Advanced Signal Processing
Introduction to Estimation Theory

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

Aims of this lecture

- To introduce the notions of: Estimator, Estimate, Estimandum
- To discuss the bias and variance in statistical estimation theory, asymptotically unbiased and consistent estimators
- Performance metric: the Mean Square Error (MSE)
- The **bias–variance dilemma** and the MSE in this context
- To derive a feasible MSE estimator
- A class of Minimum Variance Unbiased (MVU) estimators
- Extension to the vector parameter case
- Point estimators, confidence intervals, statistical goodness of an estimator, the role of noise

Role of estimation in signal processing

(try also the function specgram in Matlab)

- An enabling technology in many electronic signal processing systems

1. Radar
2. Sonar
3. Speech
4. Image analysis
5. Biomedicine
6. Communications
7. Control
8. Seismics
9. Almost everywhere ...

- Radar and sonar: range and azimuth

- Image analysis: motion estimation, segmenation

- Speech: features used in recognition and speaker verification

- Seismics: oil reservoirs

- Communications: equalization, symbol detection

- Biomedicine: various applications

Statistical estimation problem

for simplicity, consider a DC level in WGN, $x[n] = A + w[n], \ w \sim \mathcal{N}(0, \sigma^2)$

Problem statement: We seek to determine a set of parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$ from a set of data points $\mathbf{x} = [x[0], \ldots, x[N-1]]^T$ such that the values of these parameters would yield the highest probability of obtaining the observed data. In other words,

$$\max_{span\ \theta} p(\mathbf{x}; \boldsymbol{\theta}) \qquad \text{reads :} \qquad \text{“}p(\mathbf{x}) \quad \text{parametrised by} \quad \theta\text{''}$$

- The unknown parameters may be seen as deterministic or random variables

- There are essentially two alternatives to the statistical case
  - No a priori distribution assumed: Maximum Likelihood
  - A priori distribution known: Bayesian estimation

- Key problem $\looparrowright$ to estimate a group of parameters from a discrete-time signal or dataset.

Estimation of a scalar random variable

Given an N - point dataset $x[0], x[1], \ldots, x[N-1]$ which depends on an unknown parameter θ, (scalar), define an "estimator" as some function, g, of the dataset, that is

$$\hat{\theta} = g(x[0], x[1], \ldots, x[N-1])$$

which may be used to estimate θ (single parameter case).

(in our DC level estimation problem, $\theta = A$)

∘ This defines the problem of "parameter estimation"

∘ Also need to determine $g(\cdot)$

∘ Theory and techniques of statistical estimation are available

∘ Estimation based on PDFs which contain unknown but deterministic parameters is termed **classical estimation**

∘ In **Bayesian estimation**, the unknown parameters are assumed to be random variables, which may be prescribed "a priori" to lie within some range of allowable parameters (or desired performance)

The stastical estimation problem

First step: to model, mathematically, the data

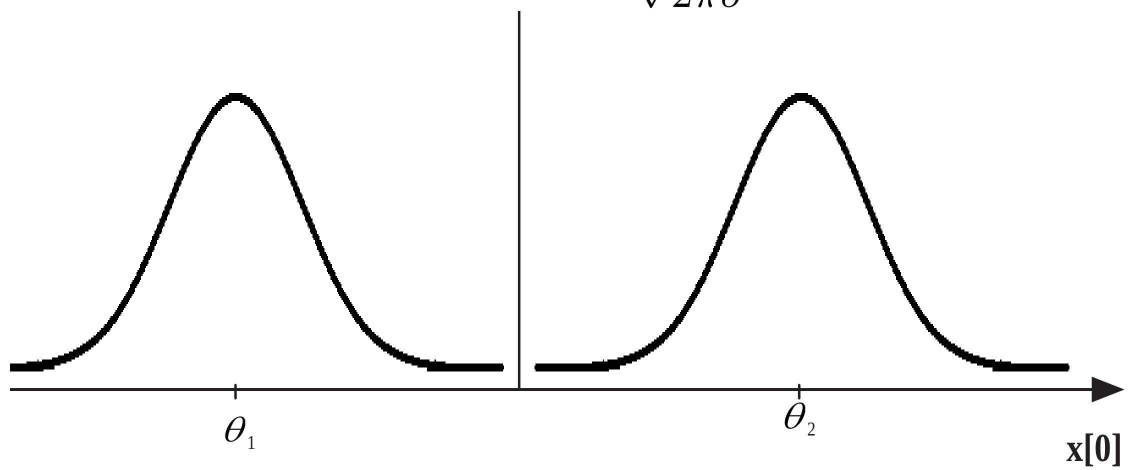
◦ We employ a Probability Desity Function (PDF) to describe the inherently random measurement process, that is

