Statistical Signal Processing & Inference

Course Introduction

Prof Danilo Mandic

room 813, ext: 46271

Department of Electrical and Electronic Engineering

Imperial College London, UK

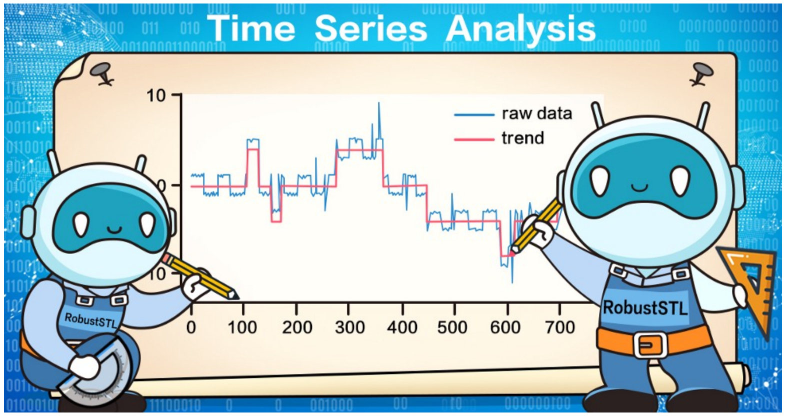
d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

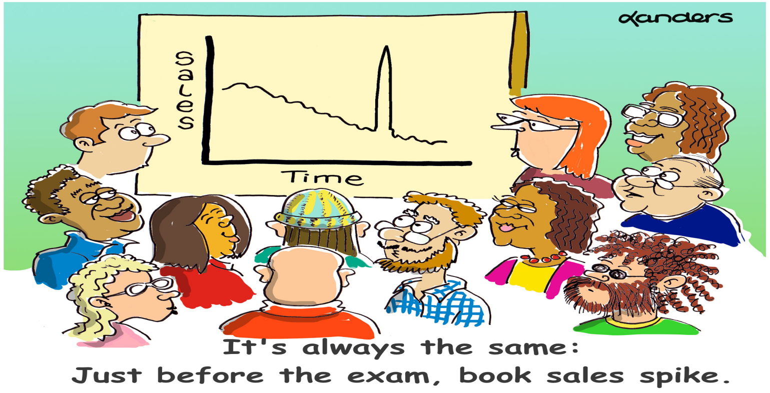
The need for Statistical Signal Processing & Inference

Q: Have you ever considered what the following tasks have in common:

- Forecasting of financial data
- Supply-demand modelling (e.g. electricity or air-ticket pricing)
- Modelling of COVID-19 spread
- Person recognition from a set of (noisy) images
- Word generation by Large Language Models such as ChatGPT

A: These are signals/images of which the signal generating mechanisms are largely unknown or untractable. We need to make sense from such data based on historical observations only $\looparrowright$ subject of **Statistical Inference.**

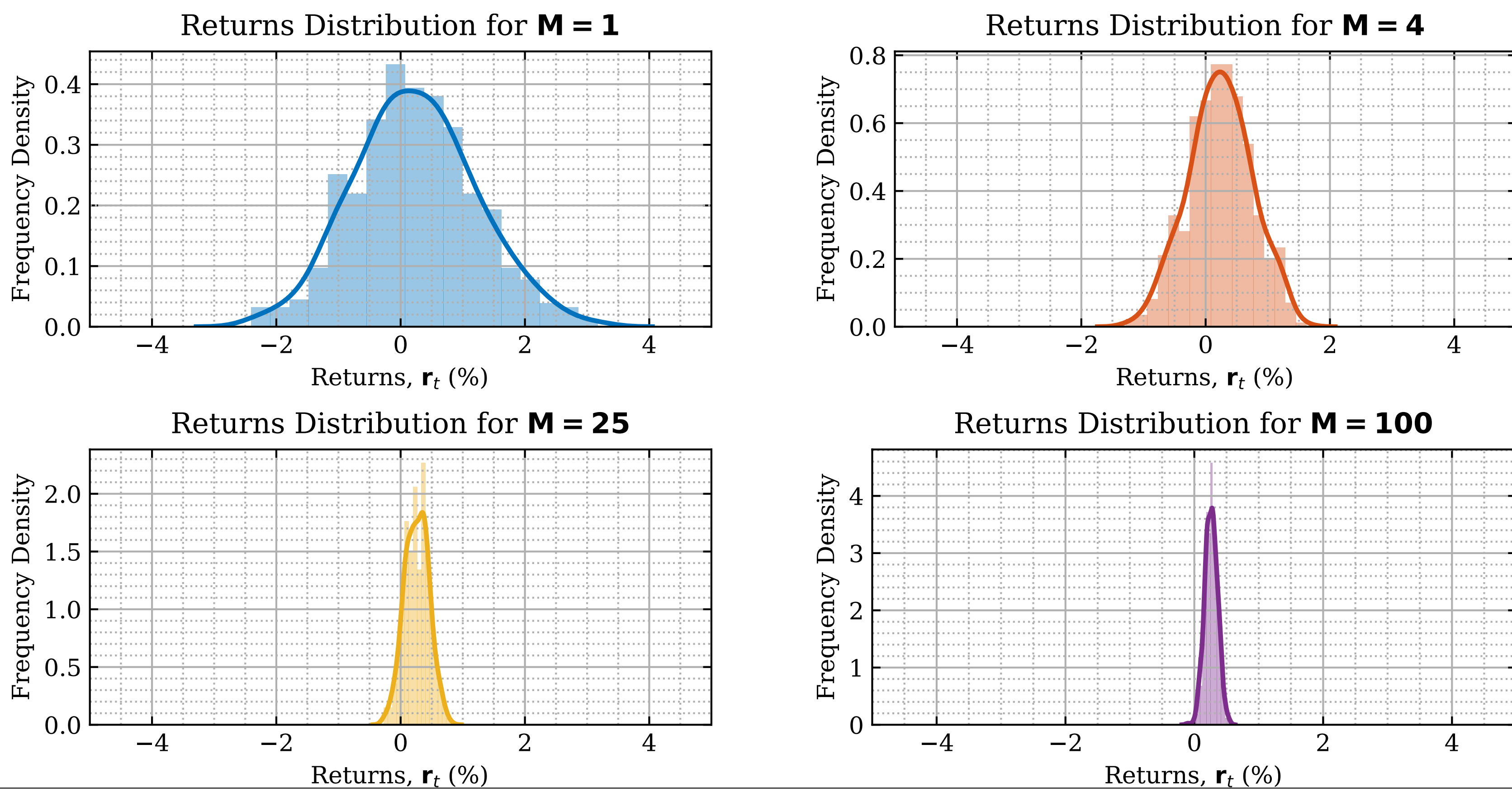




SSP&I: Use your knowledge and not brute force when designing learning machines

The need for statistical inference: Population modelling

Example from financial modelling: Risk for a single asset and a for a portfolio of uncorrelated assets. Risk is represented by the standard deviation (or the width) of the distribution curves ↬ a large portfolio ($M = 100$) can be significantly less risky than a single asset ($M = 1$).



Statistical inference

Chinese for statistics is 统计 (summarizing & counting) and

probability is 概率(论) ((theory of) randomness & chances),

Probability: Assumes perfect knowledge about the "population" of random data (through the pdf).

Typical question: There are 100 books on a bookshelf, 40 with red cover, 30 with blue cover, and 30 with green cover. What is the probability of randomly drawing a blue book from the shelf?

Statistics: No knowledge about the types of books on the shelf, **we need to infer properties** about the "population" based on random samples of "objects" on the shelf $\looparrowright$ **statistical inference**.

Typical question: A random sampling of 20 books from the bookshelf produced X red books, Y blue books and Z green books. What is the total proportion of red, blue, and green books on the shelf?

Statistical inference is applied in many different contexts under the names of: data analysis, data mining, machine learning, classification, statistical signal processing, pattern recognition, clustering, regression, classification.

Foundations of resilience: Probability vs. Statistics

For discrete RVs, $E\{X\} = \sum_{i=1}^{I} x_i P_X(x_i)$, where P_X is the probability function

Probability: A data modelling view, describes how data **will likely behave**

for example: $$average = E\{X\} = \int_{-\infty}^{\infty} x\, p_X(x)\, dx$$ no data here

Notice that there is no explicit mention of data here $\hookrightarrow$ x is a dummy variable and p_X is the pdf of a random variable X.

Statistics: A data analysis view, determines how data **did behave**

for example: $$average = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$ no pdf here

Example: Consider N coarse-quantised data points, $x[0], \ldots, x[N-1]$. The signal has $M \ll N$ possible amplitude values, $V_1, \ldots, V_M$, with the corresponding relative frequencies, $N_1, \ldots, N_M$. Calculate the mean, $\bar{x}$.

Solution:

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \quad = \quad \frac{1}{N}\sum_{m=1}^{M} V_m N_m \quad = \quad \sum_{m=1}^{M} V_m \underbrace{\frac{N_m}{N}}_{\approx\; P(x=V_m)}$$

Inference vs Artificial Intelligence

Statistics: "The art and science of gathering, analysing, and making inferences from data." Mosteller *et al.*, Probability with Statistical Applications, 1957

Popular definition of AI: "Anything that makes a decision or an action that a human used to take or helps a human decide or act."

☞ This has been subject of statistical modelling & inference since the 1990s!

Statistics starts with data. Real-world data are random, so statistical models will involve probability statements and performance bounds:

- Exploratory statistics $\looparrowright$ find patterns in data
- Confirmatory statistics $\looparrowright$ fit math models to find reproducible patterns
- Inference: Draw conclusions and/or make decisions and predictions

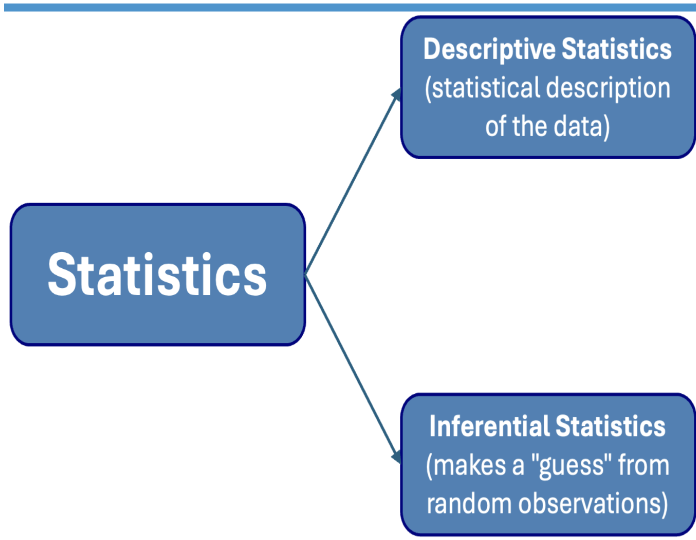
To make sense of raw data, we can calculate *summary statistics* (mean, median, variance) and visualise the data (histograms, scatterplots).

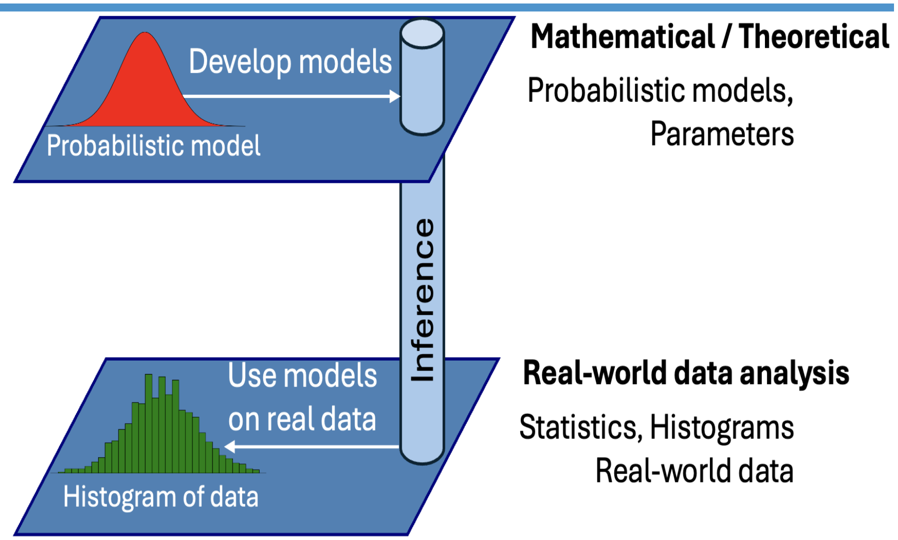
Generative modelling develops stochastic models which fit the data, in order to make inferences about the data-generating mechanisms.
Predictive modelling does not necessarily explore the underlying data generating mechanisms, and focuses on inference from historical data.

☞ Machine learning typically follows the predictive modelling culture.

Statistical Inference

From Latin *inferre*, which means "bring into, deduce, conclude"





Inferential statistics: **Statistical Estimation** and **Hypothesis Testing**

☞ In Machine Learning, the term "inference" typically indicates "prediction"

Applications:

- Adaptive learning algorithms (noise-cancelling headphones, forecasting)
- Neural Networks (e.g. classification, prediction, denoising)
- Communications, power systems, radar, sonar, biomedicine, ...
- Financial modelling (CAPM, risk estimation, confidence intervals)
- Artificial Intelligence (e.g. self-driving cars)

Inferential stats tell us "what is possible to achieve" → focus of this course

Statistical Signal Processing & Inference within AI space

Engineering solutions do not necessarily mimic the nature.

Humans provide a performance "benchmark" but mimicking human reasoning by AI???

◦ The 10^{11} neurons and 10^{15} synapses in human brain ↬ 20 W of power.

◦ A digital simulation of an ANN of same size consumes 7.9 MW of power.

M. Jordan, "AI – The revolution hasn't happened yet":

> By approaching a problem with an engineering mind, AI can be considered as a new, human-centric engineering discipline.

☞ **Strive to surpass human limitations and not to mimic humans!**

☞ **Claim: Big Data + Deep Learning → General Intelligence**

But humans learn very efficiently with little data, not Big Data!

