Statistical Signal Processing & Inference

Minimum Variance Unbiased Estimation (MVU)

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

Motivation (from Lecture 3)

A natural criterion to define **optimal estimators** is the

Mean Square Error (MSE): $$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}$$

which measures the average mean squared deviation of the estimate, $\hat{\theta}$, from the true parameter value, θ.

☞ **We desire to minimise the error power** (see Lectures 3, 6 and 7)

$$MSE\,(\hat{\theta}) = var(\hat{\theta}) + B^2(\hat{\theta})$$

MSE = VARIANCE OF THE ESTIMATOR + SQUARED BIAS

Of particular interest are unbiased estimators for which

$$\min_{\hat{\theta}} MSE(\hat{\theta}) \equiv \min_{\hat{\theta}} var(\hat{\theta})$$

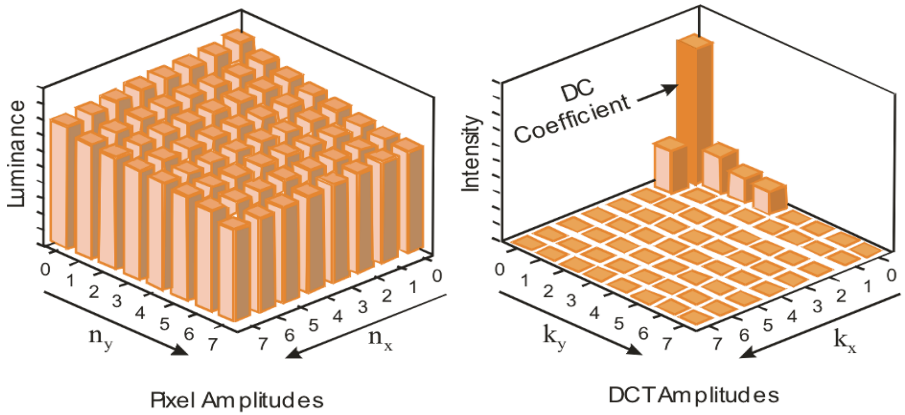
so that we can use our available degrees of freedom to minimise one performance metric (variance), instead of two (bias, variance).

☞ We now establish a theoretical bound on the optimality of MVU estimators.

Motivation: Practical example from High Definition TV

There are so many ways SSP&I can help with sustainability.

Principle of HDTV:



DFTs are everywhere:

- 4G mobile
- LANs
- OFDM & SC-OFDM
- MP3 and MPEG (audio-video source coding)
- Radar (FMCW)
- Computer axial tomography (projection theorem)
- crystralography

In HDTV, a $1000\times1000 = 10^6$ pixel image at 60 frames per sec analysed using a 8×8 DCT

$$\Rightarrow \frac{10^6\ pixels \times 60\ fps}{64\ pixels\ per\ FFT} \approx 10^6 \text{ FFT/s}$$

If 1.4% of world's 7×10^9 population watches TV at any given time, then (1.4% of $7\times10^9 = 10^8$)

$$10^6 \text{ FFT/s} \times 10^8 \text{ TVs} = 10^{14} \text{ FFT/s}$$

(the same as the number of cells in human body)

Now, there are 3600s in an hour and 30×10^6 sec in a year

$\Rightarrow$ # DFT/year = $10^{14} \times 40 \times 10^{20}$ for TV only

In a hardware implementation, the computation of a single FFT takes 52.82 nJ/FFT.

Therefore, the power consumption, per second, for FFTs for HDTV is

$52.82 \times 10^{-9} \times 10^{14}$ J = 5.282 MW.

This corresponds to the energy consumption of 20 GWh or 166 TJ per year.

Objectives

- Learn the concept of **minumum variance unbiased (MVU)** estimation
- 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
- Optimal parameter estimation, linear models, General Linear Model
- Dependence on data length (motivation for 'sufficient statistics')
- Examples:
  - DC level in WGN (frequency estimation in smart grid, bioengineering)
  - Regression as in Capital Asset Pricing Model (CAPM) in finance
  - Finding parameters of a sinusoid, e.g. in communications, radar, sonar, bioengineering (scalar case, vector case)
  - A new, statistical, view of Fourier analysis, performance bounds
  - System identification

What is the Cramer–Rao Lower Bound (CRLB)

☞ The CRLB is a lower bound on the variance **of any unbiased estimator**.

In other words, if $\hat{\theta}$ is an unbiased estimator of θ, then

$$\sigma^2_{\hat{\theta}} \geq CRLB_{\hat{\theta}}(\theta) \qquad \text{or} \qquad \sigma_{\hat{\theta}} \geq \sqrt{CRLB_{\hat{\theta}}(\theta)}$$

Therefore, the CRLB is a benchmark which tells us the best we can ever expect to be able to achieve with an unbiased estimator.

The CRLB is a must–check quantitative bound for:

- Feasibility studies (sensor relevance, if we met problem specifications)
- Assessment of the quality (goodness) of any derived estimator (we can only do as good as CRLB)
- It can sometimes provide the form of the MVU estimator (we just read it out from the CRLB theorem)
- It may be used to demonstrate the importance of physical/signal parameters to the estimation problem (e.g. optimum freq. for power)

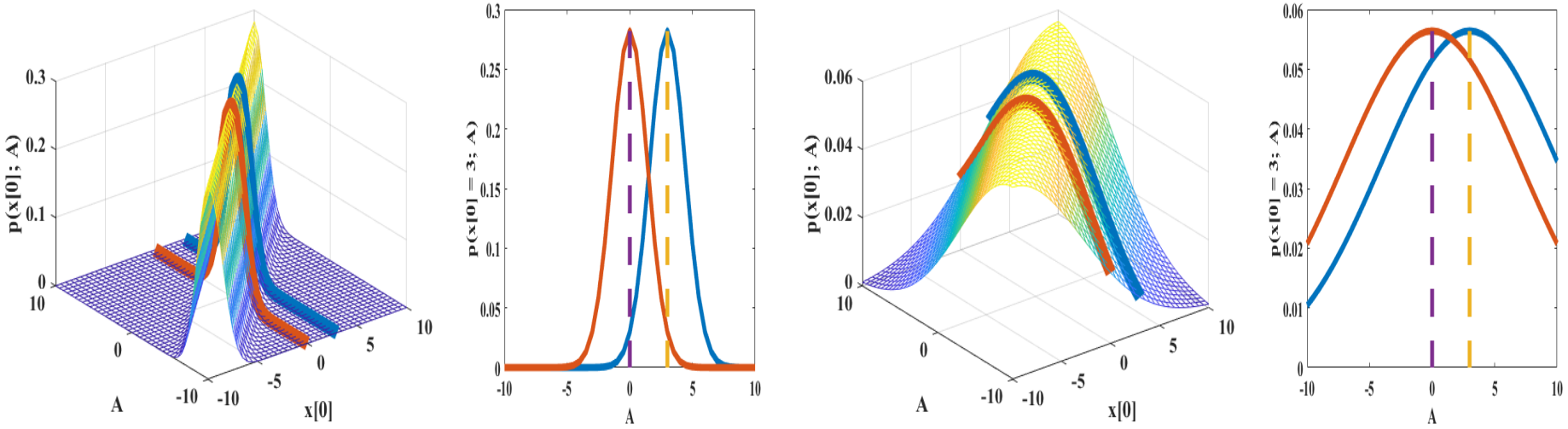
The need for the "parametrised" pdf, $p(x[0];\theta)$

$p(\mathbf{x};\theta) \leftrightarrows$ **a function of θ for fixed observed data $\mathbf{x}$ (i.e. a family of distributions)**

Q: What determines how well we estimate the unknown θ from the observed data $\mathbf{x}$?

A: Since the data $\mathbf{x}$ is a random process which depends on θ, it is the **parametrised pdf** which describes that dependence, denoted by $p(\mathbf{x};\theta)$

☞ Clearly, if $p(\mathbf{x};\theta)$ depends strongly/weakly on θ, then this implies that we should be able to estimate θ well/poorly.



Left: Strong dependence on θ **Right:** Weak dependence on θ

☞ The mean of the parametrised pdf (red & blue slices) depends on the observed point $x[0]$.

