Advanced Signal Processing
BLUE and ML Estimators

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

Overview

- It frequently occurs that the MVU estimator, even if it exists, cannot be found (mathematical tractability, violation of regularity conditions, ...)
- For instance, one typical case is that we may not know the pdf of the data, but we do know the 1st and 2nd moment (mean, variance, power) **In such cases CRLB and pdf based methods cannot be applied**
- We therefore have to resort to suboptimal solutions: impose some constraints on the estimator and data model
- If the variance of our suboptimal estimator meets our system specifications, the use of such estimators may be justified
- One common approach is to restrict the estimator to be **linear in the data**, and find a **linear estimator** that is **unbiased and has minimum variance** – best linear unbiased estimator (BLUE)
- Alternatively if the MVU estimator does not exist, we may resort to **Maximum Likelihood Estimation (MLE)**
- We first need to look at which data samples are pertinent to the estimation problem – the so called **sufficient statistics**

I'm feeling BLUE (Best Linear Unbiased Estimator): Sufficient statistics and the linearity assumption

- If an **efficient estimator does not exist**, it is still of interest to **be able to find** the MVU estimator (assuming of course that it exists)
- To achieve this, we need the concept of **sufficient statistics** and the Rao–Blackwell–Lehmann–Scheffe theorem
- Using these theories makes possible in many cases to determine the MVU estimator **by a simple inspection of the PDF**

The BLUE assumptions are also referred to as **Gauss–Markov assumptions**

They were responsible for many advances in "quantitative methodologies"

An insight into the 'sufficiency' of the data statistics

which data samples are pertinent to the est. problem? Q: ∃ a sufficient dataset?

Consider the two estimators of DC level in WGN that we addressed so far:

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n], \quad \text{var}(\hat{A}) = \frac{\sigma^2}{N} \qquad \& \qquad \tilde{A} = x[0], \quad \text{var}(\tilde{A}) = \sigma^2$$

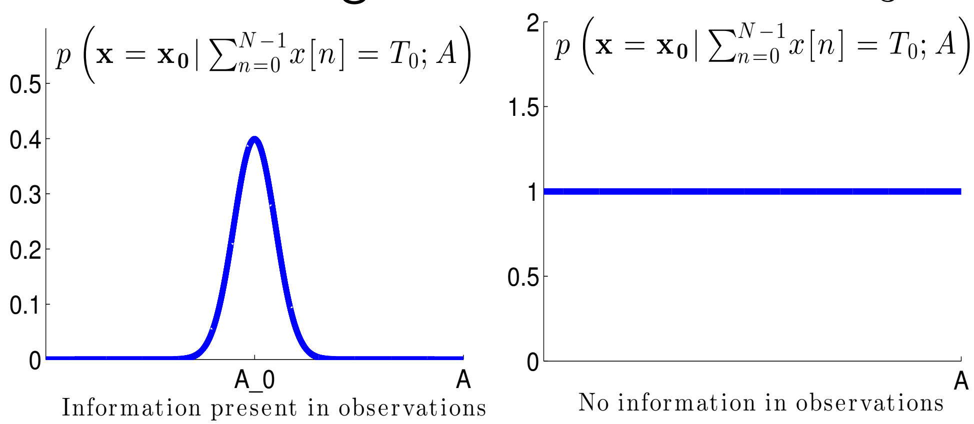
↬ Although $\tilde{A}$ is **unbiased**, its variance is much larger than the minimum. This is is due to discarding x[1],...,x[N-1] that carry information about A.

Consider now the following datasets:

$$S_1 = \{x[0], x[1], \ldots, x[N-1]\} \quad S_2 = \{x[0]+x[1], x[2], \ldots, x[N-1]\} \quad S_3 = \{\sum_{n=0}^{N-1} x[n]\}$$

The original dataset, S_1, is **always sufficient**, S_2 and S_3 are also sufficient.

⇝ In addition to being sufficient, statistics S_3 is **minimal sufficient statistics** ☺



∘ Knowledge of T_0 changes the PDF to the conditional one $p(\mathbf{x}|\sum_{n=0}^{N-1} x[n] = T_0; A)$

∘ **If the statistics is sufficient for estimating A, this condit. PDF should not depend on A** (right f.)

Sufficient statistics, for $x[n] = A + w[n],\ w \sim \mathcal{N}(0, \sigma^2)$

Split the pdf into the "data-only" and "parameter & data" parts

Sufficient statistics answers the questions:

Q1: Can we find a transformation $T(\mathbf{x})$ of lower dimension that **contains all information** about θ (the data can be very long, e.g. $\mathbf{x} \in \mathbb{R}^{N\times 1}$)

Q2: What is the lowest possible dimension of $T(\mathbf{x})$ so as to still contain all information about θ $\rightsquigarrow$ **minimal sufficient statistics**

For example, for DC level in WGN, $T(\mathbf{x}) = \sum x[n]$ **(one-dimensional)**

Solution: **Neyman-Fisher factorisation** th. which allows us to factor a pdf as

$$p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)h(\mathbf{x})$$

Then $T(\mathbf{x})$ is a sufficient statistics and the pdf can be factorised as above.

☞ For a DC level in WGN, $p(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\}$, so

$$p(\mathbf{x};A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \underbrace{\exp\Big\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\Big\}}_{h(\mathbf{x})} \underbrace{\exp\Big\{-\frac{1}{2\sigma^2}\Big[NA^2 - 2A\,\Big(\overbrace{\sum_{n=0}^{N-1}x[n]}^{T(\mathbf{x})}\Big)\Big]\Big\}}_{g(T(\mathbf{x}),A)}$$

Therefore, sufficient statistics $T(\mathbf{x}) = \sum_{n=0}^{N-1}x[n]$ (minimal & linear)

How to find the MVU from sufficient statistics?

Raw data $\mathbf{x} = [x[0], \ldots, x[N-1]]^T \in \mathbb{R}^{N\times 1} \hookrightarrow N$-dim. sufficient statistics

∘ For $T(\mathbf{x})$ to be sufficient statitics, we need $p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)h(\mathbf{x})$

How to find the MVU?

1. find **any unbiased estimator** $\bar{\theta}$ of θ and determine

$$\hat{\theta} = E[\bar{\theta}|T(\mathbf{x})] = g(\mathbf{x})$$

   (mathematically intractable)

2. find a function $\hat{\theta} = g(T(\mathbf{x}))$ s.t. $E[\hat{\theta}] = \theta$
3. if $g(\cdot)$ is unique: we have complete statistics and MVU
4. if $g(\cdot)$ is not unique: there is no MVU

How to check if $g(\cdot)$ is unique?

Rao-Blackwell-Lehmann-Scheffe

Assume that $\bar{\theta}$ is an unbiased estimator of θ and $T(\mathbf{x})$ is sufficient statistics for θ.

