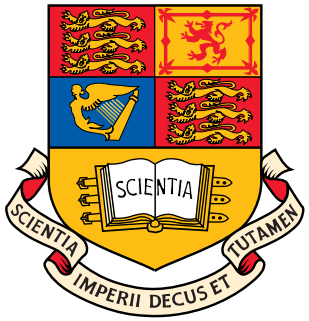
Statistical Signal Processing & Inference

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

- We are seen that it is often the case that data, which may be generated through even most complicated physical signal generating mechanisms, still admit accurate modelling based on only the available historical data
- For example, when concerned with an incredibly complex phenomenon of the generation and number of sunspots, it is sufficient to consider just historical sunspot samples in the task of Sunspot Number Prediction

☞ This highlights the need for a **unifying & rigorous framework** for the assessment of "goodness of performance" of any Data Analytics model, from the simplest "persistent" estimate, to linear ARMA processes, through to nonlinear Neural Network models $\rightsquigarrow$ a subject of this Lecture

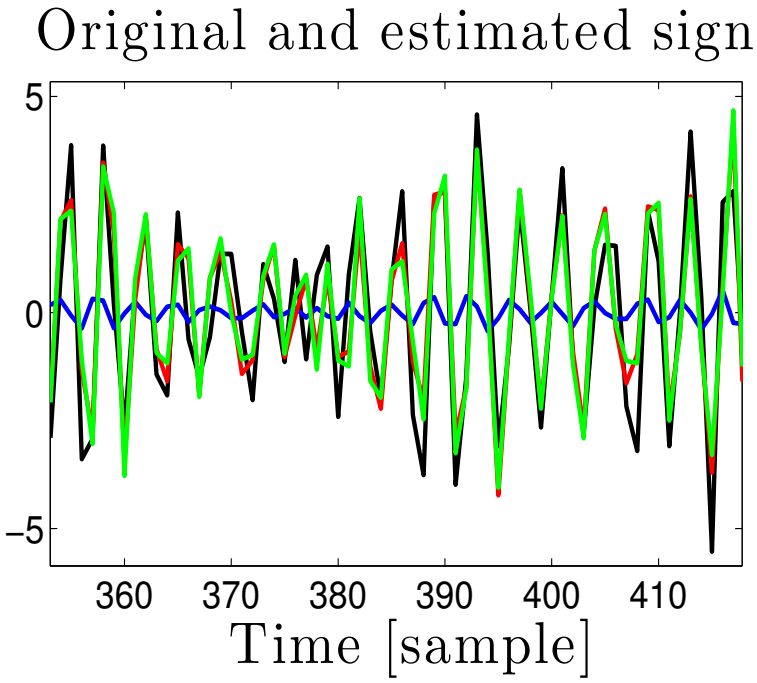
We will typically consider prediction/forecasting scenarios:

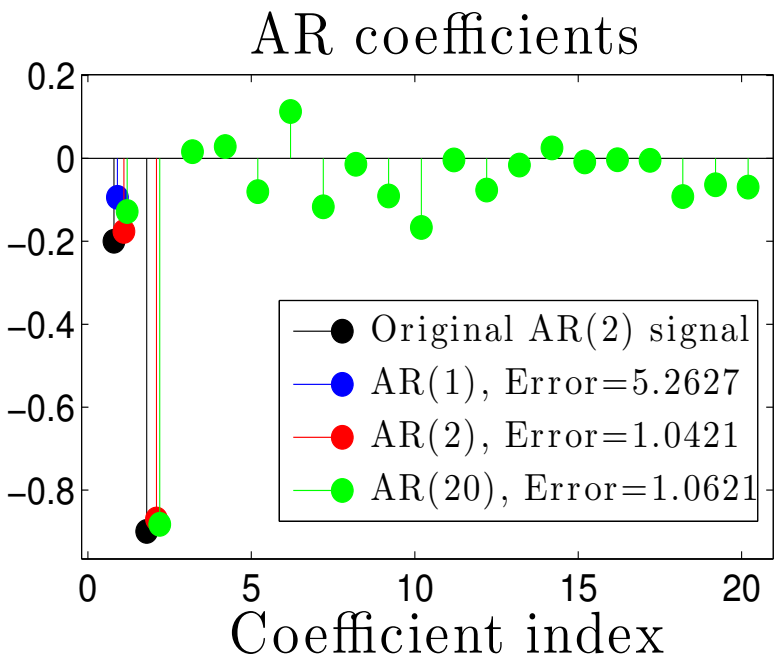
Prediction: Employs an already built model (based on in-sample data, training data) to estimate out-of-sample values (prediction, inference).

Forecasting: A type of prediction which implicitly assumes time-series, where historical data are used to predict future data. Often involves "confidence intervals" (there is 20% chance of rain at 14:00).

Example from Lecture 2: How expressive are these models (under-fitting vs. over-fitting)

Original AR(2) process $x[n] = -0.2x[n-1] - 0.9x[n-2] + w[n]$, $w[n] \sim \mathcal{N}(0,1)$, is estimated using AR(1), AR(2) and AR(20) models.





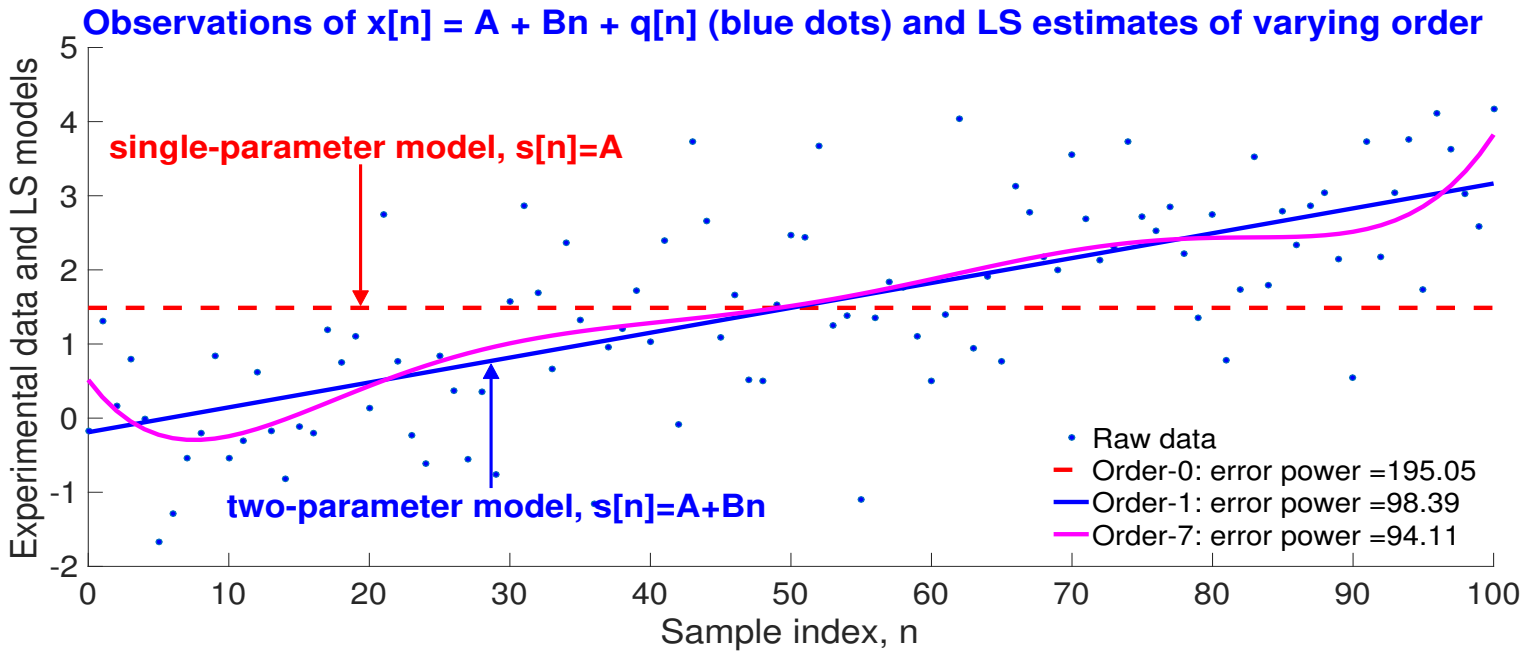
☞ Can we consider this within a bigger "estimation theory" framework?

☞ Can we quantify the "goodness" of an estimator (bias, variance, prediction error, optimality, scalability, sensitivity, sufficient statistics)?

Example from Lecture 6: Method of Least Squares (LS)

Least_Squares_Order_Selection_Ineractive, Animation_Sequential_LS

- The LS estimator of a 'noisy line' $x[n] = s[n; A, B] + w[n]$ is very sensitive to the correct model of the signal of interest, s, as shown in the figure below for the LS fit of $x[n] = A + Bn + w[n]$.
- The error power for fitting the seen (training) data monotonically decreases with the model order.
- The goodness of inference of the higher order model (extrapolation, test data) is however not adequate (overparametrisation, lack of expressivity)



Objectives: Introduction to Estimation Theory

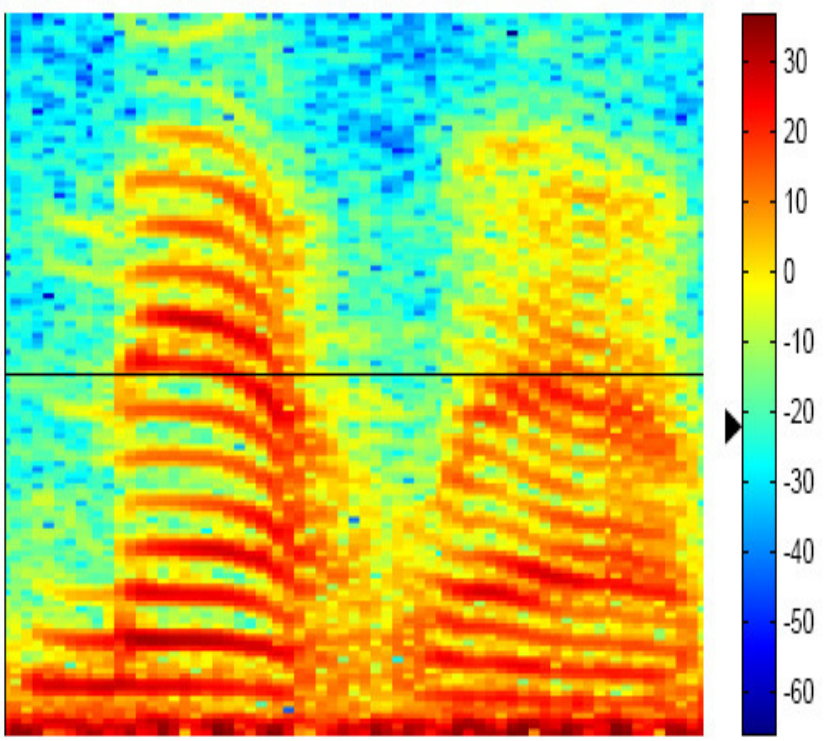
- Notions of an Estimator, Estimate, Estimandum
- The **bias** and **variance** in statistical estimation theory, asymptotically unbiased and consistent estimators
- Performance metrics, such as the Mean Square Error (MSE)
- The **bias–variance dilemma** and the MSE, feasible MSE estimators
- A class of Minimum Variance Unbiased (MVU) estimators, that is, those with the lowest possible variance of all unbiased estimators
- Extension to the vector parameter case
- Statistical goodness of an estimator, the role of noise
- **Enabling technology** for many applications: Radar and sonar (range and azimuth), image analysis (motion estimation), speech (recognition and identification), finance, seismics (oil reservoirs), communications (equalisation, symbol detection), biomedicine (ECG, EEG, respiration)

