

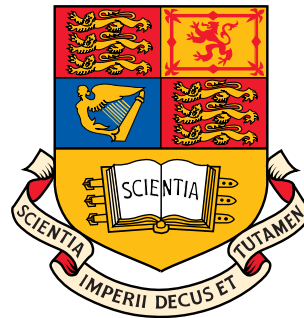
---

# Statistical Signal Processing & Inference

## 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](http://www.commsp.ee.ic.ac.uk/~mandic)

# Overview

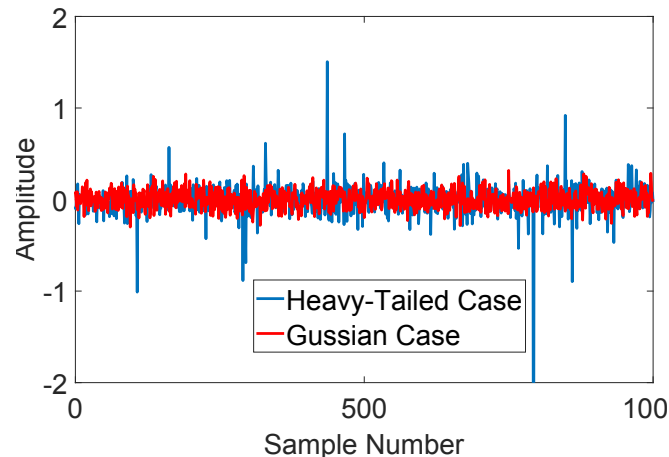
---

- 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 pdf based methods cannot be applied**
- We therefore have to resort to **suboptimal solutions**  $\leftrightarrow$  impose some constraints on the estimator and data model
- If the variance of a suboptimal estimator meets our system specifications, the use of such estimators may be justified
- The best linear unbiased estimator (BLUE)  $\leftrightarrow$  restrict the estimator to be **linear in the data**  $\leftrightarrow$  finds a **linear estimator** that is **unbiased and has minimum variance among such unbiased estimators**
- Alternatively, if the MVU estimator does not exist, or BLUE is not applicable we may resort to **Maximum Likelihood Estimation (MLE)**
- We first need to look at which data samples are pertinent to the estimation problem  $\rightsquigarrow$  the so called **sufficient statistics**

# Motivation for BLUE (Best Linear Unbiased Estimator)

## sufficient statistics and the linearity assumption

---



- In many applications, signals exhibit sharp spikes
- This results in heavy-tailed distributions (e.g.  $\alpha$ -stable)
- There may not be a general form of pdf for such distributions

- 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
- This makes it possible in many cases to determine an approximate MVU estimator **by a simple inspection of the PDF** (e.g. MLE)

The BLUE assumptions are also referred to as **Gauss–Markov assumptions**  
These have been responsible for advances in “quantitative methodologies”

# An insight into the ‘sufficiency’ of the data statistics

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

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

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

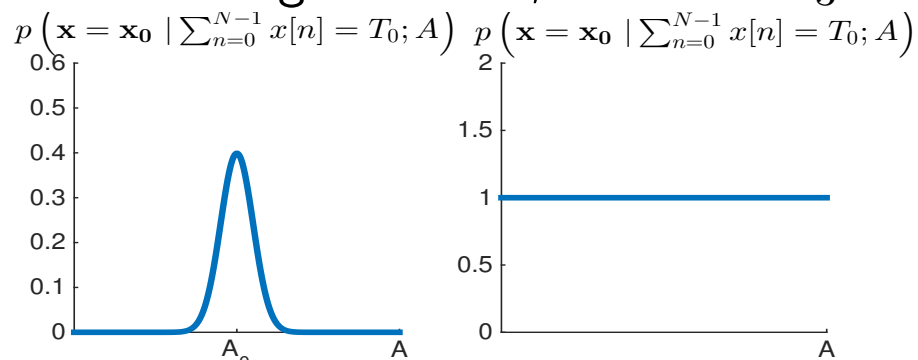
$\hookrightarrow$  Although  $\tilde{A}$  is **unbiased**, its variance is much larger than that of  $\hat{A}$ . This is due to discarding  $x[1], \dots, x[N-1]$  that carry information about  $A$ .

Consider now the following datasets:

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

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

$\rightsquigarrow$  In addition to being sufficient, statistics  $S_3$  is **minimal sufficient statistics** 😊



Information present in observations after  $T(\mathbf{x})$  observed

No information in observations after  $T(\mathbf{x})$  observed

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

o **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  $\theta$  (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  $\theta$   $\rightsquigarrow$  **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  $p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$  (function  $h$  depends only on  $\mathbf{x}$ )

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

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

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

$T(\mathbf{x})$

**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  $\mathbf{x} = [x[0], \dots, x[N-1]]^T \in \mathbb{R}^{N \times 1} \rightsquigarrow N$ -dim. sufficient statistics

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

### How to find the MVU?

1. find **any unbiased estimator**  $\bar{\theta}$  of  $\theta$  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  $\theta$  and  $T(\mathbf{x})$  is sufficient statistics for  $\theta$ .

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

- valid (not dependent on  $\theta$ )
- 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 pdf cannot be applied 😞

**Remedy:** Resort to a sub-optimal estimator  $\leadsto$  check its variance and ascertain whether it meets the required specifications (and/or CRLB)

**Common sense approach:** Assume an estimator to be:

- **linear in the data**, that is,  $\hat{\theta}_{BLUE} = \sum_{n=0}^{N-1} a_n x[n]$
- among all such linear estimators, seek for an **unbiased** one,
- then **minimise the variance**.

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

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

# The form and optimality of BLUE

Consider the data  $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^T$ , for which the *pdf*  $p(\mathbf{x}; \theta)$  depends on the unknown parameter  $\theta$ .