Then the estimator $E[\bar{\theta}|T(\mathbf{x})]$ is:

- valid (not dependent on θ)
- unbiased
- of $\leq$ variance than that of $\bar{\theta}$
- if the sufficient statistics is **complete** then it is MVU

Complete: only one function $g(T(\mathbf{x}))$ s.t. $E[g(T(\mathbf{x})] = \theta$

Best Linear Unbiased Estimator: BLUE

Motivation: When the PDF of the data is **unknown**, or **cannot be assessed**, the MVU estimator, even if it exists, cannot be found!

- In this case methods which rely on the CRLB cannot be applied ☹

Remedy: Resort to a sub-optimal estimator - check its variance and ascertain whether this meets the required specifications

Common sense approach: assume the estimator to be:

- **linear in the data**,
- **unbiased**,
- with **minimum variance**.

$\rightsquigarrow$ This estimator is termed the Best Linear Unbiased Estimator (BLUE) which **requires only knowledge of the first two moments of the PD**.

We will see that if the data are Gaussian, the BLUE and MVUEs are equivalent

The form and optimality of BLUE

Consider the data $\mathbf{x} = \left[x[0], x[1], \ldots, x[N-1]\right]^T$, whose *pdf* $p(\mathbf{x};\theta)$ depends on the unknown parameter θ.

The form of BLUE

The BLUE estimator is restricted to have the form ($\mathbf{a} = \{a_n\}$)

$$\hat{\theta} = \sum_{n=0}^{N-1} \underset{\uparrow}{a_n} x[n] = \mathbf{a}^T \mathbf{x}$$

Constants to be determined

Choice of $a_n s$ determines the nature of the estimator.

The BLUE estimator is the one which is **unbiased** and has **minimum variance.**

Optimality of BLUE

Note, the BLUE **will be optimal only when the actual MVU estimator is linear!**

For instance, when estimating the DC level in WGN

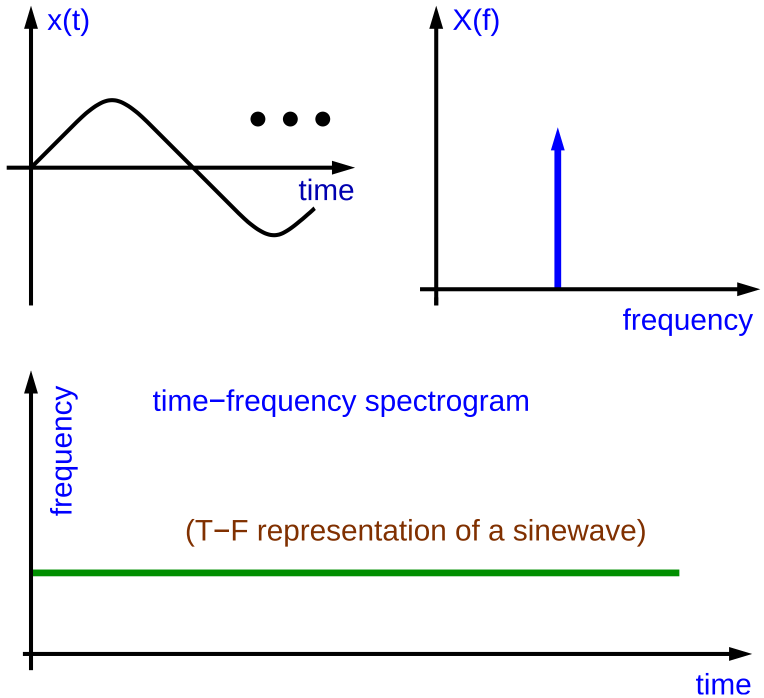
$$\hat{\theta} = \bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \qquad \left\{a_n = \frac{1}{N}\right\}$$

which is clearly **linear in the data, BLUE is an optimal MVU giving $a_n = 1/N$.**

Example 1: How useful is an estimator of DC level in noise?

In fact, very useful. It is up to us to provide correct data representation.

Sinusoidal frequency estimation

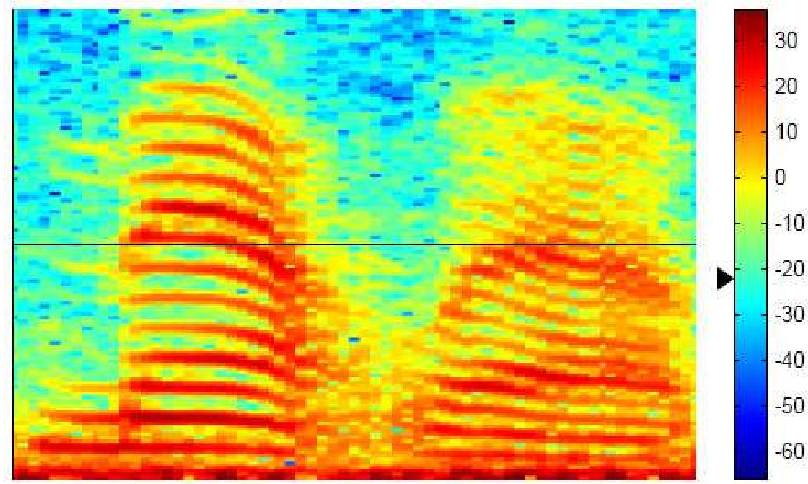


∘ Ramp in time ↬ DC level in time (via differentiation)

∘ Chirp in time ↬ ramp in T-F

Transforming other problems

time-frequency representation



horizontal: time **vertical: frequency**

This is a T-F representation of a waveform of word "matlab"

DC-level like harmonics for "a"

Example 2: Composite faces $\looparrowright$ people face averages

Can we estimate a "typical looking" person from a certain region, by taking a statistical average of a large ensemble of random faces photographed on the street?

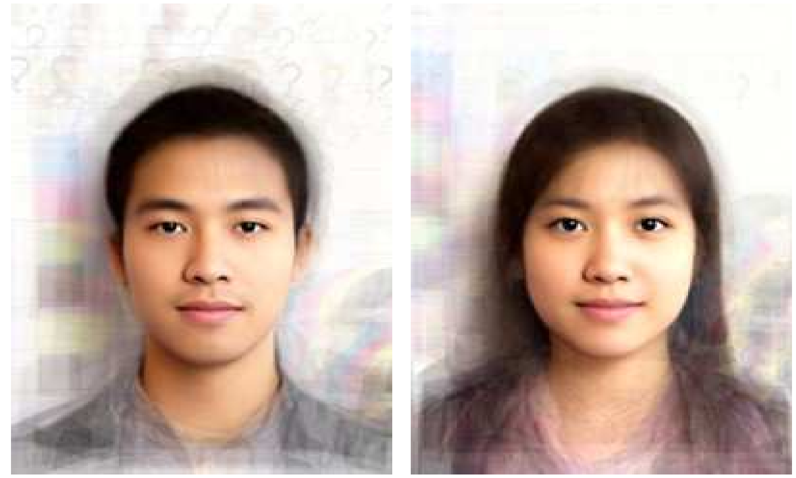
Does such an estimated face exist in real life?