Caution: We can no longer train a modern DNN on a personal computer, it would take up to 405 years. Electricity consumption for digital devices: from 3-4% today to 20% in 2050. We need a convival technology that is resilient ↬ **a real opportunity for SSP&I.**

Aims: To introduce the fundamentals of statistical estimation theory, to facilitate the design of signal processing and machine learning algorithms for inference

- The emphasis will be upon:
  - ⊛ random signals, their properties, and statistical descriptors
  - ⊛ stochastic models, to generate/describe random signals
  - ⊛ parametric (model based) and nonparametric (data driven) modelling
  - ⊛ optimal estimators for random signals, rigorous performance bounds
  - ⊛ the class of least squares methods, block & sequential LS, regression
  - ⊛ adaptive learning and estimation $\rightsquigarrow$ suitable for nonstationary data
  - ⊛ rigorous performance bounds to tell us 'what is possible to achieve'

- **Practical experience** through numerous examples on real world signals:
  - ⊛ multimedia (your own speech recorded via PC)
  - ⊛ your own physiological data, some financial data (from *yahoo finance*)

Overall: To gain the know-how and necessary expertise in **statistical inference** from random and non–stationary real world data

☞ **This material underpins in-depth understanding and interpretability statistical signal processing and machine learning tools.**

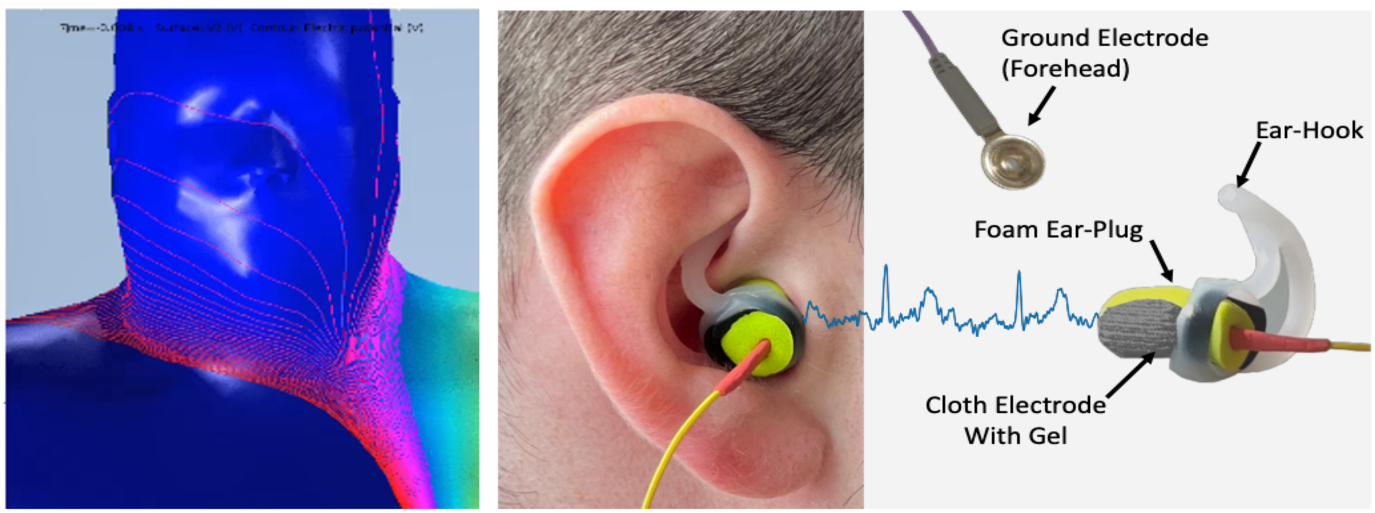
Learning from data ↬ mathematical formalism of the statistical estimation paradigm

Problem: Based on an N-point dataset $\mathbf{x} = \big[x[0], x[1], \ldots, x[N-1]\big]^T$
Find an **unknown parameter**, θ, based on the data $\mathbf{x}$, in order to define a ***statistical estimator*** (e.g. $\hat{\theta}$ can be the sinewave frequency in smart grid)

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

This is formalised as **"parameter estimation from random signals"**

Depending on the choice of g we can talk about: ⊛ linear, ⊛ nonlinear, ⊛ maximum likelihood, ⊛ minimum variance, ⊛ adaptive etc. estimation



With this we can even obtain the Electrocardiogram from the ear canal.

Course structure

The course is divided roughly into four parts:

1. Introduction to Statistical Estimation Theory
   discrete random signals, moments, bias-variance dilemma, curse of dimensionality, sufficient statistics

2. Statistical Modelling, Estimation Theory and Performance Bounds
   linear stochastic models, ARMA model, properties of estimators, Cramer Rao performance bound, minimum variance unbiased (MVU) estimator

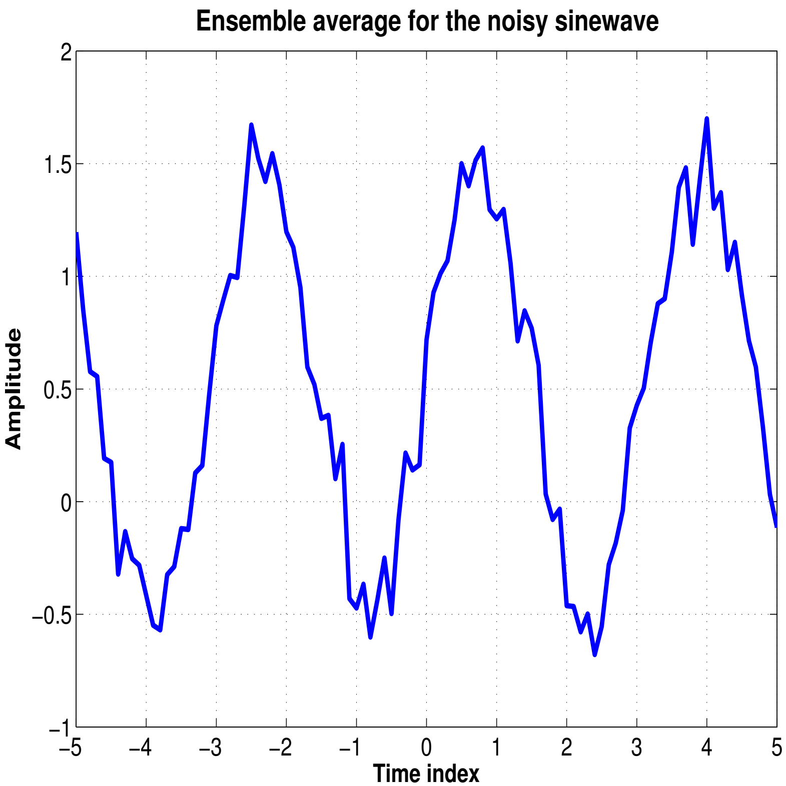
3. Statistical Inference
   best linear unbiased estimator (BLUE), maximum likelihood (ML) estimation, multivariate estimators, Bayesian estimation (optional)

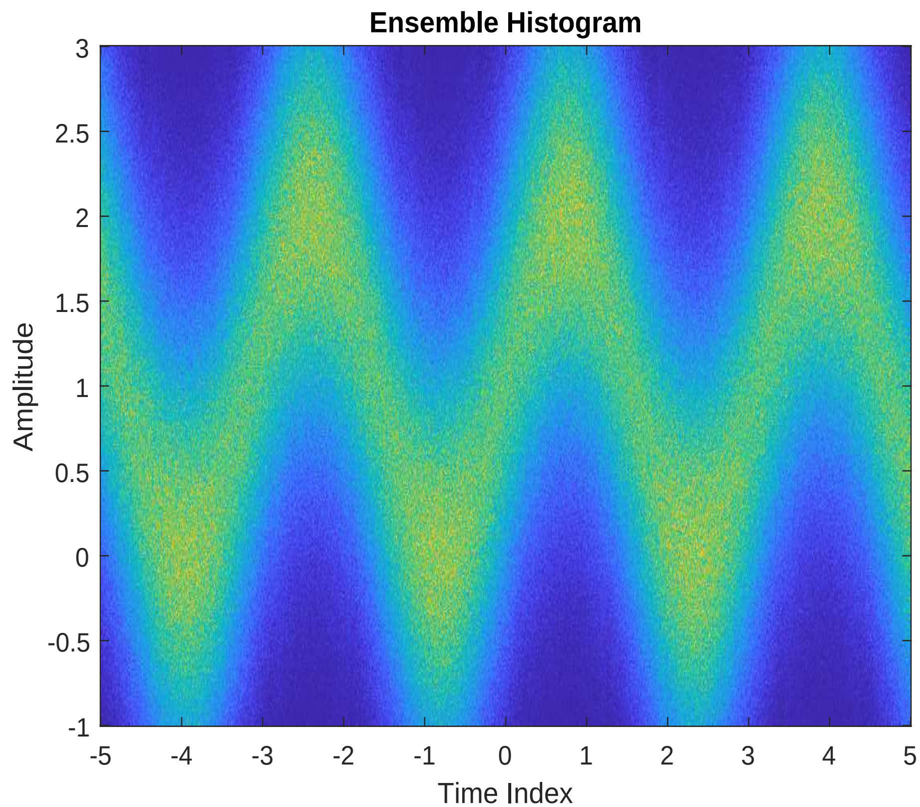
4. Mean Square Error (MSE) based learning machines
   representation bases, orthogonality principle, block and sequential Least Squares, linear and logistic regression, Wiener filter, adaptive filters, concept of an artificial neuron

Use your knowledge and not brute force when designing learning machines!

Lecture 1: Background on random signals

For illustration, consider a noisy sinusoid





The pdf at time instant n is different from that at m, in particular:

$$\mu(n) \neq \mu(m) \qquad m \neq n$$

Left & Right: Ensemble average

$$\sin(2x) + 2 * randn + 1$$

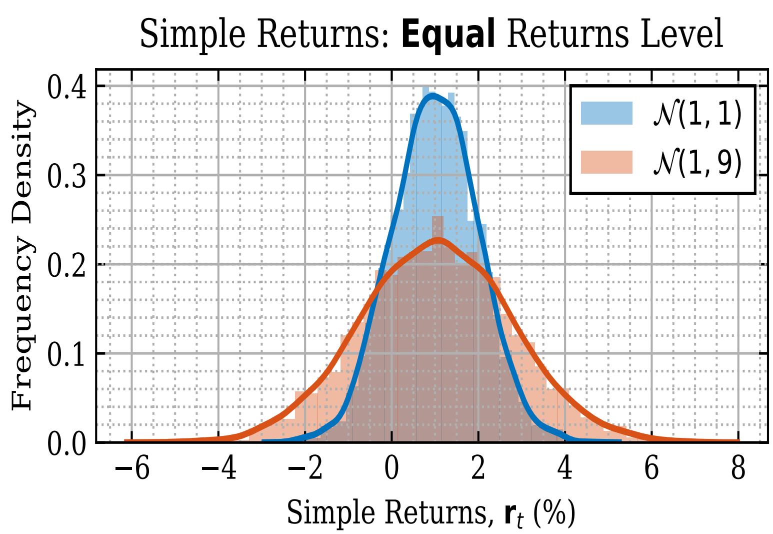
Left: 6 realisations, **Right:** 100 realisations (and the overlay plot)

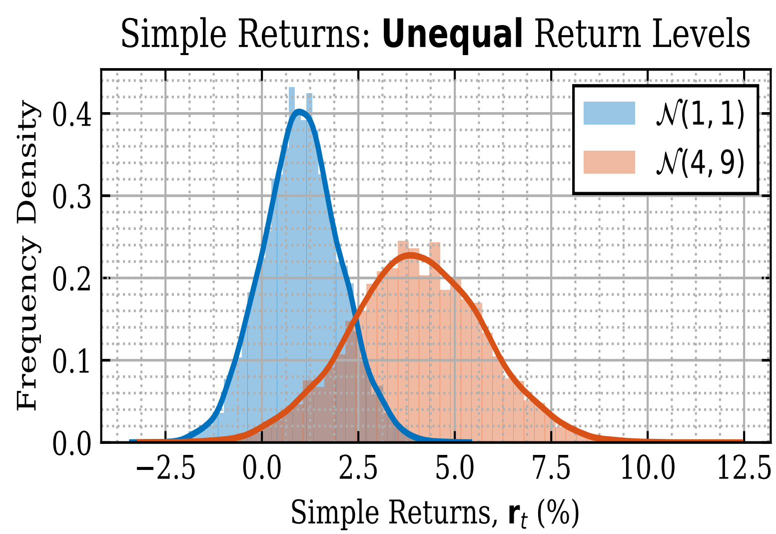
The mean and variance through the lens of statistics

Financial example: The dilemma about asset selection based on the sample mean motivates us to model the dispersion about the mean. This is naturally performed by the (co)-variance

$$Var[X] = \sigma_X^2 = E[(X - E[X])^2] \quad \text{or} \quad Var[X] \approx \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \mu_X)^2$$

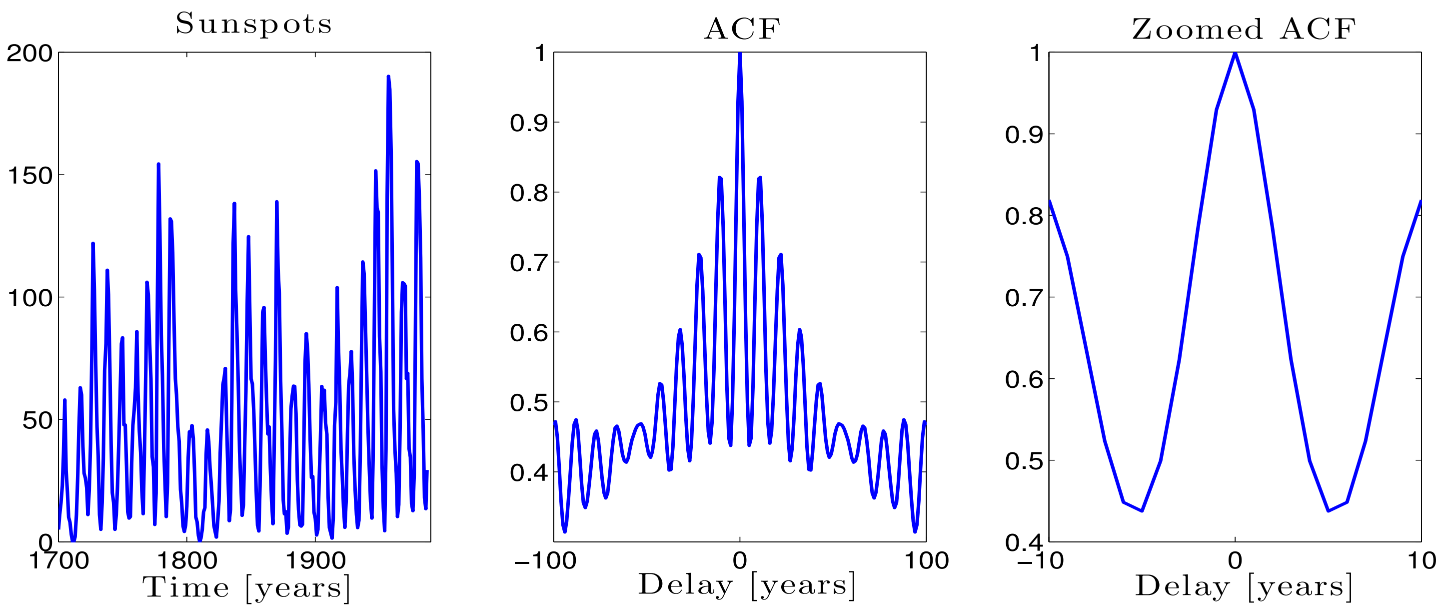
Risk Aversion: For same expected returns, choose the portfolio that minimises the volatility.