## The form of BLUE

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

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

↑  
Constants to be determined

We choose  $a_n$ s to give an unbiased est.,  $E\{\hat{\theta}\} = \theta$ . Then,  $\min(\text{var})$   
 $\leadsto$  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

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

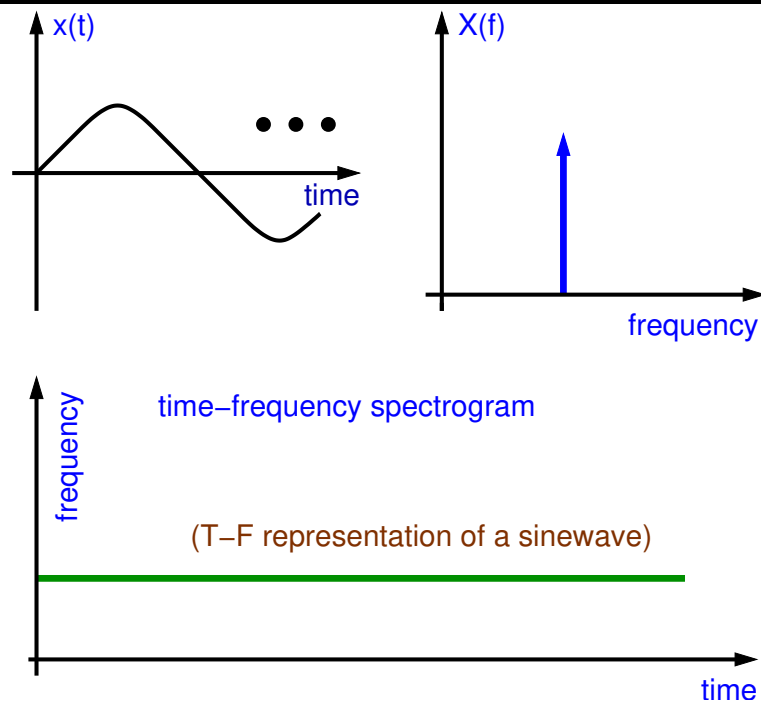
which is clearly **linear in the data**, **BLUE is an optimal MVU** giving  $a_n = 1/N$ .



# 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.

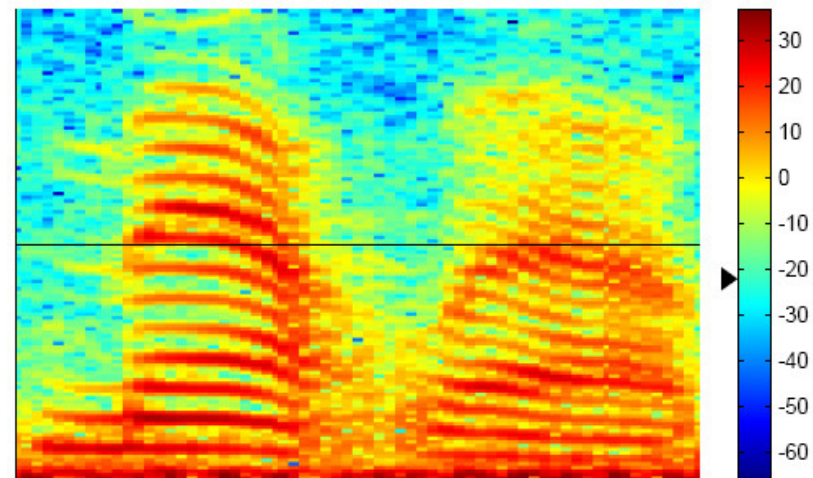
## Sinusoidal frequency estimation



- Ramp in time  $\leftrightarrow$  DC level in time (via differentiation)
- Chirp in time  $\leftrightarrow$  ramp in T-F

## Transforming other problems

time-frequency representation



horizontal: time vertical: frequency

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

**DC-level like harmonics for "a"**

## Example 2: Composite faces $\rightarrow$ 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?

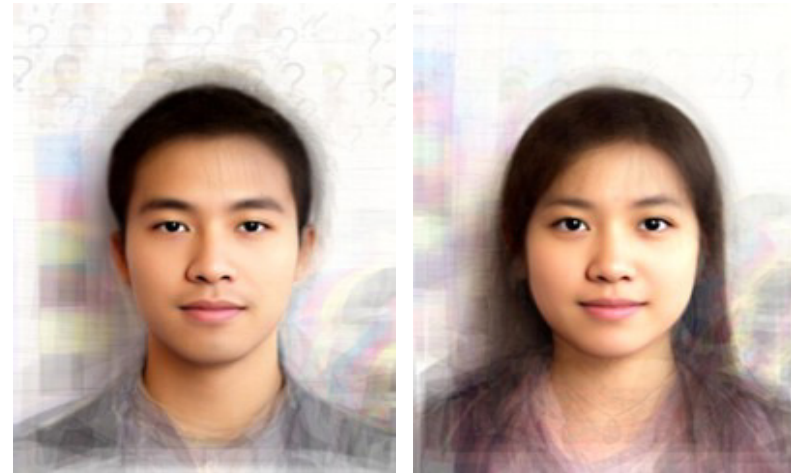


**Participants in Sydney, Australia, ranging from 0.83–93 years**

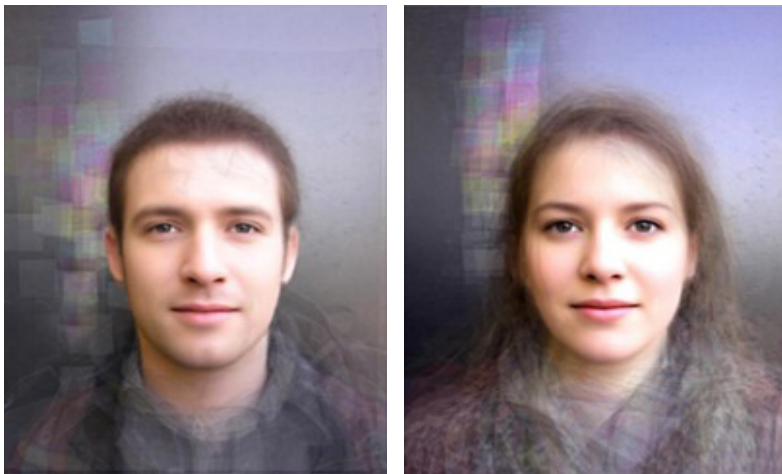
## Example 2: contd. $\rightarrow$ composite male and female faces