Participans in Sydney, Australia, ranging from 0.83–93 years

Example 2: contd. $\looparrowright$ composite male and female faces

Composite faces of Sydney

Composite faces of London

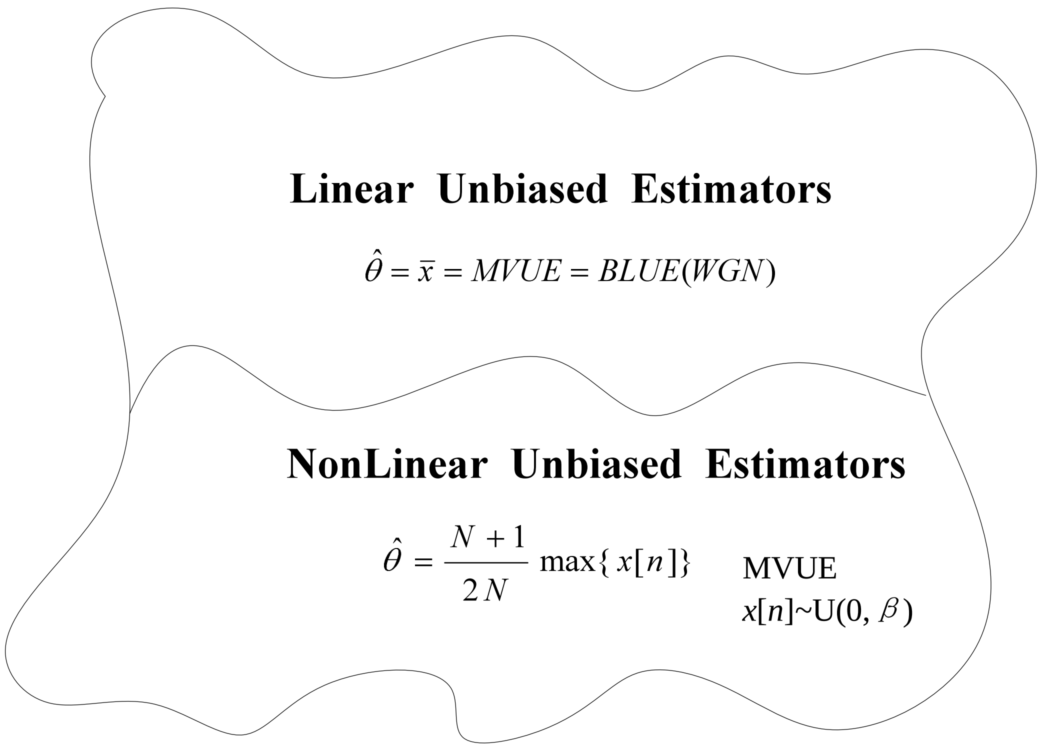


Composite faces of Hong Kong

Composite faces of Argentina

The place of BLUE amongst other estimators (e.g. for DC level in noise)

Consider the space of all unbiased estimators



- For white Gaussian noise the MVU is **linear in the data** and is given by the sample mean $\bar{x}$
- For the **uniform** noise $x[n] \sim \mathcal{U}(0, \beta)$, the MVU is **nonlinear in the data**, and is given by

$$\hat{\theta} = \frac{N+1}{2N} \max\{x[n]\}$$

$$\mathrm{var}(\hat{\theta}) = \frac{\beta^2}{12N}$$

The difference in performance between the BLUE and MVU estimators can be substantial, but can only be quantified through knowledge of the data pdf.

Example 3: Problems with BLUE. Inapropriate for nonliner problems $\looparrowright$ population dynamics example

Owing to the **linearity assumptions**, the BLUE estimator can be totally inappropriate for some estimation problems.

Power of WGN estimation

The MVU estimator $\hat{\sigma^2} = \frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$

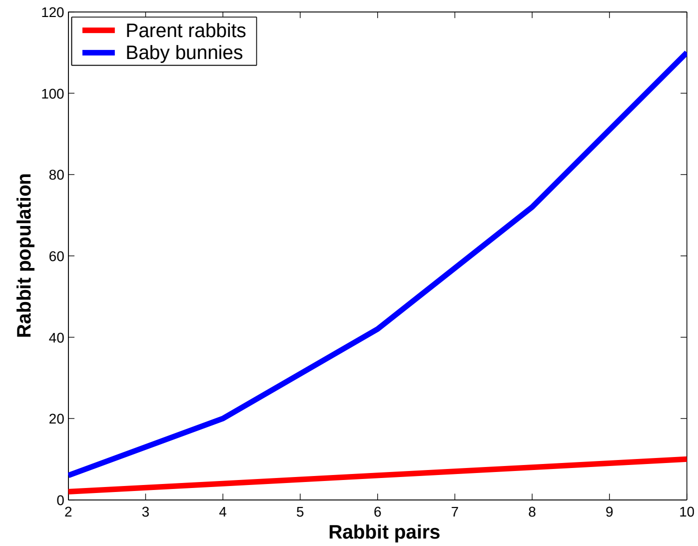
is **nonlinear** in the data. Forcing the estimator to be linear, e.g.

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=0}^{N-1} a_n x[n]$$

yields $E\{\hat{\sigma^2}\} = 0$, which is guaranteed to be biased!

A non-linear transformation of the data, i.e. $y[n] = x^2[n]$, could overcome this problem.

Example: Rabbit population



The time evolution of the rabbit population is nonlinear (exponential)

However, the number of parent pairs is linear in time $\Rightarrow$ BLUE

How to find BLUE?
Recall: BLUE is linear in data $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T \mathbf{x}$

To determine the BLUE we impose a **constraint** on $\hat{\theta}$ to be **linear and unbiased**, and find the coefficients $\{a_n\}$ which minimise the variance.

1. Unbiased constraint

$$E\{\hat{\theta}\} = \sum_{n=0}^{N-1} a_n E\{x[n]\} = \theta$$

$$\Rightarrow \quad \mathbf{a}^T \mathbf{s} = 1$$

where the **scaled data vector** (by inspection) $\mathbf{s} = [s[0]\ s[1], \ldots, s[N-1]]^T$.