☞ For the same returns level (left) we choose the less risky asset (blue). For unequal returns, the criterion is inconclusive $\rightarrow$ Sharpe Ratio $= \mu/\sigma$

Example: Sunspot number estimation using AR models



$\mathbf{a}_1 = [0.9295]$ $\quad$ $\mathbf{a}_2 = [1.4740, -0.5857]$

$\mathbf{a}_3 = [1.5492, -0.7750, 0.1284]$

$\mathbf{a}_4 = [1.5167, -0.5788, -0.2638, 0.2532]$

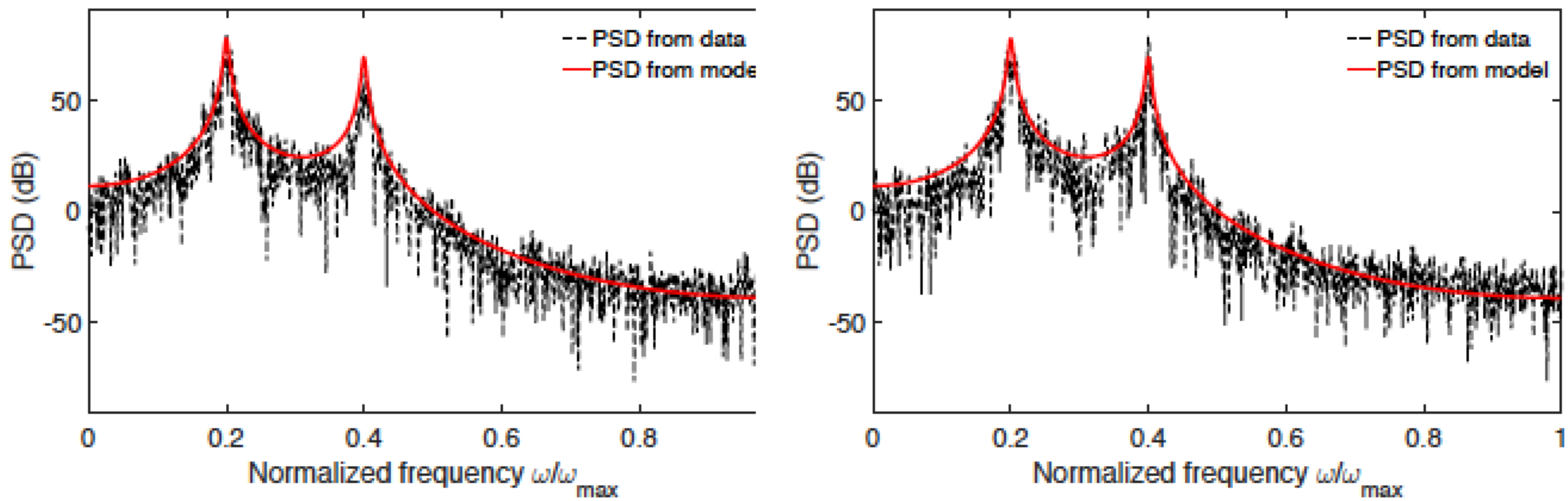
$\mathbf{a}_5 = [1.4773, -0.5377, -0.1739, 0.0174, 0.1555]$

$\mathbf{a}_6 = [1.4373, -0.5422, -0.1291, 0.1558, -0.2248, 0.2574]$

↬ The sunspots model is $\quad x[n] = 1.474\, x[n-1] - 0.5857\, x[n-2] + w[n]$

Linear Stochastic Models: Advantages

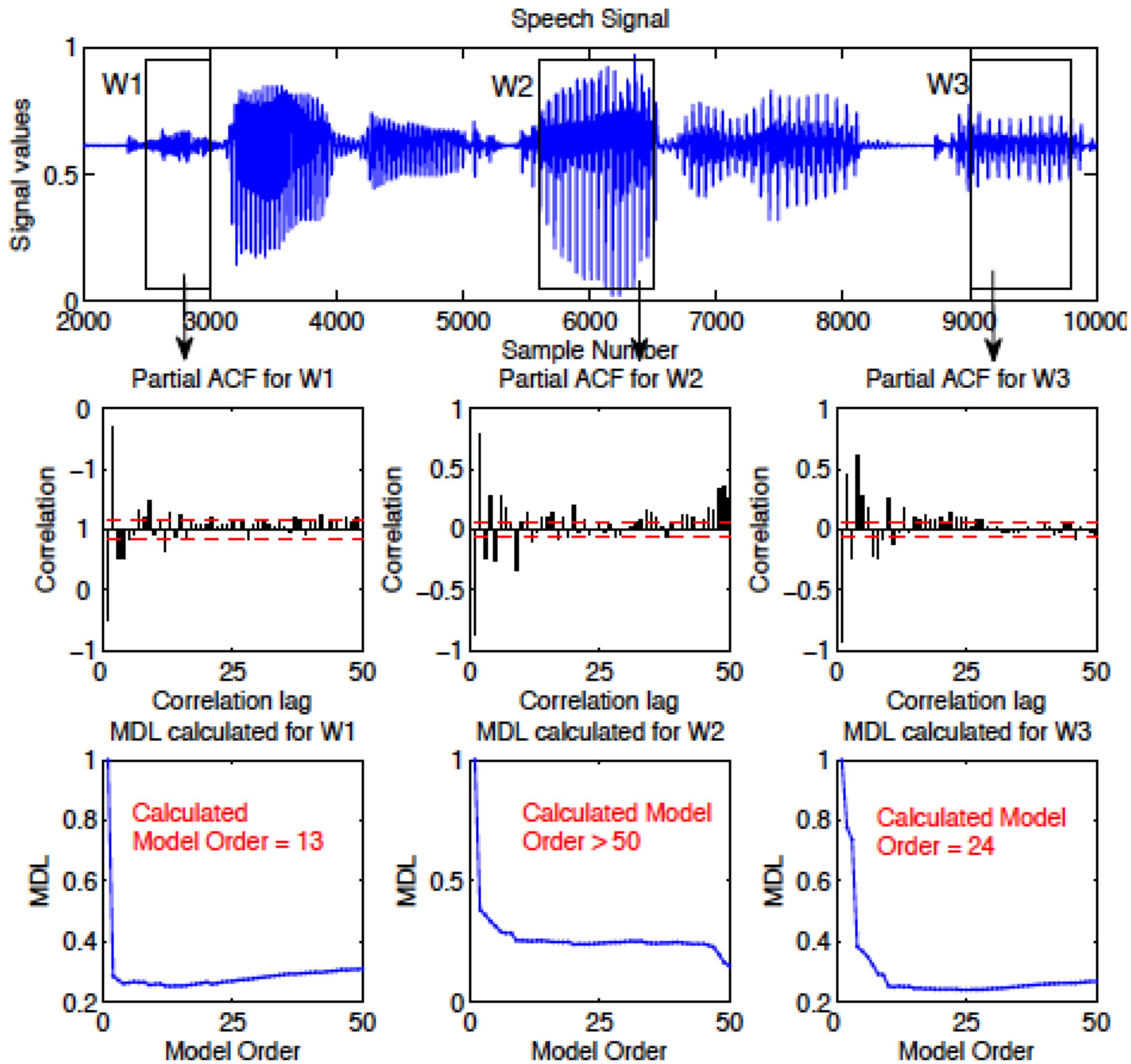
Representation of a long random signal with only a few parameters & Prediction



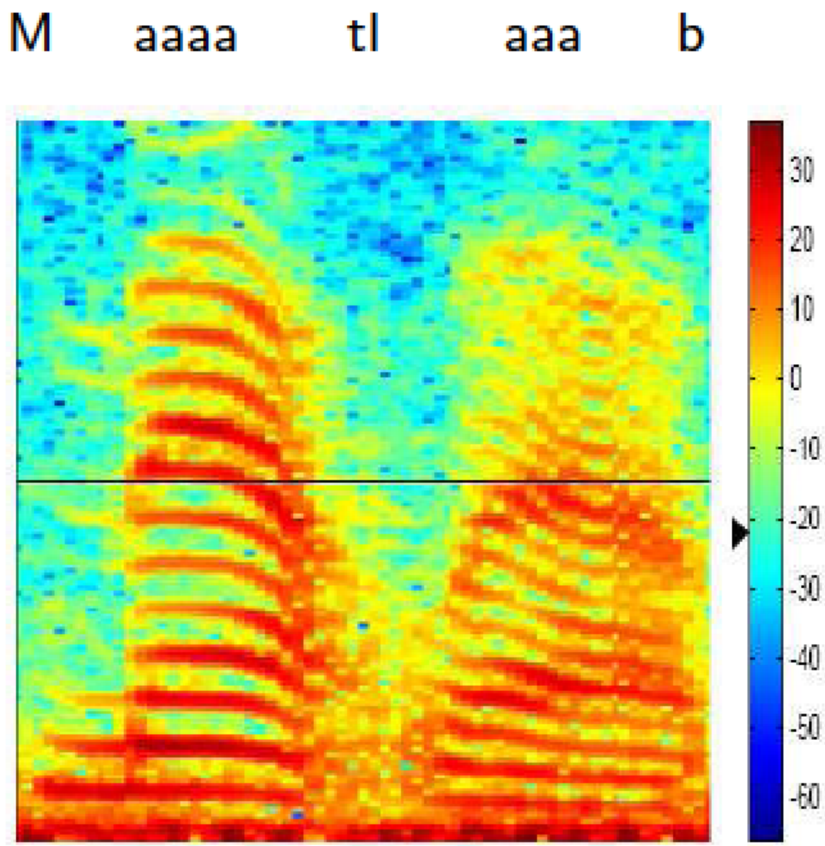
- The different realisations lead to different Emprical PSD's (in thin black)
- The theoretical PSD from the model is consistent regardless of the data (in thick red)

```
N = 1024;
w = wgn(N,1,1);
a = [2.2137, -2.9403, 2.1697, -0.9606];       % Coefficients of AR(4) process
a = [1 -a];
x = filter(1,a,w);
xacf = xcorr(x);                              % Autocorrelation of AR(4) process
dft = fft(xacf);
EmpPSD = abs(dft/length(dft)).^ 2;            % Empirical PSD obtained from data
ThePSD = abs(freqz(1,a,N,1)).^ 2 ;            % Theoretical PSD obtained from model
```

Example: Dealing with nonstationary signals



- Consider a real–world speech signal, and thee different segments with different statistical properties
- **Different AR model orders required for different segments of speech ↬ opportunity for content analysis!**
- To deal with nonstationarity we need short sliding data windows



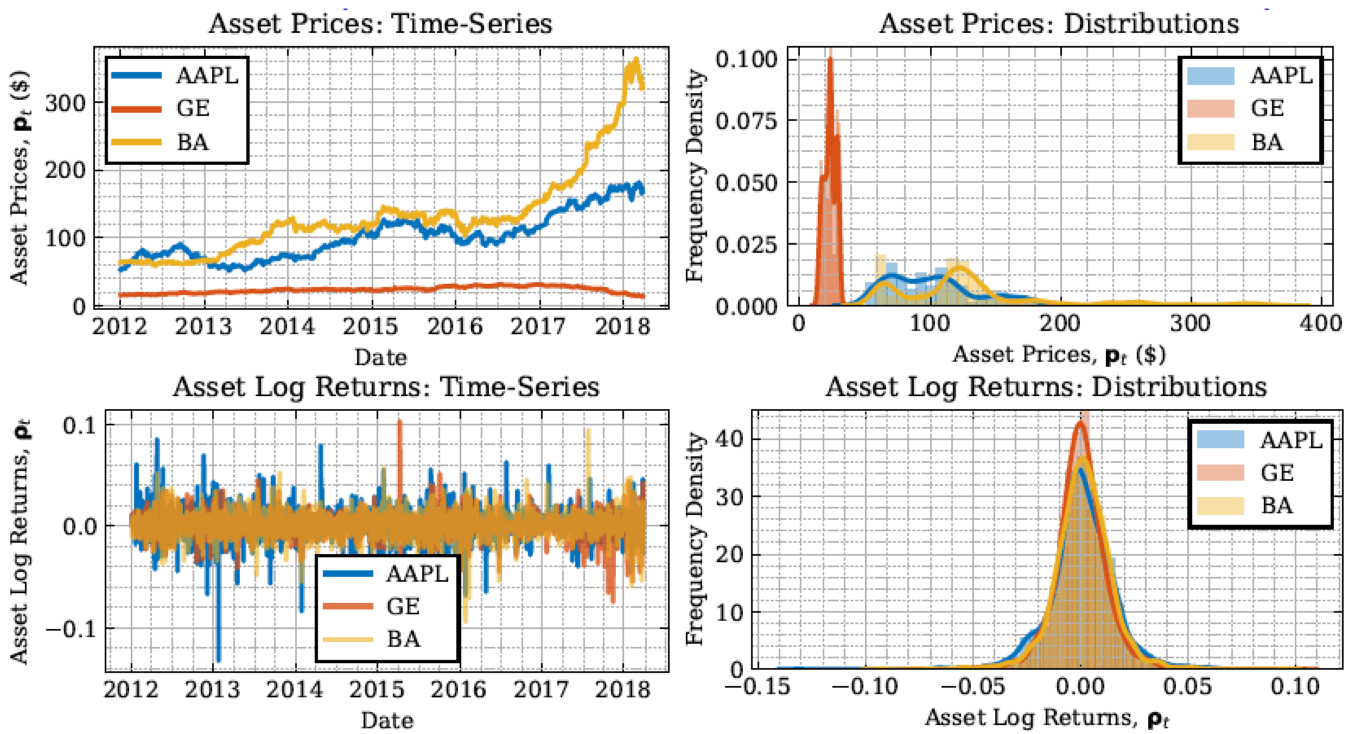
Horizontal axis: time
Vertical axis: frequency

An enabling technology in many DSP applications

- Radar and sonar: estimation of the range and azimuth
- Image analysis: motion estimation, segmentation
- Speech: features used in recognition and speaker verification
- Seismics: oil reservoirs
- Communications: equalization, symbol detection
- Biomedicine: various applications

Often we can resort to (approximately) Gaussian distrib.

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).

Importance of establishing optimal performance bounds

A typical artefact in teleconferencing, where an algorithm which provides artificial background cannot cope with movement

You will learn how to establish the optimal performance bounds in both block-based and real-time adaptive data analysis.

These will serve to:

- Indicate the quality of your algorithm/strategy against the best achievable performance for the considered class of estimators
- In many cases, this can also suggest an optimal estimation strategy
- Help identify an error in your algorithm, if its performance appears better than the optimal performance bound.