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

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

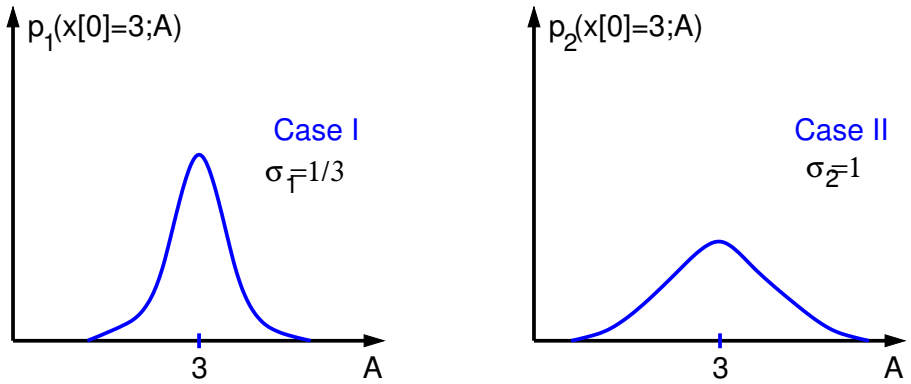
To show that the estimator accuracy improves as σ^2 decreases:

- 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 $\underbrace{x[0] = 3}$ and $i = 1, 2$ with $\sigma_1 = \frac{1}{3}$ and $\sigma_2 = 1$

fundamental step, we are fixing the data value



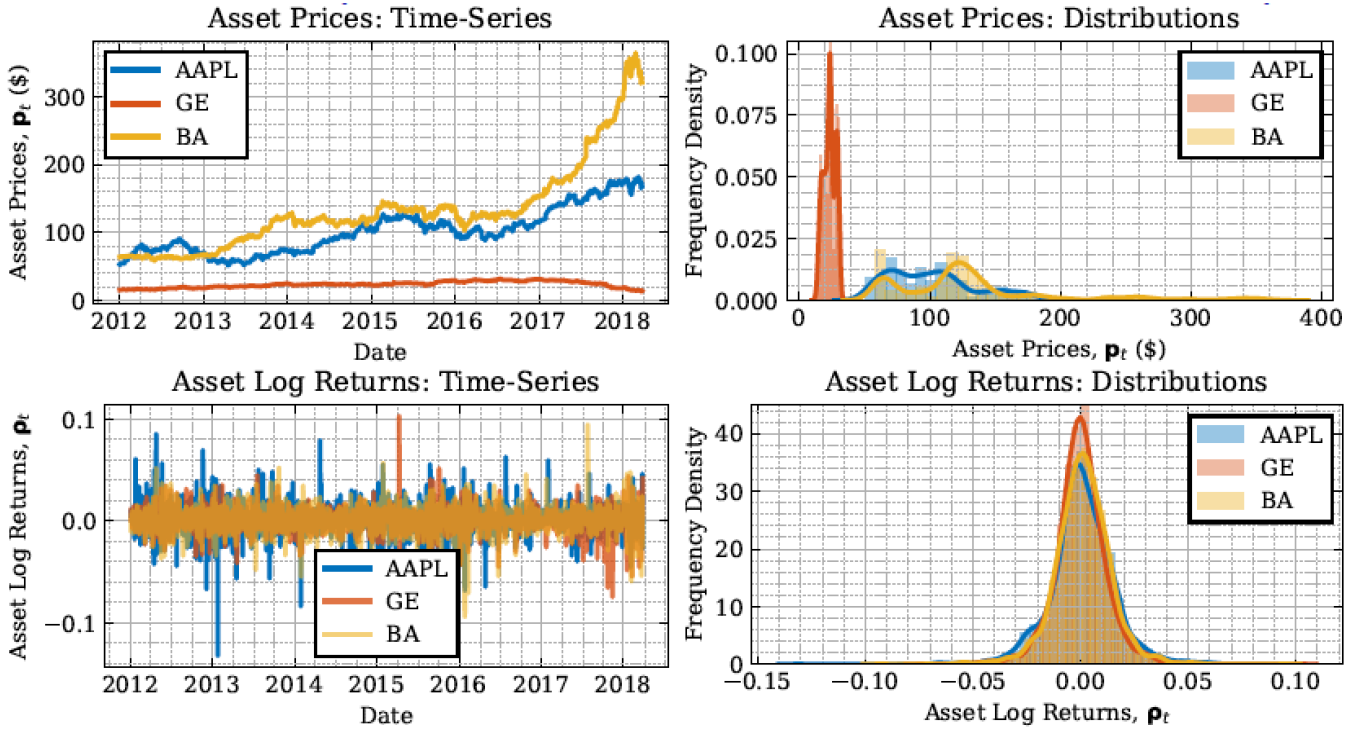
Clearly, as $\sigma_1 < \sigma_2$, the DC level A is estimated more accurately with $p_1(x[0]; A)$

Likely candidates for values of $A \in 3 \pm 3\sigma$ $\Rightarrow$ therefore [2, 4] for σ_1 and [0, 6] for σ_2.

Can we resort to (approximately) Gaussian distribution?

Yes, very often, if we re–cast our problem in an appropriate way (see Appendix 2)

Top panel. Share prices, p_n, of Apple (AAPL), General Electric (GE) and Boeing (BA) and their histogram (right). **Bottom panel.** Logarithmic returns for these assets, $ln(p_n/p_{n-1})$, that is, the log of price differences at consecutive days (left) and the histogram of log returns (right).



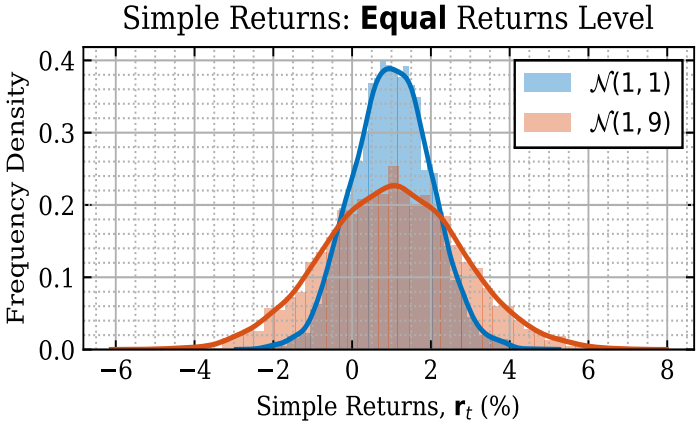
Clearly, by a suitable data transformation, we may arrive at symmetric distributions which are more amenable to analysis (bottom right).

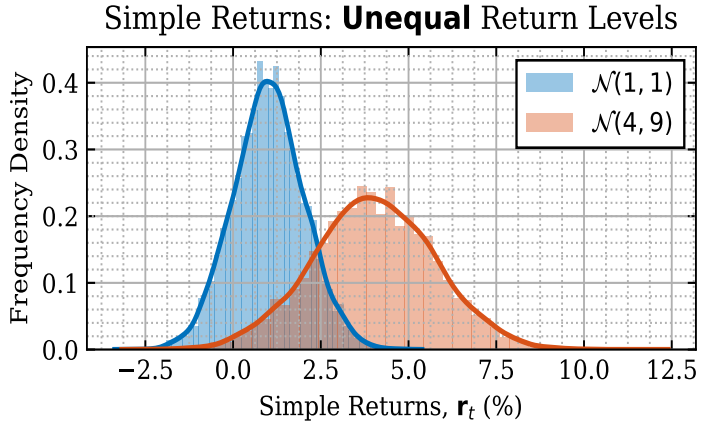
Putting this into context ↬ Sharpe ratio in finance

In financial modelling, the Sharpe Ratio (SR) models the risk-adjusted returns, whereby the volatility (risk) is designated by the variance of the distribution of returns. $return = price(t)/price(t-1)$

Sharpe ratio. The blue asset (narrower pdf) is less profitable but also less risky. To balance between the risk and profit, we can use the Sharpe ratio

$$\mathbf{SR}_{1:T} = \sqrt{T}\frac{E[\mathbf{r}_{1:T}]}{Var[\mathbf{r}_{1:T}]} \quad or\ for\ a\ single\ asset \quad SR = \frac{\mu}{\sigma}$$





Here, $SR_{blue} = \sqrt{T}\frac{1}{1}$ which is smaller than $SR_{red} = \sqrt{T}\frac{4}{3}$.

☞ We therefore choose the red asset.

Likelihood function

When a 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 at which the unknown parameter may be estimated.
- Sharpness is measured by the **"curvature"** $\looparrowright$ a negative of the second derivative of the logarithm of the likelihood function **at its peak**.

Example 2: Estimation based on 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.

☞ **Curvature $\nearrow$ $\Rightarrow$ PDF concentration $\nearrow$ $\Rightarrow$ Accuracy $\nearrow$**

Likelihood function: Curvature

Since we know that 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 one data point,** $x[0]$, and hence a **more appropriate measure of curvature is the statistical measure** (average over many random $x[0]$)

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

which measures the average curvature of the log-likelihood function

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

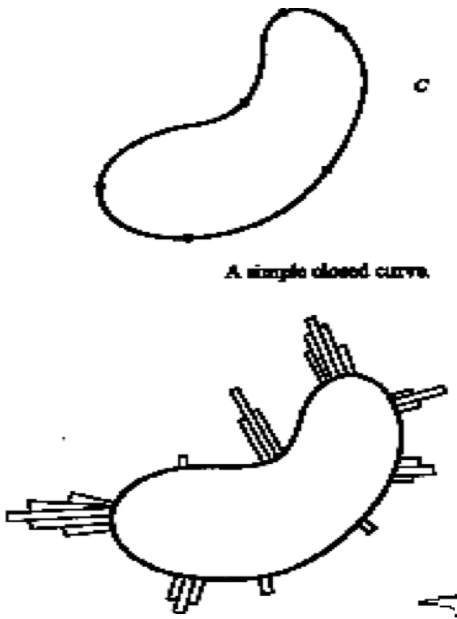
Recall: The Mean Square Error $\quad \looparrowright \quad$ MSE = Bias^2 + variance

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

Link between the curvature and human perception

In the 50s, a psychologist Fred Attneave recorded eye dwellings on objects

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



Example 3b): Read the words below ... now read letter by letter ... are you still sure?

THE
CAT

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

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

The 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

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

↑ average curvature

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

CRLB for a 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 $\mathcal{I}$

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

This 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 4: Physical relevance of CRLB

☞ Point 3 from Slide 5: "CRLB can sometimes provide the form of MVU"

Shall we therefore compare the form of regularity condition with Example 3

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

inverse of the minimum achievable variance ↑ ↑ form of the optimum est.