In other words, to satisfy the unbiased constraint for the estimate $\hat{\theta}$, $E\{x[n]\}$ must be linear in θ, or

$$E\{x[n]\} = s[n]\theta$$

2. Variance minimisation

$\hat{\theta} = \mathbf{a}^T \mathbf{x} \Rightarrow \mathrm{var}(\hat{\theta}) = E\{\mathbf{a}^T \mathbf{x}\mathbf{x}^T \mathbf{a}\}$

BLUE optimisation task. Minimise

$$\mathrm{var}(\hat{\theta}) = \mathbf{a}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{a} = \mathbf{a}^T \mathbf{C} \mathbf{a}$$

subject to the **unbiased constraint**

$$\sum_{n=0}^{N-1} a_n E\{x[n]\} = \theta \quad \Leftrightarrow \quad \mathbf{a}^T \mathbf{s}\theta = \theta$$

This is satisfied for

$$\sum_{n=0}^{N-1} a_n s[n] = 1 \quad \Leftrightarrow \quad \mathbf{a}^T \mathbf{s} = 1$$

Some remarks on variance calculation

A closer look at the variance yields

$$Var(\hat{\theta}) = E\left\{\Big(\sum_{n=0}^{N-1} a_n x[n] - E\{\sum_{n=0}^{N-1} a_n x[n]\}\Big)^2\right\} = E\left\{\left(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E\{\mathbf{x}\}\right)^2\right\}$$

With $\boldsymbol{a} \equiv \ [a_0, a_1, \ldots, a_{N-1}]^T$, $y^2 = y \times y^T$, and $(\mathbf{a}^T\mathbf{x})^T = \mathbf{x}^T\mathbf{a}$, we have

$$E\left\{\mathbf{a}^T\Big(\mathbf{x} - E\{\mathbf{x}\}\Big)\Big(\mathbf{x} - E\{\mathbf{x}\}\Big)^T\mathbf{a}\right\} \quad = \quad \mathbf{a}^T\mathbf{C}_x\mathbf{a}$$

Also assume

$$E\{x[n]\} = s[n]\theta, \quad \text{easy to show from} \quad x[n] = E\{x[n]\} + \big[x[n] - E\{x[n]\}\big]$$

$$\text{by viewing} \;\; w[n] = x[n] - E\{x[n]\}, \quad \text{we have} \;\; x[n] = \theta s[n] + w[n]$$

$\Rightarrow$ **BLUE is linear in the unknown parameter** θ, which corresponds to the amplitude estimation of known signals in noise (to generalise this, a nonlinear transformation of the data is required).

BLUE as a constrained optimisation paradigm

Also see Lecture 1 and Appendix here

Task: minimize the variance subject to the unbiased constraint

$$\underbrace{\min \{\mathbf{a}^T \mathbf{C}_x \mathbf{a}\}}_{optimisation \;\; task} \qquad \text{subject to} \qquad \underbrace{\mathbf{a}^T \mathbf{s} = 1}_{equality \;\; constraint}$$

Method of Lagrange multipliers

1. $$J = \mathbf{a}^T \mathbf{C}\mathbf{a} - \lambda(\mathbf{a}^T \mathbf{s} - 1)$$

2. Calculate
$$\frac{\partial J}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} - \lambda \mathbf{s}$$

3. Equate to zero and solve for $\mathbf{a}$
$$\mathbf{a} = \frac{\lambda}{2}\mathbf{C}^{-1}\mathbf{s}$$

Solve for the Lagrange multiplier λ

4. From the constraint equation
$$\mathbf{a}^T \mathbf{s} = \frac{\lambda}{2}\mathbf{s}^T \mathbf{C}^{-1}\mathbf{s} = 1$$
$$\Rightarrow \quad \frac{\lambda}{2} = \frac{1}{\mathbf{s}^T \mathbf{C}^{-1}\mathbf{s}}$$

5. Replace into Step 3, with the constraint satisfied for
$$\mathbf{a}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1}\mathbf{s}}$$

coefficients of BLUE!

Summary: BLUE

BLUE of an unknown parameter (our function g(x) from MVU):

$$\hat{\theta} = \frac{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} = \mathbf{a}_{opt}^T \mathbf{x} \qquad \text{where} \quad \mathbf{a}_{opt} = \frac{\mathbf{C}^{-1} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$$

BLUE variance:

$$\mathrm{Var}\left(\hat{\theta}\right) = \mathbf{a}_{opt}^T \ \mathbf{C} \ \mathbf{a}_{opt} = \frac{1}{\mathbf{s}^T \ \mathbf{C}^{-1} \ \mathbf{s}}$$

To determine the BLUE we only require knowledge of

$\mathbf{s}$ – the scaled mean

$\mathbf{C}$ – the covariance ($\mathbf{C}^{-1}$ prewhitens the data prior to averaging)

That is, for BLUE we only need to know the first two moments of the PDF

Notice that we do not need to know the functional relation of PDF ☺

Example 4: Estimation of a DC level in unknown noise

Notice that the PDF is unspecified and does not need to be known

Example: Determine the DC level in White Noise of an unspecified *pdf*

Given

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

where $\{w[n]\}$ is **any white noise with variance σ^2 (power)**.

In other words, $\{w[n]\}$ is **not necessarily Gaussian or independent** $\Rightarrow$ there may be some statistical dependence between samples (although they are uncorrelated)

Task: Estimate A.

Solution:

Since

$$E\{x[n]\} = A \qquad \text{therefore} \quad s[n] = 1 \quad \text{and} \quad \mathbf{s} = \mathbf{1} = [\, \underbrace{1, \ldots, 1}_{\text{N elements}} \,]^T = \mathbf{1}_{N\times 1}$$

Follows from $E\{x[n]\}$ being linear in θ $\Rightarrow$ $E\{x[n]\} = s[n]\theta$.

Example 4: DC level in white noise with unknown PDF, contd.

For any white noise $\{w\}$ with power σ^2, $\quad \mathbf{C}_w = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$

$\Rightarrow \quad \mathbf{C}^{-1} = \frac{1}{\sigma^2}\mathbf{I}$

The BLUE becomes

$$\hat{A} = \frac{\mathbf{1}^T \frac{1}{\sigma^2} \mathbf{I} \mathbf{x}}{\mathbf{1}^T \frac{1}{\sigma^2} \mathbf{I} \mathbf{1}} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] = \bar{x}$$

and has minimum variance (CRLB for a linear estimator)

$$\mathrm{Var}(\hat{A}) = \frac{1}{\mathbf{1}^T \frac{1}{\sigma^2} \mathbf{I} \mathbf{1}} = \frac{\sigma^2}{N}$$

- The sample mean is the BLUE independent of the PDF of the data
- BLUE is the MVU estimator if the noise $\{w\}$ is Gaussian
- If the noise is not Gaussian (e.g. uniform) the CRLB and MVU estimator may not exits, but BLUE still exists

Some help with the expressions of the type $\mathbf{a}^T \mathbf{A} \mathbf{a}$

we shall consider the expressions $\mathbf{1}^T \mathbf{I} \mathbf{1}$ and $\mathbf{1}^T \mathbf{I}\, \mathbf{x}$

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{1\text{x}N} \underbrace{\begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}}_{N\text{x}N} \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{N\text{x}1} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{1\text{x}N} \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{N\text{x}1} = N$$

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{1\text{x}N} \underbrace{\begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}}_{N\text{x}N} \underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{N\text{x}1} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \end{bmatrix}}_{1\text{x}N} \underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{N\text{x}1} = \sum x[n]$$

It is useful to visualise any type of vector–matrix expression.