$$p(x[0], x[1], \ldots, x[N-1]; \theta)$$

which is "parametrised" by the unknown parameter θ

Example: for $N = 1$, and θ denoting the mean value, a generic form of PDF for the class of Gaussian PDFs with any value of θ is given by

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



Clearly, the observed value of $x[0]$ impacts upon the likely value of θ.

Estimator vs. Estimate

The parameter to be estimated is then viewed as a **realisation of the random variable** θ

∘ Data are described by the joint PDF of the data and parameters:

$$p(x,\theta) = \underbrace{p(x \mid \theta)}_{(conditional\ \ PDF)} \underbrace{p(\theta)}_{(prior\ \ PDF)}$$

∘ An **estimator is a rule that assigns a value** of θ from each realisation of $\underline{x} = \mathbf{x} = [x[0], \ldots, x[N-1]]^T$

∘ An **estimate** of, i.e. $\hat{\theta}$ (also called 'estimandum') is the **value obtained for a given realisation** of $\mathbf{x} = [x[0], \ldots, x[N-1]]^T$ in the form $\hat{\theta} = g(\mathbf{x})$

Example: for a noisy straight line: $p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A-Bn)^2\right]}$

∘ Performance is critically dependent upon this PDF assumption - the estimator should be robust to slight mismatch between the measurement and the PDF assumption

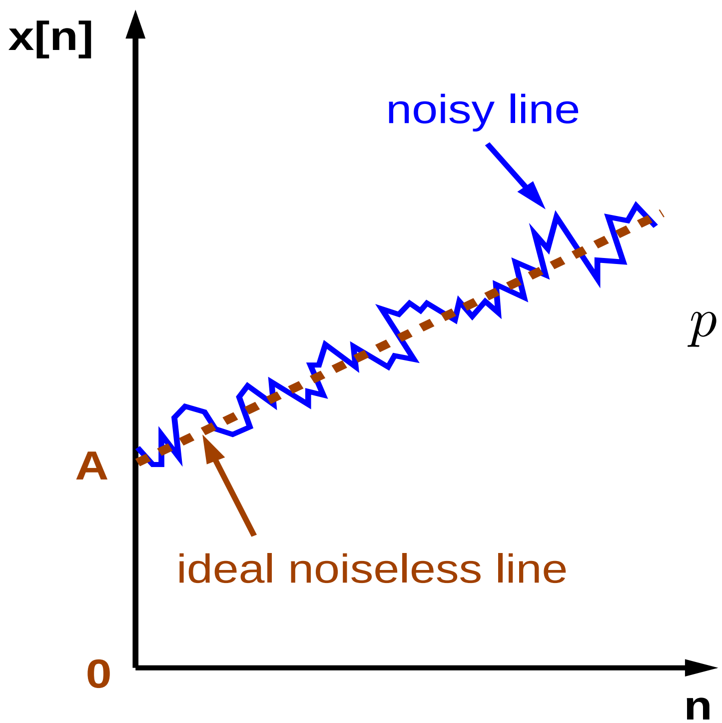
Example: finding the parameters of straight line

Specification of the PDF is critical in determining a good estimator

In practice, we choose a PDF which fits the problem constraints and any "a priori" information; **but it must also be mathematically tractable.**

Example: Assume that "on the average" the data are increasing

Data: straight line embedded in random noise $w[n] \sim \mathcal{N}(0, \sigma^2)$



$$x[n] = A + Bn + w[n]$$

$$n = 0, 1, \ldots, N-1$$

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

Unknown parameters:

$$A,\, B \Leftrightarrow \quad \boldsymbol{\theta} \equiv [A \quad B]^T$$

Careful: what would be the effects of bias in A and B?