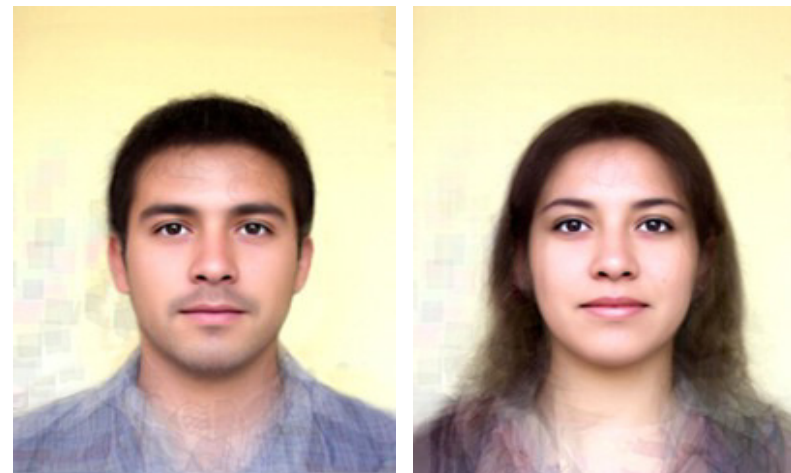
**Composite faces of Sydney**



**Composite faces of Hong Kong**



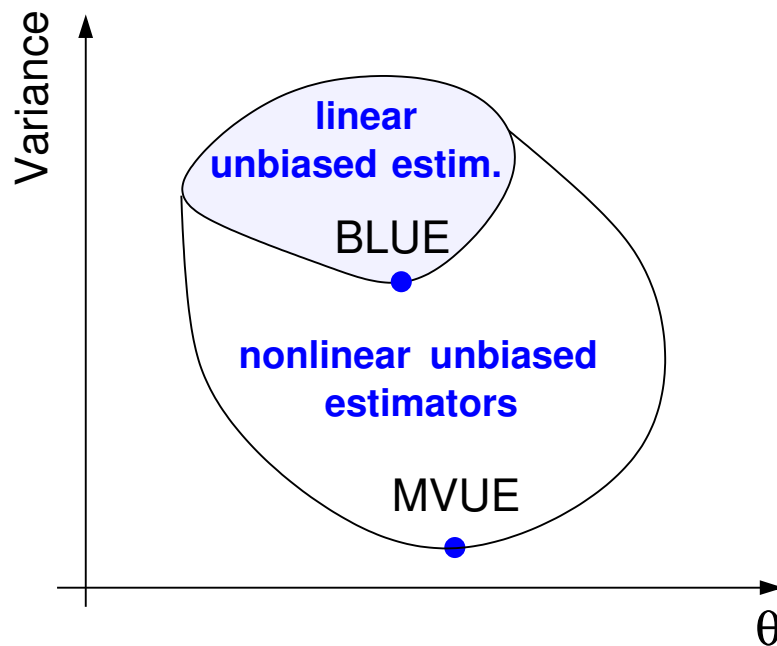
**Composite faces of London**



**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

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

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

BLUE can achieve  $\text{var}(\hat{\theta}) = \frac{\beta^2}{N}$

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

Its direct form is inappropriate for nonlinear prob. ↗ 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  $\hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$

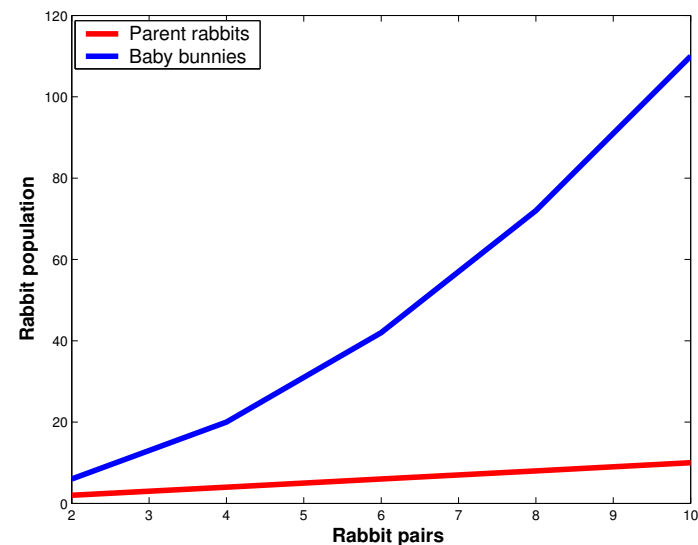
is **nonlinear** in the data. Forcing the estimator to be linear, e.g.