☞ It is now obvious that e.g. the scalar $\mathbf{a}^T \mathbf{A} \mathbf{a}$ is 'quadratic' in $\mathbf{a}$.

This is easily proven by considering $\mathbf{x}^T \mathbf{I}\, \mathbf{x}$ in the diagrams above.

Example 5: DC Level in uncorrelated zero mean noise with Var($w[n]$) $= \sigma_n^2$

Notice that now the noise variance depends on the sample number!

As before, $\mathrm{s} = 1$.

The covariance matrix of the noise

$$\mathbf{C} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}$$

and thus

$$\mathbf{C}^{-1} = \begin{bmatrix} \sigma_0^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_1^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^{-2} \end{bmatrix}$$

$\mathbf{C}^{-1}$ acts to prewhiten the data

The BLUE solution:

$$\hat{A} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{x}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} = \frac{\displaystyle\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\displaystyle\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$

- ∘ The term $\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}$ ensures that the estimator is unbiased
- ∘ BLUE weighs samples with smallest variances most heavily
- ∘ Notice that

$$var(\hat{A}) = \frac{1}{\sum_{n=0}^{N-1} 1/\sigma_n^2}$$

BLUE: Extension to vector parameter

System model: $\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n], i = 1, \ldots, p \Rightarrow \hat{\boldsymbol{\theta}} = \mathbf{A}\mathbf{x}$

Unbiased constraint:

$$E\{\hat{\theta}_i\} = \sum_{n=0}^{N-1} a_{in} E\{x[n]\} = \theta_i \quad \Rightarrow \quad E\{\hat{\boldsymbol{\theta}}\} = \mathbf{A}E\{\mathbf{x}\} = \boldsymbol{\theta}$$

Recall that for every $\theta_i \in \boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$ we have

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n], \quad i = 1, 2, \ldots, p \quad \text{and} \quad E\{\hat{\theta}_i\} = \sum_{n=0}^{N} a_n E\{x[n]\} = \theta_i$$

Recall $E\{x[n]\} = s[n]\theta \quad \Rightarrow \quad E\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta} \looparrowright$ **the constraint $\mathbf{AH} = \mathbf{I}$**

where $\mathbf{A} = [a_{in}]_{(p \times N)}$ and $\mathbf{H}$ is a vector/matrix of terms $\{s[n]\}$

The vector BLUE becomes

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

$$\text{with the covariance matrix} \quad \mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1}$$

If the data are truly Gaussian, as in

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad \text{with} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

then the BLUE also yields the Gauss-Markov theorem.

The Gauss – Markov Theorem

Given the data in a general linear model form

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

with $\mathbf{w}$ having zero mean and covariance $\mathbf{C}$, **otherwise an arbitrary PDF,**

Then, the vector BLUE of θ can be found as

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

and for every $\hat{\theta}_i \in \hat{\boldsymbol{\theta}}$, the minimum variance of $\hat{\theta}_i$ is

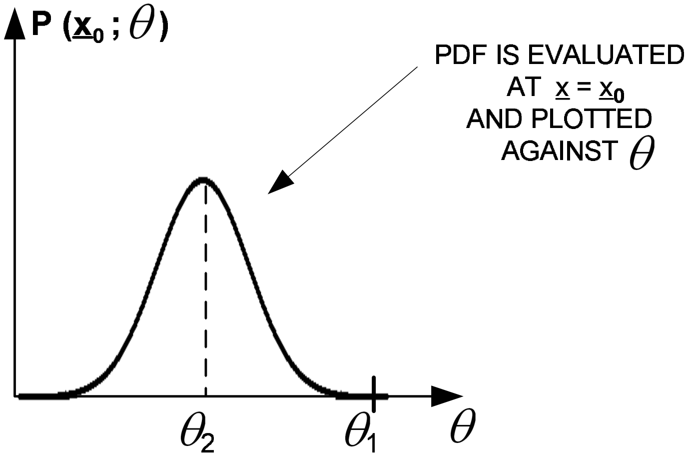
$$var(\hat{\theta}_i) = \left[\left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\right]_{ii}$$

with covariance matrix of $\hat{\boldsymbol{\theta}}$

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

Maximum Likelihood Estimation: popular for practical estimators

○ Rationale: for a given θ, the quantity $p(\mathbf{x};\theta)\,d\,\mathbf{x}$ is the probability of observing $\mathbf{x}$ in a small volume



○ The MLE = the value of θ that maximises $p(\mathbf{x};\theta)$ for $\mathbf{x}$ fixed, thus maximising the likelihood function $\forall\theta$

Alternative to an MVU estimator

○ The MLE solution is based on the ***maximum likelihood***, and is **optimal** for large enough data records

Notice that $p(\mathbf{x} = \mathbf{x}_0;\theta)d\mathbf{x}$ for each given θ gives the $p(\mathbf{x}) \in \mathbb{R}^N$, centred about $\mathbf{x}_0$ with volume $d\,\mathbf{x}$

The inference that $\theta = \theta_1$ is unreasonable because it is very unlikely that the observed value of $\mathbf{x}$ would equal $\mathbf{x}_0$

It is more "likely" that $\theta = \theta_2$, since there is a large probability that

$$\mathbf{x} = \mathbf{x}_0$$

is observed

This yields an estimator which is generally a function of $\mathbf{x}$

Maximisation performed over the allowable range of θ

Estimation theory $\looparrowright$ quick reminder
Principle of Maximum Likelihood Estimation (MLE)

Principle of estimation: We seek to determine **from a set of data, a set of parameters** such that their values would yield **the highest probability** of obtaining the observed data.

The unknown parameters may be seen as deterministic or random variable.

No a priori distribution assumed $\looparrowright$ MLE.   A priori distribution assumed $\looparrowright$ Bayesian

Principle of Maximum Likelihood Estimation (MLE): Estimate an unknown parameter such that for this value the probability of obtaining an actually observed sample is **as large as possible.**

∘ In other words: having got the observation, we look back and compute the probability that the given sample will be observed, as if the experiment is to be done again.

∘ This probability depends on a parameter which is adjusted to give it a maximum possible value.

☞ Reminds you of politicians observing the movement of the crowd and then moving to the front to lead them?