Compare with what we have derived for $x[0] = A + w[0]$ (Slide 9)

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

$\downarrow g(x[0])$

inverse of the Fisher information ↑ ↑ the unknown parameter

☞ By inspection, the optimum estimate is $\hat{A} = g([x[0]) = x[0]$

☞ From the CRLB theorem the optimum variance of this estimator: $\frac{1}{\mathcal{I}(\theta)} = \sigma^2$

Therefore: Good estimator $\Rrightarrow$ variance $\searrow$ and curvature $\nearrow$

Poor estimator $\Rrightarrow$ variance $\nearrow$ and curvature $\searrow$ (see Slide 9)

Example 5: DC level in WGN for N data points

for the validity of the Gaussian assumption, see Appendix 2 and Lecture 3 (S 21)

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 $(\theta = A)$

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

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

☞ 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 (specgramdemo)

Example 5: DC level in WGN for N data points $\hookrightarrow$ contd.

Upon taking the first derivative, we have

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

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

$\uparrow g(\mathbf{x})$, we can read out the estimator

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

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

Upon differ. again

$\downarrow$ does not depend on $\mathbf{x}$, so no $E\{\cdot\}$

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

Therefore $var(\hat{A}) = \frac{\sigma^2}{N}$ = CRLB, which implies that **the sample mean estimator attains the Cramer-Rao LB and must, therefore, be an MVU estimator of a DC level in WGN.**

Example 5: DC level in WGN (spelling out the previous slide)

Upon taking the first derivative, we have

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

$$= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n]-A\right) = \frac{N}{\sigma^2}\left(\bar{x}-A\right) \overset{\text{CRLB Th}}{=} \mathcal{I}(A)\left(g(\mathbf{x})-A\right)$$

$\mathcal{I}(A) \uparrow \quad \uparrow g(\mathbf{x})$

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

Upon differentiating again $\downarrow$ does not depend on $\mathbf{x}$, so no $E\{\cdot\}$

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

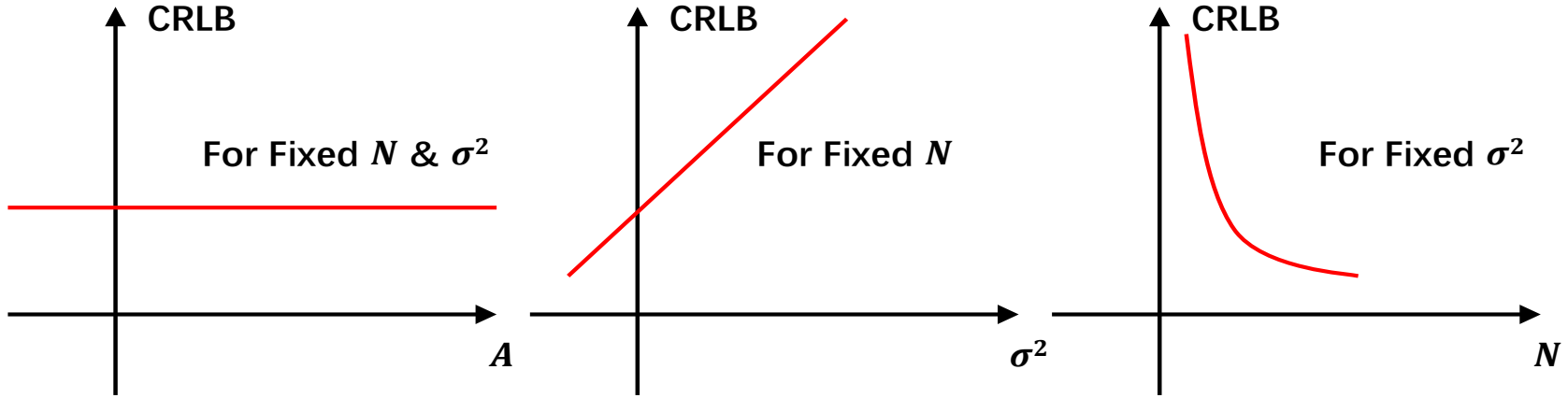
Therefore $var(\hat{A}) = \frac{\sigma^2}{N}$ = CRLB, which implies that **the sample mean estimator attains the Cramer-Rao LB and must, therefore, be an MVU estimator of A in WGN.** (for any other estimator, $\tilde{A}$, $var(\tilde{A}) \geq \sigma^2/N$)

Let us have a closer look at the CRLB for N data points

The figures below illustrate the behaviour of

$$\text{CRLB}_N = var(\hat{A}) = \frac{\sigma^2}{N} \qquad (\textit{cf. } \text{CRLB}_1 = \sigma^2)$$

with a change in the DC level, A, data length, N, and noise variance, σ^2.



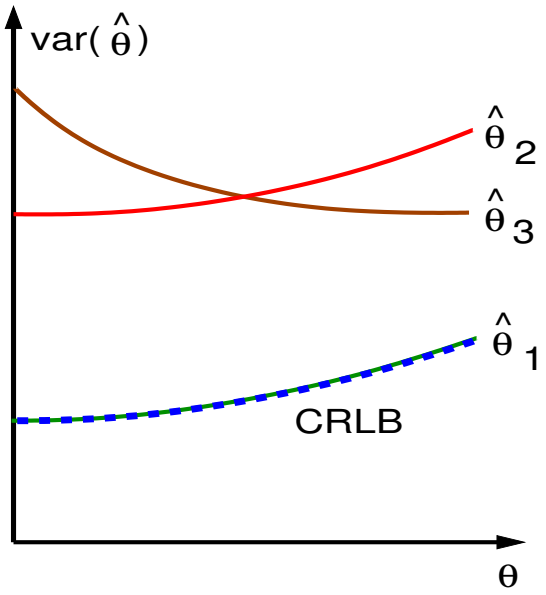
Properties of CRLB for DC level estimation from N noisy data points:

- **It does not depend on the DC level A**
- The CRLB increases linearly with the noise variance, σ^2
- The CRLB decreases as an inverse in the data length, N. For example, **doubling the data length halves the CRLB**

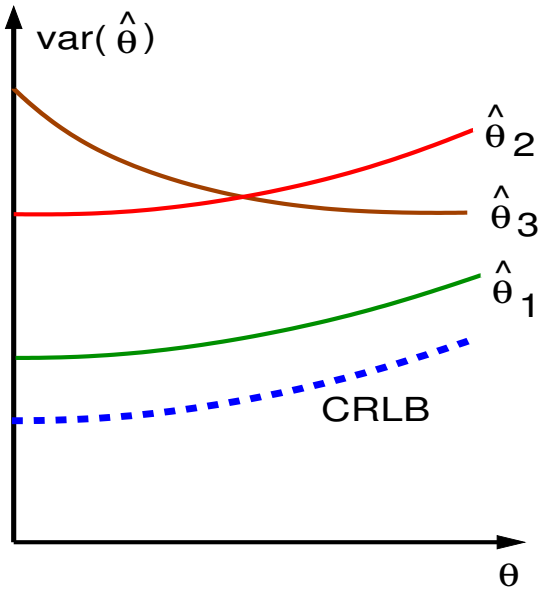
Efficient estimator ↬ concept *cf.* consistent est.

Def: An estimator which is unbiased and attains the CRLB is said to be efficient. In other words, an estimator is efficient if:

- It is an Minimum Variance Unbiased (MVU) estimator, and
- 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

☞ Not all estimators (phase est.) & not all MVU estimators are efficient

Fisher information and a general form of MVU estimator

(measures the "expected goodness" of data for making an estimate)

The term

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

in the CRLB theorem is referred to as the **Fisher information.**

Intuitively:

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

☞ **Essential properties of an information measure:**

- ☞ Non–negative
- ☞ Additive for independent observations

☞ General CRLB for **arbitrary signals** in WGN (*cf.* σ^2/N, see the next slide)

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

↑sensitivity of signal to parameter change

Accurate estimators: A signal is very sensitive to parameter change. Therefore $\frac{\partial s[n;\theta]}{\partial \theta}$ above acts as a "sensitivity" term. (see Appendix 1)

General case: Arbitrary signal in noise

CRLB via parameter sensitivity (for alternative forms, see Appendix 1)

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

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

and so

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

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

Therefore, the Fisher information

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

Example 6: Sinusoidal frequency estimation

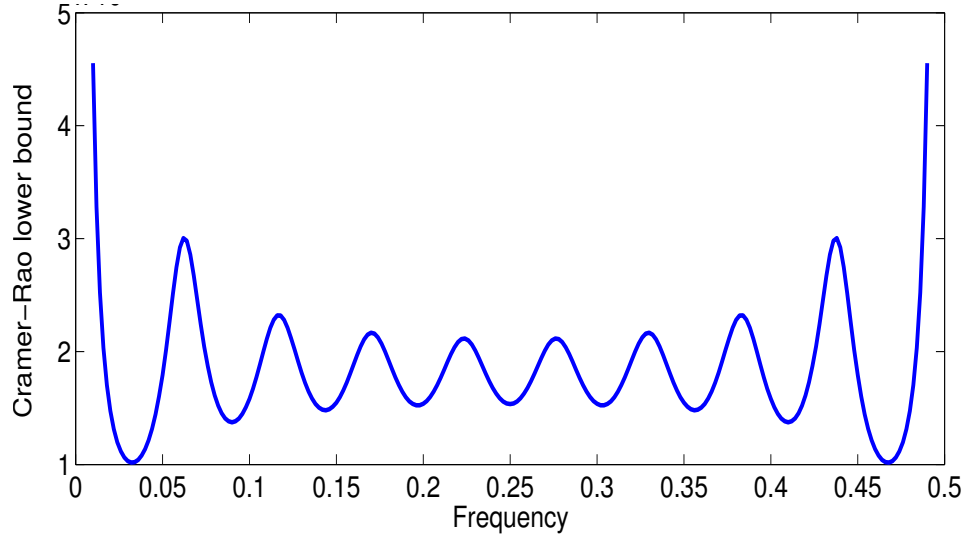
(the CRLB depends both on the unknown parameter f_0 and the data length N)

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

If only the frequency f_0 is unknown, then

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

From Slide 20: $$var(\hat{f}_0) \geq \frac{\sigma^2}{\sum_{n=0}^{N-1}\left(\frac{\partial s[n;\theta]}{\partial\theta}\right)^2} = \frac{\sigma^2}{A^2\sum_{n=0}^{N-1}\left[2\pi n\ \sin(2\pi f_0 n + \Phi)\right]^2}$$



Note the preferred frequencies, e.g.