Bias in parameter estimation

Estimation theory (scalar case): estimate the value of an unknown parameter, $\hat{\theta}$, from a set of observations of a random variable described by that parameter

$$\hat{\theta} = g\big(x[0], x[1], \ldots, x[N-1]\big)$$

Example: given a set of observations from Gaussian distribution, estimate the mean or variance from these observations.

- Recall that in linear mean square estimation, when estimating a value of random variable y from an observation of a related random variable x, the coefficents a and b in the estimation $y = ax + b$ depend upon the mean and variance of x and y as well as on their correlation.

The difference between the expected value of the estimate and the actual value θ is caleld the *bias* and will be denoted y B.

$$B = E\{\hat{\theta}_N\} - \theta$$

where $\hat{\theta}_N$ denotes estimation over N data samples, $x[0], \ldots, x[N-1]$

Asymptotic unbiasedness

If the bias is zero, then the expected value of the estimate is equal to the true value, that is

$$E\{\hat{\theta}_N\} = \theta \qquad \equiv \qquad B = E\{\hat{\theta}_N\} - \theta = 0$$

and the estimate is said to be **unbiased**.

If $B \neq 0$ then the estimator $\hat{\theta} = g(\mathbf{x})$ is said to be **biased**.

Example: Consider the **sample mean estimator** of the signal $x[n] = A + w[n], \; w \sim \mathcal{N}(0,1)$, given by

$$\hat{A} = \bar{x} = \frac{1}{N+2}\sum_{n=0}^{N-1} x[n] \qquad \text{that is} \qquad \theta = A$$

Is the above sample mean estimator of the true mean A biased?

More often: an estimator is **biased but** bias $B \rightarrow 0$ when $N \rightarrow \infty$

$$\lim_{N\rightarrow\infty} E\{\hat{\theta}_N\} = \theta$$

Such as estimator is said to be **asymptotically unbiased**.

How about the variance?

∘ It is desirable that an estimator be either unbiased or asymptotically unbiased (think about the power of estimation error due to DC offset)

∘ For an estimate to be meaningful, it is necessary that **we use the available statistics effectively**, that is,

$$Var \rightarrow 0 \qquad as \qquad N \rightarrow \infty$$

or in other words

$$\lim_{N\rightarrow\infty} var\{\hat{\theta}_N\} = \lim_{N\rightarrow\infty} \left\{ |\hat{\theta}_N - E\{\hat{\theta}_N\}|^2 \right\} = 0$$

If $\hat{\theta}_N$ is unbiased then $E\{\hat{\theta}_N\} = \theta$, and from Tchebycheff inequality $\forall\, \epsilon > 0$

$$Pr\{|\hat{\theta}_N - \theta| \geq \epsilon\} \leq \frac{var\{\hat{\theta}_N\}}{\epsilon^2}$$

$\Rightarrow$ if $Var \rightarrow 0 \quad as \quad N \rightarrow \infty$, then the probability that $\hat{\theta}_N$ differs by more than ϵ from the true value will go to zero (showing consistency).

In this case, $\hat{\theta}_N$ is said to converge to θ with probability one.

Mean square convergence

Another form of convergence, **stronger** than convergence with probability one is **mean square convergence**.

An estimate $\hat{\theta}_N$ is said to converge to θ in the mean–square sense, if

$$\lim_{N\to\infty} \underbrace{E\{|\hat{\theta}_N - \theta|^2\}}_{\text{mean square error}} = 0$$

- For an unbiased estimator this is equivalent to the previous condition that the variance of the estimate goes to zero

- An estimate is said to be **consistent** if it converges, in some sense, to the true value of the parameter

- We say that the estimator is **consistent** if it is **asymptotically unbiased** and has a **variance that goes to zero as** $N \to \infty$

Example: Assessing the performance of the Sample Mean as an estimator

Consider the estimation of a DC level, A in random noise, which could be modelled as

$$x[n] = A + w[n]$$

$w[n] \sim$ some zero mean random process.

◦ **Aim:** to estimate A given $\{x[0], x[1], \ldots, x[N-1]\}$

◦ Intuitively, the **sample mean** is a reasonable estimator

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

Q1: How close will $\hat{A}$ be to A?

Q2: Are there better estimators than the sample mean?

Mean and variance of the Sample Mean estimator

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

Estimator = f(random data), $\Rightarrow$ a random variable itself

$\Rightarrow$ **its performance must be judged statistically**

(1) What is the mean of $\hat{A}$?

$$E\left\{\hat{A}\right\} = E\left\{\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\} = \frac{1}{N}\sum_{n=0}^{N-1} E\{x[n]\} = A \quad \looparrowright \quad \text{unbiased}$$

(2) What is the variance of $\hat{A}$?

Assumption: The samples of $w[n]$s are uncorrelated

$$E\left\{\hat{A}^2\right\} = Var\left\{\hat{A}\right\} = Var\left\{\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\}$$

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

Notice the variance $\rightarrow 0$ as N $\rightarrow \infty$ $\quad \looparrowright \quad$ **consistent** (see your P&A sets)

Minimum Variance Unbiased (MVU) estimation

Aim: to establish "good" estimators of unknown deterministic parameters

Unbiased estimator $\looparrowright$ "on the average" yields the true value of the unknown parameter independent of its particular value, i.e.

$$E(\hat{\theta}) = \theta \qquad a < \theta < b$$

where (a, b) denotes the range of possible values of θ

Example: Unbiased estimator for a DC level in White Gaussian Noise (WGN). If we are given

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

where A is the unknown, but deterministic, parameter to be estimated which lies within the interval $(-\infty, \infty)$, then the sample mean can be used as an estimator of A, namely

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

Careful: the estimator is parameter dependent!

An estimator may be unbiased for certain values of the unknown parameter but not all, such an estimator is not unbiased

Consider another sample mean estimator:

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

Therefore: $\quad E\left\{\hat{\hat{A}}\right\} = 0 \quad$ when $A = 0 \quad$ **but**

$E\left\{\hat{\hat{A}}\right\} = \frac{A}{2} \quad$ when $A \neq 0 \quad$ **(parameter dependent)**

Hence $\hat{\hat{A}}$ is **not an unbiased estimator**

◦ A biased estimator introduces a **"systemic error"** which should not generally be present

◦ Our goal is to avoid bias if we can, as we are interested in stochastic signal properties and bias is largely deterministic

Remedy

(also look in the Assignment dealing with PSD in your CW)

Several unbiased estimates of the same quantity may be averaged together, i.e. given the L independent estimates

$$\left\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_L\right\}$$

We may choose to average them, to yield

$$\hat{\theta} = \tfrac{1}{L}\textstyle\sum_{l=1}^{L} \hat{\theta}_l$$

Our assumption was that the individual estimators are unbiased, with equal variance, and uncorrelated with one another.

Then **(NB: averaging biased estimators will not remove the bias)**

$$E\left\{\hat{\theta}\right\} = \theta$$

and

$$Var\left\{\hat{\theta}\right\} = \tfrac{1}{L^2}\textstyle\sum_{l=1}^{L} Var\left\{\hat{\theta}_l\right\} = \tfrac{1}{L} Var\left\{\hat{\theta}_l\right\}$$

Note, as $L \rightarrow \infty, \hat{\theta} \rightarrow \theta$ (consistent)

Effects of averaging for real world data

Problem 3.4 from your P/A sets: heart rate estimation

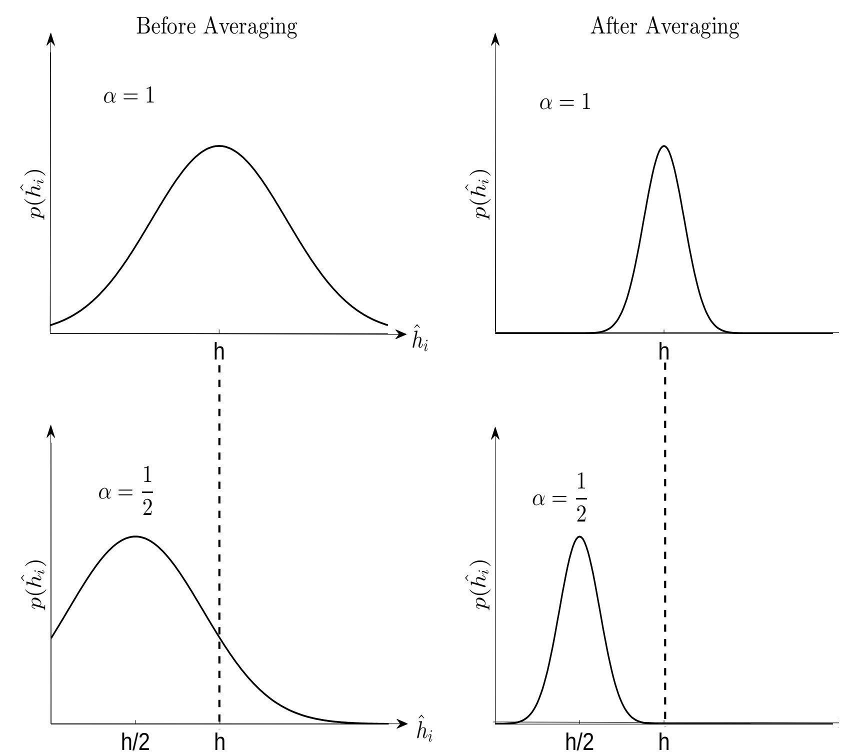
The heart rate, h, of a patient is automatically recorded by a computer every $100ms$. In one second the measurements $\left\{\hat{h_1}, \hat{h_2}, \ldots, \hat{h_{10}}\right\}$ are averaged to obtain $\hat{h}$. Given than $E\left\{\hat{h_i}\right\} = \alpha h$ for some constant α and $var(\hat{h_i}) = 1$ for all i, determine whether averaging improves the estimator if $\alpha = 1$ and $\alpha = 1/2$.

$$\hat{h} = \frac{1}{10}\sum_{i=1}^{10}\hat{h}_i[n],$$

$$E\left\{\hat{h}\right\} = \frac{\alpha}{10}\sum_{i=1}^{10} h = \alpha h$$

If $\alpha = 1$, unbiased, if $\alpha = 1/2$ it will not be unbiased unless the estimator is formed as $\hat{h} = \frac{1}{5}\sum_{i=1}^{10}\hat{h}_i[n]$.

$$var\left\{\hat{h}\right\} = \frac{1}{L^2}\sum_{i=1}^{10} var\left\{\hat{h}_i\right\}$$



Minimum variance criterion

$\Rightarrow$ An optimality criterion is necessary to define an optimal estimator

Mean Square Error (MSE)

$$MSE(\hat{\theta}) = E\left\{\left(\hat{\theta} - \theta\right)^2\right\}$$

measures the average mean squared deviation of the estimator from the true value.

This criterion leads, however, to unrealisable estimators - namely, ones which are not solely a function of the data

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

$$= Var(\hat{\theta}) + E\left\{(\hat{\theta}) - \theta\right\}^2 = Var(\hat{\theta}) + B^2(\hat{\theta})$$

$\Rightarrow$ **MSE = VARIANCE OF THE ESTIMATOR + SQUARED BIAS**

Example: An MSE estimator with 'gain factor'

Consider the following estimator for DC level in WGN

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

Task: Find a which results in minimum MSE

Given

$E\left\{\hat{A}\right\} = aA$ and

$Var(\hat{A}) = \frac{a^2\sigma^2}{N}$

we have

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

Of course, the choice of $a = 1$ removes the mean and minimises the variance

Continued: MSE estimator with 'gain'

But, can we find a analytically? Differentiating with respect to a yields

$$\frac{\partial MSA}{\partial a}(\hat{A}) = \frac{2a\sigma^2}{N} + 2(a-1)A^2$$

and setting the result to zero gives the optimal value

$$a_{opt} = \frac{A^2}{A^2 + \frac{\sigma^2}{N}}$$

but we do not know the value of A

◦ The optimal value depends upon A which is the unknown parameter

◦ Comment - any criterion which depends on the value of the unknown parameter to be found is likely to yield unrealisable estimators

◦ Practically, the minimum MSE estimator needs to be abandoned, and the estimator must be constrained to be unbiased

A counter-example: A little bias can help (but the estimator is difficult to control)

Q: Let $\{y[n]\}, n = 1, \ldots, N$ be iid Gaussian variables $\sim \mathcal{N}(0, \sigma^2)$. Consider the following estimate of σ^2

$$\hat{\sigma}^2 = \frac{\alpha}{N} \sum_{n=1}^{N} y^2[n] \quad \alpha > 1$$

Find α which minimises the MSE of $\hat{\sigma}^2$.