Discrete–time statistical estimation problem

(try also the function specgram in Matlab ↬ it produces the TF diagram below)

Time–Freq. spectrogram of speech

M aaaa tl aaa b



↑ Frequency → Time

Observe also mathematical artefacts

Consider e.g. the estimation of a fundamental frequency, f_0, of a speaker from this TF spectrogram

Signal $s[n; f_0, \Phi_0]$ is burried in noise

$$x[n] = s[n; f_0, \Phi_0] + w[n]$$

◦ Each time we observe $x[n]$ it contains the desired $s[n]$ but also a different realisation of noise $w[n]$

☞ The estimated frequency, $\hat{f}_0$, and phase, $\hat{\Phi}_0$, are random variables

Our goal: Find an estimator which maps the data $\mathbf{x}$ to the estimates $\hat{f}_0 = g_1(\mathbf{x})$ and $\hat{\Phi}_0 = g_2(\mathbf{x})$.

The RVs $\hat{f}_0, \hat{\Phi}_0$ are best described via a prob. model which depends on: structure of $s[n]$, pdf of $w[n]$, and form of $g(\mathbf{x})$.

Statistical estimation problem (learning from data)

What captures all the necessary statistical information for an estimation problem?

Problem statement: Given an N-point dataset, $x[0], x[1], \ldots, x[N-1]$, which depends on an unknown scalar parameter, θ, an **estimator** is defined as a function, $g(\cdot)$, of the dataset $\{x\}$, that is

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

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

Vector case: Analogously to the scalar case, we seek to determine a set of parameters, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$, from data samples $\mathbf{x} = [x[0], \ldots, x[N-1]]^T$ such that the values of these parameters would yield the highest probability of obtaining the observed data. This can be formalised as

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

There are essentially two alternatives to estimate the unknown θ

- **Classical estimation:** Unknown parameter(s) is deterministic with no means to include *a priori* information about θ (minimum var., ML, LS)
- **Bayesian estimation:** Parameter θ is a random variable, which allows us to use *prior* knowledge on θ (Wiener and Kalman filters, adaptive SP)

The need for a PDF of the data, parametrised by θ

(really, just re–phrasing the previous slide)

Mathematical statement of the general estimation problem:

From the measured data $\mathbf{x} = \big[x[0], x[1], \ldots, x[N-1]\big]^T$

↑ an N-dimensional random vector

Find the unknown (vector) parameter $\theta = [\theta_1, \theta_2, \ldots, \theta_p]^T$

↑ $\boldsymbol{\theta}$ is not random

Q: What captures all the statistics needed for successful estimation of $\boldsymbol{\theta}$

A: It has to be the N-dimensional PDF of the data, **parametrised by $\boldsymbol{\theta}$**

So, it is $p(\mathbf{x}\,;\boldsymbol{\theta})$ that contains all the information needed

↑ we will use $p(\mathbf{x}\,;\boldsymbol{\theta})$ to find $\hat{\boldsymbol{\theta}} = g(\mathbf{x})$

☞ When we know this PDF, we can design optimal estimators

☞ In practice, this PDF is not given, and our goal is to choose a model which:

- ◦ Captures the essence of the signal generating physical model;
- ◦ Leads to a mathematically tractable form of an estimator.

Random variable (RV), some general observations

Random variable $\looparrowright$ quantifies the outcome of a random event.

For example, "heads" or "tails" on a coin or a blue square on Rubik's cube are not random variables per se, but can be made random variables *through numerical characterisation*.

☞ We therefore do not know how to determine the value of a RV, but can specify the probability of occurrence of a certain value of a RV.

A random var. X with the pdf

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is called a Gaussian RV.

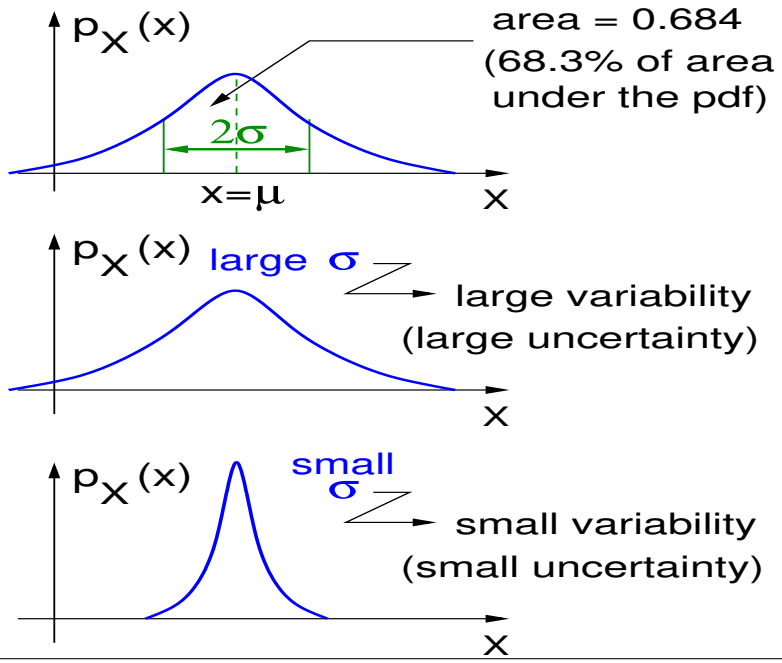
$\rightarrow$ μ is the mean of a RV X

$\rightarrow$ σ is the standard deviation of a RV X, and $\sigma > 0$

$\rightarrow$ σ^2 is the variance of a RV X

So, we can write $X \sim \mathcal{N}(\mu, \sigma^2)$

Variance effect on Gaussian pdf



Conditional pdf

"slice and normalise" the joint pdf $p(x, y)$

Formal definition

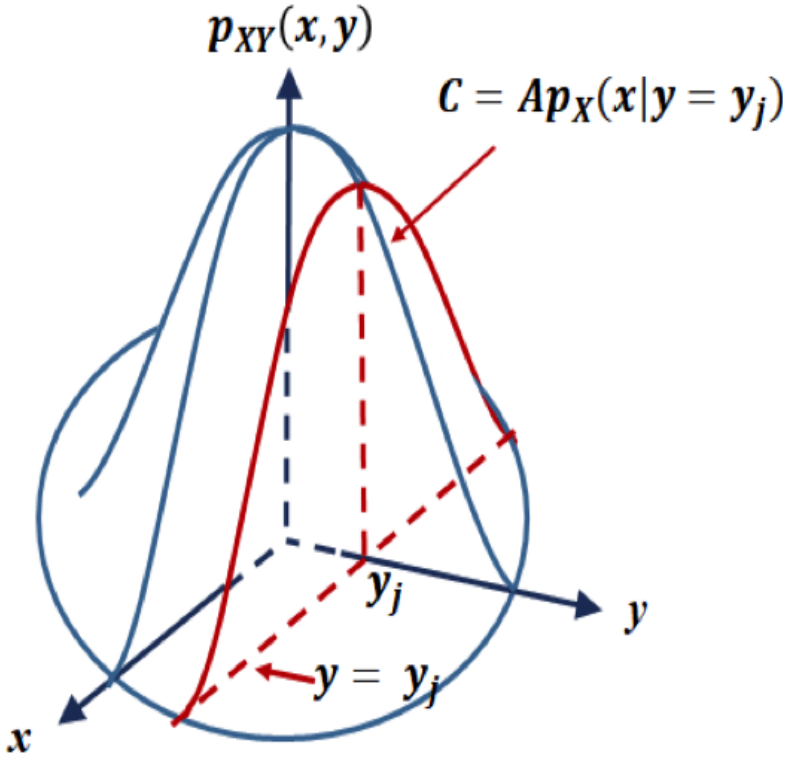
$$p_{Y|X}(y|x) = \begin{cases} \frac{p_{XY}(x,y)}{p_X(x)}, & p_X(x) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

x is held fixed

or more often

$$p(x|y) = \begin{cases} \frac{p(x,y)}{p(y)}, & p(y) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

y is held fixed



Conditional pdf $p(x|y)$

Depends on joint pdf $p(x, y)$ because there are two rand. variables, x and y.

Example: Length of holidays, X, conditioned on the salary $Y =$£60k?

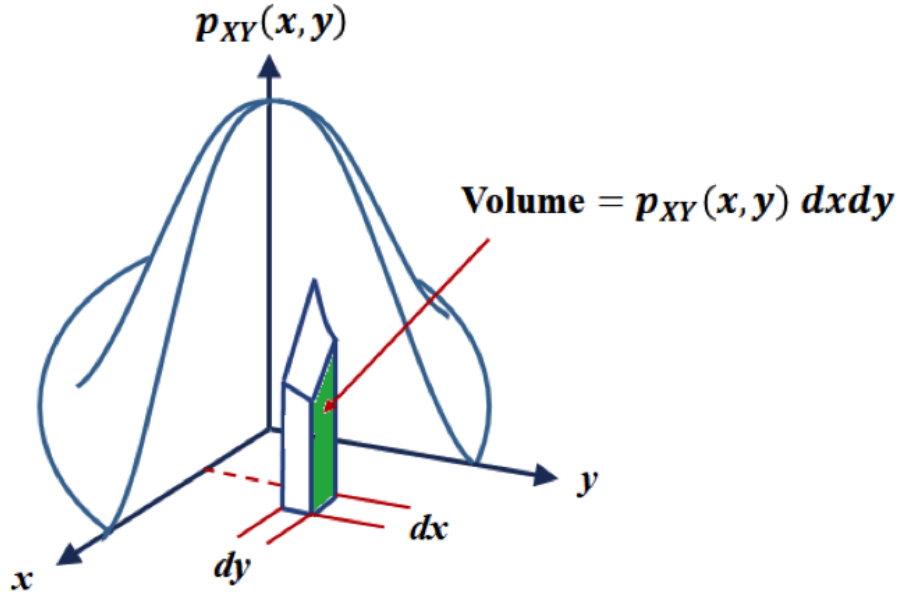
Ans: Find all people who make exactly £60k, how is holiday length distributed?

We therefore:

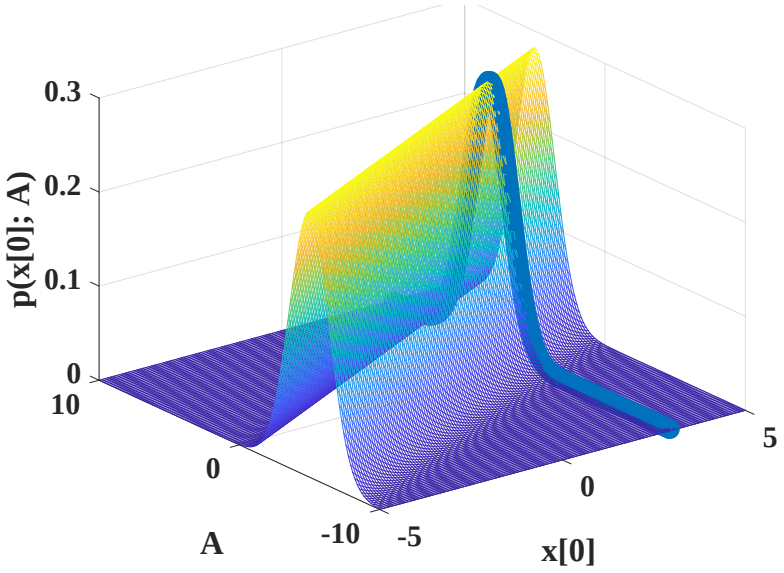
- **slice** the joint $p(x, y)$ at $Y = £60k$
- **normalise** by $p_Y(60,000)$ so that $p(x|y)=p(x, 60k)/p_Y(60k)$ is valid pdf

Joint pdf $p_{XY}(x,y)$ versus parametrised pdf $p(\mathbf{x};\boldsymbol{\theta})$
We will use $p(\mathbf{x};\boldsymbol{\theta})$ to find $\hat{\boldsymbol{\theta}} = g(\mathbf{x})$

Joint pdf $p(x,y)$

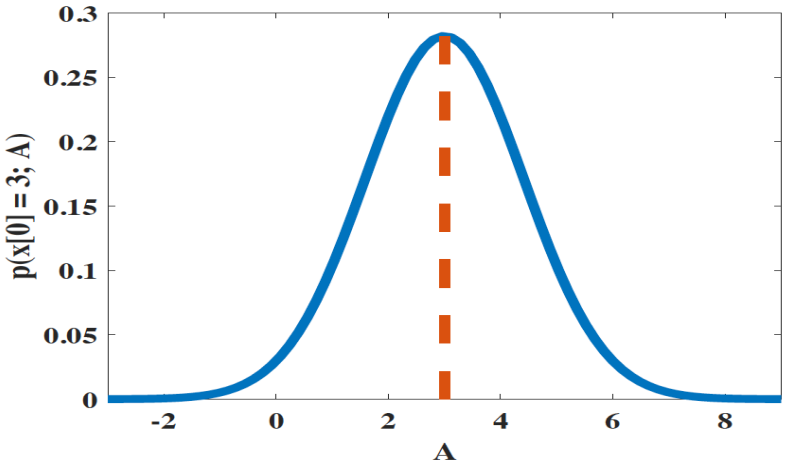


Parametrised pdf $p(x[0];A)$



The parametrised $p(\mathbf{x};\boldsymbol{\theta})$ should be looked at as a function of $\boldsymbol{\theta}$ for a fixed value of observed data $\mathbf{x}$

Right: For $x[0]=A+w[0]$, if we observe $x[0]=3$, then $p(x[0]=3;A)$ is a slice of the parametrised $p(x[0];A)$ for a fixed $x[0]=3$.



The statistical estimation problem

First step: To model, mathematically, the data

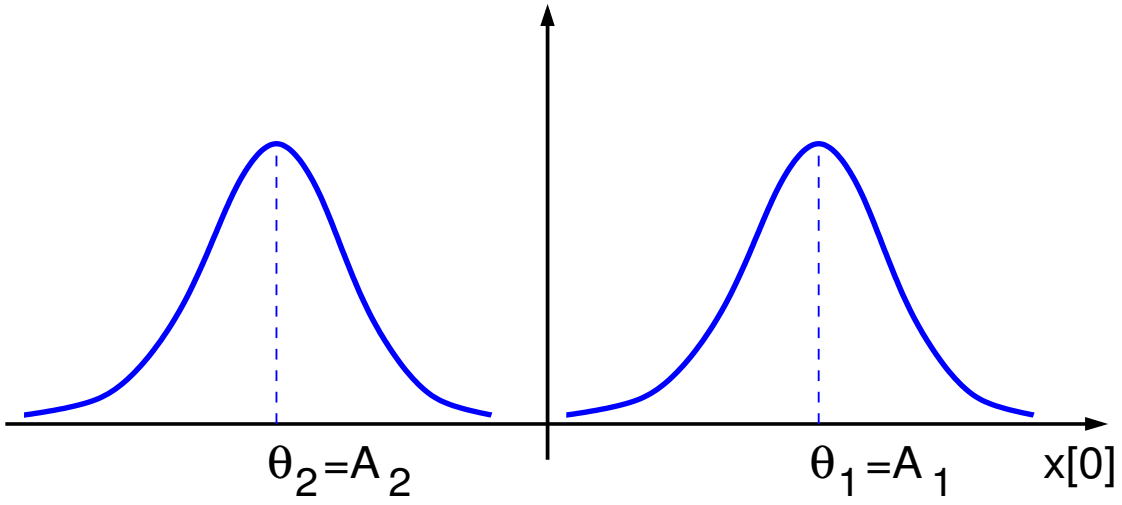
Consider a single observation of a DC level, A, in WGN, w (that is, $\theta = A$)

$$x[0] = A + w[0] \qquad \text{where} \quad w[0] \sim \mathcal{N}(0, \sigma^2). \quad \text{Then} \quad x[0] \sim \mathcal{N}(A, \sigma^2).$$

The "parametrised" pdf, $p(x[0]; \theta) = p(x[0]; A)$, is obviously Gaussian with the mean of $A \rightsquigarrow$ parametrisation affects the mean of $p(x[0]; A)$.

Example 1: For $N = 1$, and with θ denoting the mean value, a generic form of $p(x[0]; \theta)$ for the class of Gaussian parametrised PDFs is given by

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



☞ Clearly, the observed value of $x[0]$ critically impacts upon the likely value of the parameter θ (here, the DC level A)

Estimator vs. Estimate

specification of the PDF is critical to determining a good estimator

An estimator is a rule, $g(\mathbf{x})$, that assigns a value to the parameter θ from each realisation of $\underline{x} = \mathbf{x} = [x[0], \ldots, x[N-1]]^T$.

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

☞ Upon establishing the parametrised $p(x;\theta)$, the estimate $\hat{\theta} = g(\mathbf{x})$ itself is then viewed as a **random variable** and has a pdf of its own, $p(\hat{\theta})$.

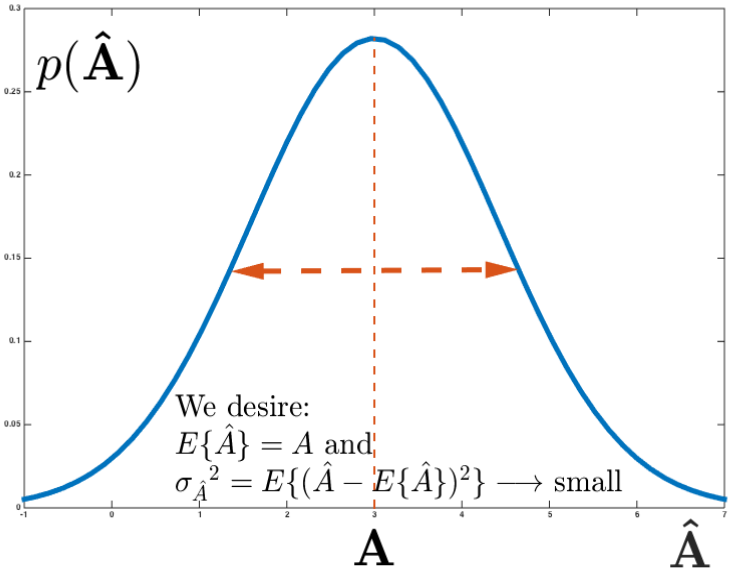
Example 2: Estimate a DC level A in WGN.

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

- The mean of $p(\hat{A})$ measures the centroid
- The variance of $p(\hat{A})$ measures the spread of the pdf around the centroid

PDF concentration ↑ $\Longrightarrow$ Accuracy ↑

This pdf displays the quality of performance.

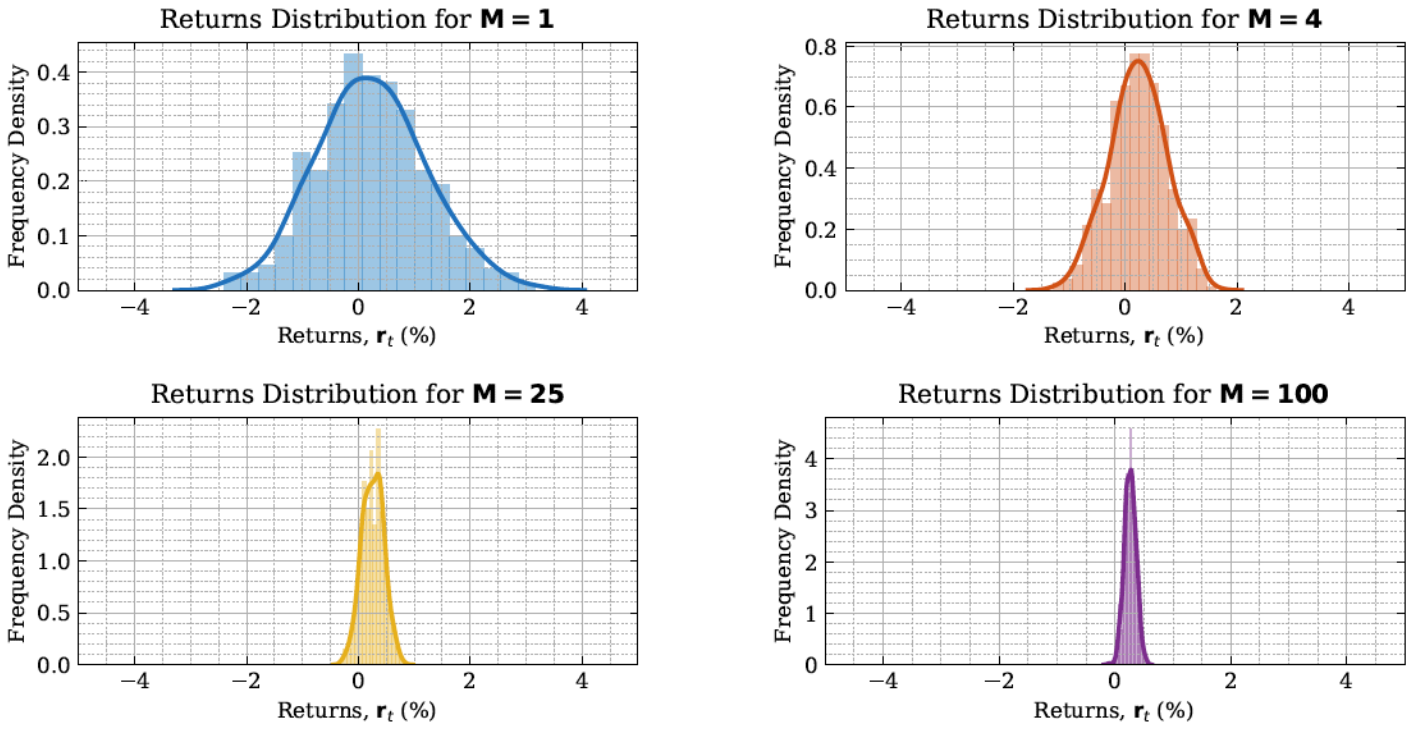


Example 3a: Need for pdf concentration in real–world

An application of risk assessment in finance

In finance, the risk of investment is minimised through "diversification", that is, the investment in many assets (a portfolio) as opposed to single–asset investment. Here, "returns", $R_t = p_t/p_{t-1}$, with p a price.

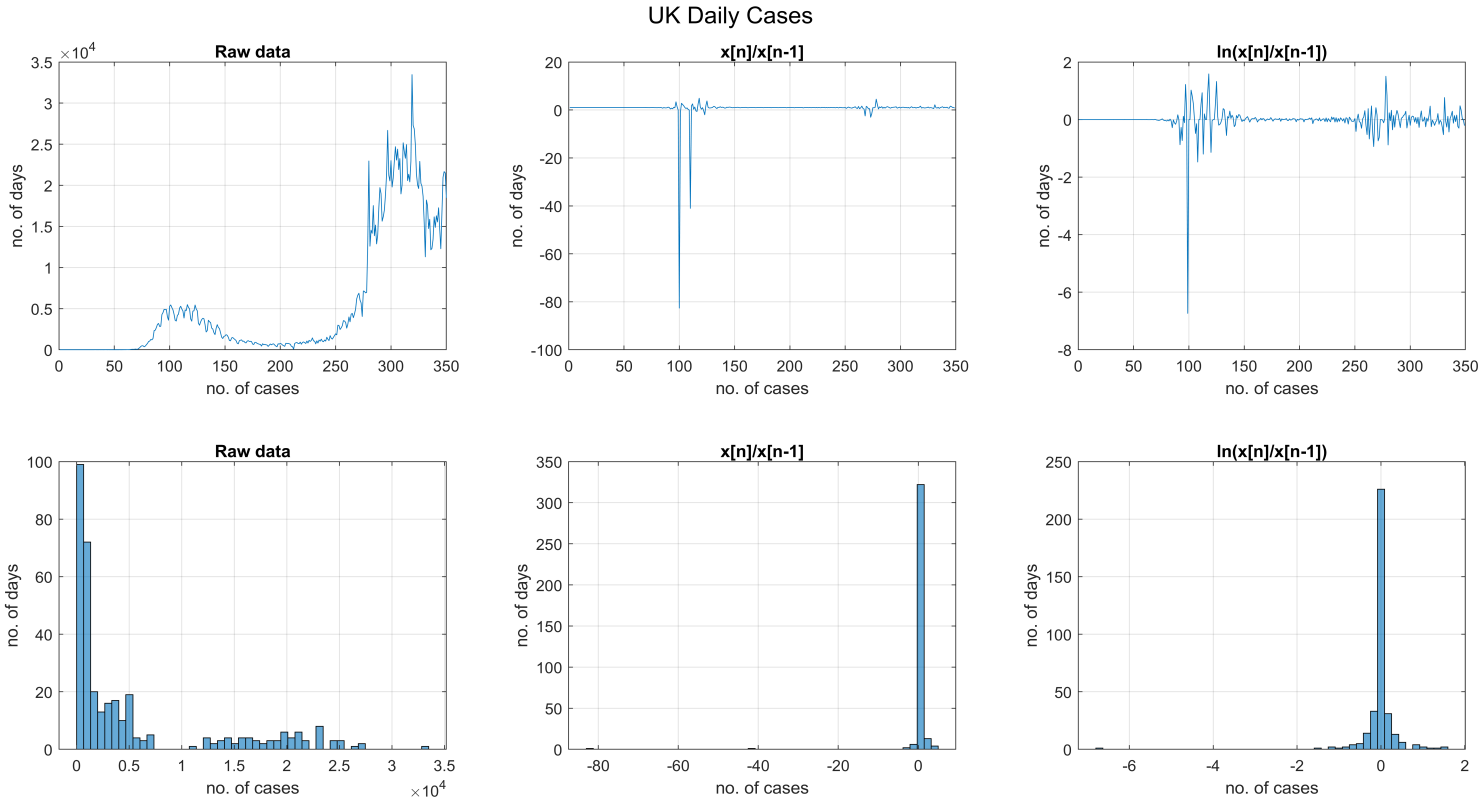
- Risk for a single asset and a number of uncorrelated portfolios. Risk is represented by the standard deviation or the width of the distribution curves, illustrating that a large portfolio ($M = 100$) can be significantly less risky than a single asset ($M = 1$).



Example 3a (contd): Justification for a Gaussian pdf

the closer a pdf to a Gaussian one the more appropriate a 2nd order linear model

An application to COVID-19 infection rate prediction in the UK



- The raw data have asymmetric distribution (non–Gaussian)
- The x[n]/x[n-1] transformed data have a more concentrated distribution
- The ln(x[n]/x[n-1]) transformation $\rightarrow$ pdf is closer to a Gaussian one

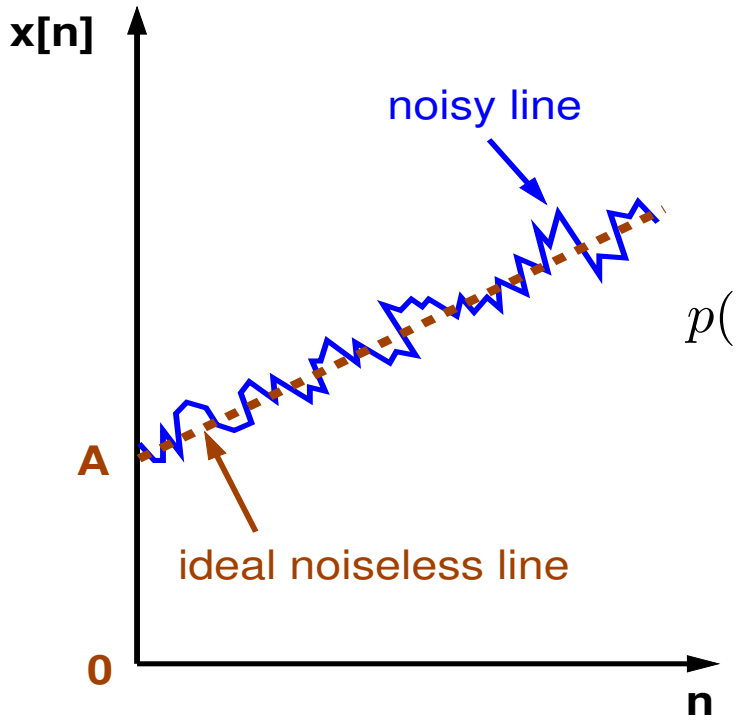
Example 3b: Finding the parameters of a straight line

recall that we have $n = 0, \ldots, N-1$ observed points in the vector $\mathbf{x}$

In practice, the chosen PDF should fit the problem set–up and incorporate any "prior" information; **it must also be mathematically tractable.**

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

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



$$x[n] = A + Bn + w[n]$$
$$= s[n; A, B] + w[n]$$

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

Unknown parameters:

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

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

Bias in parameter estimation

Our goal: Estimate the value of an unknown parameter, θ, from a set of observations of a random variable described by that parameter

$$\hat{\theta} = g\big(x[0], x[1], \ldots, x[N-1]\big) \qquad (\hat{\theta} \text{ is a RV too})$$

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

- Recall that in linear mean square estimation, when estimating the value of a random variable y from an observation of a related random variable x, the coefficients A and B within the estimator $y = Ax + B$ 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, $\hat{\theta}$, and the actual value, θ, is called the *bias* and will be denoted by B.

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

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

Example 4: When estimating a DC level in noise, $x[n] = A + w[n]$, the estimator, $\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} |\, x[n] \,|$, is **biased for A < 0.** (see Appendix)

Now that we have a statistical estimation set–up how do we measure "goodness" of the estimate?

☞ Noise $\mathbf{w}$ is usually assumed white with i.i.d. samples (independent, identically distributed)

⇝ whiteness often does not hold in real–world scenarios

⇝ Gaussianity is more realistic, due to validity of Central Limit Theorem

⇝ zero–mean noise is a nearly universal assumption, and it is realistic since

$$w[n] = w_{zm}[n] + \mu$$

non–zero–mean noise ↑ ↑ zero–mean–noise μ is the mean

Good news: We can use these assumptions to find a bound on the performance of "optimal" estimators.

More good news: Then, the performance of any practical estimator and for any noise statistics will be bounded by that theoretical bound!

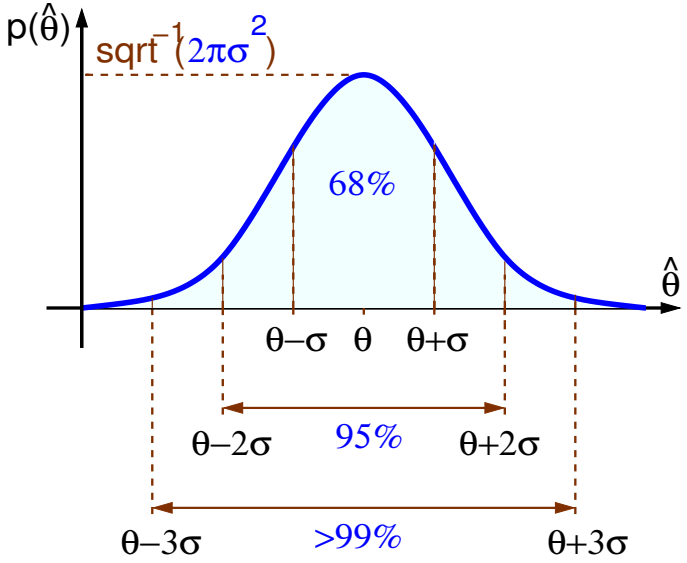
◦ Variance of noise does not always have to be known to make an estimate

◦ But, we must have tools to assess the "goodness" of the estimate

◦ Usually, the goodness analysis is a function of noise variance, σ_w^2, expressed in terms of SNR = signal to noise ratio. (noise sets SNR level)

Assessing the performance of an estimator

Recall that the estimate $\hat{\theta} = g(\mathbf{x})$ is a random variable. As such, it has a pdf of its own, and this pdf completely depicts the quality of the estimate.



We can only assess performance when the value of θ is known

The quality (goodness) of an estimator is typically captured through the **mean** and **variance** of $\hat{\theta} = g(\mathbf{x})$.

We desire: $\mu_{\hat{\theta}} = E\{\hat{\theta}\} = \theta$

and $\sigma^2_{\hat{\theta}} = E\{(\hat{\theta} - E\{\hat{\theta}\})^2\}$ $\looparrowright$ small

◦ In an ideal scenario, we would like to always be able to theoretically analyse the problem to assess its goodness (bias and variance). This also shows how performance depends on problem specification.

◦ Sometimes, we have to make use of simulations: i) to verify theoretical analysis, ii) if theoretical results cannot be found.

An equivalent assessment via the estimation error

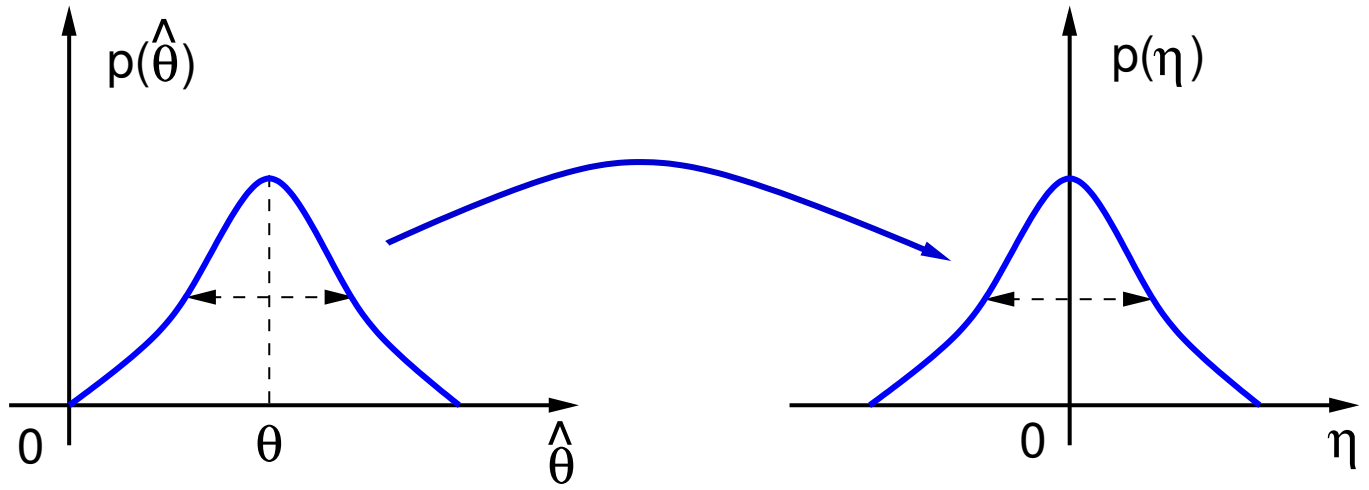
Since $\hat{\theta}$ is a RV, it has a PDF of its own (more in the next lecture on CRLB)

Given that $\hat{\theta} = g(\mathbf{x})$ then $\hat{\theta} = \theta + \eta$ (η is the estimation error)

$\hat{\theta}$: random variable (RV) ↑; θ: not random variable ↑; η: ↑ random variable

Since $\hat{\theta}$ is a random variable (RV), the estimation error η is also a RV

$$\eta = \hat{\theta} - \theta \quad \Longrightarrow \quad \eta = 0 \text{ indicates an unbiased estimator}$$



☞ Quality of the estimator is completely described by the error PDF $p(\eta)$

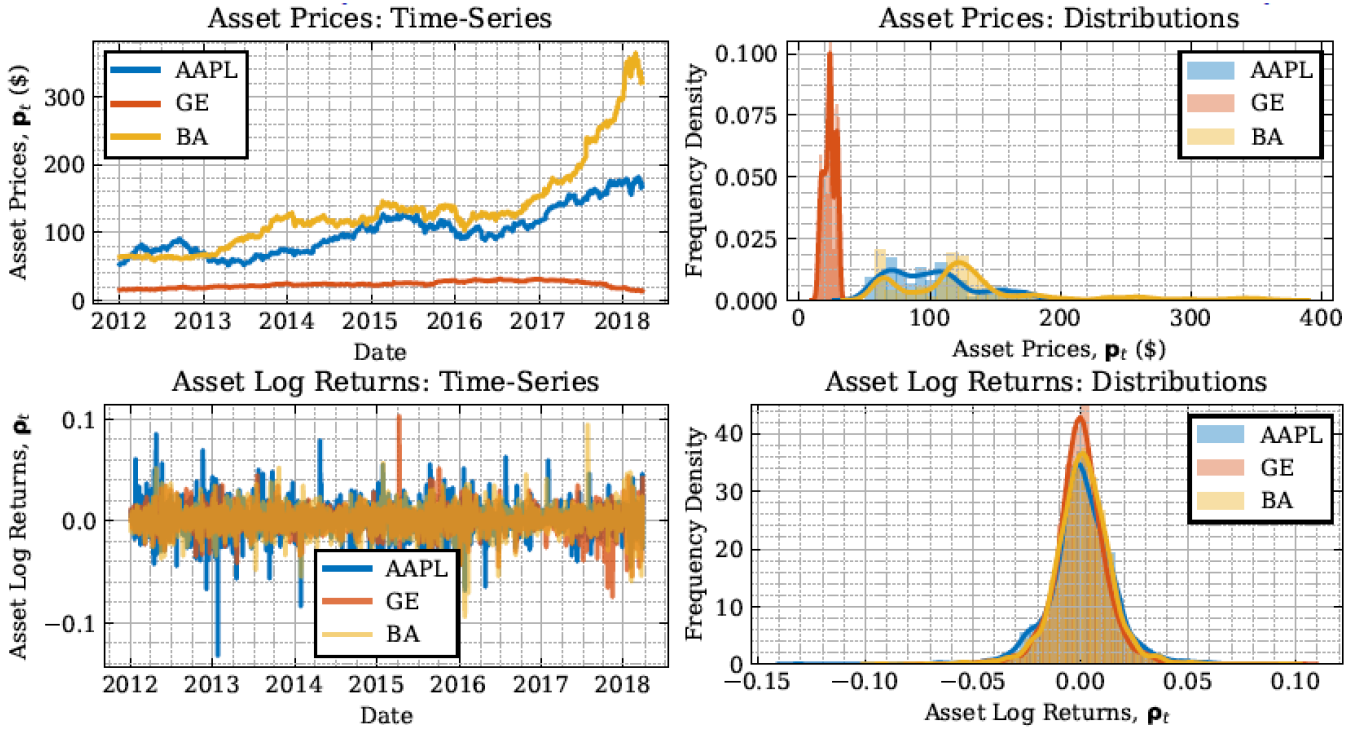
We desire: 1) an unbiased estimator, that is, $E\{\eta\} = 0$

2) minimum variance, $var(\eta) = E\{(\eta - E\{\eta\})^2\} \longrightarrow$ small

Can we resort to (approximately) Gaussian distribution?

Yes, very often, if we re–cast our problem in an appropriate way

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



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

Asymptotic unbiasedness

If the bias is zero, then for sufficiently many observations of $x[n]$ (N large), the expected value of the estimate, $\hat{\theta}$, is equal to its 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 5: Consider the **sample mean estimator** of a DC level, A, in WGN, $x[n] = A + w[n],\ w \sim \mathcal{N}(0,1)$, 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?

Observe: This estimator is **biased but** the bias $B \to 0$ when $N \to \infty$

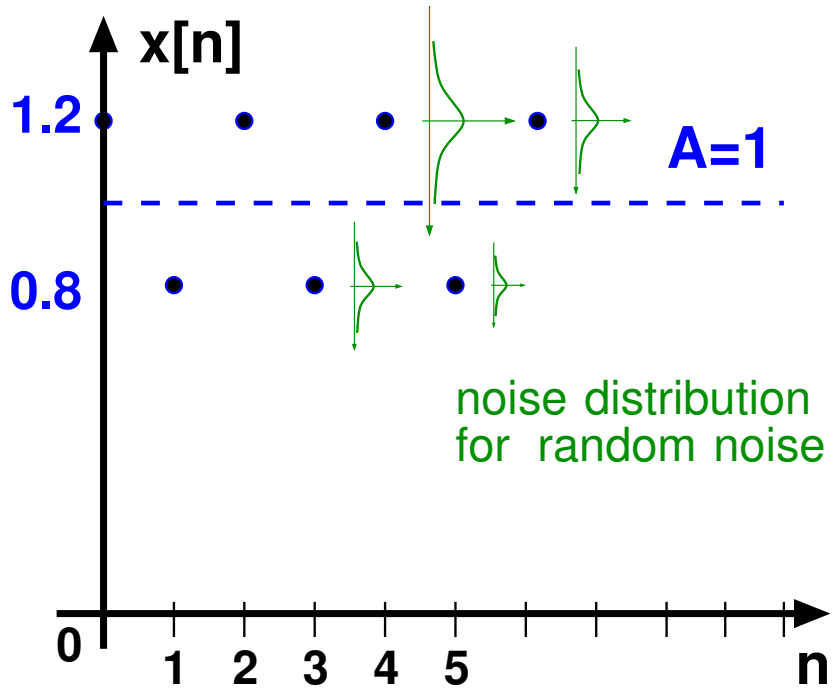
$$\lim_{N\to\infty} E\{\hat{\theta}_N\} = \theta$$

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

Example 6: Asymptotically unbiased estimator of DC level in noise

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

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



For "deterministic" noise where $w[n] \in \{-0.2, 0.2\}$

$$\hat{A}_1 = \frac{1}{1+2}\, 1.2 = 0.4$$
$$\hat{A}_2 = \frac{1}{2+2}\, (1.2 + 0.8) = 0.5$$
$$\hat{A}_3 = \frac{1}{3+2}\, 3.2 = 0.64$$
$$\vdots \quad \vdots \quad \vdots$$
$$\hat{A}_8 = \frac{1}{8+2}\, 8 = 0.8$$
$$\vdots \quad \vdots \quad \vdots$$
$$\hat{A}_{100} = \frac{1}{100+2}\, 100 = 0.98$$

How about the variance of the estimate $\hat{\theta}$

◦ 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,

$$var(\hat{\theta}) \to 0 \qquad as \qquad N \to \infty$$

or in other words

$$\lim_{N\to\infty} var\{\hat{\theta}_N\} = \lim_{N\to\infty} \Big\{ E\big[\big|\hat{\theta}_N - E\{\hat{\theta}_N\}\big|^2\big] \Big\} = 0$$

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

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

☞ If $var\{\hat{\theta}_N\} \to 0 \quad as \quad N \to \infty$, 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. (see Appendix)

Mean square convergence

NB: Mean square error criterion is very different from the variance criterion

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

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

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

- This is different from the previous slide, as θ **is now assumed to be known**, in order to be able to measure the performance
- **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, or more formally:

☞ 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 7: Assessing the performance of the Sample Mean as a statistical estimator

Consider the estimation of a DC level, A, in random noise, $w[n]$, whereby the measured signal, $x[n]$, can be modelled as

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

where $w[n] \sim$ some zero-mean random i.i.d. process.

Goal: To estimate DC level, A, from the data $\{x[0], x[1], \ldots, x[N-1]\}$

◦ Intuitively, the **sample mean** is a reasonable estimator, and has the form

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

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

Q2: Are there better estimators than the sample mean?

Example 7 (contd.): Mean and variance of the Sample Mean estimator

Estimator = f(random data) $\quad\leadsto\quad$ it is a random variable itself

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

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

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

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

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

$$var\left\{\hat{A}\right\} = \underbrace{E\{\left[\hat{A} - E\{\hat{A}\}\right]^2\}}_{\text{variability around the mean}} = var\left\{\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\}$$

$$= \frac{1}{N^2}\sum_{n=0}^{N-1} var\left\{x[n]\right\} = \frac{1}{N^2} N\,\sigma^2 = \frac{\sigma^2}{N} \qquad \text{(as noise is white i.i.d.)}$$

Since $var\{\hat{A}\} \to 0$ as $N \to \infty$ $\quad\leadsto\quad$ **consistent estimator** (see P&A sets)

Some intricacies which are often not fully spelled–out

In our example, each random data sample has the **same mean**, namely A

probability theory ↑

and the mean, A, is exactly the quantity we are trying to estimate

☞ We are estimating A using the **sample mean**, $\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$

statistics ↑

◦ We desire to be always able to perform theoretical analysis to find the bias and variance of the estimator (measure of its goodness)

⇝ theoretical results show how estimates depend on problem spec.

◦ Sometimes it is necessary to make use of simulations

⇝ to verify correctness of theoretical results

⇝ when we cannot find theoretical results (e.g. Monte Carlo simulations)

⇝ when estimators have no optimality properties, but do work in practice

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, independently 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 8: Consider an unbiased estimator for a DC level in white Gaussian noise (WGN), observed as

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

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 for all values; such an estimator is biased

Example 9: Consider another sample mean estimator of a DC level:

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

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

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

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

◦ A biased estimator introduces a **"systemic error"** which should not be present if at all possible

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

Effects of averaging for real world data

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

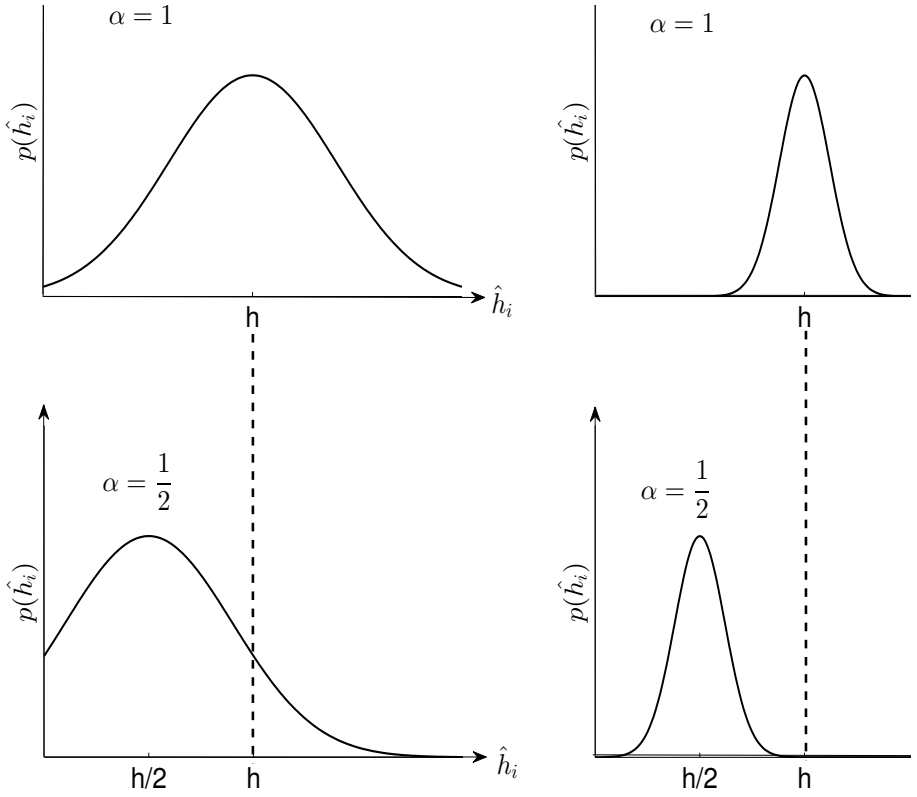
The heart rate, h, of a patient is automatically recorded by a computer every $100\,ms$. One second of the measurements $\left\{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{10}\right\}$ are averaged to obtain $\hat{h}$. Given that $E\left\{\hat{h}_i\right\} = \alpha h$ for some constant α and $var(\hat{h}_i) = 1$ for all i, determine whether averaging improves this estimator, for $\alpha = 1$ and $\alpha = 1/2$.

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

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

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

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



Remedy: How about averaging? Averaging data segments vs averaging estimators? Also look in your CW Assignment dealing with PSD.

Several unbiased estimators of the same quantity may be averaged together. For example, 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

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

Our assumption was that the individual estimates, $\hat{\theta}_l = g(\mathbf{x})$, are unbiased, with equal variances, and mutually uncorrelated.

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

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

and

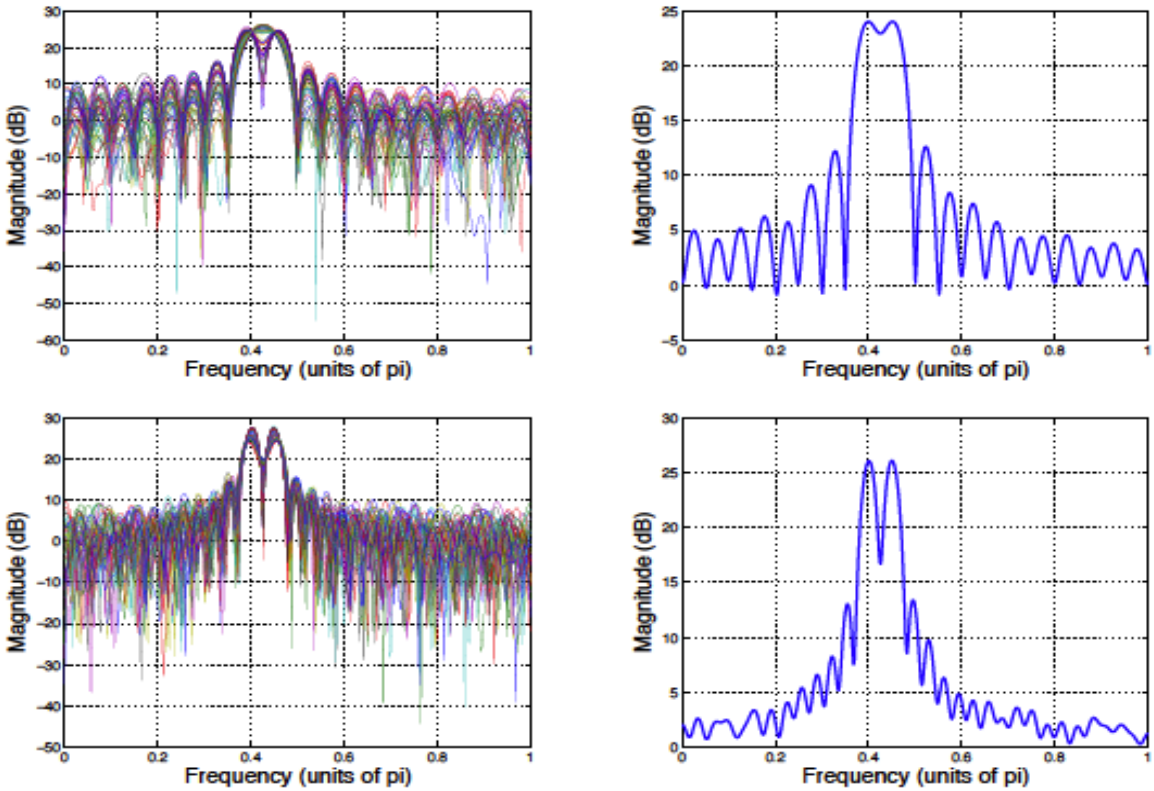
$$var\left\{\hat{\theta}\right\} = \frac{1}{L^2}\sum_{l=1}^{L} var\left\{\hat{\theta}_l\right\} = \frac{1}{L} var\left\{\hat{\theta}_l\right\}$$

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

Example 10: Effect of averaging in spectral estimation

Averaging power spectra of 50 independent realisations of a mixture of two sinewaves in noise, $x[n] = sin(0.4\pi n) + sin(0.45\pi n) + w[n]$

N=40: Overlay of 50 periodograms periodogram average



N=64: Overlay of 50 periodograms periodogram average

Mean square error criterion & bias – variance dilemma

☞ An optimality criterion is necessary to define an optimal estimator

One such natural criterion is the **Mean Square Error (MSE)**, given by

$$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\} \qquad \looparrowright \qquad E\{error^2\} = \text{error power}$$

which measures the average mean squared deviation of the estimate, $\hat{\theta}$, from the true value (error power).

$$MSE\,(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\} = E\{\left[(\hat{\theta} - E\{\hat{\theta}\}) + (\underbrace{E\{\hat{\theta}\} - \theta}_{=bias,\ B(\hat{\theta})})\right]^2\}$$

$$= E\left\{\left[\hat{\theta} - E\{\hat{\theta}\}\right]^2\right\} + 2\,B(\hat{\theta})\,\underbrace{E\{\hat{\theta} - E\{\hat{\theta}\}\}}_{=0} + B^2(\hat{\theta})$$

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

MSE = VARIANCE OF THE ESTIMATOR + SQUARED BIAS

Example 11: An MSE estimator with a 'gain factor' (motivation for unbiased estimators)

Consider the following estimator for DC level in WGN

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

Task: Find the value of a which results in the minimum MSE.

Solution:

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

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

so that we have

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

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

Example 11 (continued): An MSE estimator with a 'gain' (is a biased estimator feasible?)

Can we find an optimum a analytically? Differentiate wrt a to yield

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

and set the result to zero arrive at the optimal value

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

☞ **but we do not know the value of $\theta = A$** ☜

☞ Although MSE makes sense, estimates usually rely on the unknown θ

Without any constraints, this criterion leads to **unrealisable estimators** ↬ those which are not solely a function of the data (see Example 6).

☞ **Practically, the minimum MSE (MMSE) estimator needs to be abandoned, and the estimator must be constrained to be unbiased.**

Minimum variance estimation & MSE criterion, together

Basic idea of MVU: Out of all possible unbiased estimators, find the one with the lowest variance.

If the Mean Square Error (MSE) is used as a criterion, this means that

$$MSE\,(\hat{\theta}) = var(\hat{\theta}) + \underbrace{B^2(\hat{\theta})}_{=0\ for\ MVU}$$

☞ By constraining the bias to be zero, our task is much easier, that is, to find an estimator that minimises the variance.

◦ In this way, the feasibility problem of MSE is completely avoided.

Have you noticed:

MVU estimator = Minimum mean square error unbiased estimator

We will use the acronym MVUE for minimum variance unbiased estimator.

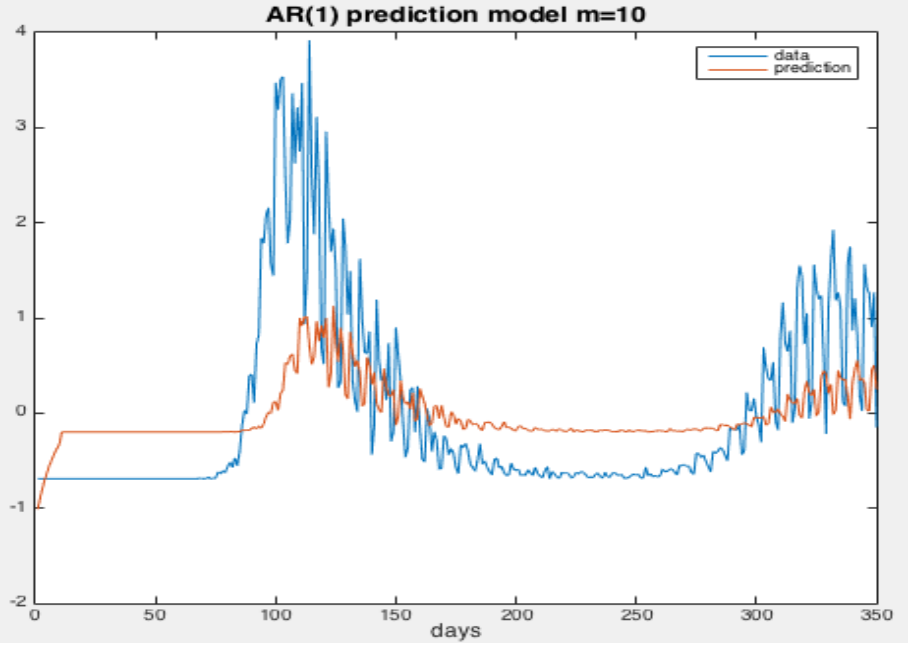
☞ **Course goal: To find optimal statistical estimators and inference**

(see the Appendix for an alternative relation between the error function and the quality (goodness) of an estimator)

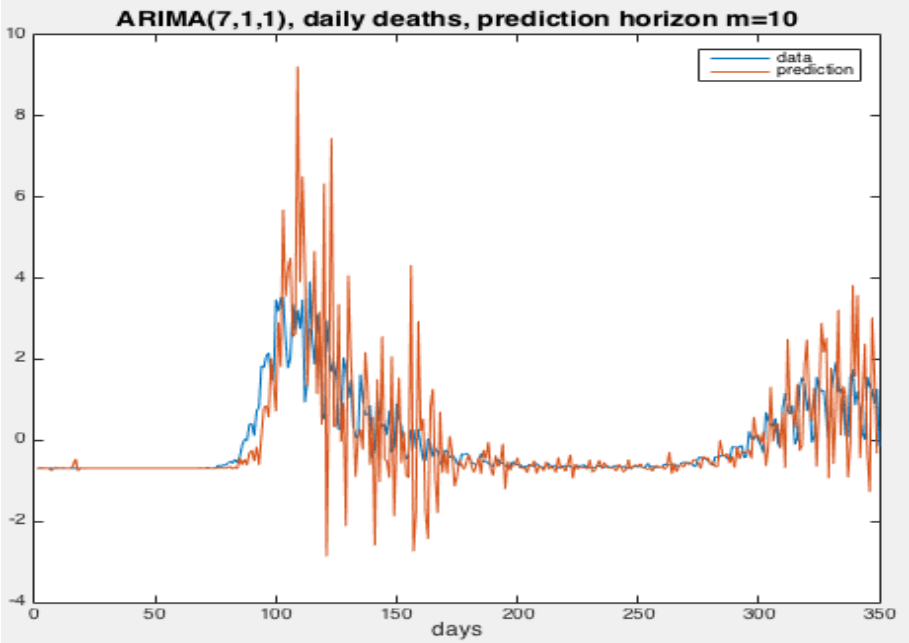
Bias–variance illustration: ARIMA prediction of COVID-19 death data

Consider the prediction of COVID-19 death rates in the UK.

AR(1) model, 10 days ahead



ARIMA(7,1,1) 10 days ahead

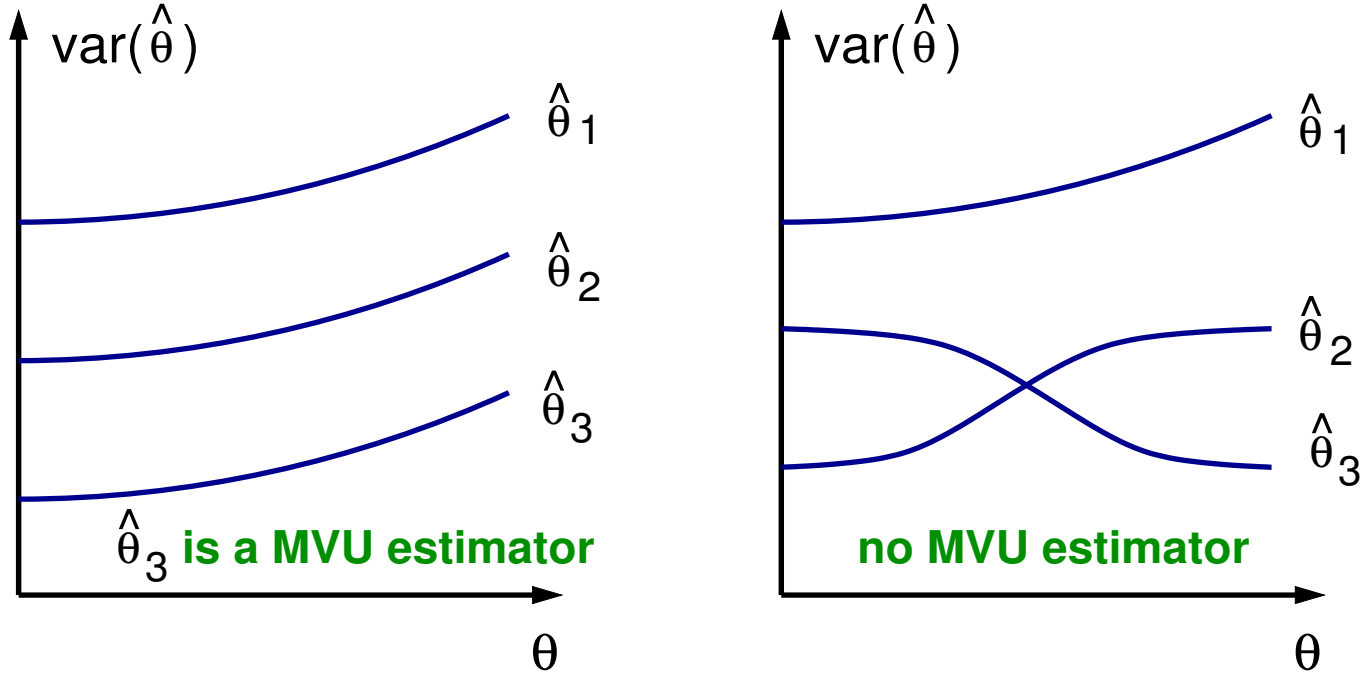


- The AR(1) prediction exhibits bias, as the mean of the predicted data (in red) is "off-set" from the mean of true data (in blue) for most of the plot
- The ARIMA(7,1,1) prediction coincides with the original data in terms of the mean for the whole plot, but exhibits large variability (which do you prefer)

Desired: Minimum variance unbiased (MVU) estimator

Minimising the variance of an unbiased estimator concentrates the PDF of the error about zero $\looparrowright$ 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, $\hat{\theta}_3$ on the plot above

Extensions to the vector parameter case

- If $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T \in \mathbb{R}^{p \times 1}$ is a vector of unknown parameters, then a vector parameter estimator is said to be unbiased if

$$E(\hat{\theta}_i) = \theta_i \qquad \text{that is, every } \theta_i \text{ is unbiased, for } i = 1, 2, \ldots, p$$

By defining

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

an unbiased vector parameter estimator has the property

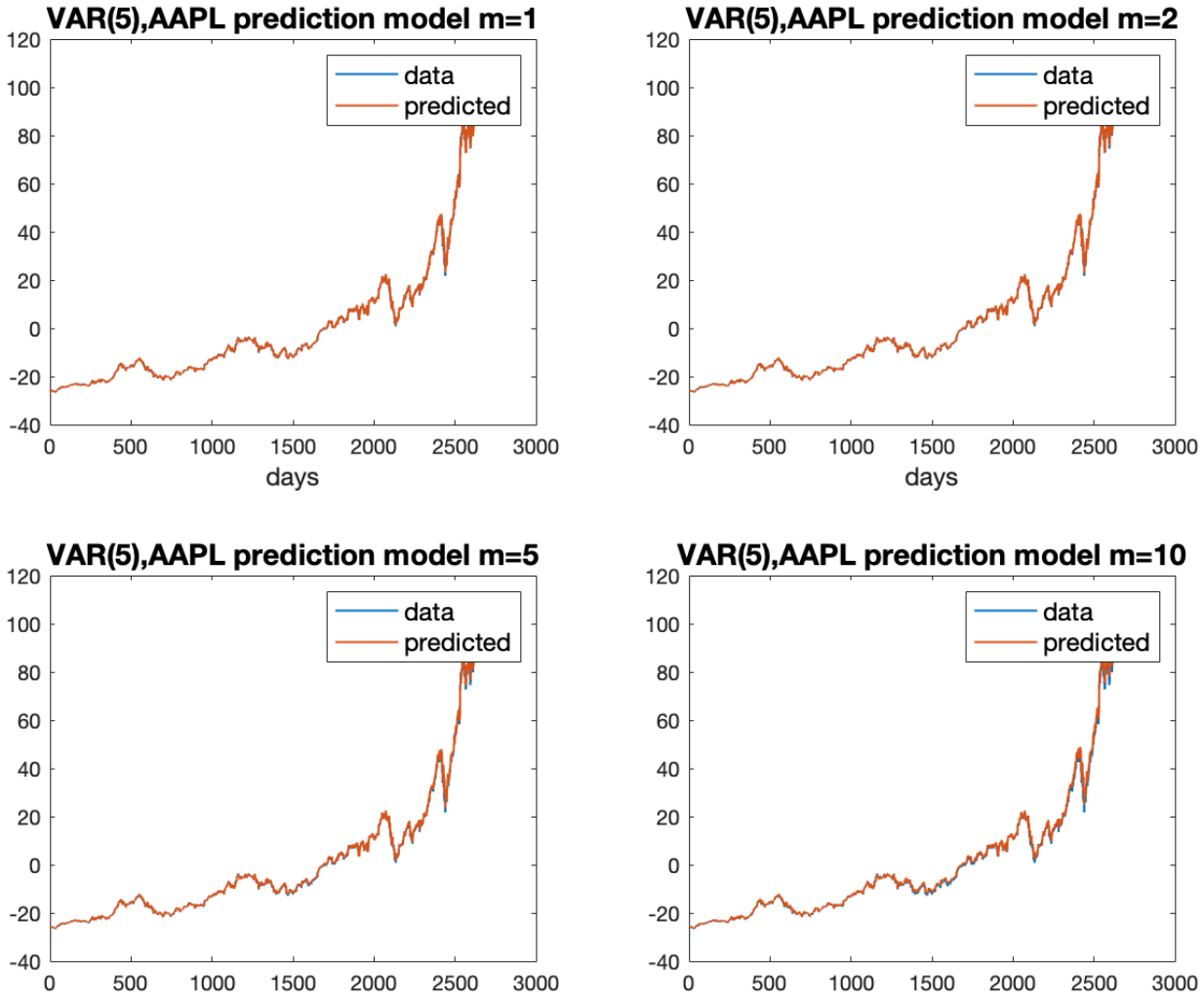
$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$$

within the p–dimensional space of parameters spanned by $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$.

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

Multivariate inference often helps

For a rigorous account of multivariate inference, see Lecture 4



Apple stock prediction using a vector autoregressive VAR(5) model (Apple as one variate and 4 other stocks from S&P 500 as other variates)

Methods to find the MVU estimator

The MVU estimator **may not always exist**, for example, when:

- There are no unbiased estimators $\rightsquigarrow$ a search for the MVU is futile
- None of the unbiased estimators has uniformly minimum variance, as in the right hand side figure on the previous slide

If the MVU estimator (MVUE) exists, we may not always be able to find it. While there is no general **"turn-the-crank"** method for this purpose, the approaches to finding the MVUE employ the following procedures:

- Determine the Cramer-Rao lower bound (CRLB) and find some estimator which satisfies the so defined MVU criteria (Lecture 4)
- Apply the Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem (rare in pract.)
- Restrict the class of estimators to be not only unbiased, but also linear in the parameters, this gives MVU for linear problems (Lecture 5)
- Employ optimisation and prior knowledge about the model (Lecture 6)
- Drop all assumptions, employ real–time adaptive estimation schemes and perform on-line estimation on streaming data (Lecture 7)

Summary

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

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

◦ Even a simple Sample Mean estimator is an example of the power 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.

◦ In our SSPI course, we will address linear, best linear unbiased, maximum likelihood, least squares, sequential least squares, and adaptive estimators.

Homework: Check another proof for the MSE expression

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

Let us now evaluate these terms:

$$(1) \qquad var(\hat{\theta}-\theta) = var(\hat{\theta})$$

$$(2) \qquad E[\hat{\theta}-\theta]^2 = \mathrm{MSE}$$

$$(3) \qquad \big[E[\hat{\theta}-\theta]\big]^2 = \big[E[\hat{\theta}] - E[\theta]\big]^2 = \big[E[\hat{\theta}-\theta]\big]^2 = \mathrm{bias}^2(\hat{\theta})$$

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

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

Recap: Unbiased estimators

Due to the linearity property of the statistical expectation operator, $E\{\cdot\}$, that is

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

the sample mean estimator can be 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.

For example, the value of an electronic component such as an inductor, capacitor or resistor would be positive (prior knowledge).

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

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

Appendix: Some usual assumptions in the analysis

How realistic are the assumptions on the noise?

- Whiteness of the noise is quite realistic to assume, unless the evidence or physical insight suggest otherwise
- The independent identically distributed (i.i.d.) assumption is straightforward to implement through e.g. the weighting matrix $\mathbf{W} = diag(1/\sigma_0^2, \ldots, 1/\sigma_{N-1}^2)$ (see Lectures 5 and 6)
- In real world scenarios, we often deal with e.g. bandpass or correlated noise (e.g. pink or $1/f$ noise in physiological recordings)
- The assumption of Gaussianity is often realistic to keep, due to e.g. the validity of Central Limit Theorem, or an appropriate data transformation

Is the zero–mean assumption realistic? Yes, as even for non–zero mean noise, $w[n] = w_{zm}[n] + \mu$, where $w_{zm}[n]$ is zero–mean noise, the mean of the noise μ can be incorporated into the signal model.

Do we always need to know noise variance? In principle no, but when assessing performance (goodness), variance is needed to measure the SNR.

Appendix. Example 12: A counter-example $\rightarrow$ 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

$$\hat{\sigma}^2 = \frac{\alpha}{N} \sum_{n=1}^{N} y^2[n] \quad \alpha > 0$$

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

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

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

The MMSE is obtained for $\alpha_{min} = \frac{N}{N+2}$ and is $MMSE(\hat{\sigma}^2) = \frac{2\sigma^4}{N+2}$.

Given that the corresponding $\hat{\sigma}^2$ of an optimal 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.

Appendix (full analysis of Example 4)

Biased estimator:

$$\tilde{A} = \frac{1}{N} \sum_{n>1}^{N} \big|x[n]\big|$$

Therefore,

- if $A \geq 0$, then $\big|x[n]\big| = x[n]$, and $E\{\tilde{A}\} = A$
- if $A < 0$, then $E\{\tilde{A}\} \neq A$

$$\Rightarrow \quad \text{Bias} = \begin{cases} = 0, & A \geq 0 \\ \neq 0, & A < 0 \end{cases}$$

Appendix: Tschebycheff (or Chebyshev) inequality, Slide 24

We do not need full knowledge of pdf, just knowledge of the mean and variance

It provides an upper bound to the probability of the absolute deviation of a random variable (RV) exceeding a given threshold (probability of rare events). One way to prove it is through the Markov inequality.

Markov inequality: For a positive RV, X, and a positive threshold, ϵ, the following holds:

$$P(X \geq \epsilon) \leq \frac{E\{X\}}{\epsilon} = \frac{\mu}{\epsilon}$$

For a positive X, this property holds **almost surely** (with probability one).

Example 13: An average salary in a company is 40,000 GBP. What is the probability of a given person having salary greater than 100,000 GBP?
Answer: Markov's inequality gives the upper bound on this probability

$$P(X \geq 100,000) \leq \frac{40,000}{100,000} = \frac{1}{2.5}$$

☞ **Markov's inequality can be used to prove that mean square convergence implies convergence in probability, and also to prove Chebyshev's inequality.** (see Lecture 1 for more detail)

Appendix: Tschebycheff (or Chebyshev) inequality, Slide 24

We do not need full knowledge of pdf, just knowledge of the mean and variance

Chebyshev's inequality: Consider a random variable, X, with a finite mean μ and a finite variance σ^2, and a positive number ϵ. Then

$$P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

To arrive to the Chebyshev inequality on Slide 24, substitute $X \to \hat{\theta}, \mu \to \theta, \sigma^2 \to var\{\hat{\theta}_N\}$ into the Markov inequality. The proof follows immediately.

Example 14: An average salary in a company is 40,000 GBP, with a stand. dev. of 20,000 GBP. What is the probability of a given person having salary which is either less than 10,000 GBP or greater than 70,000 GBP?
Answer: This probability cannot be computed exactly, however, Chebyshev's inequality will give an upper bound to this probability. We are looking for the bound on $|X-\mu| \geq \epsilon$, with $\mu = 40,000$ and $\epsilon = 30,000$. The probability of this happening is then

$$P(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} = \frac{400,000,000}{900,000,000} = \frac{4}{9}$$

For more details, examples, and proofs, see Lecture 1.

Notes

Notes