$f \approx 0.03$, and that

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

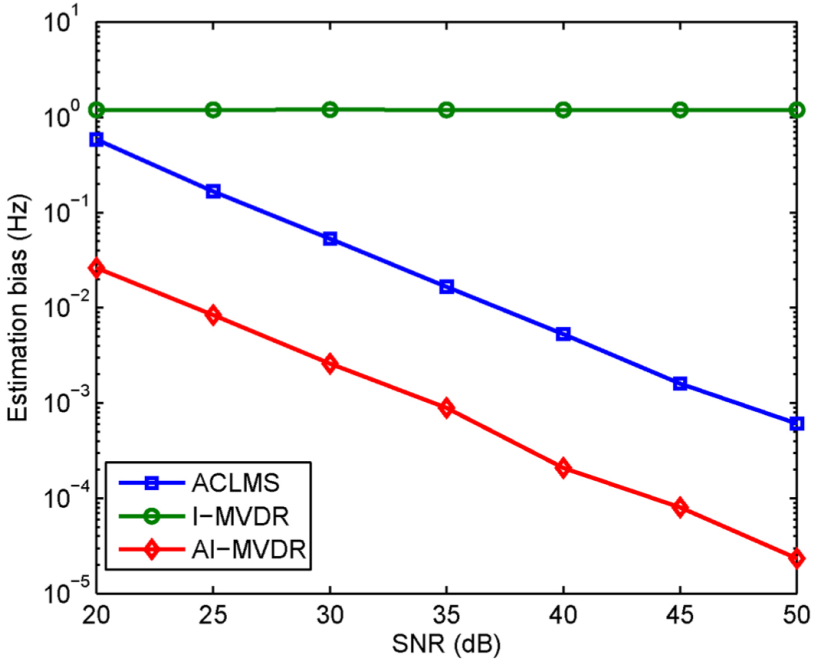
Parameters: N = 10, $\Phi = 0$, SNR = 1

Example 6: Sinusiodal frequency estimation (contd.)

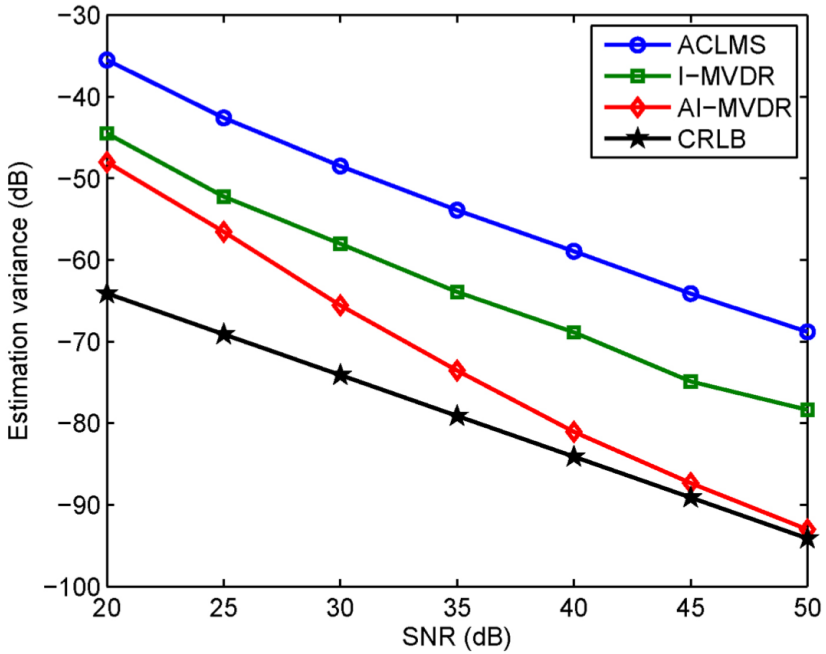
Practical context: Frequency estimation in power (smart) grids, $f_s = 1000$ Hz

To illustrate the bias–variance–consistency, consider some recent frequency estimation algorithms (see Lecture Supplement). For convenience, the performance was evaluated against the signal to noise ratio (SNR).

Observe that both bias, variance and CRLB are a function of SNR.



Left: Bias in frequency estimation



Right: Variance against the CRLB

The AI-MVDR algorithm was asymptotically unbiased and also consistent, as it approached the CRLB for frequency est. with an increase in SNR.

CRLB Theorem: Extension to a vector parameter

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

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**

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

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

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

- 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 matrix

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 ↪ 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 (P&As)

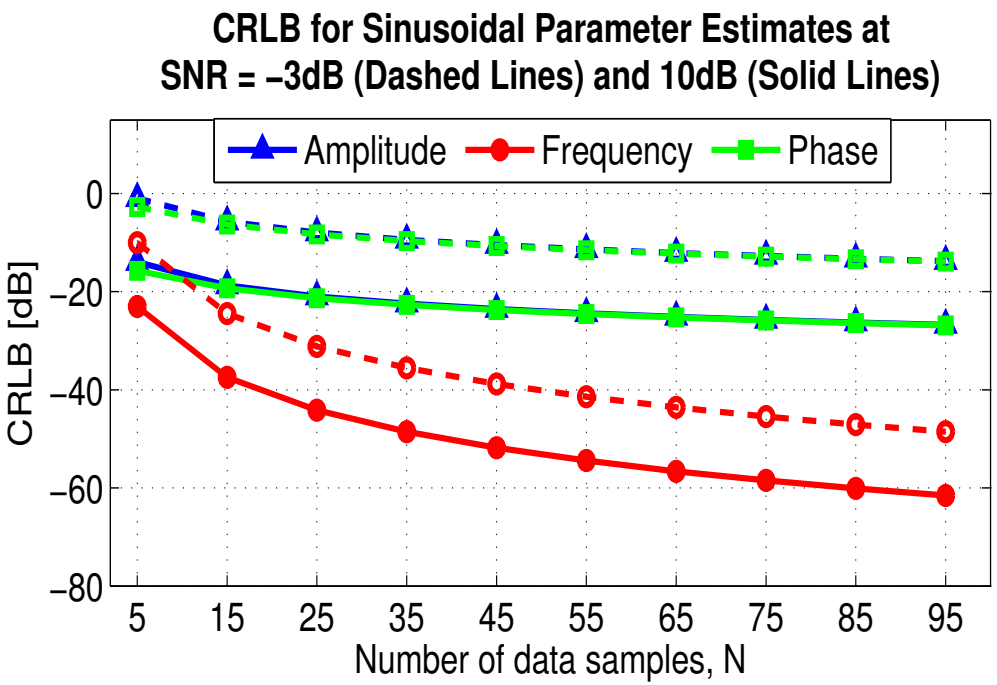
$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} N/2 & 0 & 0 \\ 0 & 2A^2\pi^2\sum_{n=0}^{N-1} n^2 & \pi A\sum_{n=0}^{N-1} n \\ 0 & \pi A\sum_{n=0}^{N-1} n & \frac{NA^2}{2} \end{bmatrix}$$

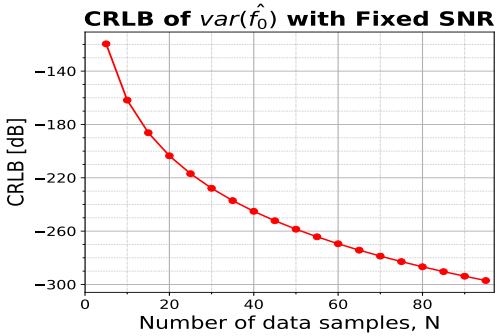
Example 7: Sinusoid parameter estimation ↬ continued

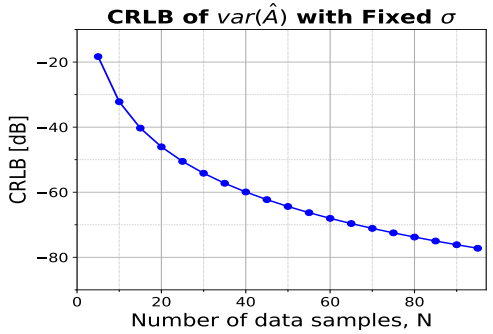
since $\mathrm{C}_{\hat{\theta}} = \mathcal{I}^{-1}(\theta)$ (see Slide 22) make an inverse of the FIM

After inversion of $\mathcal{I}(\boldsymbol{\theta})$, its diagonal components are ($\eta = \frac{A^2}{2\sigma^2}$ is SNR):