Lecture 4: Bias-variance dilemma $\rightsquigarrow$ Minimum Variance Unbiased estimation, rigorous performance bounds CRLB

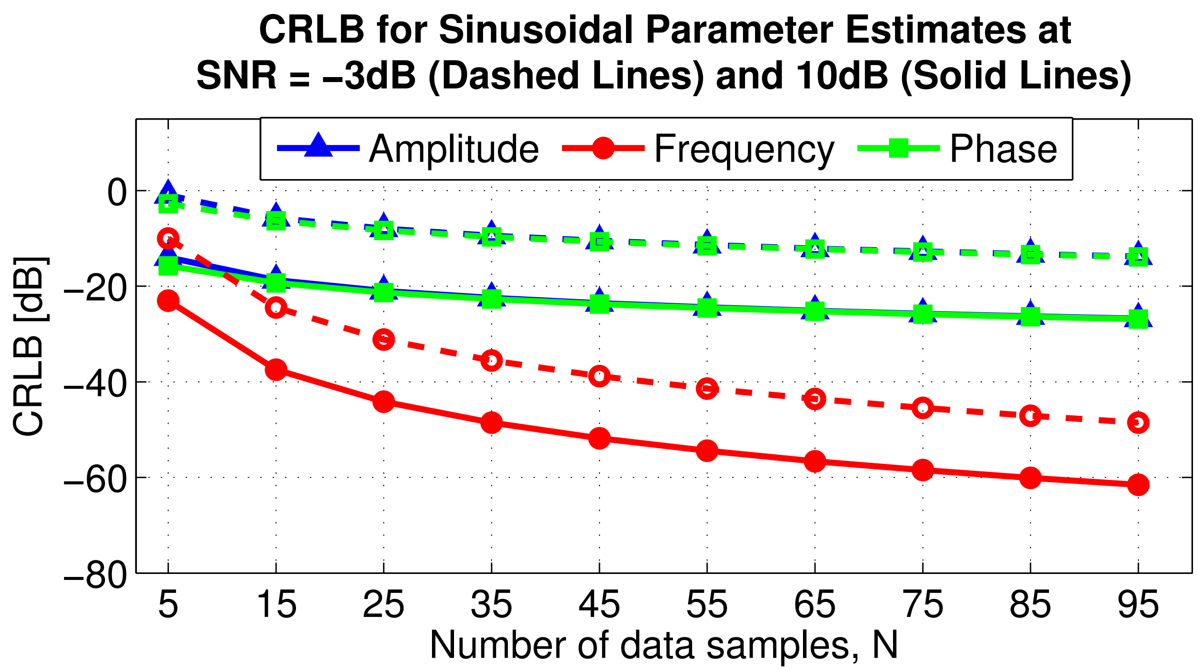
☞ **variance of the estimated parameters is sensitive to data length**

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

Task: Find the parameters A, f_0, Φ, from the noisy measurements $x[n]$

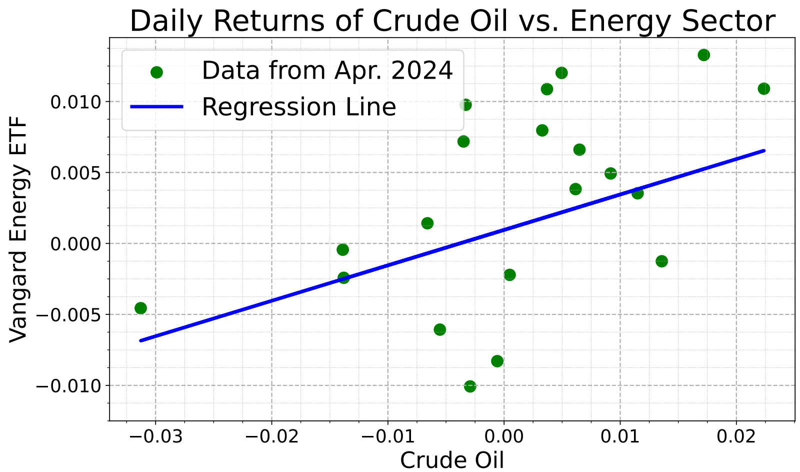
We will show that the optimal estimators obey (where $\eta = \frac{A^2}{2\sigma^2}$ is SNR):

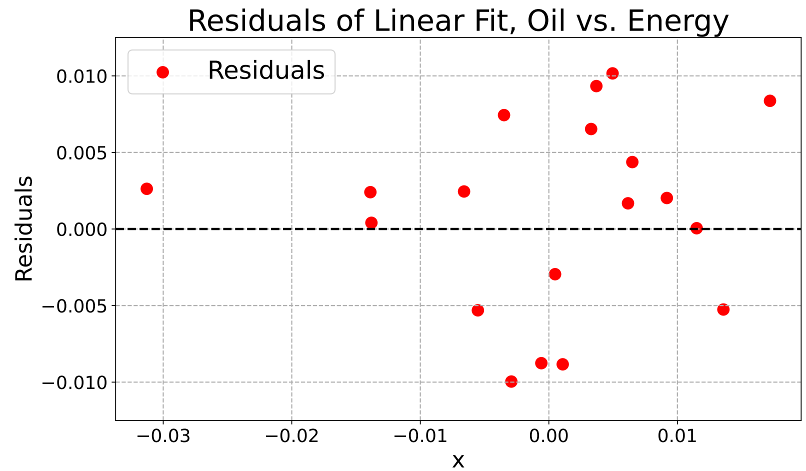
$$var(\hat{A}) \geq \frac{2\sigma^2}{N} \qquad var(\hat{f}_0) \geq \frac{12}{(2\pi)^2 \eta N(N^2-1)} \qquad var(\hat{\Phi}) \geq \frac{2(2N-1)}{\eta N(N+1)}$$

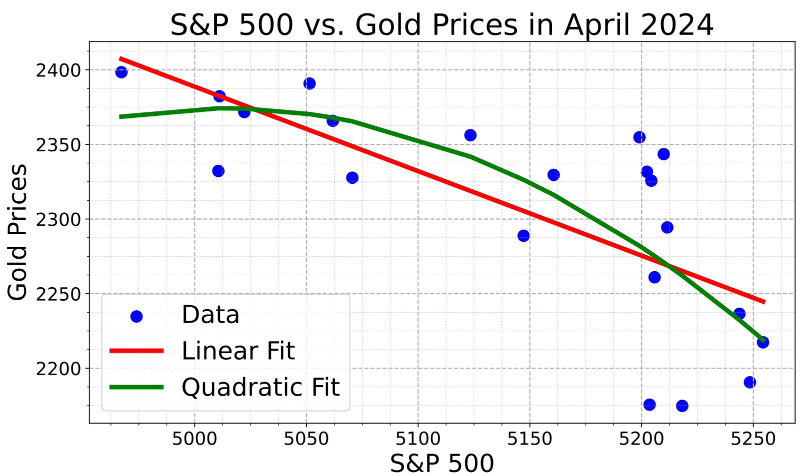


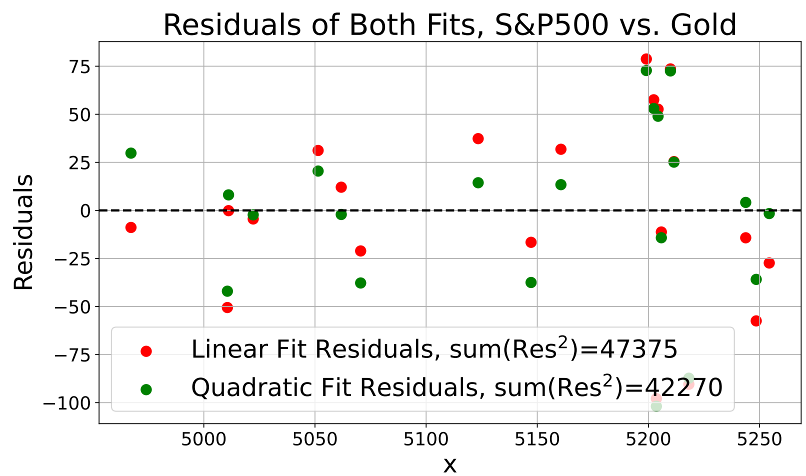
A rigorous account of Linear Models (regression models)

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



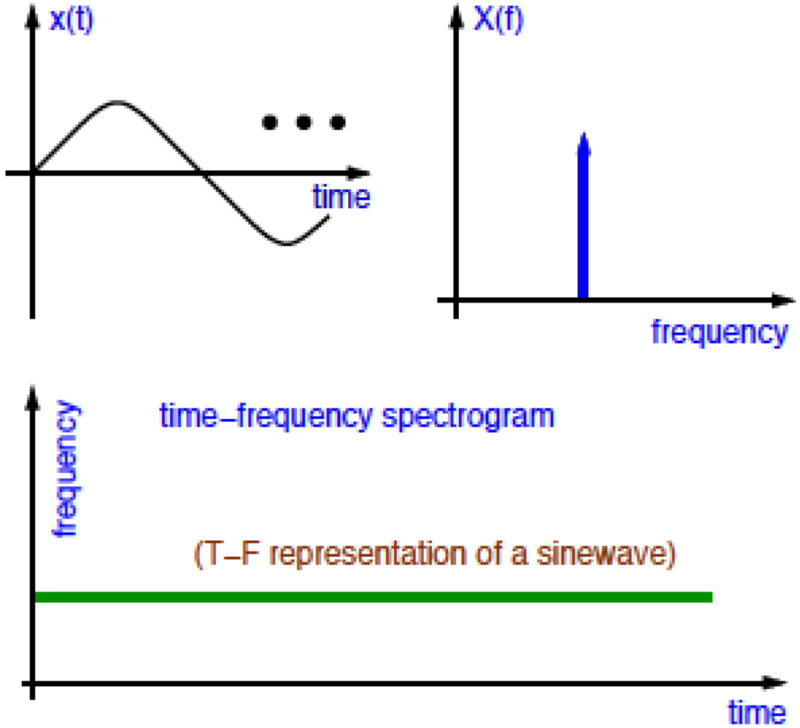






☞ Even a nonlinear models is often "linear in the parameters"

Sinusoidal frequency estimation

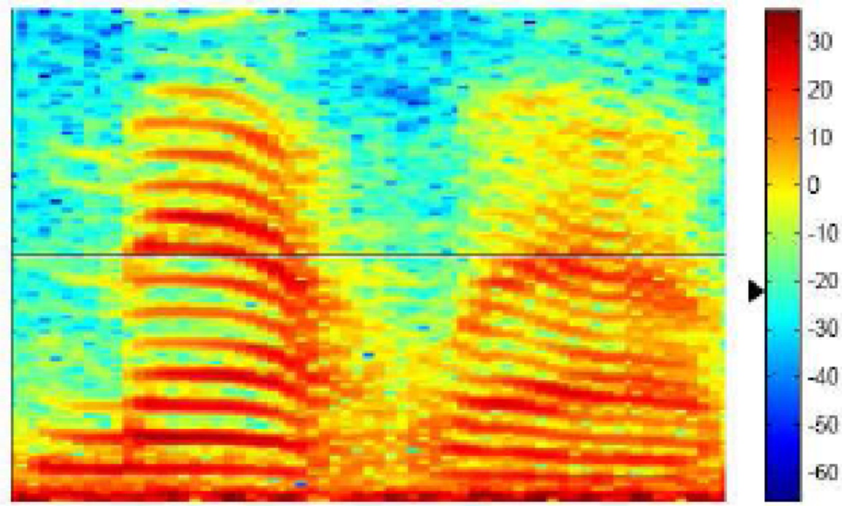


∘ Ramp in time $\not\leftrightarrow$ DC level in time (via differentiation)

∘ Chirp in time $\not\leftrightarrow$ ramp in T-F

Transforming other problems

time-frequency representation



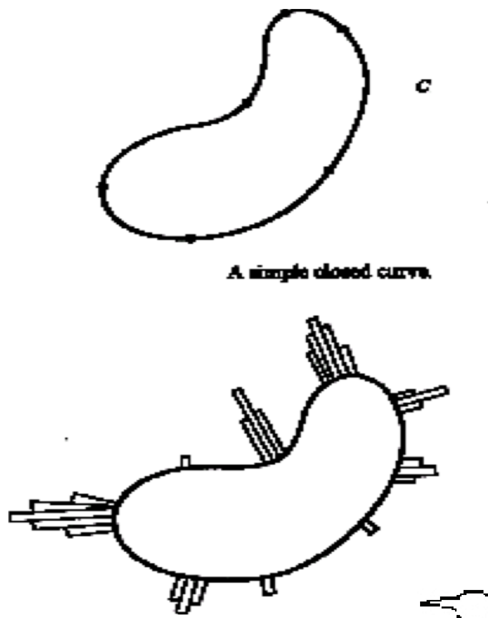
horizontal: time **vertical: frequency**

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

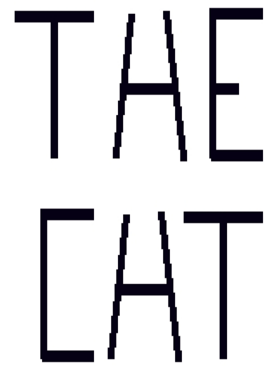
DC-level like harmonics for "a"

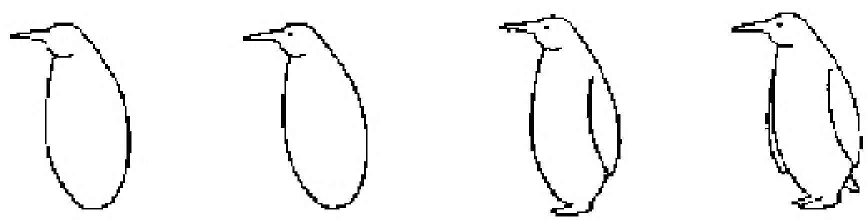
Sufficient statistics, goodness of an estimator

Example 7a: The drawing of a bean (top) and the histogram of eye dwellings (bottom)



Example 7b: Read the words below ... now read letter by letter ... are you still sure?



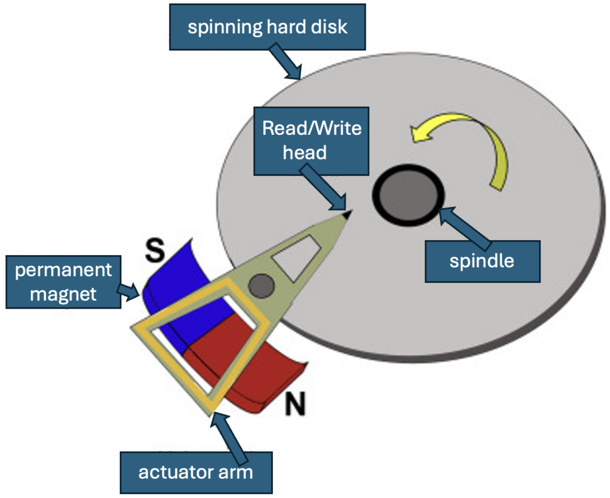


Example 7c: Is the drawing on the left still a penguin?

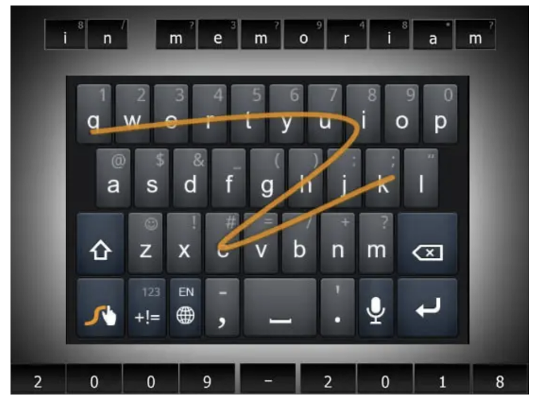
So, what is the sufficient information to 'estimate' an object?

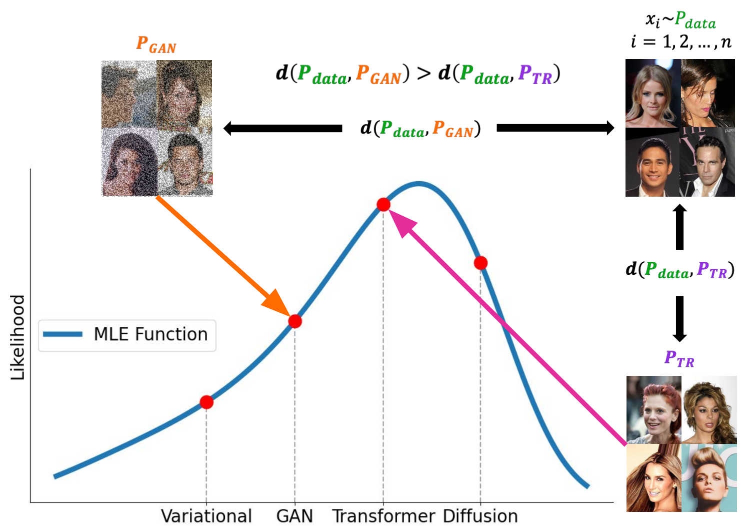
Motivation for Maximum Likelihood Estimation (MLE)
What do you think these applications have in common?