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

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

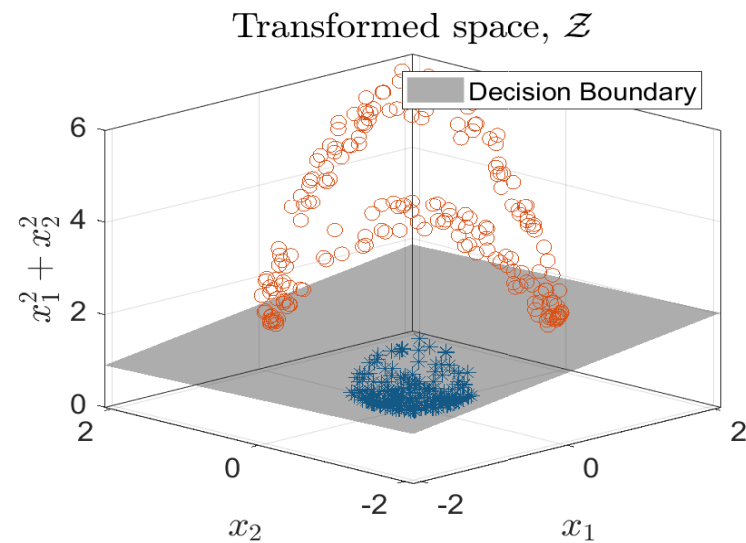
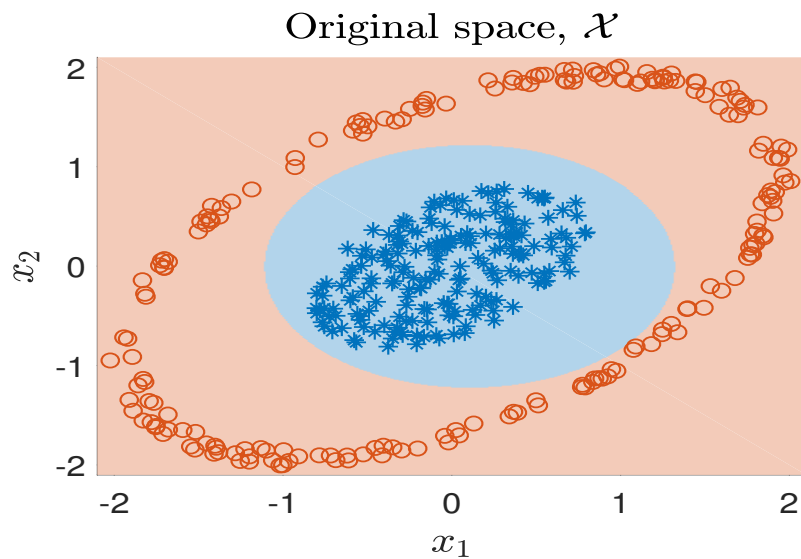
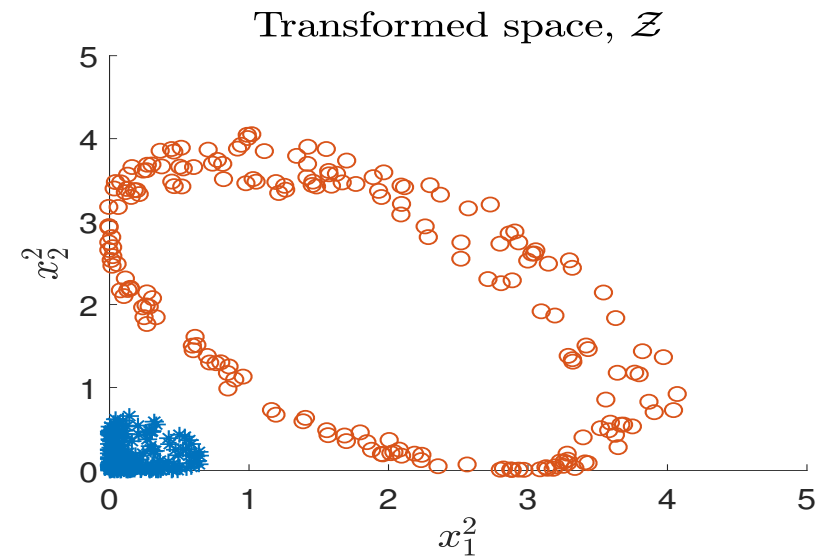
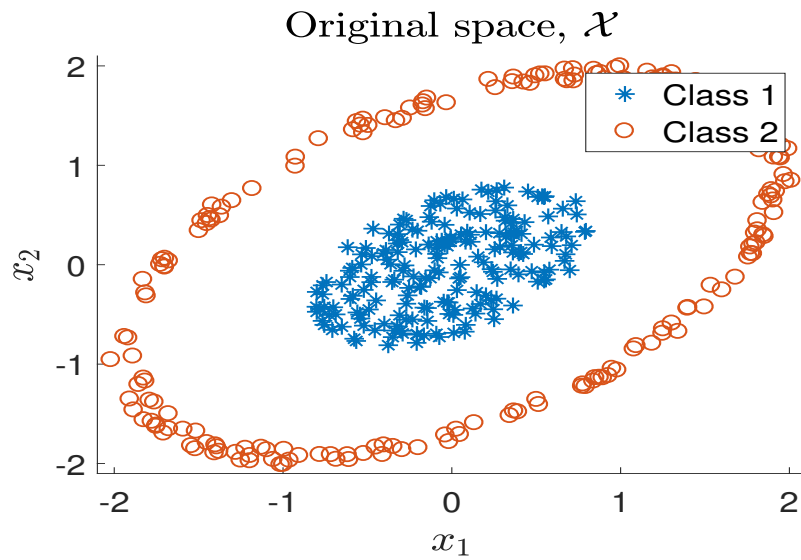


The time evolution of the rabbit population is nonlinear (exponential)

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

# Example 3 cont.: Nonlinear transformation of data

Left: Original data (nonlin. sep.) Right: Linear separability after nonlinear transf.



# How to find BLUE?

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


Consider a scalar linear observation  $x[n] = \theta s[n] + w[n] \Rightarrow E\{x[n]\} = \theta s[n]$

and notice that  $E\{\hat{\theta}\} = \theta \sum_{n=0}^{N-1} a_n s[n]$   $s[n] \rightsquigarrow$  scaled mean

## 1. Unbiased constraint

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

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

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

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

## 2. Variance minimisation

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

### BLUE optimisation task

**Minimise:**

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

subject to the **unbiased constraint**

$$\mathbf{a}^T \mathbf{s} \theta = \theta \Leftrightarrow \mathbf{a}^T \mathbf{s} = 1$$

## Some remarks on variance calculation

---

A closer look at the variance yields

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

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

$$E \left\{ \mathbf{a}^T (\mathbf{x} - E\{\mathbf{x}\}) (\mathbf{x} - E\{\mathbf{x}\})^T \mathbf{a} \right\} = \mathbf{a}^T \mathbf{C} \mathbf{a} \quad \text{like } \text{var}(aX) = a^2 \text{var}(X)$$

Also assume

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

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



**BLUE is linear in the unknown parameter  $\theta$** , 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

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

## 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  $\lambda$

4. From the constraint equation

$$\begin{aligned} \mathbf{a}^T \mathbf{s} &= \frac{\lambda}{2} \mathbf{s}^T \mathbf{C}^{-1} \mathbf{s} = 1 \\ \Rightarrow \frac{\lambda}{2} &= \frac{1}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} \end{aligned}$$

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

Recall that  $\theta = \mathbf{a}^T \mathbf{x}$        $\text{var}(\hat{\theta}) = \mathbf{a}^T \mathbf{C} \mathbf{a}$

---

**BLUE of an unknown parameter (our function  $g(\mathbf{x})$  from MVU):**

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

**BLUE variance:**

$$\text{var}(\hat{\theta}) = \mathbf{a}_{opt}^T \mathbf{C} \mathbf{a}_{opt} = \frac{1}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$$

To determine the BLUE we only require knowledge of

$\mathbf{s}$      $\leftrightarrow$     the scaled mean

$\mathbf{C}$      $\leftrightarrow$     the covariance matrix      ( $\mathbf{C}^{-1}$  pre-whitens the data prior to averaging, see Slide 42 in Lecture 4)

**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

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N - 1$$

where  $\{w[n]\}$  is **any white noise with variance  $\sigma^2$  (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

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

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

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

Recall that  $\mathbf{a}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$ ,  $\text{var}(\hat{\theta}) = \frac{1}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$ , and  $\hat{\theta} = \mathbf{a}_{opt}^T\mathbf{x}$

For any i.i.d. white noise  $\{w\}$  with power  $\sigma^2$ ,

$$\mathbf{C} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I} \quad \Rightarrow \quad \mathbf{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix} = \frac{1}{\sigma^2}\mathbf{I}$$