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



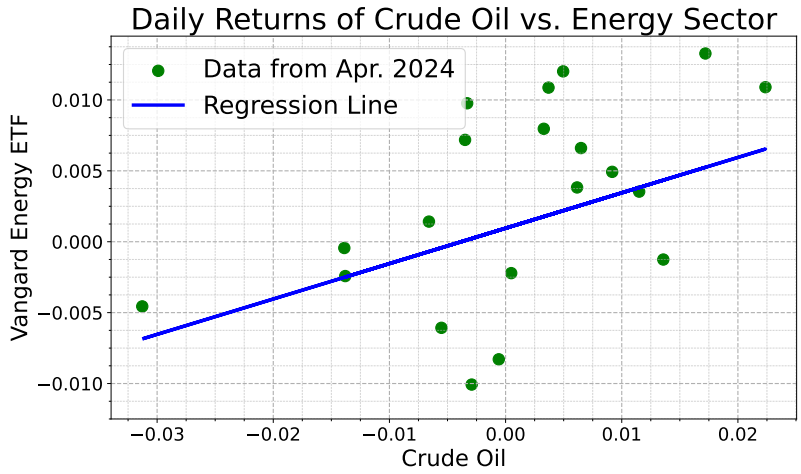


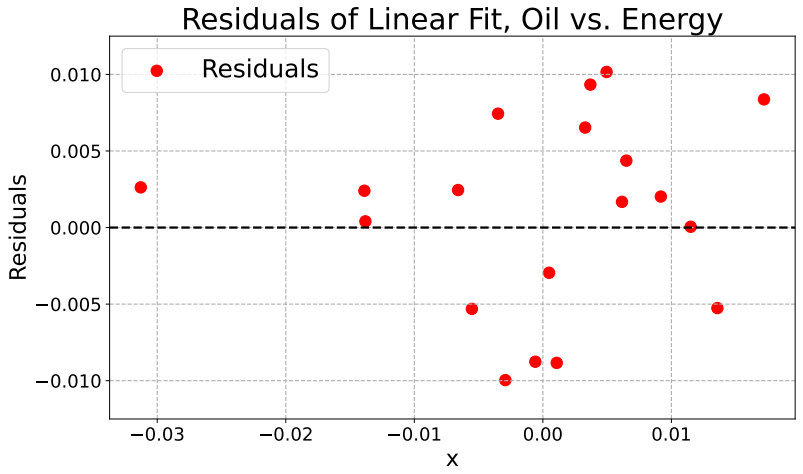


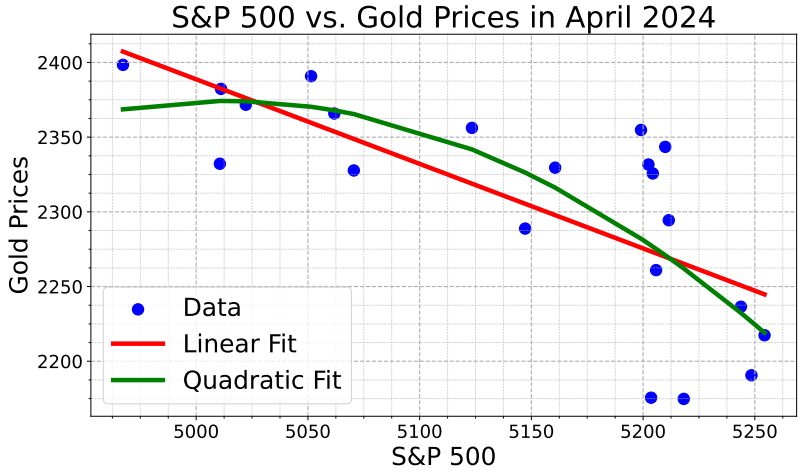
☞ The variance of the estimated parameters of a sinusoid behaves $\propto 1/\eta$ and $\propto 1/N^3$, thus **exhibiting strong sensitivity to data length**

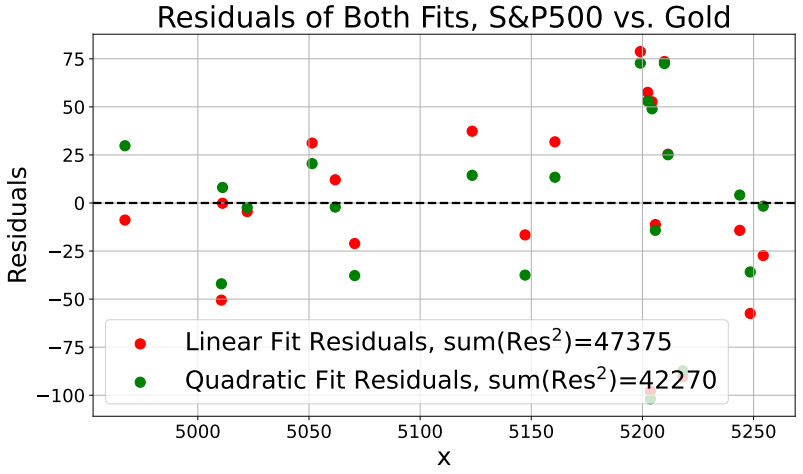
Need for Linear Models (regression models) (see Lecture 6)

These underpin many areas e.g. the CAPM and Fama-French models in finance









Linear Models

Generally, it is difficult to determine the MVU estimator.

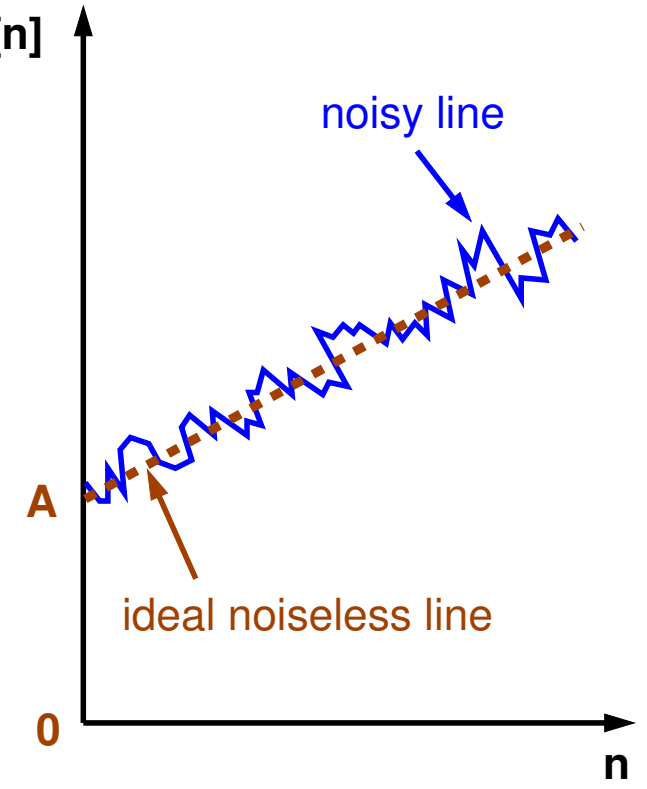
- In practice, however, a **linear data model** can often be employed ↬ straightforward to determine the MVU estimator.

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 the matrix notation as

known ↓ ↙ known pdf

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

where observed ↗ ↑ unknown

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

and

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

From a scalar to the vector/matrix notation

The "spelled out" form of the likelihood function for $\boldsymbol{\theta} = [A, B]^T$ is

$$p(\mathbf{x}; A, B) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(\underbrace{x[n] - A - Bn}_{w[n]}\right)^2} \qquad (*)$$

To arrive at the vector form, swap the variables as $w[n] = x[n] - A - Bn$.

Then, with $\mathbf{w} = [w(0), \ldots, w(N-1)]^T$, the term $\sum_{n=0}^{N-1} w^2(n)$, which appears in the above likelihood function can be written as

$$\sum_{n=0}^{N-1} w^2(n) = \mathbf{w}^T\mathbf{w}.$$

This applies to any vector, so that for $\mathbf{w} = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$ we have

This gives

$$\sum_{n=0}^{N-1} \underbrace{\left(x[n] - A - Bn\right)^2}_{w^2[n]} = \underbrace{\left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)^T}_{\mathbf{w}^T}\underbrace{\left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)}_{\mathbf{w}}$$

$$p(\mathbf{x}; A, B) = p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})}{2\sigma^2}} \qquad \text{equivalent to } (*)$$

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 then 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

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

Then, for the Linear Model

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

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

Linear models: Some useful matrix/vector derivatives

the derivations are given in Lecture Supplement

Use the identities (remember that both $\mathbf{b}^T\boldsymbol{\theta}$ and $\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}$ are scalars)

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

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

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

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

Then, the form of the partial derivative from the previous slide becomes

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

Linear models: Cramer-Rao lower bound

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

Similarly to the vector CRLB, $\hookrightarrow$ recall that $(\mathbf{H}^T\mathbf{H})^T = \mathbf{H}^T\mathbf{H}$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\frac{\partial^T}{\partial \boldsymbol{\theta}}\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H} \qquad \leftarrow \text{does not depend on data}$$

Therefore

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

By inspection, the **linear MVU estimator** is then given by

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

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

Back to Example 8

Start from $p(\mathbf{x};\boldsymbol{\theta}) = p(\mathbf{x};A,B) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n]-A-Bn\right)^2}$

To find the elements of the Fisher Information Matrix (FIM), start from

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

$$\frac{\partial^2\, ln\, p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2\, ln\, p(\mathbf{x};\boldsymbol{\theta})}{\partial A\partial B} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n \qquad \frac{\partial^2\, ln\, p(\mathbf{x};\boldsymbol{\theta})}{\partial B^2} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^2$$

Then, the Fisher Information Matrix, $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$, is given by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix} N & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} n^2 \end{bmatrix} = \begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix} \rightarrow \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) = \sigma^2\begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N-1)} \\ -\frac{6}{N(N-1)} & \frac{12}{N(N^2-1)} \end{bmatrix}$$

It now follows that the CRLB is

$$\text{var}(\hat{A}) = \frac{2(2N-1)}{N(N+1)}\sigma^2 \qquad \text{var}(\hat{B}) = \frac{12}{N(N^2-1)}\sigma^2$$

☞ B is easier to estimate as its CLRB is decreasing as $1/N^3$ as opposed to the $1/N$ dependence for the CRLB of A.