HDD microcontroller

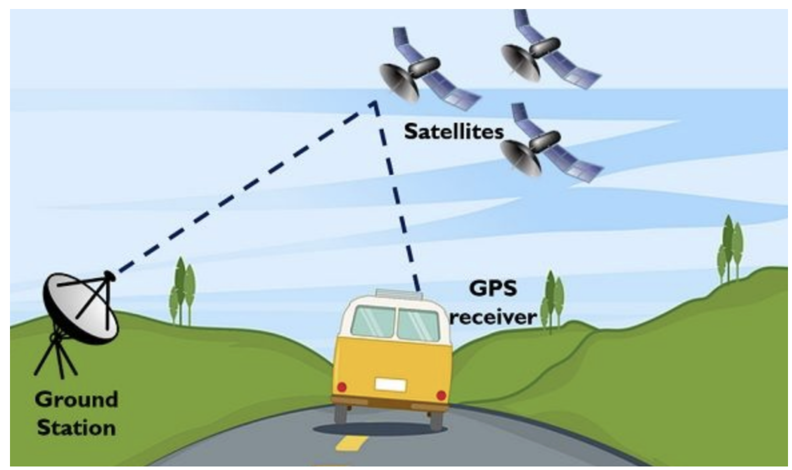


Gesture (Swype) keyboard





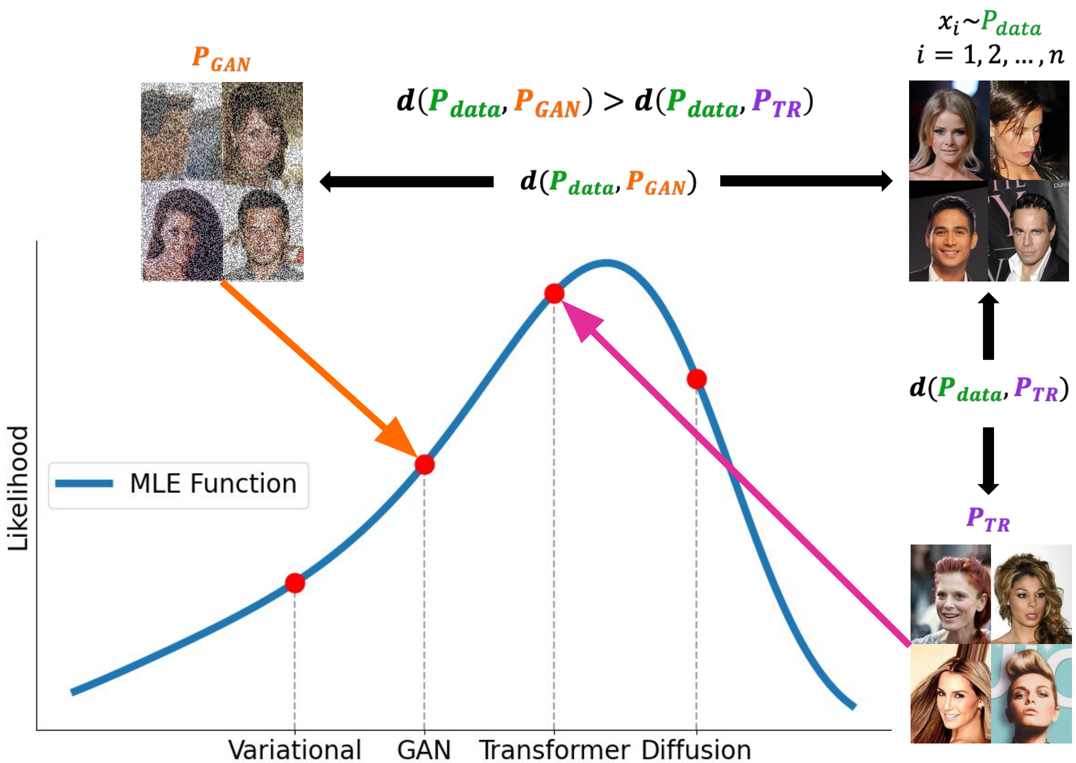
Generative AI (density estimation)



Global Positioning System

Example: MLE in Generative Artificial Intelligence

We often have a limited amount of samples of the dataset of interest, e.g. we do not know the true distribution of all male and female face images.



○ Generative models aim to generate "new" data based on the available samples of a dataset of interest.

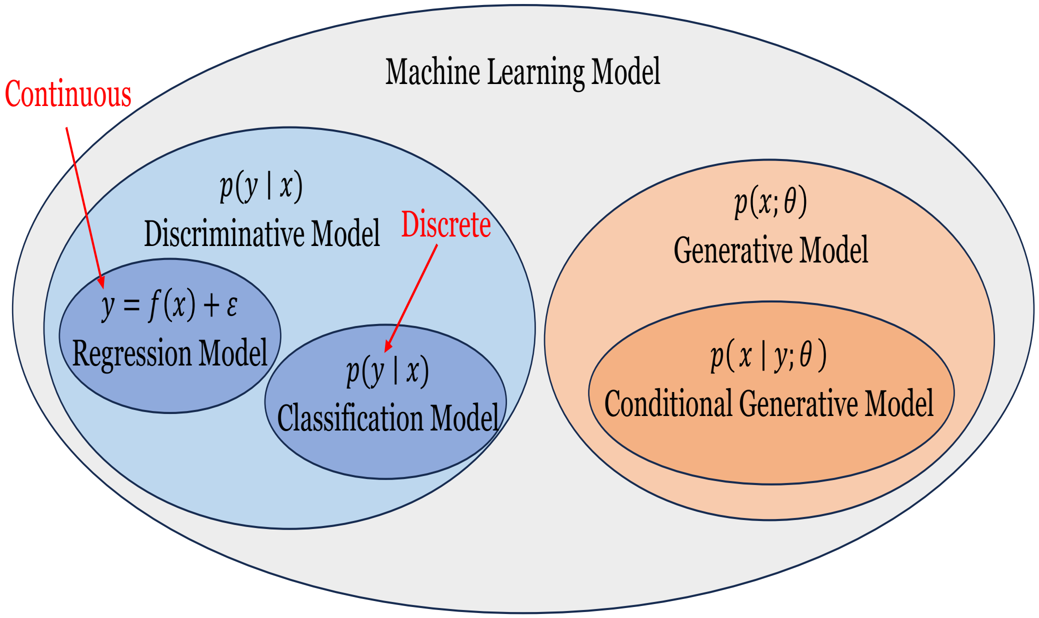
○ Generated data should approximate the "true distribution" of unseen data, p_{data}, as best as possible in some statistical sense, e.g.

$\min \text{distance}(p_{data}, p_{model})$.

☞ We examine the likelihood of the model, given the dataset ($\equiv$ MLE).

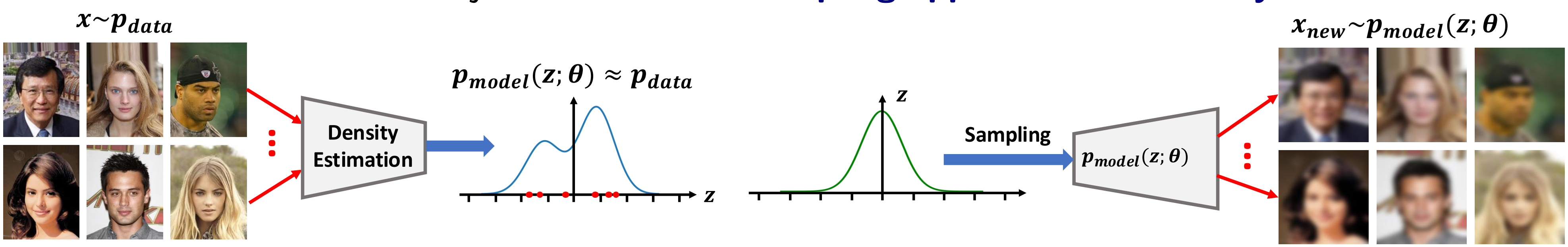
☞ This boils down to **maximising the likelihood** that the generated data will have similar distribution to true data of interest ↬ a backbone of Gen-AI

Example 12d: Big picture of learning data distributions
Most important general cases



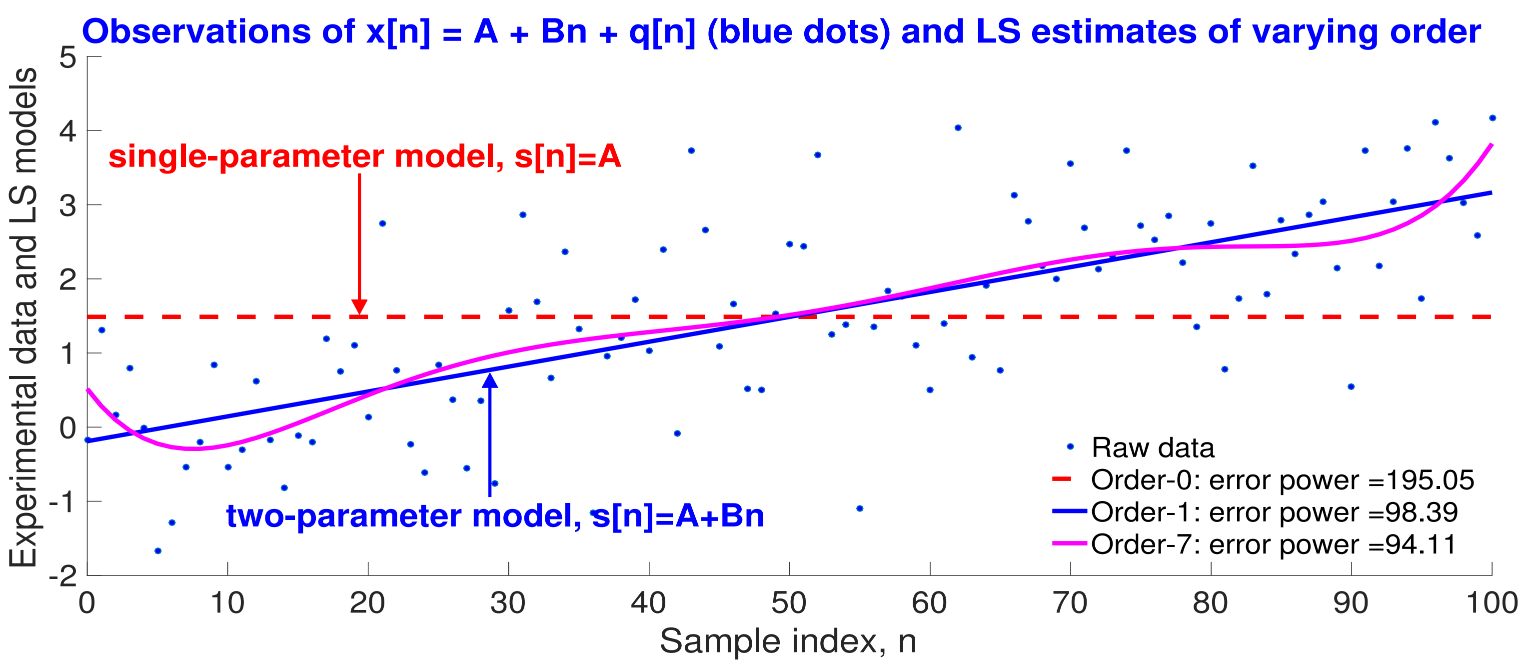
- ◦ $p(y|x;\theta) \looparrowright$ classification (discriminative model)
- ◦ $p(y|x;\theta) \looparrowright$ regression
- ◦ $p(x;\theta) \looparrowright$ generative model (e.g. VAE, GANN)
- ◦ $p(x|y;\theta) \looparrowright$ conditional generative model

☞ Generative models learn a joint distribution: **sampling applications** or **density estimation**



☞ $\arg\min_{p_\theta} D_{KL}(p_{data}||p_{model}) \equiv$ **Max. Likelihood Est.** $\arg\max_{p_\theta} \log p_{model}(\mathbf{x};\boldsymbol{\theta})$

- The LS approach can be interpreted as the problem of approximating a data vector $\mathbf{x} \in \mathbb{R}^N$ by another vector $\hat{\mathbf{s}}$ which is a linear combination of vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_p\}$ that lie in a p-dimensional subspace $S \in \mathbb{R}^p \subset \mathbb{R}^N$
- The problem is solved by choosing $\hat{\mathbf{s}}$ so as to be an orthogonal projection of $\mathbf{x}$ on the subspace spanned by $\mathbf{h}_i, i = 1, \ldots, p$
- The LS estimator is very sensitive to the correct deterministic model of $\mathbf{s}$, as shown in the figure below for the LS fit of $x[n] = A + Bn + q[n]$.



Least Squares Regression (LSR): A brief summary

Linear regression $\looparrowright$ relationship between two variables based on a line of best fit

Consider a line fit: $\boldsymbol{y} = \beta \boldsymbol{x} + \boldsymbol{e} \iff y_i = \beta x_i + e_i \quad i \in \{1, \dots, N\}$

○ Least Squares regression (LSR) aims to minimise the sum of the squares of the differences between the observed and predicted values

$$\underset{\beta}{\operatorname{argmin}} \, ||\boldsymbol{y} - \beta \boldsymbol{x}||_2^2 \iff \underset{\beta}{\operatorname{argmin}} \, ||\boldsymbol{e}||_2^2$$

○ We say that we regress y onto x, with β as the regression coefficient.

Common terminologies for Least Squares Regression

	Econometrics	Statistics	Machine Learning
$\boldsymbol{y}$	Dependent Var., Estimate	Explained V., Response, Regressand	True Label, Criterion
β	Coefficients	Coefficients	Parameters
$\boldsymbol{x}$	Independent Var., Predictor	Explanatory Var. Regressor	Features, Predictors
$\boldsymbol{e}$	Residual	Error	Prediction Error

LS Regression: Capital Asset Pricing Model (CAPM)

We here employ a block-LS approach, over blocks of 22 days

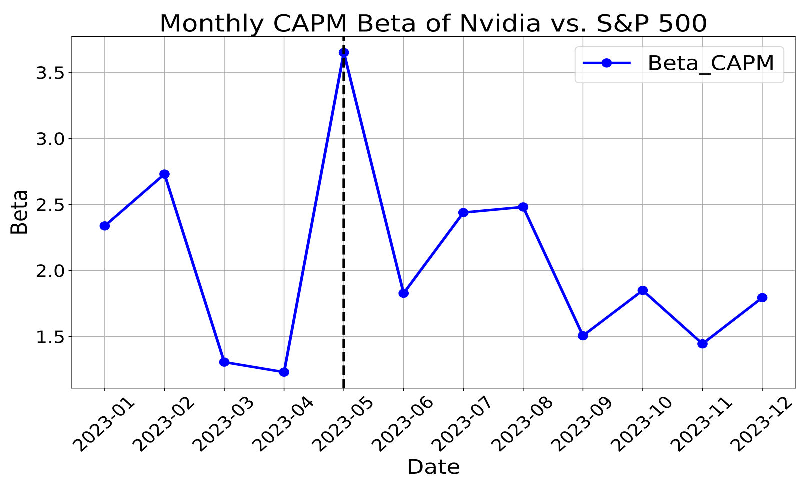
Asset return, R_i, risk-free interest rate, R_f, and market return, R_m, (S&P500 return) are all known. We consider log-returns.