Maximum likelihood principle in a nutshell

Assumptions: The joint pdf of m sample random variables evaluated at each the sample point $x_1, x_2, \ldots, x_m$ is given as

$$l(\theta, x_1, x_2, \ldots, x_m) = l(\theta, \mathbf{x}) = \prod_{i=1}^{m} p_x(x_i|\theta)$$

The above is known as the likelihood of the sampled obserevation

Assum. 1 A random variable x has a probability distribution dependent on a parameter θ. The parameter θ lies in a space of all possible parameters θ

Assum. 2 Let $p_x(x|\theta)$ be the probability density function of x. Assume the the mathematical form of p_x is known but not θ

The likelihood function is a therefore function of the unknown parameter θ for a fixed set of observations.

☞ **The Maximum Likelihood Principle requires us to select that value of θ which maximises the likelihood function.**

Example 6: MLE of a DC level in noise

D.C. level in WGN, $w[n] \sim \mathcal{N}(0, \sigma^2)$

$$x[n] = A + w[n] \qquad\qquad n=0,1,\ldots,N\text{-}1$$

$\uparrow$
A to be estimated

Step 1: Start from the PDF

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

Step 2: Take the derivative of the log-likelihood function

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

Step 3: Set the result to zero to yield the MLE (in general, no optimality)

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

$\looparrowright$ **clearly this is an MVU estimator which yields the CRLB (efficient)**

MLE: Observations

- If an efficient estimator exists, the maximum likelihood procedure will produce it
- When an efficient estimator does not exist, the MLE has the desirable feature that it yields "an asymptotically efficient" estimator. For sufficiently large datasets, such an estimator is
  - unbiased
  - achieves the CRLB
  - has a Gaussian PDF, $\hat{\theta}^{asy} \sim \mathcal{N}(\theta, \mathcal{I}^{-1}(\theta))$
- Provided the PDF $p(\mathbf{x};\theta)$ satisfies the regularity conditions:
  - the derivatives of the log-likelihood function exist
  - and the Fisher information is non-zero

In other words, if θ is the parameter to be estimated and $\mathbf{x}$ is the observation, then the MLE estimator $\hat{\theta}_{mle}$ is found as

$$\hat{\theta}_{mle} = \arg\max \;\; p(\mathbf{x};\theta) \quad \text{for} \quad \text{fixed} \quad (\text{given}) \quad \mathbf{x}$$

that is, $\hat{\theta}_{mle}$ is the argument of $p(\mathrm{x};\theta)$ that maximises its value.

Example 7: MLE sinusoidal phase estimator

MLE of sinusoidal phase. No single sufficient statistics exists for this case. The sufficients statistics are:

$$T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n) \qquad T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$$

The observed data:

$x[n] = A\cos(2\pi f_0 n + \Phi) + w[n] \qquad n = 0, 1, \ldots, N-1 \quad w[n] \sim \mathcal{N}(0, \sigma^2)$

Task: Find the MLE estimator of Φ by maximising

$$p(\mathbf{x};\Phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2\Big]$$

or, equivalently, minimise

$$J(\Phi) = \sum_{n=0}^{N-1}\big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2$$

Example 7: MLE sinusoidal phase estimator

To find the minimum, differentiate wrt the unknown parameter Φ to yield

$$\frac{\partial J(\Phi)}{\partial \Phi} = -2 \sum_{n=0}^{N-1} \big(x[n] - A\cos(2\pi f_0 n + \Phi)\big) A \sin(2\pi f_0 n + \Phi)$$

and set the result to zero, to give

$$(SP1) \quad \sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\Phi}) = A \sum_{n=0}^{N-1} \sin(2\pi f_0 n + \hat{\Phi}) \cos(2\pi f_0 n + \hat{\Phi})$$

Notice, however (you should verify)

$$(SP2) \quad \frac{1}{N} \sum_{n=0}^{N-1} \sin(2\pi f_0 n + \hat{\Phi}) \cos(2\pi f_0 n + \hat{\Phi}) = \frac{1}{2N} \sum_{n=0}^{N-1} \sin(4\pi f_0 n + 2\hat{\Phi}) \approx 0$$

provided f_0 is not near 0 or $\frac{1}{2}$.

Example 7: MLE sinusoidal phase estimator

Thus the LHS of (SP1) when divided by N and set equal to zero will yield an approximation of MLE

$$\text{MLE of the phase } \Phi \qquad \leftrightsquigarrow \qquad \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n + \hat{\Phi}) = 0$$

Upon expanding, this yields

$$\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)\cos\hat{\Phi} = -\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)\sin\hat{\Phi}$$

so that the MLE

$$\hat{\Phi} = -\arctan\frac{\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)}{\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)}$$

The MLE $\hat{\Phi}$ is clearly a function of the sufficient statistic!

Example 7: Sinusoidal phase $\looparrowright$ numerical results

The expected asymptotic PDF of the phase estimator: $\hat{\Phi}^{asy} \sim \mathcal{N}(\Phi, \mathcal{I}^{-1}(\Phi))$

$\looparrowright$ so that the **asymptotic variance** $\text{var}(\hat{\Phi}) = \dfrac{1}{\frac{NA^2}{2\sigma^2}} = \dfrac{1}{\eta N}$

where $\eta = \dfrac{P_{signal}}{P_{noise}} = \dfrac{A^2/2}{\sigma^2} \quad (SNR)$ is the **"signal-to-noise-ratio"**

∘ **Below:** Simulation results with A=1, $f_0 = 0.08$, $\Phi = \pi/4$ and $\sigma^2 = 0.05$

Data record length	Mean, E($\hat{\Phi}$)	$N_x \times$ variance, N var($\hat{\Phi}$)
10	0.732	0.0978
40	0.746	0.108
60	0.774	0.110
80	0.789	0.0990
Theoretical asymptotic values	Φ=0.785	$\frac{1}{\eta} = 0.1$

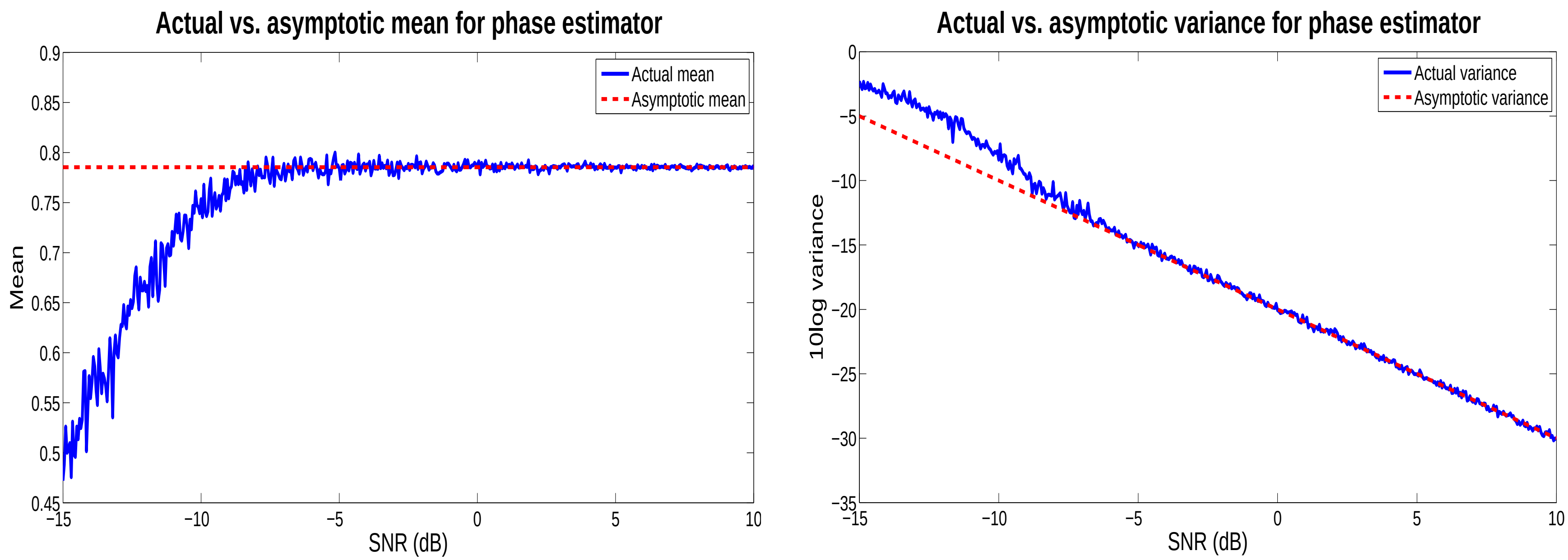
☞ For shorter data records the MLE estimate is considerably biased. Part of this bias is due to the assumption (SP2)

Example 7: MLE of sinusoidal phase phase $\looparrowright$ asymptotic mean and variance (performance vs. SNR for a fixed N)

- Data length was fixed at $N = 80$, SNR was varied from -15 to +10 dB
- The asymptotic variance (or CRLB) is

$$10\log_{10}\mathrm{var}(\hat{\Phi}) = 10\log_{10}\frac{1}{N\eta} = -10\log_{10}N - 10\log_{10}\eta$$

- Mean and variance are also functions of SNR
- Asymptotic mean attained for SNRs > -10dB



Observe that the minimum data length to attain CRLB also depends on SNR

MLE: Extension to vector parameter

☞ A distinct advantage of the MLE is that we can always find it for a given dataset numerically, as the MLE is a maximum of a known function.

∘ For instance, a grid search of $p(\mathbf{x}; \boldsymbol{\theta})$ can be performed over a finite interval [a, b].

∘ If the grid search cannot be performed (e.g. infinite range of θ) then we may resort to **iterative maximisation**, such as the Newton-Raphson method, the scoring approach, and the expectation-maximisation (EM) approach. Good MLE for good initial guess.

∘ Since the likelihood function to be maximised **is not known a priori** and it changes for each dataset, we effectively maximise a **random function**.

☞ **Extension to a vector parameter** is straightforward: The MLE for a vector parameter $\boldsymbol{\theta}$ is the value that maximises the likelihood function $p(\mathbf{x}; \boldsymbol{\theta})$ over the allowable domain of $\boldsymbol{\theta}$. The MLE is found from

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad \text{with usual regularity conditions} \quad \hat{\boldsymbol{\theta}} \overset{\mathbf{asy}}{\sim} \mathcal{N}\big(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\big)$$

Example 8: MLE of a DC level in WGN. Both the DC level A and the noise variance (power) σ^2 are unknown

Consider the data $x[n] = A + w[n], \quad n = 0, 1, \ldots, N-1, \quad w[n] \sim \mathcal{N}(0, \sigma^2)$

The vector parameter $\boldsymbol{\theta} = [A,\, \sigma^2]^T$ is to be estimated.

Solution:

$$\begin{aligned}
\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \\
\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} (x[n] - A)^2
\end{aligned}$$

From first equation solve for A, from second equation solve for σ^2 to obtain

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \bar{x} \\ \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \end{bmatrix} \stackrel{N \rightarrow \infty}{\longrightarrow} \begin{bmatrix} A \\ \frac{\sigma^2}{N} \end{bmatrix} \qquad \text{asymptotic CRLB}$$

where $\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$.

Example 9: Sinusoidal parameter estimation with three unknown parameters $\hookrightarrow$ A, f_0, and Φ

Since the likelihood function

$$p(\mathbf{x};\Phi)=\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\big(x[n]-A\cos(2\pi f_0 n+\Phi)\big)^2\Big]$$

For $A>0,\ 0<f_0<\frac{1}{2}$, the MLE of $\boldsymbol{\theta}=[A,f_0,\Phi]^T$ is found by minimising

$$\begin{aligned}J(A,f_0,\Phi) &= \sum_{n=0}^{N-1}\big(x[n]-A\cos(2\pi f_0 n+\Phi)\big)^2\\ &= \sum_{n=0}^{N-1}(x[n]-A\cos\Phi\cos 2\pi f_0 n+A\sin\Phi\sin 2\pi f_0 n)^2\end{aligned}$$

☞ Begin by transforming A and Φ to a quadratic function, where

$$\alpha_1=A\cos\Phi,\qquad \alpha_2=A\sin\Phi$$

N.B. $\quad A=\sqrt{\alpha_1^2+\alpha_2^2}\quad \&\quad \Phi=\tan^{-1}(\frac{\alpha_2}{\alpha_1})$

Exampe 9: Sinusoidal parameter estimation of three unknown parameters, cont.

$$\text{Let} \qquad \mathbf{c} = \big[1, \cos 2\pi f_0, \ldots, \cos 2\pi f_0(N-1)\big]^T$$
$$\mathbf{s} = \big[0, \sin 2\pi f_0, \ldots, \sin 2\pi f_0(N-1)\big]^T \qquad \text{to yield}$$

$$J'(\alpha_1, \alpha_2, f_0) = \big(\mathbf{x} - \alpha_1 \mathbf{c} - \alpha_2 \mathbf{s}\big)^T \big(\mathbf{x} - \alpha_1 \mathbf{c} - \alpha_2 \mathbf{s}\big) = \big(\mathbf{x} - \mathbf{H}\boldsymbol{\alpha}\big)^T \big(\mathbf{x} - \mathbf{H}\boldsymbol{\alpha}\big)$$

We therefore obtain a **linear estimator** where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^T$, $\mathbf{H} = [\mathbf{c}\ \mathbf{s}]$.