☞ CRLB always increases as we estimate more parameters (compare the CRLB of σ^2/N for the estimation of DC level A only, with the CRLB for the intercept A here).

Theorem: CRLB for linear models

- **We have seen that the MVU estimator for the linear model is efficient $\looparrowright$ it attains the CRLB**

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

Theorem: (Minimum Variance Unbiased Estimator for the Linear Model)

If the observed data can be modelled as

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

where

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

Theorem: CRLB for linear models (contd.)

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 has the form

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

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

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

End of Theorem □

Example 9: Fourier analysis as a linear estimator

Recall that we need to calculate $\hat{\theta} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$

The data model is given by $(n = 0, 1, \ldots, N-1,\ w[n] \sim \mathcal{N}\left(0, \sigma^2\right))$

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

where the Fourier coefficients, a_k and b_k, that is, the amplitudes of the cosine and sine terms, are to be estimated.

- Frequencies are multiples of the fundamental $f_1 = \frac{1}{N}$, that is, $f_k = \frac{k}{N}$.
- Then, the parameter vector is $\boldsymbol{\theta} = [a_1, a_2, \ldots, a_M, b_1, b_2, \ldots, b_M]^T$

and the observation matrix $\mathbf{H}$ is $N \times \underbrace{2M}_{p}$–dimensional, and takes the form

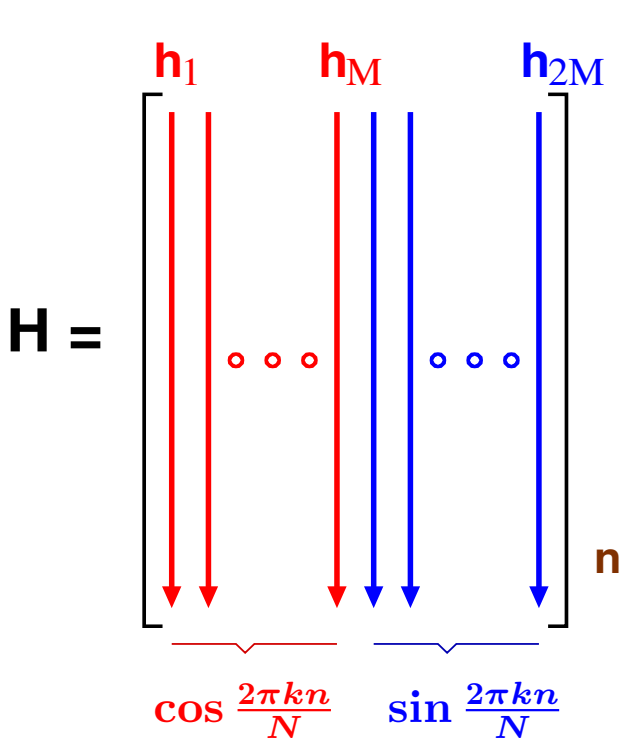
$$\mathbf{H} = \begin{bmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ \cos\frac{2\pi}{N} & \ldots & \cos\frac{2\pi M}{N} & \sin\frac{2\pi}{N} & \ldots & \sin\frac{2\pi M}{N} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos\frac{2\pi(N-1)}{N} & \ldots & \cos\frac{2\pi M(N-1)}{N} & \sin\frac{2\pi(N-1)}{N}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

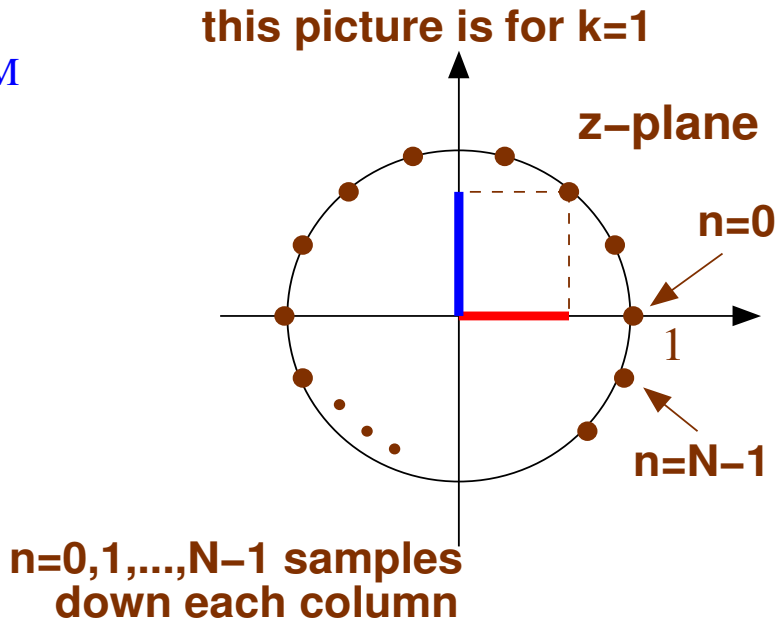
Example 9: Fourier analysis, geometric view

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

↑ parameters to estimate ↑ WGN ↑

Parameter vector: $\boldsymbol{\theta} = [a_1, a_1, \ldots, a_M, b_1, b_2, \ldots, b_M]^T$ (Fourier coeffs.)





Example 9: Fourier analysis ↬ continued

(see also Lecture 6)

For $\mathbf{H}$ not to be under-determined, it has to satisfy $N > p \quad \Rightarrow M \quad < \frac{N}{2}$

☞ For **mathematical convenience**, the columns of $\mathbf{H}$ should be **orthogonal**

This is because the **columns of $\mathbf{H}$ form a basis of a new representation space**, which is obvious if we rewrite the measurement matrix in the form

$$\mathbf{H} = \left[\mathbf{h}_1 \,|\, \mathbf{h}_2 \,|\, \cdots \,|\, \mathbf{h}_{2M} \right]$$

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

Then, for a large enough number of data points, N, due to the orthogonality properties of products of sines and cosines of different frequencies, we have

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

In other words, $\mathbf{h}_i \perp \mathbf{h}_j$, that is, the columns of matrix $\mathbf{H}$ are orthogonal

Example 9: Fourier analysis ↬ contd. contd.

The **orthogonality of the columns of $\mathbf{H}$** (for large N) follows from the properties of sines and cosines of different frequencies:

$$\sum_{n=0}^{N-1} \cos\left(\frac{2\pi in}{N}\right) \cos\left(\frac{2\pi jn}{N}\right) = \frac{N}{2}\delta_{ij} \qquad \text{(as power of a sinusoid is } A^2/2)$$

$$\sum_{n=0}^{N-1} \sin\left(\frac{2\pi in}{N}\right) \sin\left(\frac{2\pi jn}{N}\right) = \frac{N}{2}\delta_{ij} \qquad \text{(as power of a sinusoid is } A^2/2)$$

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

where the Kronecker delta

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

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 $\rightarrow$ measurement matrix

Therefore (orthogonality)

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \,|\, \cdots \,|\, \mathbf{h}_{2M} \end{bmatrix} = \begin{bmatrix} \frac{N}{2} & 0 & \cdots & 0 \\ 0 & \frac{N}{2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{N}{2} \end{bmatrix} = \frac{N}{2}\mathbf{I} \quad \rightarrow \quad (\mathbf{H}^T\mathbf{H})^{-1} = \frac{2}{N}\mathbf{I}$$

and the MVU estimator of the Fourier coefficients is given by

$$\hat{\boldsymbol{\theta}} = \underbrace{\left(\mathbf{H}^T\mathbf{H}\right)^{-1}}_{=\frac{2}{N}\mathbf{I}}\mathbf{H}^T\mathbf{x} \qquad \Rightarrow \qquad \hat{\boldsymbol{\theta}}_{MVU} = \frac{2}{N}\mathbf{H}^T\mathbf{x}$$

$$\hat{\boldsymbol{\theta}} = \frac{2}{N}\mathbf{H}^T\mathbf{x} = \frac{2}{N}\begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{2M}^T \end{bmatrix}\mathbf{x} = \begin{bmatrix} \frac{2}{N}\mathbf{h}_1^T\mathbf{x} \\ \vdots \\ \frac{2}{N}\mathbf{h}_{2M}^T\mathbf{x} \end{bmatrix} = \begin{bmatrix} \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right]\cos\left(\frac{2\pi\times 1\times n}{N}\right) \\ \vdots \\ \frac{2}{N}\sum_{n=0}^{N-1} x\left[n\right]\sin\left(\frac{2\pi\times 2M\times n}{N}\right) \end{bmatrix}$$

☞ Fourier coefficients of a "signal + WGN" are MVU estimates of the Fourier coefficients of the noise–free signal.