The BLUE for the estimation of DC level in noise then becomes

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

and has minimum variance (CRLB for a linear estimator)

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

- 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 exist, 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  $\mathbf{1}^T \mathbf{I} \mathbf{x}$

$$\begin{array}{c} \boxed{1 \ 1 \ 1 \ \dots \ 1} \\ 1 \times N \end{array} \begin{array}{c} \boxed{\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & 0 \\ & & & \ddots \\ 0 & & & \ddots \\ & & & & 1 \end{array}} \\ N \times N \end{array} \begin{array}{c} \boxed{\begin{array}{c} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{array}} \\ N \times 1 \end{array} = \begin{array}{c} \boxed{1 \ 1 \ 1 \ \dots \ 1} \\ 1 \times N \end{array} \begin{array}{c} \boxed{\begin{array}{c} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{array}} \\ N \times 1 \end{array} = N$$

$$\begin{array}{c} \boxed{1 \ 1 \ 1 \ \dots \ 1} \\ 1 \times N \end{array} \begin{array}{c} \boxed{\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & 0 \\ & & & \ddots \\ 0 & & & \ddots \\ & & & & 1 \end{array}} \\ N \times N \end{array} \begin{array}{c} \boxed{\begin{array}{c} x[0] \\ x[1] \\ \cdot \\ \cdot \\ \cdot \\ x[N-1] \end{array}} \\ N \times 1 \end{array} = \begin{array}{c} \boxed{1 \ 1 \ 1 \ \dots \ 1} \\ 1 \times N \end{array} \begin{array}{c} \boxed{\begin{array}{c} x[0] \\ x[1] \\ \cdot \\ \cdot \\ \cdot \\ x[N-1] \end{array}} \\ N \times 1 \end{array} = \sum x[n]$$

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 $\text{var}(w[n]) = \sigma_n^2$ (de-emphasising bad samples)

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

As before,  $s = 1$ .

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:

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

- 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

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

## BLUE: Extension to vector parameter

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

---

**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  $\theta_i \in \boldsymbol{\theta} = [\theta_1, \dots, \theta_p]^T$  we have

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

Recall  $E\{x[n]\} = s[n]\theta \quad \Rightarrow \quad E\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta} \quad \Leftrightarrow \quad \text{the constraint } \mathbf{A}\mathbf{H} = \mathbf{I}$

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

**The vector BLUE becomes**

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

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

**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 Minimum Variance Unbiased estimator.**

## The Gauss – Markov Theorem

---

Consider the observed data in the form of a general linear model

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

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

**Then, the vector BLUE of  $\boldsymbol{\theta}$  can be found as**

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

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

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

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

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



## Example 6: Sinusoidal phase estim. (DSB, PSK, QAM)

### Motivation for Maximum Likelihood Estimation (MLE)

**Signal model:**  $x[n] = A \cos(2\pi f_0 n + \Phi) + w[n]$   $w \sim \mathcal{N}(0, \sigma^2)$

**Signal to noise ratio (SNR):**  $SNR = \frac{P_{signal}}{P_{noise}} = \frac{A^2}{2\sigma^2}$

**Parametrised pdf:**  $p(\mathbf{x}; \Phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{\sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \Phi))^2}{2\sigma^2}}$

**Regularity condition within CRLB:**  $\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = I(\theta) [g(\mathbf{x}) - \theta]$

**In our case:** (see Example 8, slide 32)

$$\frac{\partial \ln p(\mathbf{x}; \Phi)}{\partial \Phi} = -\frac{A}{\sigma^2} \sum_{n=0}^{N-1} (x[n] \sin(2\pi f_0 n + \Phi) - \frac{A}{2} \sin(4\pi f_0 n + 2\Phi))^2$$

 **The regularity condition is not satisfied, and an efficient estimator for sinusoidal phase estimation does not exist**

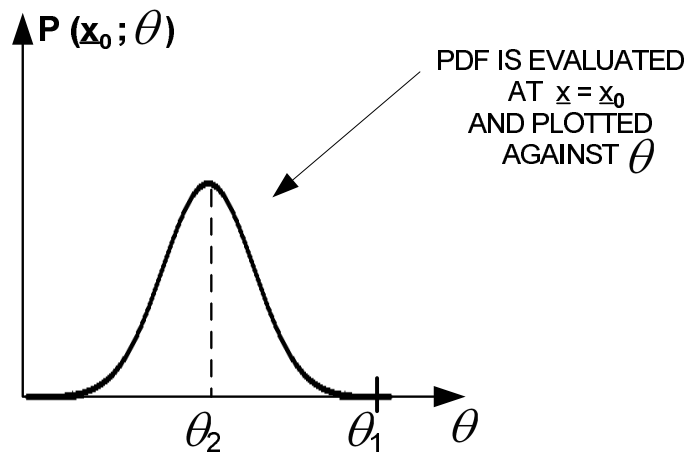
**Remedy:** Using MLE, we can still obtain an  $\approx$  CRLB for freq. far from 0 and 1/2

**Approximate CRLB:**  $var(\Phi) \geq \frac{1}{N \times SNR}$  (see Example 8)

# Maximum Likelihood Estimation: popular for practical estimators

Effectively, we treat  $\theta$  as a variable, not as a parameter,  $\hat{\theta} = \arg \max_{\theta} p(\mathbf{x}; \theta)$

- Rationale:**
- MVU estimator often does not exist or cannot be found
  - BLUE may not be applicable (e.g.  $\mathbf{x} \neq \mathbf{H}\theta + \mathbf{w}$ )
  - If the pdf is known, then MLE can always be used!



- The MLE = the value of  $\theta$  that maximises  $p(\mathbf{x}; \theta)$  for  $\mathbf{x}$  fixed, thus maximising the likelihood function  $\forall \theta$ .
- ↪ **Alternative to an MVU estimator**

**Notice that  $p(\mathbf{x} = \mathbf{x}_0; \theta)d\mathbf{x}$  for each given  $\theta$  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.

**In other words, pick  $\hat{\theta}_{ML}$  so that  $p(\mathbf{x}; \hat{\theta}_{ML})$  is largest.**

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

Maximisation performed over the allowable range of  $\theta$ .

## Estimation theory $\leadsto$ 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 a deterministic or a random variable.