The minimum by, inspection, is $\hat{\boldsymbol{\alpha}} = \big(\mathbf{H}^T\mathbf{H}\big)^{-1}\mathbf{H}^\mathbf{T}\mathbf{x}$, so that

$$\hat{\boldsymbol{\alpha}} \approx \frac{2}{N}\begin{bmatrix} \mathbf{c}^\mathbf{T}\mathbf{x} \\ \mathbf{s}^\mathbf{T}\mathbf{x} \end{bmatrix} = \begin{bmatrix} \frac{2}{N}\sum x[n] \cos 2\pi \hat{f}_0 n \\ \frac{2}{N}\sum x[n] \sin 2\pi \hat{f}_0 n \end{bmatrix} \qquad \hat{\Phi} = -\arctan \frac{\sum_{n=0}^{N-1} x[n] \sin(2\pi \hat{f}_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi \hat{f}_0 n)}$$

$$\text{and} \qquad \hat{A} = \sqrt{\alpha_1^2 + \alpha_2^2} = \frac{2}{N}\Big|\sum_{n=0}^{N-1} x[n] \exp(-j2\pi \hat{f}_0 n)\Big|$$

Theorem: Optimality of MSE for a linear model

Theorem: If the observed data can be described by the general linear model

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

where $\mathbf{H}$ is a known $N \times p$ matrix with $N > p$ and of rank p, $\boldsymbol{\theta}$ is a $p \times 1$ parameter vector to be estimated, and $\mathbf{w}$ is a noise vector with PDF $\mathcal{N}(\mathbf{0}, \mathbf{C})$, then the MLE of $\boldsymbol{\theta}$ is

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

In addition, $\hat{\boldsymbol{\theta}}$ is also an efficient estimator in that it attains the CRLB and hence it is the MVU estimator. The PDF of $\hat{\boldsymbol{\theta}}$ is given by

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

Summary: BLUE and MLE

NB: the optimal MVU estimator and CRLB do not always exist or are impossible to find

Best Linear Unbiased Estimator

- It operates even when the *pdf* of data is unknown
- Restricts the estimates to be linear in the data (e.g. DC level in noise)
- Produces unbiased etimates
- Minimises the variance of such unbiased estimates
- Requires knowledge of only the mean and variance of the data, and not the full *pdf*
- BLUE may be used more generally if the data model is linearised

Maximum Likelihood Estimator

- Can always be applied if the *pdf* is known, and does not restrict the data model (cf. BLUE)
- It is asymptotically optimal (for large data size)
- Can be computationally complex (numerical methods)
- **The basic idea:** in the *pdf* $p(\mathbf{x};\theta)$, θ is regarded as a variable and not as a parameter!
- **ML estimate:** the value of θ that maximises the likelihood funct. $\ln p(\mathbf{x};\theta)$ $\looparrowright$ found by different. wrt θ and setting to 0

Appendix: Some observations about BLUE

- BLUE is applicable to amplitude estimation of known signals in noise, where to satisfy the unbiased constraint, $E\{x[n]\}$ must be linear in the unknown parameter θ, or in other words, $E\{x[n]\} = s[n]\theta$
- **Counter-example:** if $E\{x[n]\} = \cos\theta$, which is not linear in θ, then from the unbiased assumption we have $\sum_{n=0}^{N-1} a_n \cos\theta = \theta$. Clearly, there are no $\{a_n\}$ that satisfy this condition
- For the vector parameter BLUE, the unbiased constraint generalises from the scalar case as

$$E\{x[n]\} = s[n]\theta \quad \rightarrow \quad \mathbf{a}^T\mathbf{s} = 1 \qquad \Rightarrow \qquad E\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta} \quad \rightarrow \quad \mathbf{AH} = \mathbf{I}$$

Since the **unbiased constraint:**

$$E\{\hat{\theta}_i\} = \sum_{n=0}^{N-1} a_{in} E\{x[n]\} = \theta_i \quad \Rightarrow \quad E\{\hat{\boldsymbol{\theta}}\} = \mathbf{A}E\{\mathbf{x}\} = \boldsymbol{\theta}$$

this is equivalent to $\mathbf{a}_i^T\mathbf{h}_j = \delta_{ij}$ (=0 for i≠j, = 1 for i=j)

Lecture supplement: Constrained optimisation using Lagrange multipliers

Consider a two-dimensional problem:

$$\text{maximize} \qquad \underbrace{f(x,y)}_{function\ to\ max/min}$$

$$\text{subject to} \qquad \underbrace{g(x,y) = c}_{constraint}$$

↬ **we look for point(s) where curves f & g touch (but do not cross).**

In those points, the tangent lines for f and g are parallel $\Rightarrow$ so too are the gradients $\boldsymbol{\nabla_{x,y} f \parallel \lambda \nabla_{x,y} g}$, where λ is a scaling constant.

Although the two gradient vectors are parallel they can have different magnitudes!

Therefore, we are looking for $\max$ or $\min$ points (x,y) of $f(x,y)$ for which

$$\nabla_{x,y} f(x,y) = -\lambda \nabla_{x,y} g(x,y) \quad \text{where } \nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \text{ and } \nabla_{x,y} g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}\right)$$

We can now combine these conditions into one equation as:

$$F(x,y,\lambda) = f(x,y) - \lambda\big(g(x,y) - c\big) \quad \text{and solve} \quad \nabla_{x,y,\lambda} F(x,y,\lambda) = \mathbf{0}$$

Obviously, $\nabla_\lambda F(x,y,\lambda) = 0 \quad \Leftrightarrow \quad g(x,y) = c$

The method of Lagrange multipliers in a nutshell

max/min of a function $f(x, y, z)$ where x, y, z are coupled

Since x, y, z **are not independent** there exists a constraint $g(x, y, z) = c$

Solution: Form a new function

$F(x, y, z, \lambda) = f(x, y, z) - \lambda\big(g(x, y, z) - c\big)$ and calculate $F'_x, F'_y, F'_z, F'_\lambda$

Set $F'_x, F'_y, F'_z, F'_\lambda = 0$ and solve for the unknown x, y, z, λ.

Example 10: Economics Two factories, A and B make TVs, at a cost

$$f(x, y) = 6x^2 + 12y^2 \qquad \text{where} \quad x = \#TV\ in\ A \quad \& \quad y = \#TV\ in\ B$$

Task: Minimise the cost of producing 90 TVs, by finding optimal numbers of TVs, x and y, produced respectively at factories A and B.

Solution: The constraint $g(x, y)$ is given by (x+y=90), so that

$$\boldsymbol{F(x, y, \lambda) = 6x^2 + 12y^2 - \lambda(x + y - 90)}$$

Then: $F'_x = 12x - \lambda, \quad F'_y = 24y - \lambda, \quad F'_\lambda = -x - y + 90$, and we need to set $\nabla F = \mathbf{0}$ in order to find $\min / \max$.

☞ Upon setting $[F'_x, F'_y, F'_\lambda] = \mathbf{0}$ we find $x = 60, y = 30, \lambda = 720$

Notes:

○

Notes:

-