Example 9: Finally ↬ Fourier coefficients (Fourier coefficients of "signal + AWGN" are MVU estimates of Fourier coeff. of noise–free signal)

Therefore, the Fourier analysis represents a linear MVU estimator, given by

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

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

where the a_k and b_k are the **discrete Fourier transform coefficients.**

From CRLB for Linear Model, the covariance matrix of this estimator is

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

↖ decreases with N

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

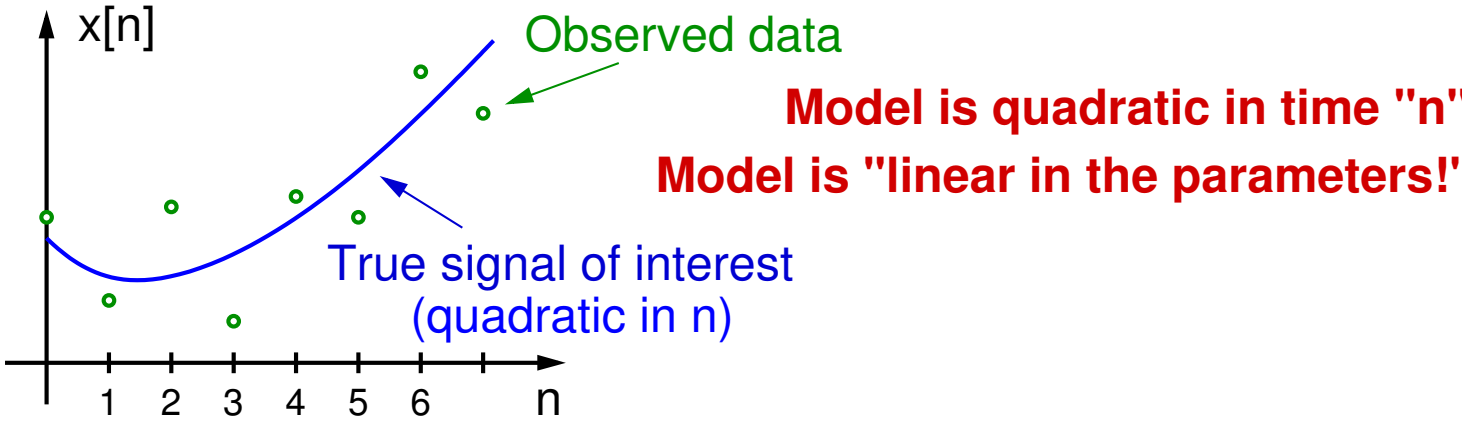
ii) The orthogonality of the columns of $\mathbf{H}$ is fundamental in the computation of the MVU estimator (invertible parsimonious basis);

iii) For accuracy, the measurement matrix $\mathbf{H}$ is desired to be a **tall matrix with orthogonal columns.**

Example 10: The concept of "linear in the parameters" models (e.g. like neural networks)

(see also Lecture 8)

☞ Recall that the notion "linear" in the term "Linear Models" **does not arise from fitting straight lines to data!**



Observations: $x[n] = \underbrace{\theta_0 + \theta_1 n + \theta_2 n^2}_{\text{linear in parameters } \theta} + w[n] \quad \Rightarrow \quad \mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$

where $\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$ $\mathbf{H} = \begin{bmatrix} 1 & 0 & 0^2 \\ 1 & 1 & 1^2 \\ 1 & 2 & 2^2 \\ \vdots & \vdots & \vdots \\ 1 & N-1 & (N-1)^2 \end{bmatrix}$

The need for a General Linear Model (GLM)

We shall now consider a general case where:

1) the observed signal may contain a known but non-white component, $\mathbf{s}$

2) the observation noise, $\mathbf{w}$, may be non-white, that is, $\mathbf{C} \neq \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Case 1) Often in practical applications (e.g. in radar), the observed signal consists of some known signal, $\mathbf{s}$, and another signal whose components are not known, $\mathbf{H}\boldsymbol{\theta}$, so that the linear model of the observed signal becomes

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{s} + \mathbf{w} \qquad \text{(here, noise is assumed to be white)}$$

The MVU estimator is determined immediately from $\mathbf{x}' = \mathbf{x} - \mathbf{s}$, so that

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

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

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

An example may be a DC level observed in random white noise, but also with a known sinusoidal interference (e.g. from the mains).

Incorporating coloured (correlated) noise into GLM

Case 2) For coloured noise, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C} \neq \sigma^2\mathbf{I}$, that is, $\mathbf{C}$ is not a scaled identity matrix!

To this end, we can use a **whitening** approach as follows:

Since $\mathbf{C}$ is +ve semidefinite, so too is $\mathbf{C}^{-1}$, $\looparrowright$ it can be factored as

$$\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}, \qquad \mathbf{D}_{N\times N} \text{ is invertible}$$

Now, $\mathbf{D}$ acts as a **whitening transform** when applied to $\mathbf{w}$, since

$$E\big[(\mathbf{Dw})(\mathbf{Dw})^T\big] = E\big[\mathbf{Dww}^T\mathbf{D}^T\big] = \mathbf{DCD}^T = \mathbf{DD}^{-1}\mathbf{D}^{T^{-1}}\mathbf{D}^T = \mathbf{I}$$

☞ This allows us to transform the general linear model

from $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ to $\mathbf{x}' = \mathbf{Dx} = \mathbf{DH}\boldsymbol{\theta} + \mathbf{Dw} = \mathbf{H}'\boldsymbol{\theta} + \mathbf{w}'$

☞ The noise is now whitened, as $\mathbf{w}' = \mathbf{Dw} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ $\looparrowright$ use Linear Model

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}'^T\mathbf{H}'\right)^{-1}\mathbf{H}'^T\mathbf{x}' = \left(\mathbf{H}^T\mathbf{D}^T\mathbf{DH}\right)^{-1}\mathbf{H}^T\mathbf{D}^T\mathbf{Dx} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

In a similar fashion, for the variance of this estimator we have

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}'^T\mathbf{H}'\right)^{-1} \qquad \text{and finally} \qquad \mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}$$

☞ For $\mathbf{C} = \sigma^2\,\mathbf{I}$ we have our previous results for standard Linear Estimator

Theorem: MVU Estimator for the General Linear Model

Upon combining Case 1 and Case 2 above (non-white noise + known component)

i) General linear data model: $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta} + \mathbf{s}$

known ↓ ↓ some known signal

$$\mathbf{x} = \mathbf{H}\,\boldsymbol{\theta} \;+\; \mathbf{s} \;+\; \mathbf{w} \qquad\qquad \mathbf{w} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{C}\big)$$

observed ↑ ↑ unknown ↑ known statistics, can be non-white

ii) Then, the MVU estimator has the form

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

with the covariance matrix

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

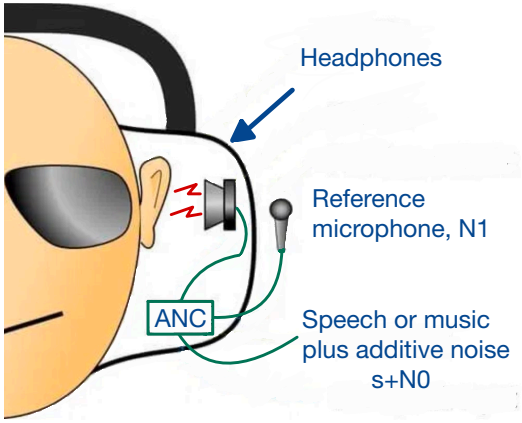
and attains the Cramer Rao Lower Bound (CRLB).

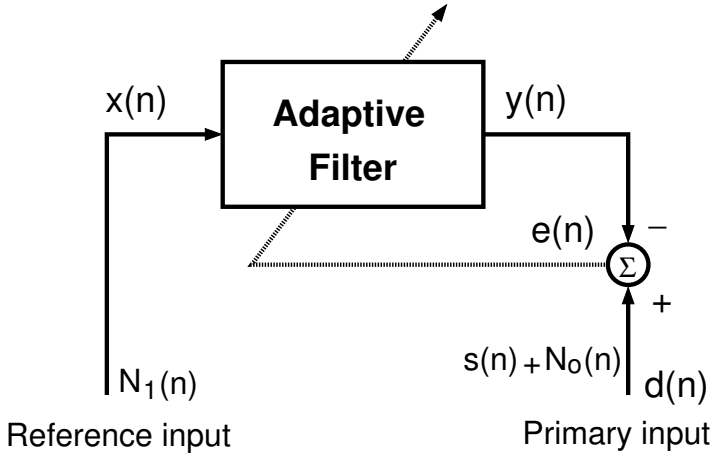
☞ We must assume that $\mathbf{H}$ is full rank, otherwise for any $\mathbf{s}$ there exist some $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ which both give $\mathbf{s}$, that is, $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}_1 = \mathbf{H}\boldsymbol{\theta}_2$ (no uniqueness)

Example 11: Adaptive noise cancellation with reference

e.g. noise-cancelling headphones **(see Example 11 in Appendix & Lecture 7)**

Physical intuition behind "MVU estimator is a ratio of the input-output cross-correlation to the input autocorrelation" on a noise cancel. example.





Input: The cockpit noise, $x(n) = N_1(n)$, that is, the noise for Reference Microphone. The only requirement is that N_1 is correlated with the noise, N_0, which you hear through the headphones, but not with the music signal, $s(n)$. The filter aims to estimate N_0 from N_1, that is, $y = \hat{N}_0$.

☞ Based on the input-output cross-correlation, the filter output can only produce and estimate of the noise you hear, that is, $y(n) = \hat{N}_0(n)$, as cockpit noise, N_1, is not correlated with the music, s.