No *a priori* distribution assumed  $\leadsto$  MLE.    *A priori* distribution assumed  $\leadsto$  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.*

☞ MLE is a **turn-the-crank** method which is optimal for large enough data. It can be computationally complex and may require numerical methods.

☞ Makes the data you **did observe** the most likely data **you have observed!**

# Maximum likelihood principle in a nutshell

---

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

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

The above is known as the likelihood of the sampled observation.

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

The likelihood function is a therefore function of the unknown parameter  $\theta$  for a fixed set of observations.

 **The Maximum Likelihood Principle requires us to select that value of  $\theta$  which maximises the likelihood function.**

## Example 7: MLE of a DC level in noise

---

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

$$x[n] = A + w[n] \quad n=0,1,\dots,N-1$$



A to be estimated

**Step 1:** Start from the PDF

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

**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)

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



Clearly this is an MVU estimator which yields the CRLB (efficient)

## MLE: Observations so far

---

- If an efficient estimator exists, the maximum likelihood procedure will produce it (see Example 7)
- When an efficient estimator does not exist, the MLE has the desirable feature that it yields “an asymptotically efficient” estimator (shown in Example 8). 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  $\theta$  is the parameter to be estimated and  $\mathbf{x}$  is the observation, then the MLE estimator  $\hat{\theta}_{mle}$  is found as

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

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

## Example 8: MLE sinusoidal phase estimator (cf. Ex. 6)

Recall the Neyman-Fisher factorisation:  $p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$

---

**MLE of sinusoidal phase.** No single sufficient statistics exists for this case. The sufficients statistics are (see also Slide 5):

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

**The observed data:**

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

**Task:** Find the MLE estimator of  $\Phi$  by maximising

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

or, equivalently, minimise

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

## Example 8: MLE sinusoidal phase estimator (cf. Ex. 6)

Recall the Neyman-Fisher factorisation:  $p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$

---

To find the minimum, differentiate wrt the unknown parameter  $\Phi$  to yield

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

and set the result to zero, to give

$$(SP1) \quad \sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\Phi}) = A \sum_{n=0}^{N-1} \underbrace{\sin(2\pi f_0 n + \hat{\Phi}) \cos(2\pi f_0 n + \hat{\Phi})}_{\text{inner product of sine and cosine}}$$

Notice, however (use  $\sin(2a) = 2\sin(a)\cos(a)$ , see also Example 9 in Lecture 4)

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

provided  $f_0$  is not near 0 or  $\frac{1}{2}$ , and for a large enough  $N$ .



## Example 8: MLE sinusoidal phase estimator (cf. Ex. 6)

Recall the Neyman-Fisher factorisation:  $p(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$

---

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

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

Upon expanding  $\sin(2\pi f_0 n + \hat{\Phi})$ , 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 statistics**, which are

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

## Example 8: Sinusoidal phase $\varphi \rightarrow$ numerical results


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

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

where  $\eta = \frac{P_{signal}}{P_{noise}} = \frac{A^2/2}{\sigma^2}$  (*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, $N \text{var}(\hat{\Phi})$
10	0.732	0.0978
40	0.746	0.108
60	0.774	0.110
80	0.789	0.0990
<b>Theoretical asymptotic values</b>	$\Phi=0.785$	$\frac{1}{\eta} = 0.1$

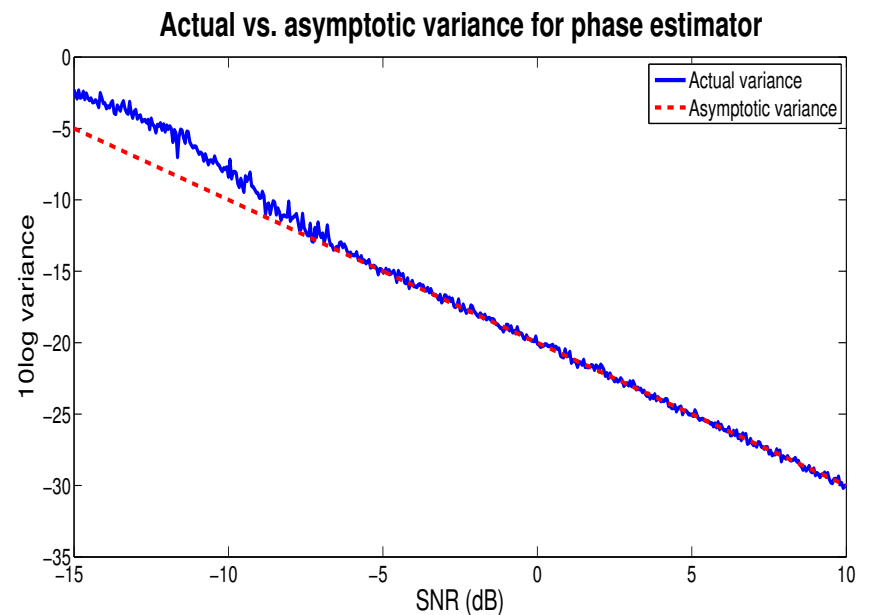
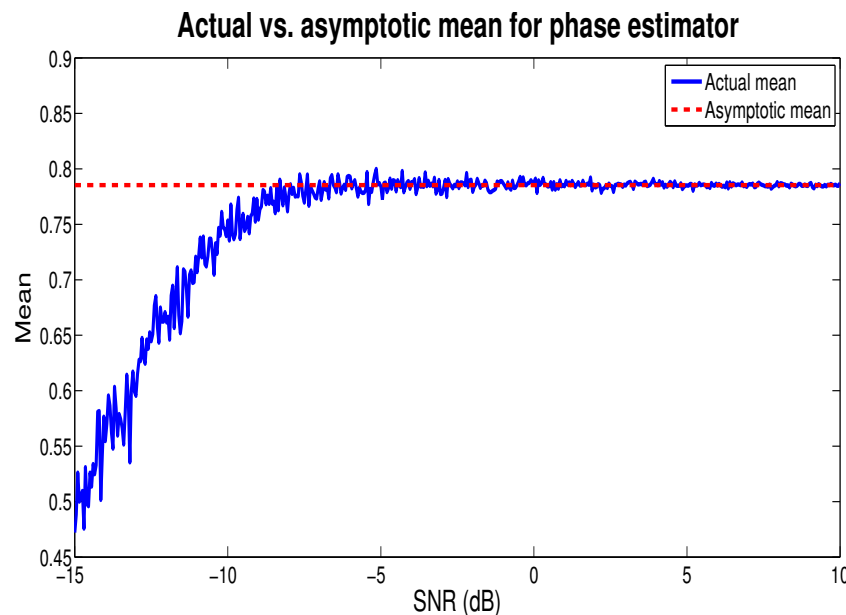
 For shorter data records the MLE estimate is considerably biased. Part of this bias is due to the assumption (SP2)

## Example 8: MLE of sinusoidal phase $\varphi \rightarrow$ asymptotic mean and variance (performance vs. SNR for a fixed $N$ )

- For a fixed data length at  $N = 80$ , SNR was varied from -15 to +10 dB
- The asymptotic variance (or CRLB) then becomes

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

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



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

## Asymptotic properties of MLE

---

We can now formalise the asymptotic properties of  $\hat{\theta}_{ML}^{asy}$  (see the previous slide).