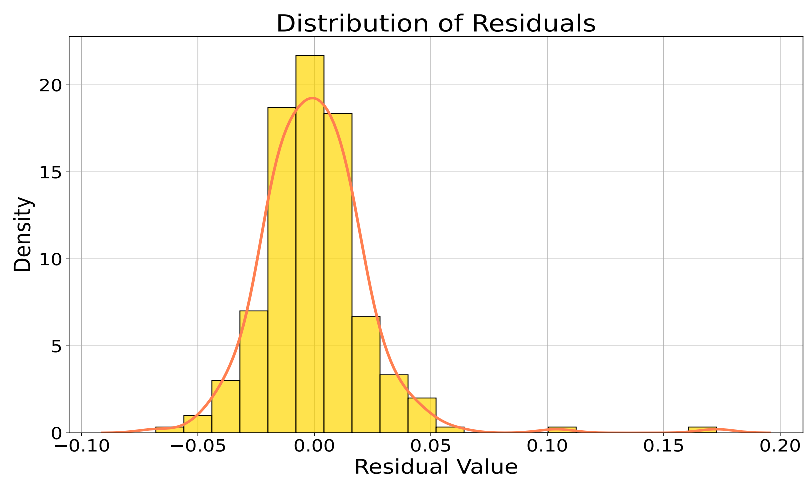
☞ We can now perform LS regression to obtain the value of β.

Each month has 22 trading days. Then, the CAPM states that

$$\begin{bmatrix} R_{i;day1} - R_f \\ R_{i;day2} - R_f \\ \vdots \\ R_{i;day22} - R_f \end{bmatrix} = \beta \begin{bmatrix} R_{m;day1} - R_f \\ R_{m;day2} - R_f \\ \vdots \\ R_{m;day22} - R_f \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{22} \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{r_i} = \beta\, \boldsymbol{r_m} + \boldsymbol{e}$$

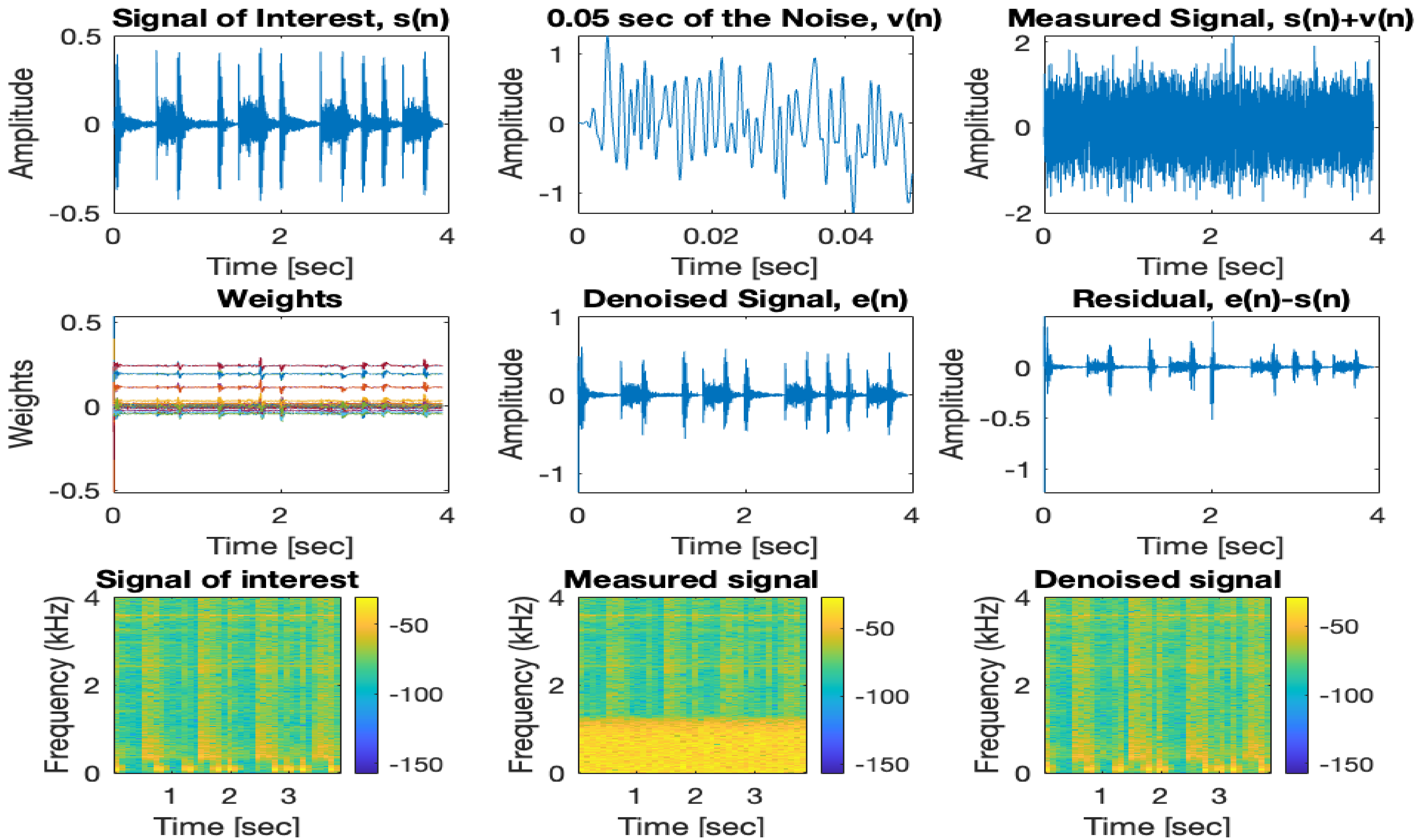
Therefore, the LS estimate: $\hat{\beta} = (\boldsymbol{r_m}^{\mathbf{T}} \boldsymbol{r_m})^{-1} \boldsymbol{r_m}^{\mathbf{T}} \boldsymbol{r_i}$





Sequential LS for streaming data: Noise cancelling headphones ($\lambda = 0.99$)

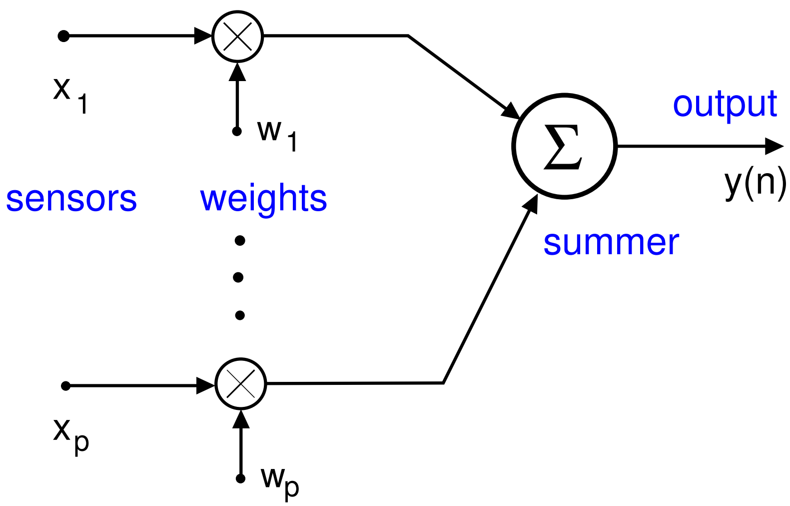
Denoising_SLS_GUI.m

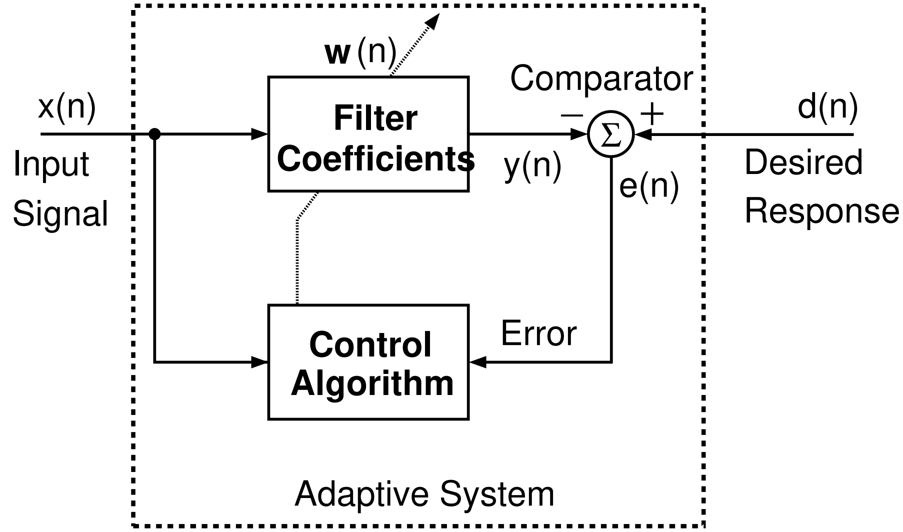


Lecture 7: Adaptive Learning and Inference

from fixed imp. resp. $\mathbf{h}$ in digital filters to a time-varying $\mathbf{w}(n)$ in adaptive filters

Consider a set of p sensors at different points in space (filter order p)





- The sensor signals are weighted by the corresponding set of **time–varying** filter parameters $\mathbf{w}(n) = [w_1(n), \ldots, w_p(n)]^T$ (weights)
- The weighted signals are then summed to produce the output

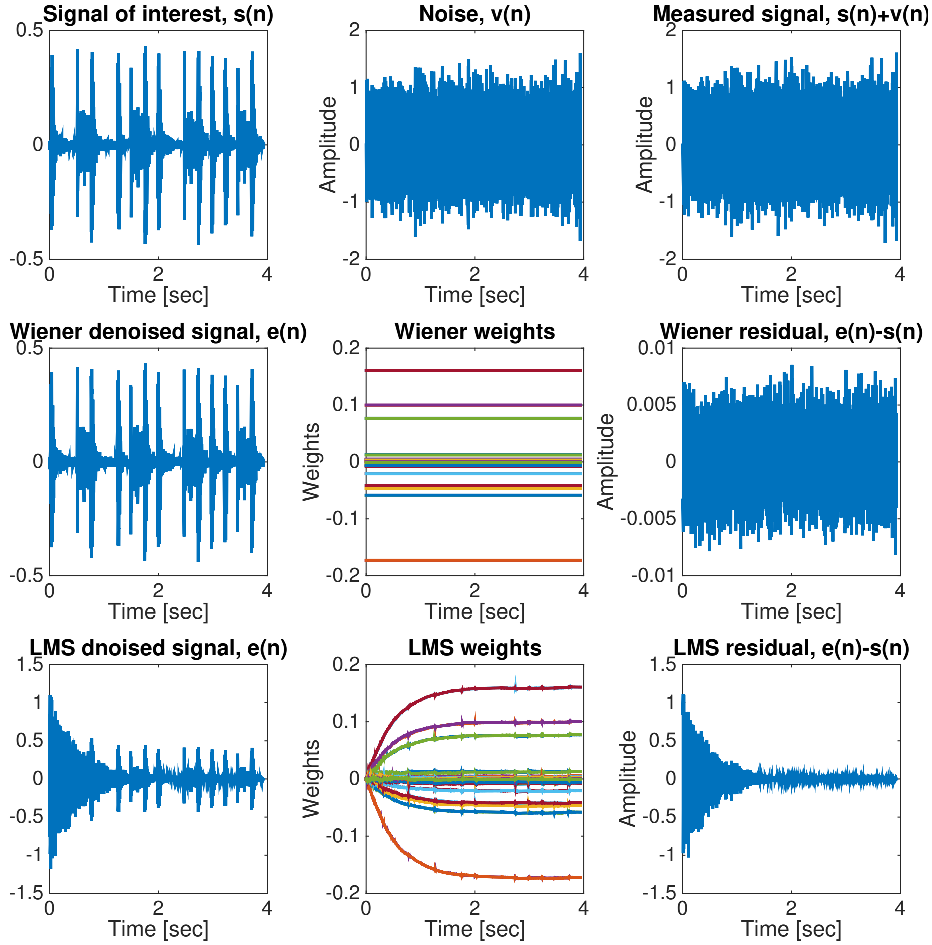
$$y(n) = \sum_{i=1}^{p} w_i(n)x_i(n) = \mathbf{w}^T(n)\mathbf{x}(n) = \mathbf{x}^T(n)\mathbf{w}(n) \qquad n = 0, 1, 2, \ldots$$

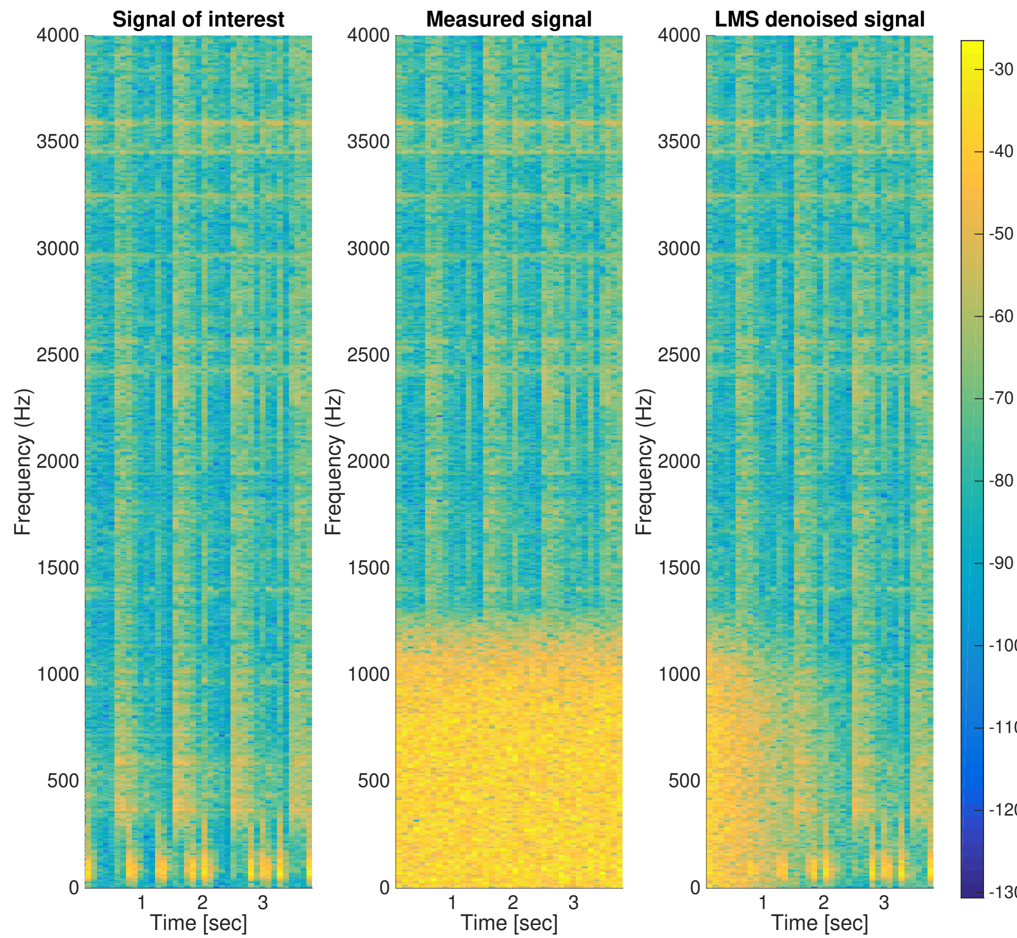
where $\mathbf{x}^T(n) = [x_1(n), \ldots, x_p(n)], \quad \mathbf{w}^T(n) = [w_1(n), \ldots, w_p(n)]$

Example: Adaptive filter for cancellation of cockpit noise

ALE_Handel, Denoising_Reference_Drum_WienerAndLMS

Consider an adaptive noise cancellation problem, like that in noise cancelling headphones when you are listening to music on the plane.



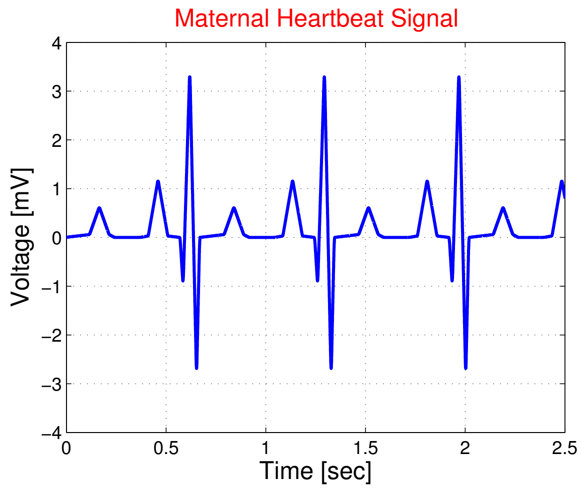


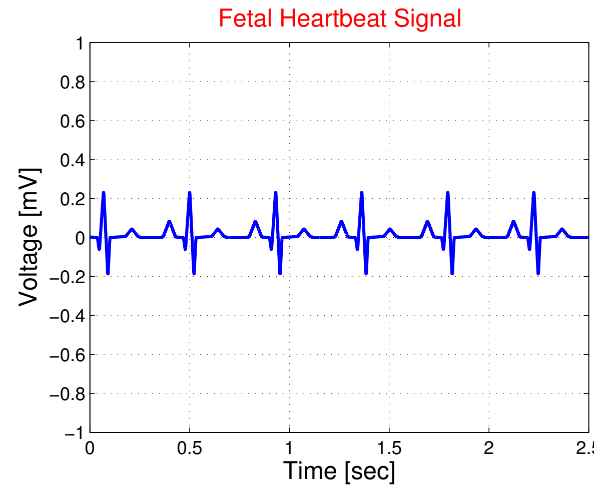
Right: The time–frequency representation of the performance of the LMS algorithm

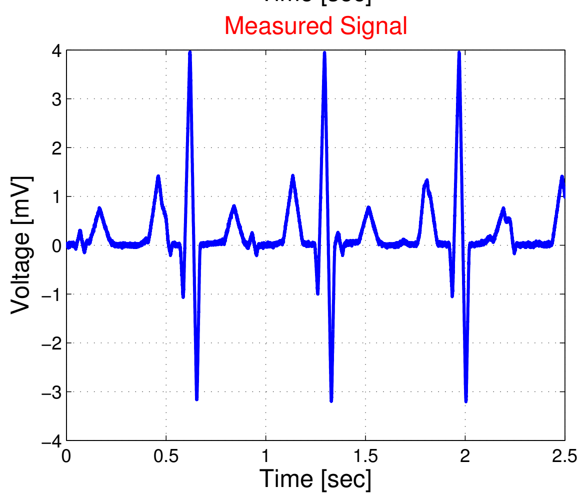
Adaptive noise cancellation: A biomedical example

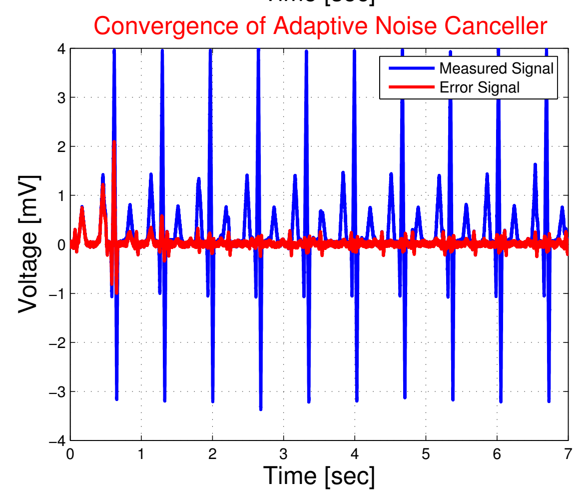
Maternal ECG signal

Foetal heartbeat





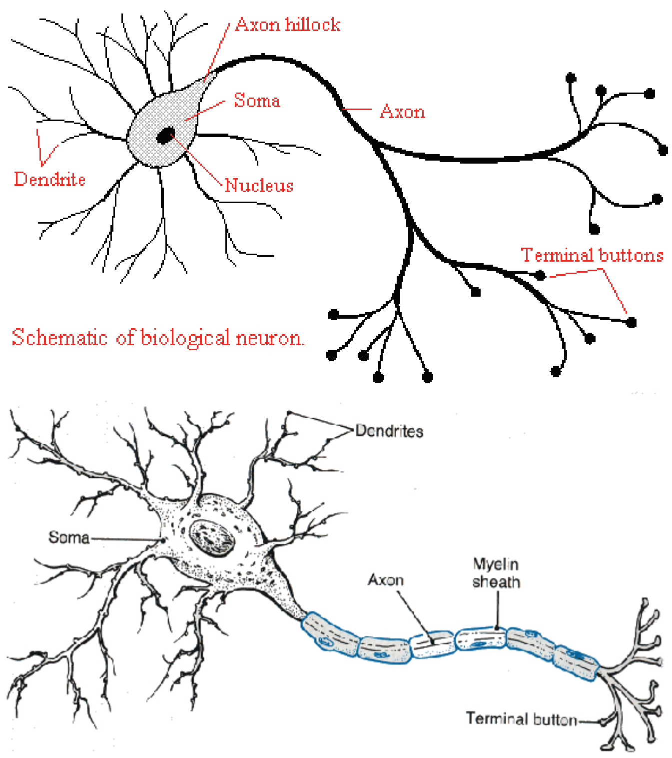




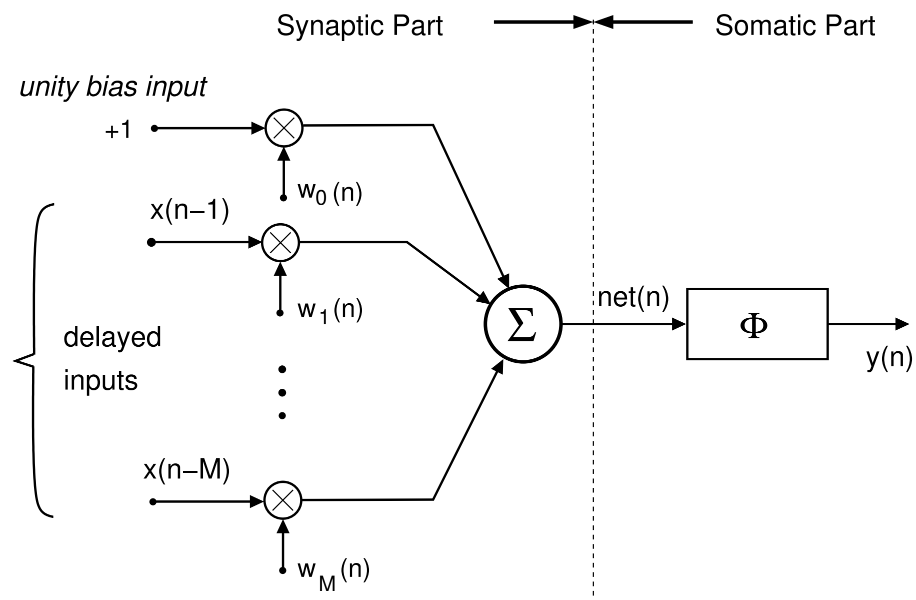
Measured foetal ECG

Maternal and foetal ECG

Artificial neuron: Introduction to neural adaptive filters



Biological neuron



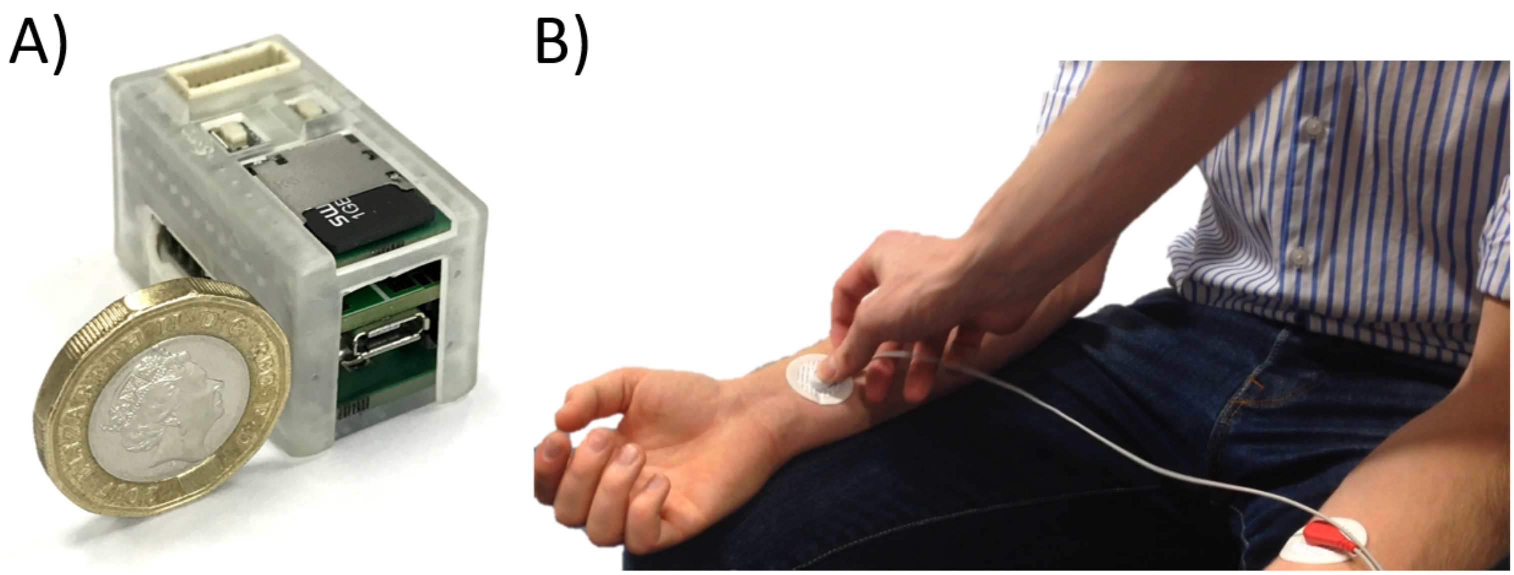
Model of an artificial neuron

- delayed inputs $\mathbf{x}$
- bias input with unity value
- sumer and multipliers
- output nonlinearity

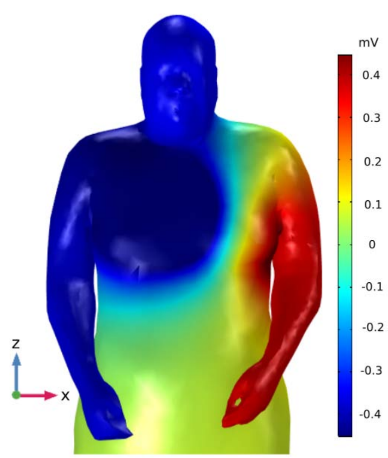
Coursework: Your own speech and biosignal recordings

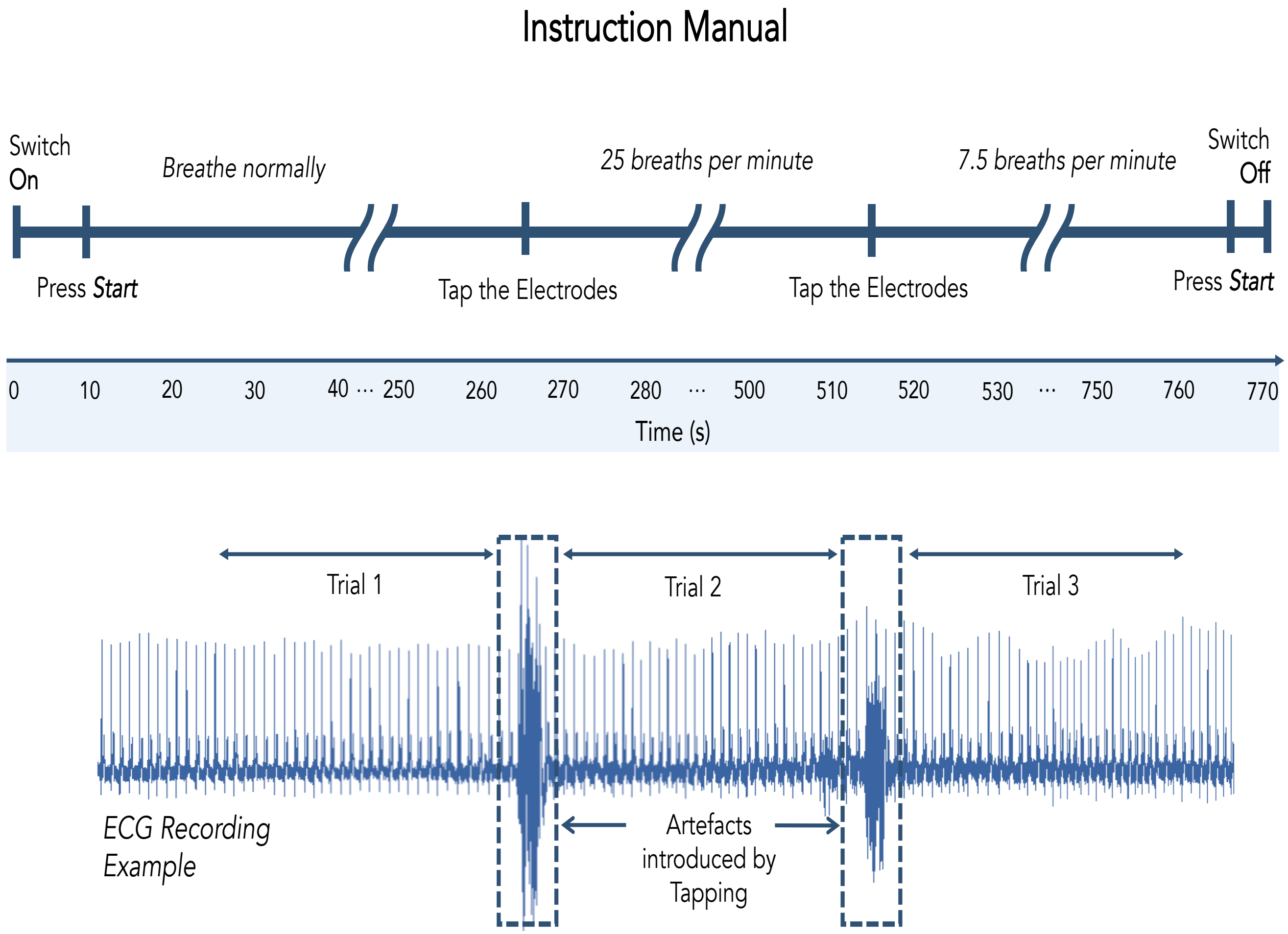
- Our own custom-made portable signal acquisition device – the BioBoard – is designed to record any biopotentials, such as the Electrocardiogram (ECG), Electroencephalogram (EEG), from up to eight channels