Therefore we hear $d(n) - y(n) = s(n) + N_0(n) - \hat{N}_0(n) \approx s(n)$

What 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 should be abandoned $\looparrowright$ we need to consider alternative approaches
- **Remedy:** Constrain the bias to zero and find an estimator which minimises the variance $\looparrowright$ 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 $\looparrowright$ **this makes it easier to perform the analysis**

Things to remember about CRLB

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 $\looparrowright$ 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 infeasible 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 in the examples on the estimation of Fourier coefficients and a quadratic curve in noise

Some "rule of thumb" practical hints with CRLB

1. Start from the log–likelihood parametrised PDF function, which depends on the unknown parameter θ, that is, $\ln p(\mathbf{x};\theta)$
2. Fix $\mathbf{x}$ and take 2nd partial derivative of the log-likelihood function, that is, $\partial^2 \ln p(\mathbf{x};\theta)/\partial\theta^2$
3. If the result still depends on $\mathbf{x}$, then fix the θ and take the expected value with respect to $\mathbf{x}$. Otherwise, this step is not needed.
4. Should the result still depend on θ, then evaluate at every specific value of θ
5. For the CRLB, perform the reciprocal and negate

☞ **Transformation of parameters:** If we know the CRLB for θ, we can easily obtain it for any function of θ, e.g. $\alpha = g(\theta)$. (see Appendix 3)

For some problems, an efficient estimator may not exist, for example the estimation of sinusoidal phase (see your P& A sets)

Appendix 1: An alternative form of CRLB (via the sensitivity of $p(\mathbf{x};\theta)$ to θ)

Sometimes, it is easier to find CRLB as

$$var(\hat{\theta}) \geq \frac{1}{E\left\{\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right]^2\right\}} \qquad cf. \text{ the original } \quad var(\hat{\theta}) \geq \frac{1}{-E\left\{\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right\}}$$

Motivation: Sensitivity analysis, ease of interpretation

For an increment in θ, i.e. $\theta \rightarrow \theta + \Delta\theta \quad \Rightarrow \quad p(\mathbf{x};\theta) \rightarrow p(\mathbf{x};\theta+\Delta\theta)$

Then, the sensitivity of $p(\mathbf{x};\theta)$ to that change is

$$\tilde{S}_{\theta}^{p}(\mathbf{x}) = \frac{\left[\frac{\Delta p(\mathbf{x};\theta)}{p(\mathbf{x};\theta)}\right]}{\left[\frac{\Delta\theta}{\theta}\right]} = \frac{\%\text{ change in } \; p(\mathbf{x};\theta)}{\%\text{ change in } \; \theta} = \left[\frac{\Delta p(\mathbf{x};\theta)}{\Delta\theta}\right]\left[\frac{\theta}{p(\mathbf{x};\theta)}\right]$$

$$\text{For } \Delta\theta \rightarrow 0 \quad S_{\theta}^{p}(\mathbf{x}) = \lim_{\Delta\theta\rightarrow 0} \tilde{S}_{\theta}^{p}(\mathbf{x}) = \left[\frac{\partial p(\mathbf{x};\theta)}{\partial\theta}\right]\left[\frac{\theta}{p(\mathbf{x};\theta)}\right] = \theta\frac{\partial\, ln\, p(\mathbf{x}'\theta)}{\partial\theta}$$

(recall the derivative rules of a $\log$ function, $\frac{\partial\, ln\, f(x)}{\partial x} = \frac{1}{f(x)}\frac{\partial f(x)}{\partial x}$)

Appendix 1: An alternative form of CRLB (contd.) (via the sensitivity of $p(\mathbf{x};\theta)$ to θ)

Therefore (Gardner, IEEE Transactions on Information Theory, July 1979)

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

Interpretation: This is an inverse mean square sensitivity of $p(\mathbf{x};\theta)$ to θ.

- Modelling and estimation are obviously intertwined
- Unknown parameters may have a physical interpretation, such as e.g. direction in beamforming, delay in radar, ...
- Otherwise, parameters may be part of an imposed model, such as e.g. the fixed sine-cosine bases in Fourier analysis

Appendix 2: The validity of Gaussian assumption

(The Gaussian data assumption leads to the largest Cramer-Rao bound)

- When there is no information about the distribution of observations, Gaussian assumption appears as the most conservative choice
- This follows from the fact that the Gaussian distribution minimises the Fisher information (inverse of the CRLB), or in other words **the Gaussian distribution maximises the CRLB**
- Indeed, it leads to the largest CRLB in quite a general class of data distributions and for a significant set of parameter estimation problems
- Therefore, any optimisation based on the CRLB under the Gaussian assumption is min-max optimal in the sense of minimising the largest CRLB (they yield the best CRB-related performance in the worst case, and over a large class of data distributions which satisfy the regularity condition)
- Also, the Gaussian random vector maximises a differential entropy, and also the worst additive noise lemma

For more detail see: S. Park, E. Serpedin, and K. Qaraqe, "Gaussian assumption: The least favourable but the most useful", *IEEE Signal Processing Magazine*, May 2013, pp. 183–186 and the references therein

Appendix 3: Transformation of parameters

Suppose that there is a parameter θ for which we know the CRLB, denoted by $CRLB_\theta$.

Our task is the estimate another parameter α which is a function of θ, i.e.

$$\alpha = g(\theta)$$

Then, it can be shown that (see S. Kay's book on Statistical Signal Processing)

$$var(\alpha) \geq CRLB_\alpha = \left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 CRLB_\theta$$

↖ sensitivity of α to θ

☞ Therefore, a large sensitivity $\frac{\partial g(\theta)}{\partial \theta}$ means that a small error in θ gives a large error in α. This, in turn, increases the CRLB (that is, worsens accuracy).

It can be shown that if $g(\theta)$ has an affine form, that is, $g(\theta) = a\theta + b$, then $\hat{\alpha} = g(\hat{\theta})$ is efficient.

Otherwise, for any other form of $g(\theta)$, the result is asymptotically efficient for $N \to \infty$.

Appendix 4: Modelling vs. Estimation

◦ Oftentimes parameters we wish to estimate have some physical significance (e.g. heart rate, or delay in the time of arrival of the back-scattered signal in radar).

◦ It is also common that the parameters of interest arise from a non-physical model which is imposed onto data (e.g. Fourier analysis).

◦ However, even then, the Fourier coefficients for a signal in AWGN are the MVU estimates of the Fourier coefficients in the noise-free case!

◦ Similar reasoning applies to ARMA modelling, the coefficients may or may not have physical meaning.

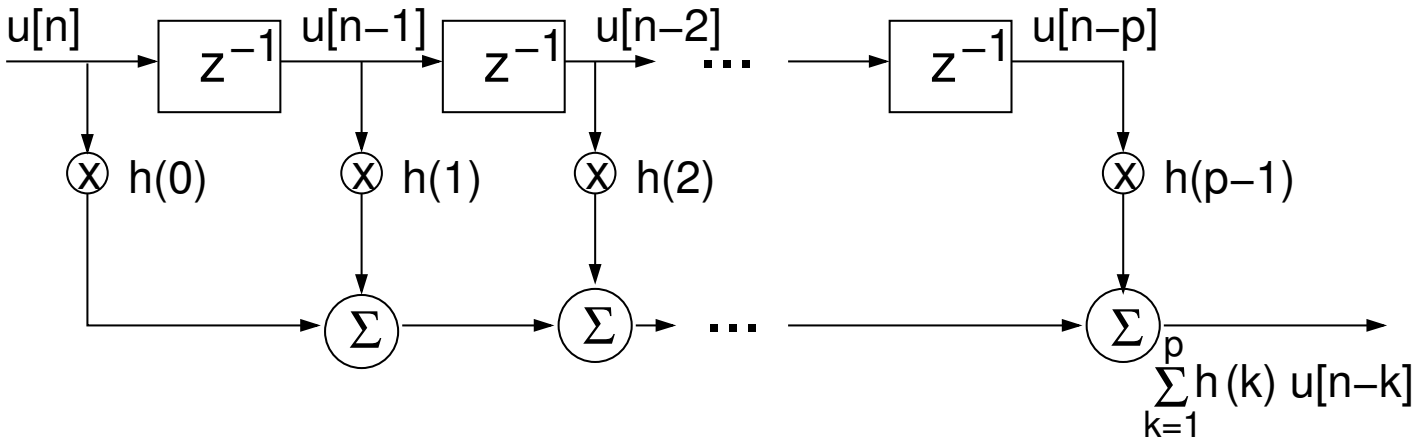
◦ Model $\not\mapsto$ related to data generation (e.g. a generative model)

◦ Estimation $\not\mapsto$ related to both model accuracy (bias/variance) and when using a model to e.g. future values of a signal (inference).

☞ **Modelling and Estimation/Inference are intertwined. It is our goal to understand the bounds on the best achievable performance for a certain paradigm, and use this as a domain knowledge for inference.**

App. 5 ↬ Example 11: System Identification (SYS ID)

Aim: To identify the model of a system (filter coefficients $\{h\}$) from input/output data. Assume an FIR filter system model given below



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

App. 5 ↬ Example 11: SYS ID ↬ data model in noise $w \sim \mathcal{N}(0, \sigma^2)$

Data model

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

The equivalent matrix–vector form is

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

that is

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

Then, 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

App. 5 ↬ Example 11: SYS ID ↬ more about $\mathbf{H}$

Now, $\mathbf{H}^T\mathbf{H}$ becomes a symmetric Toeplitz autocorrelation matrix, given by

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

where

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

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

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

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

App. 5 ↬ Example 11: SYS ID ↬ 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

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

and

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

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

☞ Thus, the MVU estimator is a ratio of the input-output cross-correlation to the input autocorrelation (makes perfect physical sense).

☞ **Compare with the Wiener filter in Lecture 7 (adaptive inference)**

Notes:

Notes:

Notes: