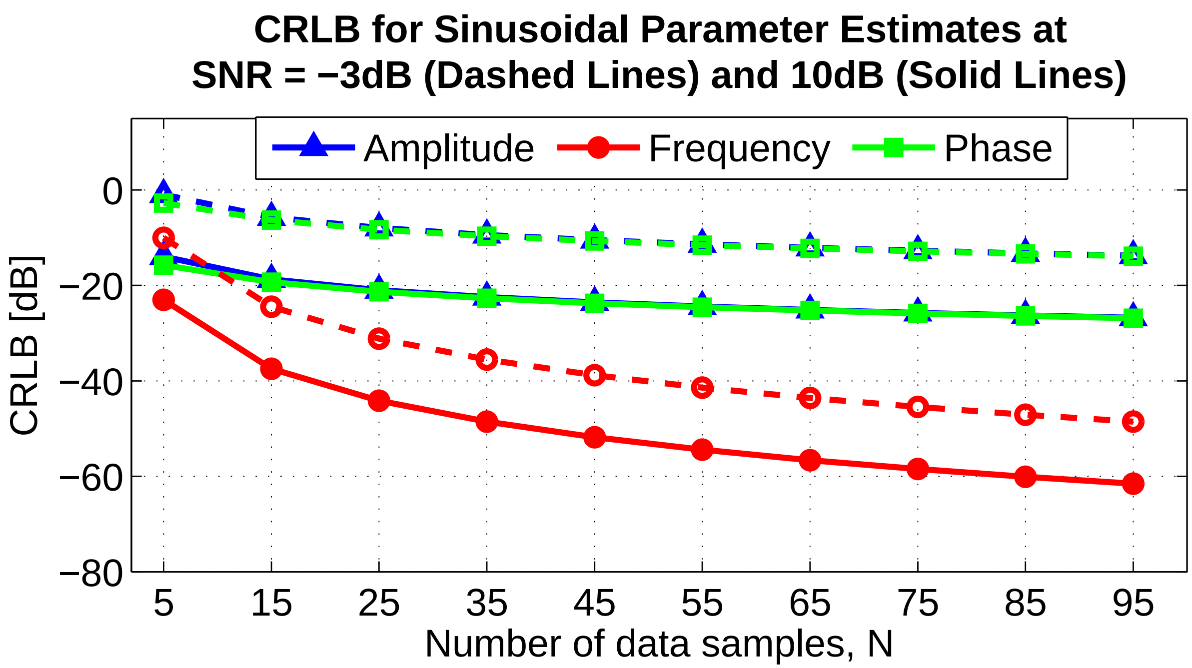
Solution: The elements of the Fisher Information Matrix become

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} N/2 & 0 & 0 \\ 0 & 2A^2\pi^2\sum_{n=0}^{N-1} n^2 & \pi A\sum_{n=0}^{N-1} n \\ 0 & \pi A\sum_{n=0}^{N-1} n & \frac{NA^2}{2} \end{bmatrix}$$

Example 7: Sinusoid parameter estimation $\looparrowright$ continued

After inversion, the diagonal components yield (where $\eta = \frac{A^2}{2\sigma^2}$ is SNR):

$$Var(\hat{A}) \geq \frac{2\sigma^2}{N} \qquad Var(\hat{f}_0) \geq \frac{12}{(2\pi)^2 \eta N(N^2-1)} \qquad Var(\hat{\Phi}) \geq \frac{2(2N-1)}{\eta N(N+1)}$$



☞ the variance of the estimated parameters of a sinusoid behaves $\propto 1/\eta$ and $\propto 1/N^3$, thus **exhibiting strong sensitivity to data length**

Linear models

Generally it is difficult to determine the MVU estimator.

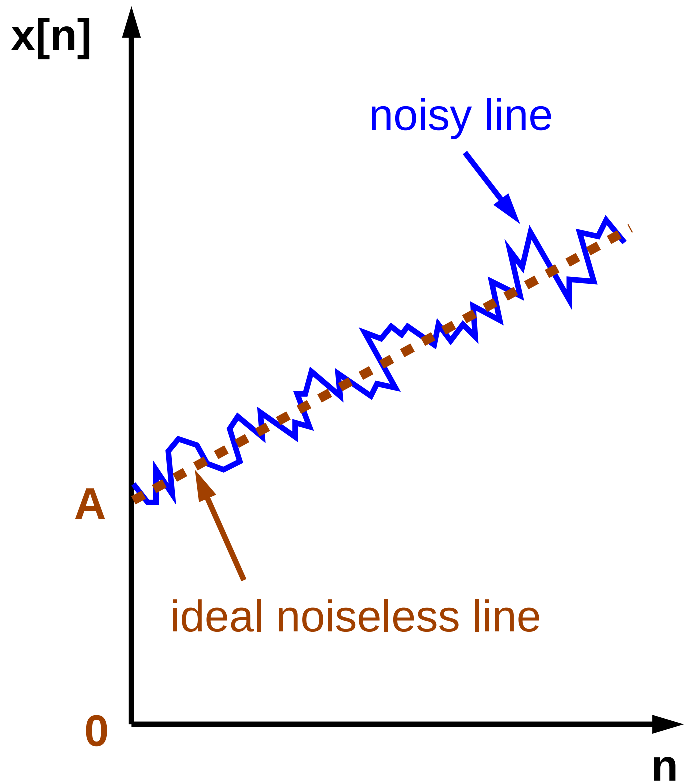
- In signal processing, however, a **linear data model** can often be employed $\Rightarrow$ straightforward to determine the MVU estimator.

Example 8: Linear model of a straight line in noise

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

where

- $w[n] \sim \mathcal{N}(0, \sigma^2)$,
- B - slope and
- A - intercept.



Linear models: Compact notation (Example 8 contd.)

This data model can be written more compactly in matrix notation as

$$\underline{x} = H\underline{\theta} + \underline{w} \qquad \text{or} \qquad \mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \big[x[0], x[1], \dots, x[N-1]\big]^T \qquad \mathbf{H} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}$$

and

$$\begin{aligned} \boldsymbol{\theta} &= [A \quad B]^T \\ \mathbf{w} &= [w[0], w[1], \dots, w[N-1]]^T \\ \mathbf{w} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \\ \mathbf{I} &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} \quad = \quad \text{diag}(1, 1, \dots, 1) \end{aligned}$$

Linear models: Fisher information matrix

NB: $p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})}{2\sigma^2}}$

The CRLB theorem can be used to obtain the MVU estimator for $\boldsymbol{\theta}$

The MVU estimator, $\hat{\boldsymbol{\theta}} = g(\mathbf{x})$, will satisfy

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

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the **Fisher information matrix**, whose elements are

$$[\boldsymbol{\mathcal{I}}]_{ij} = -E\Big[\frac{\partial^2 \ \ln \ p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\Big]$$

Applying the Linear Model

$$\begin{aligned}
\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\big(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\big)^T\big(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\big)\right] \\
&= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}\big[\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\big]
\end{aligned}$$

Note that the only quadratic term in θ involves the matrix $\mathbf{H}$

Linear models: Some useful matrix/vector derivatives

Use the identities

$$\frac{\partial \mathbf{b}^T \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \mathbf{b} \quad \looparrowright \quad \frac{\partial \mathbf{x}^T \mathbf{H} \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = (\mathbf{x}^T \mathbf{H})^T = \mathbf{H}^T \mathbf{x}$$

$$\frac{\partial \boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta} \quad \looparrowright \quad \frac{\partial \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{H}^T \mathbf{H} \boldsymbol{\theta}$$

(which you should prove for yourself), that is, the follow the rules of vector/matrix differentiation.

As a rule of thumb, watch for the position of the $(\cdot)^T$ operator

Then, for $\mathbf{A}$ symmetric:

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}\left[\mathbf{H}^T \mathbf{x} - \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}\right]$$

Linear models: Cramer-Rao lower bound

Find the MVU estimator: $\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\big(g\left(\mathbf{x}\right) - \boldsymbol{\theta}\big)$

Similarly (recall that $(\mathbf{H}^T\mathbf{H})^T = \mathbf{H}^T\mathbf{H}$),

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\frac{\partial^T}{\partial \boldsymbol{\theta}}\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}$$

Therefore

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \underbrace{\frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}}_{\mathcal{I}(\boldsymbol{\theta})}\Big[\underbrace{\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}}_{\mathbf{g}(\mathbf{x})} - \boldsymbol{\theta}\Big]$$

By inspection, the **linear estimator**

$$\hat{\boldsymbol{\theta}} = g\left(\mathbf{x}\right) = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

provided $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$ is invertible (it is, as $\mathbf{H}$ is full rank, with orthogonal rows and columns).

The covariance matrix of $\hat{\boldsymbol{\theta}}$ now becomes $\quad \mathbf{C}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{\mathcal{I}}^{-1}\left(\boldsymbol{\theta}\right) = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$

CRLB for linear models – continued

- **The MVU estimator for the linear model is efficient - it attains the CRLB**

- The columns of $\mathbf{H}$ must be **linearly independent** for $\left(\mathbf{H}^T\mathbf{H}\right)$ to be invertible

Theorem: (Minimum Variance Unbiased Estimator for the Linear Model)
If the observed data can be modeled as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

$\mathbf{x}$	is an N$\times$1 "observation vector"
$\mathbf{H}$	is an N$\times$p "observation matrix" of rank p
$\boldsymbol{\theta}$	is a p$\times$1 "parameter vector"
$\mathbf{w}$	is an N$\times$1 "noise vector" $\sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$

CRLB for linear models: Theorem

Then, the MVU estimator is given by

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

for which the covariance matrix

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$$

Note the statistical performance of $\hat{\boldsymbol{\theta}}$ is completely satisfied because $\hat{\boldsymbol{\theta}}$ is a **linear transformation** of a Gaussian vector $\mathbf{x}$, i.e.

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\right)$$

Example 9: Fourier analysis

Data model ($n = 0, 1, \ldots, N-1$, where $w[n] \sim \mathcal{N}\left(0, \sigma^2\right)$)

$$x[n] = \sum_{k=1}^{M} a_k \; cos\left(\frac{2\pi k n}{N}\right) + \sum_{k=1}^{M} b_k \; sin\left(\frac{2\pi k n}{N}\right) + w[n]$$

where the amplitudes a_k, b_k of the cosines and sines are to be estimated.

◦ Frequencies **harmonically related**, i.e. $f_1 = \frac{1}{N}$, and $f_k = \frac{k}{N}$.

◦ Parameter vector $\boldsymbol{\theta} = [a_1, a_2, \ldots, a_M, b_1, b_2, \ldots, b_M]^T$

Observation matrix **H** ($N \times \underbrace{2M}_{P}$–dimensional)

$$\mathbf{H} = \begin{bmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ \cos\frac{2\pi}{N} & \ldots & \cos\frac{2\pi M}{N} & \sin\frac{2\pi}{N} & \ldots & \sin\frac{2\pi M}{N} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos\frac{2\pi(N-1)}{N} & \ldots & \cos\frac{2\pi M(N-1)}{N} & \sin\frac{2\pi(N-1)}{N} & \ldots & \sin\frac{2\pi M(N-1)}{N} \end{bmatrix}_{N\times 2M}$$

Example 9: Fourier analysis $\looparrowright$ continued

For $\mathbf{H}$ to satisfy $N > P \Rightarrow M < \frac{N}{2}$.

Columns of $\mathbf{H}$ have to be **orthogonal**.

If we write

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{2M}]$$

where $\underline{h}_i = \mathbf{h}_i$ is the i-th column of $\mathbf{H}$

then

$$\mathbf{h}_i^T \mathbf{h}_j = 0 \qquad for \quad i \neq j$$

That is, $\mathrm{h}_i \perp \mathrm{h}_j$ and the columns of matrix H are orthogonal

Example 9: Fourier analysis $\looparrowright$ contd. contd.

The **orthogonality of the columns** follows from the discrete Fourier transform (DFT) relationships

$$\begin{aligned}
\sum_{n=0}^{N-1} \cos\left(\frac{2\pi i n}{N}\right) \cos\left(\frac{2\pi j n}{N}\right) &= \frac{N}{2}\delta_{ij} \\
\sum_{n=0}^{N-1} \sin\left(\frac{2\pi i n}{N}\right) \sin\left(\frac{2\pi j n}{N}\right) &= \frac{N}{2}\delta_{ij} \\
\sum_{n=0}^{N-1} \cos\left(\frac{2\pi i n}{N}\right) \sin\left(\frac{2\pi j n}{N}\right) &= 0 \qquad \forall i, j, \quad \text{s.t. } i, j = 1, 2, \ldots, M < \frac{N}{2}
\end{aligned}$$

where

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

In other words: (i) $\cos i\alpha \perp \sin j\alpha, \ \forall i, j$, (ii) $\cos i\alpha \perp \cos j\alpha, \ \forall i \neq j$, (iii) $\sin i\alpha \perp \sin j\alpha, \ \forall i \neq j$

Example 9: Fourier analysis $\rightarrow$ observation matrix

Therefore

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix} \left[\mathbf{h}_1, \ldots, \mathbf{h}_{2M}\right] = \begin{bmatrix} \frac{N}{2} & 0 & \ldots & 0 \\ 0 & \frac{N}{2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{N}{2} \end{bmatrix} = \frac{N}{2}\mathbf{I}$$

The MVU estimator of the amplitudes is

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \qquad \text{that is}$$

$$\hat{\boldsymbol{\theta}} = \frac{2}{N}\mathbf{H}^T\mathbf{x} = \frac{2}{N}\begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix}\mathbf{x} = \begin{bmatrix} \frac{2}{N}\mathbf{h}_1^T\mathbf{x} \\ \vdots \\ \frac{2}{N}\mathbf{h}_{2M}^T\mathbf{x} \end{bmatrix}$$

and finally the estimates of Fourier coefficients become

$$\hat{a}_k = \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right]\cos\left(\frac{2\pi kn}{N}\right) \qquad \hat{b}_k = \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right]\sin\left(\frac{2\pi kn}{N}\right)$$

Example 9: Finally $\looparrowright$ Fourier coefficients

$$\hat{a}_k = \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right] cos\left(\frac{2\pi kn}{N}\right)$$
$$\hat{b}_k = \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right] sin\left(\frac{2\pi kn}{N}\right)$$

namely the **discrete Fourier transform coefficients.**

Their covariance matrix is:

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1} = \frac{2\sigma^2}{N}\mathbf{I}$$

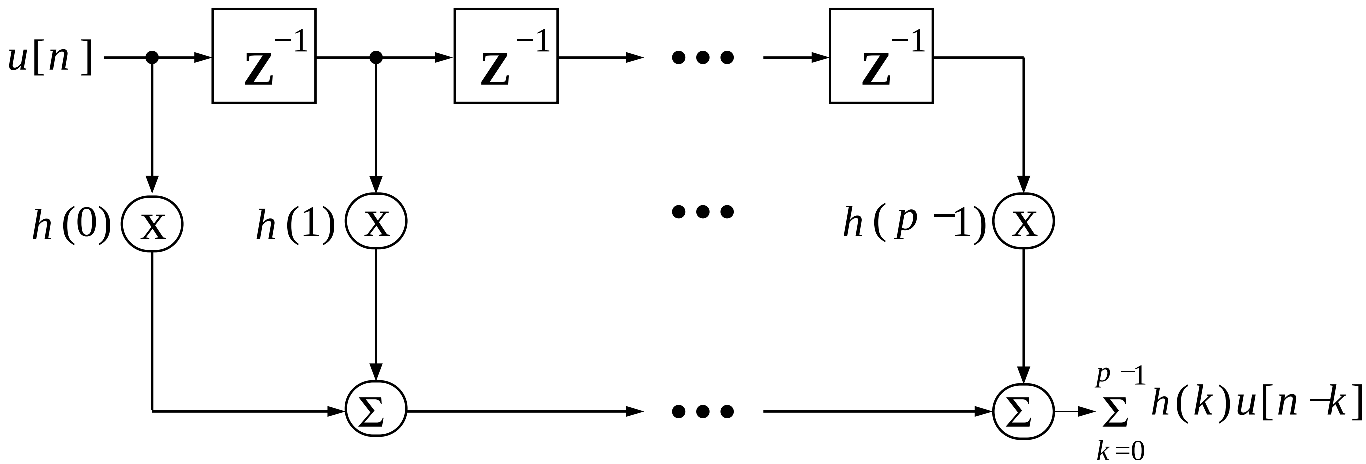
i) Note, as $\hat{\boldsymbol{\theta}}$ is Gaussian and the covariance matrix is diagonal, the amplitude estimates are statistically independent;

ii) The orthogonality of the columns of $\mathbf{H}$ is fundamental in the computation of the MVU estimator.

Example 10: System Identification (SYS ID)

Aim: to identify the model of a system from input/output data

Assume an FIR filter system model given in the figure below



- The input $u[n]$ "probes" the system, then the output of the FIR filter is given by $x[n] = \sum_{k=0}^{p-1} h(k)x[n-k]$
- We wish to estimate the filter coefficients [h(0),...,h(p-1)]^T from x(n)
- In practice, the output is corrupted by additive WGN

Example 10: SYS ID $\looparrowright$ data model in noise w

Data model

$$x[n] = \sum_{k=0}^{p-1} h(k)u[n-k] + w[n] \qquad n = 0, 1, \ldots, N-1$$

Equivalently, in the matrix–vector form

$$\underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{\text{obs. vec. } \mathbf{x}} = \underbrace{\begin{bmatrix} u[0] & 0 & \ldots & 0 \\ u[1] & u[0] & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \ldots & u[N-p] \end{bmatrix}}_{\text{measurement matrix } \mathbf{H}} \underbrace{\begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(p-1) \end{bmatrix}}_{\text{coeff. vec. } \boldsymbol{\theta}} + \underbrace{\begin{bmatrix} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{bmatrix}}_{\text{noise vec. } \mathbf{w}}$$

that is

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \qquad \text{where} \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

The MVU estimator

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \qquad \text{with} \qquad \mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2 \left(\mathbf{H}^T\mathbf{H}\right)^{-1}$$

This representation also lends itslef to state-space modelling

Example 10: SYS ID $\looparrowright$ more about $\mathbf{H}$

With this assumption $\mathbf{H}^T\mathbf{H}$ becomes a symmetric Toeplitz autocorrelation matrix

$$\mathbf{H}^T\mathbf{H} = N \begin{bmatrix} r_{uu}(0) & r_{uu}(1) & \dots & r_{uu}(p-1) \\ r_{uu}(1) & r_{uu}(0) & \dots & r_{uu}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_{uu}(p-1) & r_{uu}(p-2) & \dots & r_{uu}(0) \end{bmatrix}$$

where, e.g.

$$r_{uu}(0) = \frac{1}{N} \sum_{n=0}^{N-1-k} u[n]u[n+k]$$

For $\mathbf{H}^T\mathbf{H}$ to be diagonal, $r_{uu}(k) = 0$ for $k \neq 0$, which holds for a pseudorandom (PRN) sequence.

Finally, when $\qquad \mathbf{H}^T\mathbf{H} = N\, r_{uu}(0)\mathbf{I}$

$$Var\big(\hat{h}(i)\big) = \frac{\sigma^2}{N\, r_{uu}(0)}, \qquad i = 0, 1, \dots, p-1$$

Example 10: SYS ID $\looparrowright$ MVU estimator

For a PRN sequence, the MVU estimator becomes

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

Then

$$\hat{h}(i) = \frac{1}{N\, r_{uu}(0)} \sum_{n=0}^{N-1} u[n-i]x[n]$$

and

$$\frac{r_{ux}(i)}{r_{uu}(0)} = \frac{\frac{1}{N}\sum_{n=0}^{N-1-i} u[n]x[n+i]}{r_{uu}(0)}$$

$$i = 0, 1, \ldots, p-1$$

Thus the MVU estimator is the ratio of the input-output cross-correlation to the input autocorrelation.

Compare with the Wiener filter in Lecture 7.

Theorem: The MVU Estimator for a General Linear Model (GLM)

i) Data model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \underbrace{\mathbf{s}}_{\text{known signal}} + \mathbf{w}$$

$$\mathbf{w} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{C}\big)$$

ii) Then, the MVU estimator

$$\hat{\boldsymbol{\theta}} = \big(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\big)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\big(\mathbf{x} - \mathbf{s}\big)$$

iii) with covariance matrix

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \big(\mathbf{H}^T\mathbf{C}\mathbf{H}\big)^{-1}$$

Things to remember about MVU estimators

- **An estimator is a random variable** and as such its performance can only be described statistically by its PDF
- The use of computer simulations for assessing the performance of an estimator is **rarely conclusive**
- Unbiased estimators **tend to have symmetric PDFs** centred about the true value of θ
- The minimum mean square error (MMSE) criterion is natural to search for optimal estimators, but it most often leads to unrealisable estimators (those that cannot be written solely as a function of data)
- Since MSE$=$ Bias^2 $+$ variance, any criterion that depends on bias ought to be abandoned – we need to consider an alternative approach
- **Remedy:** Constrain the bias to zero and find an estimator which minimises the variance – the minimum variance unbiased (MVU) estim.
- Minimising the variance of an unbiased estimator also has the effect of concentrating the PDF of the estimation error, $\hat{\theta} - \theta$, about zero $\looparrowright$ **estimation error less likely to be large**

A few things about CRLB to remember

Even if the MVU estimator exists, there is no "turn of the crank" procedure to find it.

The CRLB sets a lower bound on the variance of any unbiased estimator!

This can be extremely useful in several ways:

- If we find an estimator that achieves the CRLB $\Rightarrow$ we known we have found an MVU estimator
- The CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator
- The CRLB enables us to rule out impossible estimators. **It is physically impossible to find an unbiased estimator that beats the CRLB**
- We may require the estimator to be linear, which is not necessarily a severe restriction, as shown on the example of the estimation of Fourier coefficients

Notes:

○

Notes:

○

Notes:

◦