- It consists of an analogue-to-digital converter (ADC), a microcontroller, a secure digital (SD) card slot to store the data, and Bluetooth link



Coursework: Recording your own ECG

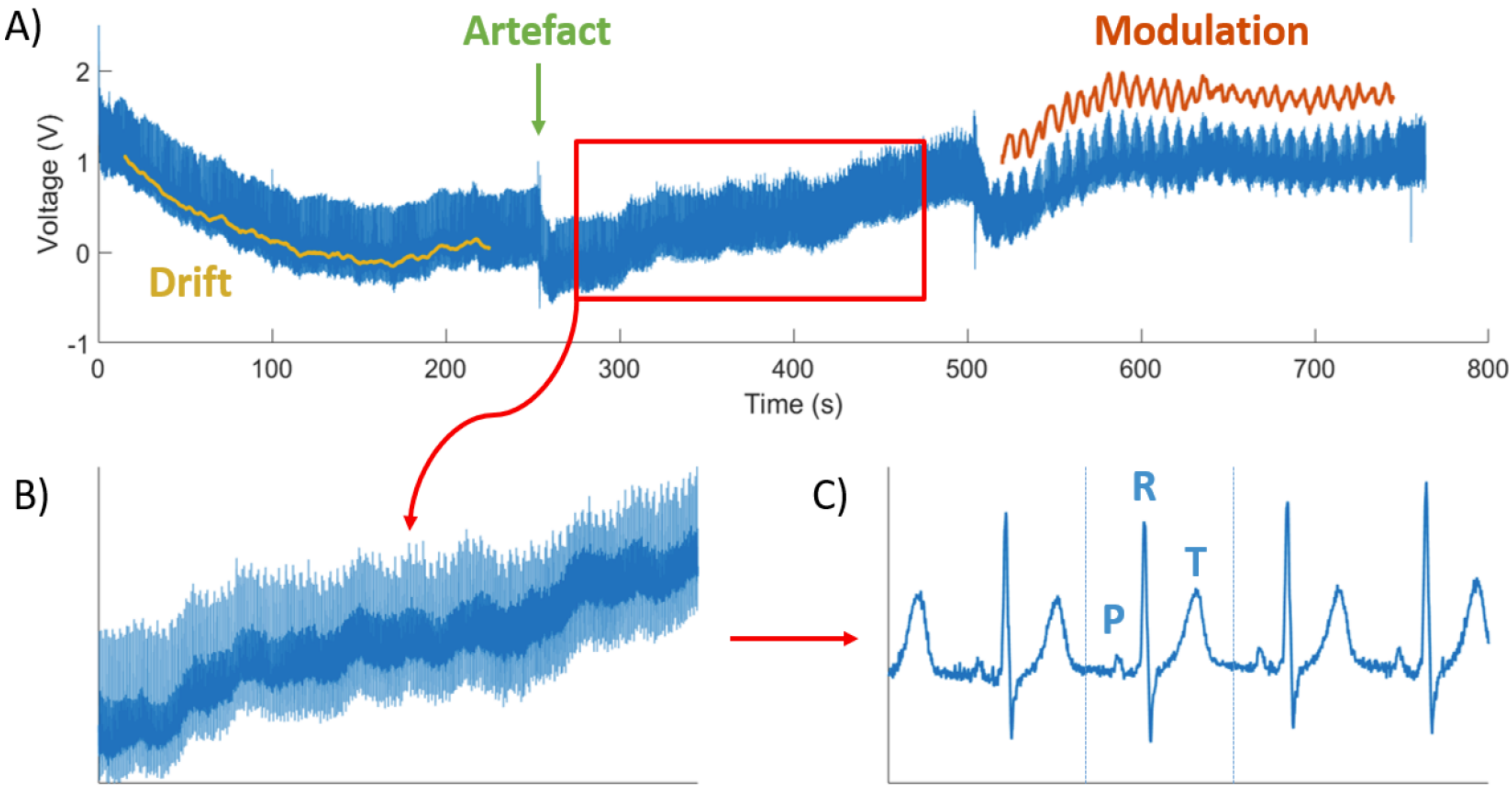




Left: Electric heart potentials on human body. Right: Experiment protocol

Coursework: Gain experience with real–world data

Example relevant for eHealth: Estimate your own ECG from your wrists.



Course format

Lecture notes with problem/answer sets and coursework.

- Coursework involves the implementation of the algorithms we discuss in the class
- We will regularly discuss coursework and Matlab implementation

  Prerequisites:

  - ⊛ There are no prerequisites, although DSP and basic probability would be useful
  - ⊛ The course is aimed to be self-contained
  - ⊛ Due to algorithm implementation, knowledge of Matlab is important

Assessment:

100% Coursework assignments. There are 5 Assignments (from random signals to audio denoising) ↬ Matlab based

Feedback ↬ after completing Assignment 1

Reference material

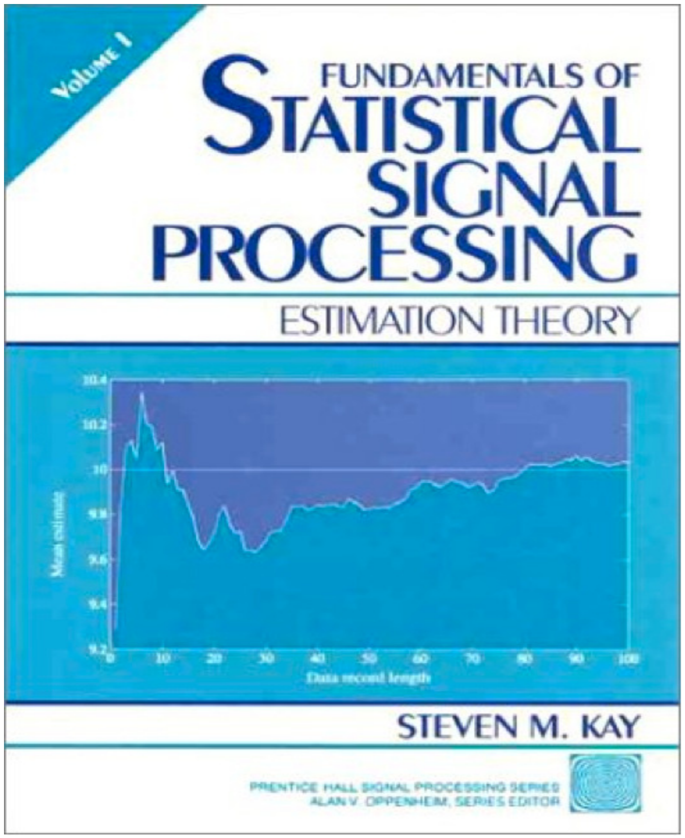
- There is no single textbook that covers all the material in the course
- We will use S. Kay's book for the first part of the course (an excellent text, covers most of the estimation theory, well worked-out examples, highly recommended, has many editions)
- For parametric modelling we will use the Box & Jenkins book (a 'bible' for time series analysis, easy to read, excellent examples, used by people in engineering, physics, finance, has many editions)
- For the least squares part, we will use M. Hayes' book (wider scope than Kay's book, less detailed derivations, a must have for practitioners)
- For further reading, the book by S. Haykin (Adaptive Filters) and D. Mandic & J. Chambers (Recurrent Neural Networks)

The course is self-contained: Most of the material is already in course notes
Course web page: www.commsp.ee.ic.ac.uk/∼mandic/Teaching

Lectures, additional reading, homework, problem sets, and other material will be put on the course webpage

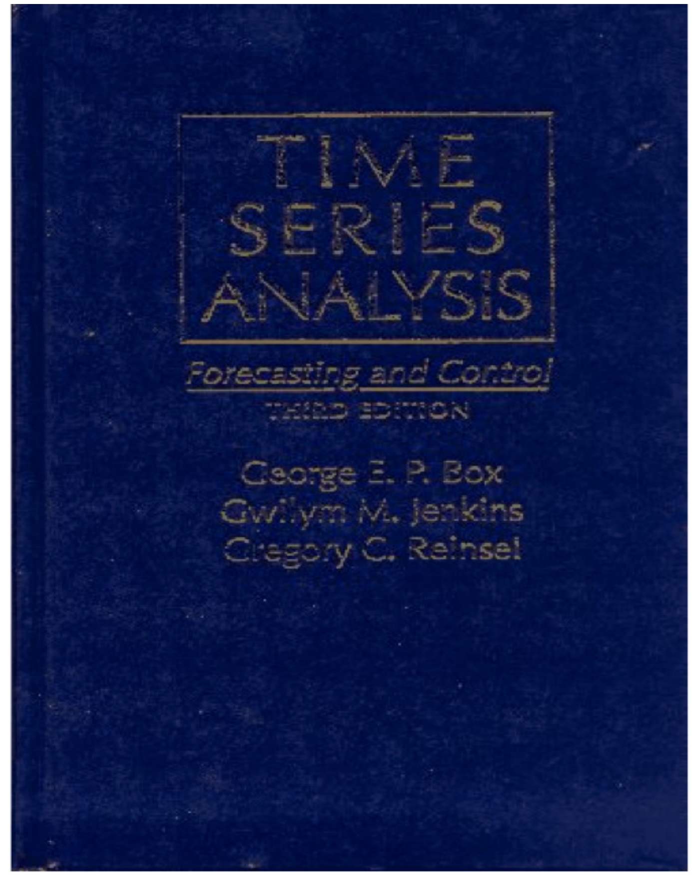
Textbooks: Recommended

S. Kay (*Estimation Theory*, several editions)



a comprehensive account of estimation theory

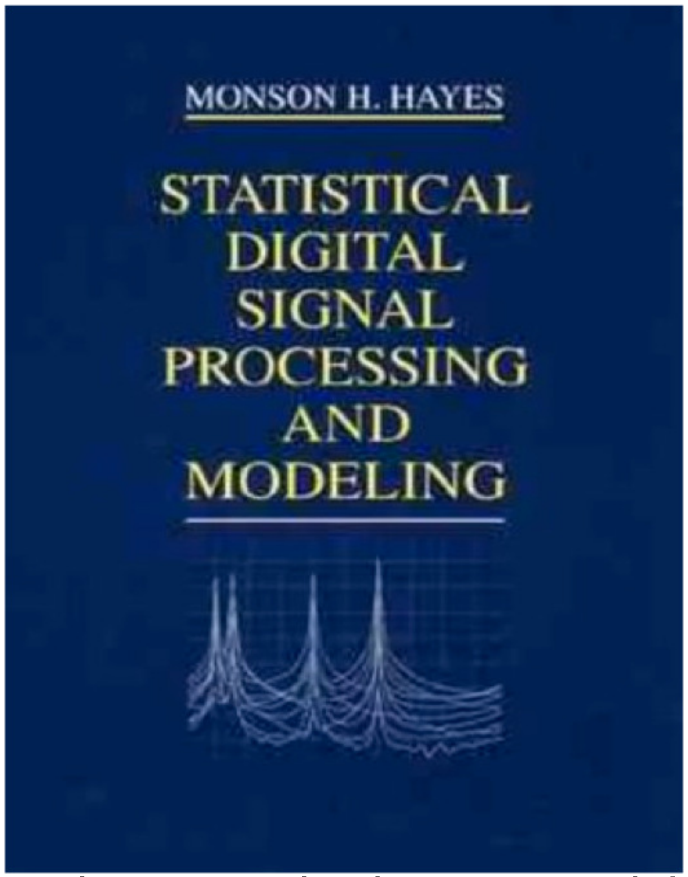
G. Box and G. Jenkins (*Time Series Analysis*, several editions)



linear stochastic models

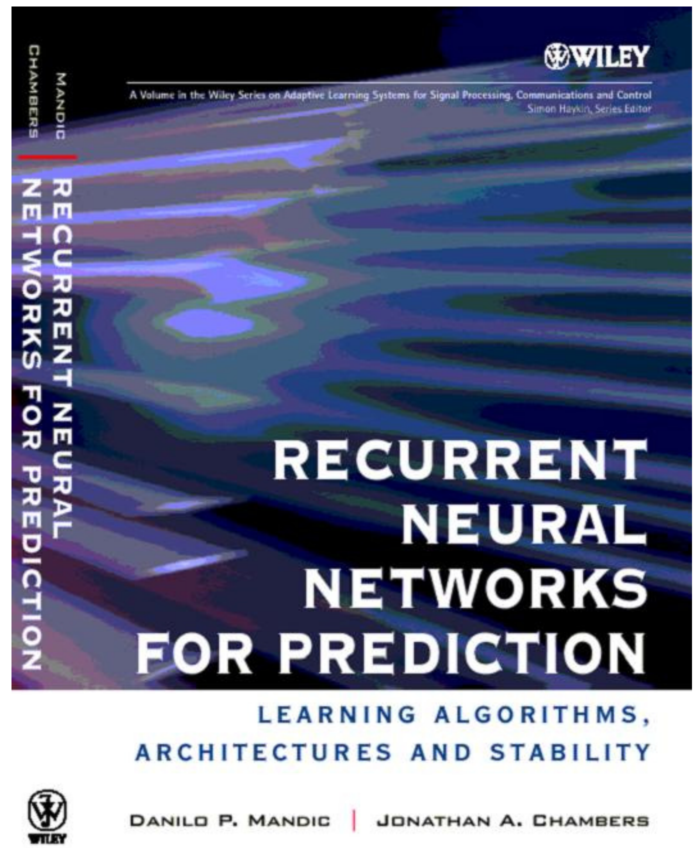
Textbooks: Additional reading (optional)

M. Hayes (*Statistical Signal Processing and Modeling*, several editions)



stochastic and adaptive models

D. Mandic and J. Chambers (*Recurrent Neural Networks*, Wiley, 2001)



(what can I say) - neural models

Statistical Sig Proc & Inference $\looparrowright$ A stealth technology

- There will always be signals
- They always need processing
- There will always be new mathematics for processing them

☞ The future is bright $\looparrowright$ a lot to do for all of us!

SSP&I: Use your knowledge and not brute force when designing learning machines

Notes:

○

Notes:

○