S: It is straightforward to show that $E\{\sigma^2\} = \alpha\sigma^2$ and

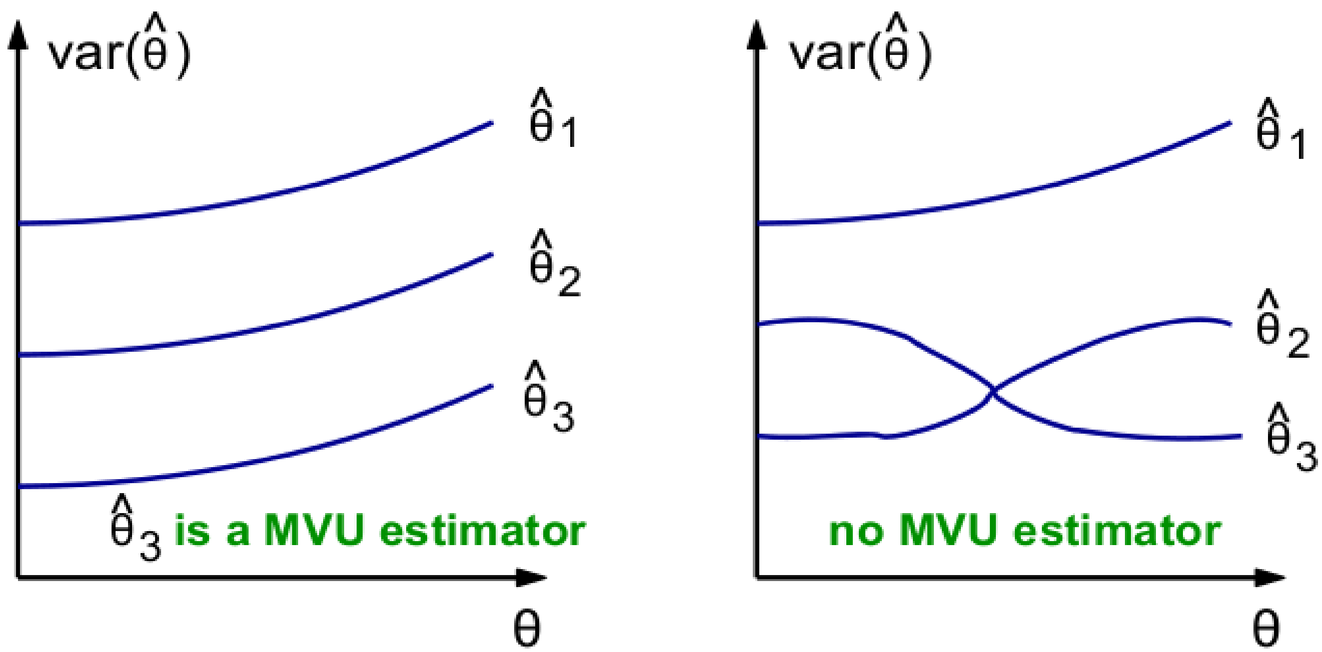
$$\begin{aligned} MSE(\hat{\sigma}^2) &= E\{\left(\hat{\sigma}^2 - \sigma^2\right)^2\} = E\{\hat{\sigma}^4\} + \sigma^4(1 - 2\alpha) \\ &= \frac{\alpha^2}{N^2} \sum_{n=1}^{N} \sum_{s=1}^{N} E\{y^2[n]y^2[s]\} + \sigma^4(1 - 2\alpha) \\ &= \frac{\alpha^2}{N^2} \left(N^2\sigma^4 + 2N\sigma^4\right) + \sigma^4(1 - 2\alpha) = \sigma^4\left[\alpha^4(1 + \frac{2}{N}) + (1 - 2\alpha)\right] \end{aligned}$$

The MMSE is obtained for $\alpha_{min} = \frac{N}{N+2}$ and has the value $\min MSE(\hat{\sigma}^2) = \frac{2\sigma^4}{N+2}$. **Given that the minimum variance of an unbiased estimator (CRLB, later) is $2\sigma^4/N$, this is an example of a biased estimator which obtains a lower MSE than the CRLB.**

Desired: minimum variance unbiased (MVU) estimator

Minimising the variance of an unbiased estimator concentrates the PDF of the error about zero $\Rightarrow$ estimation error is therefore less likely to be large

∘ Existence of the MVU estimator



The MVU estimator is an unbiased estimator with minimum variance for all θ, that is, θ_3 on the graph.

Methods to find the MVU estimator

◦ The MVU estimator **may not always exist**

◦ A **single unbiased estimator may not exist** – in which case a search for the MVU is fruitless!

1. Determine the Cramer-Rao lower bound (CRLB) and find some estimator which satisfies
2. Apply the Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem
3. Restrict the class of estimators to be not only unbiased, but also linear (BLUE)
4. Sequential vs. block estimators
5. Adaptive estimators

Extensions to the vector parameter case

- If $\boldsymbol{\theta} = \left[\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right]^T \in \mathbb{R}^{p\times 1}$ is a vector of unknown parameters, an estimator is unbiased if

$$E(\hat{\theta}_i) = \theta_i \qquad a_i < \theta_i < b_i$$

for $i = 1, 2, \ldots, p$

and by defining

$$E(\boldsymbol{\theta}) = \begin{bmatrix} E(\theta_1) \\ E(\theta_2) \\ \vdots \\ E(\theta_p) \end{bmatrix}$$

an unbiased estimator has the property $\quad E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$

within the $p-$ dimensional space of parameters

- An MVU estimator has the additional property that $Var(\hat{\theta}_i)$ for $i = 1, 2, \ldots, p$ is **minimum among all unbiased estimators**

Summary and food for thoughts

◦ We are now equipped with performance metrics for assessing the goodnes of any estimator (bias, variance, MSE)

◦ Since MSE $= \mathrm{var} + \mathrm{bias}^2$, some biased estimators may yield low MSE. However, we prefer the minimum variance unbiased (MVU) estimators

◦ Even a simple Sample Mean estimator is a very rich example of the advantages of statistical estimators

◦ The knowledge of the parametrised PDF p(data;parameters) is very important for designing efficient estimators

◦ We have introduced statistical "point estimators", would it be useful to also know the "confidence" we have in our point estimate

◦ In many disciplines it is useful to design so called "set membership estimates", where the output of an estimator belongs to a pre-definined bound (range) of values

◦ We will next address linear, best linear unbiased, maximum likelihood, least squares, sequential least squares, and adaptive estimators

Homework: Check another proof for the MSE expression

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= var(\hat{\theta}) + \text{bias}^2(\theta) \\
\text{Note}: \quad var(x) &= E[x^2] - \big[E[x]\big]^2 \qquad (*) \\
\textbf{Idea}: \quad \text{Let} \quad x &= \hat{\theta} - \theta \qquad \rightarrow \quad \text{substitute into} \quad (*) \\
\text{to give} \quad \underbrace{var(\hat{\theta} - \theta)}_{\text{term} \quad (1)} &= \underbrace{E[(\hat{\theta} - \theta)^2]}_{\text{term} \quad (2)} - \underbrace{\big[E[\hat{\theta} - \theta]\big]^2}_{\text{term} \quad (3)} \qquad (**)
\end{aligned}$$

Let us now evaluate these terms:

$$\begin{aligned}
&(1) \qquad var(\hat{\theta} - \theta) = var(\hat{\theta}) \\
&(2) \qquad E[\hat{\theta} - \theta]^2 = \text{MSE} \\
&(3) \qquad \big[E[\hat{\theta} - \theta]\big]^2 = \big[E[\hat{\theta}] - E[\theta]\big]^2 = \big[E[\hat{\theta} - \theta]\big]^2 = \text{bias}^2(\hat{\theta})
\end{aligned}$$

Substitute (1), (2), (3) into (**) to give

$$var(\hat{\theta}) = \text{MSE} - \text{bias}^2 \qquad \Rightarrow \qquad \text{MSE} = var(\hat{\theta}) + \text{bias}^2(\hat{\theta})$$

Recap: Unbiased estimators

Due to the linearity properties of the $E\{\cdot\}$, that is

$$E\{a+b\} = E\{a\} + E\{b\}$$

the sample mean operator can be simply shown to be **unbiased**, i.e.

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

◦ In some applications, the value of A may be constrained to be positive

a component value such as an inductor, capacitor or resistor would be

positive (prior knowledge)

◦ For N data points in random noise, unbiased estimators generally have symmetric PDFs centred about their true value, i.e.

$$\hat{A} \sim \mathcal{N}(A, \sigma^2/N)$$

Notes

Notes