**Theorem (asymptotic properties of MLE):** If  $p(\mathbf{x}; \theta)$  satisfies some “regularity” conditions, then the MLE is **asymptotically distributed** as

$$\hat{\theta}^{asy} \sim \mathcal{N}(\theta, \mathcal{I}^{-1}(\theta))$$

where “regularity” refers to the existence of the derivative of the log-likelihood function (as well as Fisher information being non-zero), and  $\mathcal{I}$  is the Fisher Information evaluated at the true value of the unknown parameter  $\theta$ .

👉 The Maximum Likelihood Estimator is therefore **asymptotically:**


- unbiased
- efficient (that is, achieves the CRLB)


👉 For a small  $N$ , there is no guarantee how the MLE behaves

We use **Monte Carlo simulations** to answer “*how large an  $N$  do we need?*” (see Appendix for more detail)

## 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  $\theta$ ) 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  $\boldsymbol{\theta}$ .

**Asymptotic properties:** If  $\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$  then  $\hat{\boldsymbol{\theta}}^{\text{asy}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta}))$

## Example 9: MLE of a DC level in WGN. Both the DC level $A$ and the noise variance (power) $\sigma^2$ are unknown

Consider the data  $x[n] = A + w[n]$ ,  $n = 0, 1, \dots, N - 1$ ,  $w[n]$  is zero-mean  
The vector parameter  $\boldsymbol{\theta} = [A, \sigma^2]^T$  is to be estimated ( $\text{var}(w)$  is unknown too)

**Solution:** (our  $p(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{x}; A, \sigma^2)$  has the usual form)

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

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

From first equation solve for  $A$ , from second equation solve for  $\sigma^2$  to obtain

$$\hat{\boldsymbol{\theta}} = \left[ \begin{array}{c} \bar{x} \\ \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \end{array} \right] \xrightarrow{N \rightarrow \infty} \left[ \begin{array}{c} A \\ \sigma^2 \end{array} \right] \quad \text{asymptotic CRLB}$$

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

 **Amazing, as we only knew the type the PDF, but not the variance!**


## Example 10: Sinusoidal parameter estimation with three unknown parameters $\vartheta \rightarrow A, f_0,$ and $\Phi$

Now,  $\boldsymbol{\theta} = [A, f_0, \Phi]^T$ , and

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \underbrace{(x[n] - A \cos(2\pi f_0 n + \Phi))^2}_{\text{we need this as } (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})} \right]$$

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

$$\begin{aligned} J(A, f_0, \Phi) &= \sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \Phi))^2 \\ &= \sum_{n=0}^{N-1} (x[n] - \underbrace{A \cos \Phi}_{\alpha_1} \cos 2\pi f_0 n + \underbrace{A \sin \Phi}_{-\alpha_2} \sin 2\pi f_0 n)^2 \end{aligned}$$

 **The function  $J(A, f_0, \Phi)$  is non-quadratic in  $A$  and  $\Phi$** , and thus hard to minimise. To this end, we can transform the multiplicative terms involving  $A$  and  $\Phi$  to new “linear terms”  $\alpha_1 = A \cos \Phi$ ,  $\alpha_2 = A \sin \Phi$

with the inverse mapping  $A = \sqrt{\alpha_1^2 + \alpha_2^2}$  &  $\Phi = \tan^{-1}(\frac{-\alpha_2}{\alpha_1})$


## Example 10: Sinusoidal parameter estimation of three unknown parameters, cont. (see Linear Models in Lecture 4)

For convenience of notation, we shall now introduce the vectors of sampled cos and sin (containing the unknown frequency  $f_0$ ) in the form

$$\mathbf{c} = [1, \cos 2\pi f_0, \dots, \cos 2\pi f_0(N-1)]^T \quad \mathbf{s} = [0, \sin 2\pi f_0, \dots, \sin 2\pi f_0(N-1)]^T$$

to yield the function  $J'(\alpha_1, \alpha_2, f_0)$  which is **quadratic in**  $\alpha = [\alpha_1, \alpha_2]^T$

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

 We arrive at a **linear estimator** of the vector parameter  $\alpha = [\alpha_1, \alpha_2]^T$ , where  $\mathbf{H} = [\mathbf{c} \mid \mathbf{s}]$  (see Example 9 in Lecture 4)

This function can be minimised over  $\alpha$ , exactly as in the linear model (with  $\mathbf{C} = \mathbf{I}$ ), to yield (Slide 33, Lecture 4)

$$\hat{\theta} = \hat{\alpha} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \quad \rightarrow \quad \text{insert into } (*)$$

to yield  $J'(\alpha_1, \alpha_2, f_0) = (\mathbf{x} - \mathbf{H}\hat{\alpha})^T (\mathbf{x} - \mathbf{H}\hat{\alpha}) = \mathbf{x}^T (\mathbf{I} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T) \mathbf{x}$



## Example 10: Sinusoidal parameter estimation of three unknown parameters, cont. cont.

Hence, to find  $\hat{f}_0$  we need to minimise  $J'$  over  $\hat{f}_0$  or, equivalently

$$\text{maximise } \mathbf{x}^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

👉 Using the definition of  $\mathbf{H}$ , the MLE for frequency  $\hat{f}_0$  is **the value that maximises the power spectrum estimate** (see your P&A sets)

$$\begin{bmatrix} \mathbf{c}^T \mathbf{x} \\ \mathbf{s}^T \mathbf{x} \end{bmatrix}^T \begin{bmatrix} \mathbf{c}^T \mathbf{c} & \mathbf{c}^T \mathbf{s} \\ \mathbf{s}^T \mathbf{c} & \mathbf{s}^T \mathbf{s} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}^T \mathbf{x} \\ \mathbf{s}^T \mathbf{x} \end{bmatrix} = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j2\pi f_0 n} \right|^2 \quad \leftarrow \text{periodogram}$$

$\swarrow \mathbf{x}^T \mathbf{H}$        $\swarrow (\mathbf{H}^T \mathbf{H})^{-1}$        $\swarrow \mathbf{H}^T \mathbf{x}$

Use this expression to find  $\hat{f}_0$ , and proceed to find  $\hat{\alpha}$  (Example 9, Lect. 4)

$$\hat{\alpha} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \approx \frac{2}{N} \begin{bmatrix} \mathbf{c}^T \mathbf{x} \\ \mathbf{s}^T \mathbf{x} \end{bmatrix} = \begin{bmatrix} \frac{2}{N} \sum x[n] \cos 2\pi \hat{f}_0 n \\ \frac{2}{N} \sum x[n] \sin 2\pi \hat{f}_0 n \end{bmatrix} \quad \hat{\Phi} = -\arctan \frac{\sum_{n=0}^{N-1} x[n] \sin(2\pi \hat{f}_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi \hat{f}_0 n)}$$

$$\text{and } \hat{A} = \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2} = \frac{2}{N} \left| \sum_{n=0}^{N-1} x[n] \exp(-j2\pi \hat{f}_0 n) \right|$$

# MLE for transformed parameters (invariance property)

This invariance property of MLE is another big advantage of MLE


---

Following the above example, we can now state the **invariance property** of MLE (also valid for the scalar case).

**Theorem (invariance property of MLE):** The MLE of a vector parameter  $\alpha = f(\theta)$ , where the pdf  $p(\mathbf{x}; \theta)$  is parametrised by  $\theta$ , is given by

$$\hat{\alpha} = f(\hat{\theta})$$

where  $\hat{\theta}$  is the MLE of  $\theta$ .

 Since MLE of  $\hat{\theta}$  is obtained by maximising  $p(\mathbf{x}; \theta)$ , if  $f$  is a one-to-one function this is obvious, and the MLE of the transformed parameter is found by substituting the MLE of the original parameter into the transformation.

For example, if  $x[n] = A + w[n]$ ,  $w \in \mathcal{N}(0, \sigma^2)$ , but we wish to find the MLE of  $\alpha = \exp(A)$ . The resulting log-likelihood is still parametrised by  $A$ , and by using  $\ln \alpha = A$  as a transform, the resulting MLE is obtained as

$$\hat{\alpha} = \exp(\hat{A}) \quad \text{see also your P \& A sets}$$

## Theorem: Optimality of MLE for a linear model

---

**Theorem:** Assume that 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  $p$  (tall matrix),  $\boldsymbol{\theta}$  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  $\boldsymbol{\theta}$  takes the form

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

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

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

# 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 estimates
- 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*  $p(\mathbf{x}; \theta)$ ,  $\theta$  is regarded as a variable and not as a parameter!
- **ML estimate:** the value of  $\theta$  that maximises the likelihood funct.  $\ln p(\mathbf{x}; \theta) \rightarrow$  found by different. wrt  $\theta$  and setting to 0

# Motivation and Pro's and Con's of BLUE

---

**Motivation for BLUE:** Except for the Linear Model (Lecture 4), the optimal MVU estimator might:

- Not even exist,
- Be difficult or even impossible to find.



BLUE is one such sub-optimal estimator.

## Idea behind BLUE:

- Restrict the estimate to be **linear in data  $x$** ,
- Restrict the estimate to be **unbiased**,
- Find the **best** among such estimates, **that is, the one with minimum variance**

**Advantages of BLUE:** It needs only the 1st and 2nd moments of PDF (mean and variance)

**Disadvantages of BLUE:** 1) In general it is sub-optimal, and 2) It may be totally inappropriate for some problems (see the next slide).

## 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  $\theta$ , or in other words,  $E\{x[n]\} = s[n]\theta$
- **Counter-example:** if  $E\{x[n]\} = \cos \theta$ , which is not linear in  $\theta$ , 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

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

Since the **unbiased constraint yields:**

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

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

# Appendix: Monte Carlo simulations

Use computer simulations to evaluate performance of any estimation method

---

The MC simulations are illustrated here for a determin. sig.  $s[n, \theta]$  in AWGN

## 1. Data collection

- Select a true parameter value,  $\theta_{true}$  (usually performed over a range of values of  $\theta$ )
- Generate signal having  $\theta_{true}$  as a parameter
- Generate WGN with unit variance and form measurement  $x = s + w$
- Choose  $\sigma$  to obtain the desired SNR value and perform one MC simulation for one SNR value (usually you run many simulations over a range of SNR values)

## 2. Statistical evaluation

- Compute bias,  $B = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}_m - \theta_{true})$
- Compute error RMS,  $RMS = \sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{\theta}_m - \theta_{true})^2}$
- Compute error variance,  $var = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}_m - (\frac{1}{M} \sum_{m=1}^M \hat{\theta}_M))^2$
- Plot histogram or scatter plot (if needed)

## 3. Explore (via plots)

How bias, RMS, variance vary with the value of  $\theta$ , SNR, number of data points,  $N$ , etc.      **Q:** Is bias = 0, is RMS = CRLB<sup>1/2</sup>, etc.

# Lecture supplement: Constrained optimisation using Lagrange multipliers

---

Consider a two-dimensional problem:

$$\begin{array}{ll} \text{maximize} & \underbrace{f(x, y)}_{\text{function to max/min}} \\ \text{subject to} & \underbrace{g(x, y) = c}_{\text{constraint}} \end{array}$$

↪ **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  $\nabla_{x,y}f \parallel \lambda \nabla_{x,y}g$ , where  $\lambda$  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:

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

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



## 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

$F(x, y, z, \lambda) = f(x, y, z) - \lambda(g(x, y, z) - c)$  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  $x, y, z, \lambda$ .

#### Example 10: Economics

Two factories, A and B make TVs, at a cost

$f(x, y) = 6x^2 + 12y^2$  where  $x = \#TV$  in A &  $y = \#TV$  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

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

Then:  $F'_x = 12x - \lambda$ ,  $F'_y = 24y - \lambda$ ,  $F'_\lambda = -x - y + 90$ , 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:

---

○

# Notes:

---

○