3F3 Statistical Signal Processing

Howard Mei

January 12, 2021

1 Probability Space

1.1 Notation

- $x \in \mathbf{A}$ x is an element of $\mathbf{A}$ "Set membership"
- $\mathbf{A} \subseteq \mathbf{\Omega}$ $\mathbf{A}$ is a subset of $\mathbf{\Omega}$
- $\mathbf{A} \subset \mathbf{\Omega}$ $\mathbf{A}$ is a proper subset of $\mathbf{\Omega}$
- $\mathbf{A} \cup \mathbf{B}$ Union of two sets
- $\mathbf{A} \cap \mathbf{B}$ Intersection of two sets
- $\mathbf{A^c}$ Complementary Set
- $\mathbf{A} \backslash \mathbf{B}$ $\mathbf{A} \cap \mathbf{B^c}$ intersection of A with not B
- $\emptyset$ Empty set

1.2 Probability Space

- **Random experiment** is used to describe any situation which has a set of possible outcomes, each of which occurs with a particular probability.
- **Sample space** Ω is the set of all possible outcomes of the **random experiment**.
- **Event** any subset $\mathbf{A} \subseteq \mathbf{\Omega}$
- **Probability** P mapping/function from events to a number in the interval $[0, 1]$. Therefore, specify $\{P(\mathbf{A}), \mathbf{A} \subset \mathbf{\Omega}\}$
- **Probability Space** defined as: (Ω, P)
- **Indicator function** for a set or event E defined as:

$$\mathbb{I}_E(t) = \begin{cases} 1 \text{ if } t \in E, \\ 0 \text{ if } t \notin E \end{cases}$$

- Examples:
  - Toss a coin twice. $\Omega = \{HH, HT, TH, TT\}$ - Finite set

– The temperature is a perturbation of seasonal average. $\Omega = (-\infty, \infty)$ - Real line

– Toss a coin n times. One elementary outcome is $\omega = (o_1, o_2, ..., o_n)$

$$\Omega = \{\omega = (o_1, o_2, ..., o_n) : o_i \in \{H, T\}\}.$$

– Toss a coin n times, the event **E** that the first head Occurs on third toss is:

$$\mathbf{E} = \{\omega = (T, T, H, o_4, o_5, ..., o_n) : o_i \in \{H, T\} \text{ for } i > 3\}.$$

$$P(\mathbf{E}) = (1/2)^3$$

1.3 Axioms of probability

A probability P assigns each event **E**, $\mathbf{E} \subset \Omega$, a number in [0,1] and P must satisfy following properties:

- $P(\Omega) = 1$
- For events A,B such that $\mathbf{A} \cap \mathbf{B} = \emptyset$ (i.e. disjoint) then $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$
- if $A_1, A_2...$ are disjoint then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.
- The third one implies the second one.

Examples:

(i) Show that, if event $\mathbf{A} \subset \mathbf{B}$ then $P(A) \leq P(B)$.

$$B = (B \cap A^c) \cup A = (B \backslash A) \cup A$$

$$P(B) = P(B \backslash A) + P(A) \leq P(A)$$

(ii) Show that, $P(A^c) = 1 - P(A)$

$$\Omega = A \cup A^c$$

$$P(\Omega) = P(A) + P(A^c) = 1$$

(iii) Defining P: Ω is a finite discrete set, i.e. $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$. Let $p_1, p_2, ..., p_n$ be non negative numbers that add to 1. For any event A, set,

$$P(A) = \sum_{i=1}^{n} \mathbb{I}_A(\omega_1) P_i$$

Let $P_i = 1/n$. Then

$$P(\{\omega_i\}) = p_i = 1/n$$

i.e. each outcome is equally likely. This is the *uniform probability distribution.*

1.4 Conditional Probability

- Definition: The conditional probability of event A occurring given that event B has occurred :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ for } P(B) > 0$$

  - Think of $P(A|B)$ as the fraction of times A occurs among those in which B occurs.
  - AB is shorthand for $A \cup B$
  - Example: Verify any set given set G is a probability i.e. $P(\cdot|G)$ is a probability

$$\text{Firstly, } P(\Omega|G) = P(\Omega \cup G)/p(G) = 1$$

$$\begin{aligned} \text{Secondly, for disjoint events A and B } P(A \cap B|G) &= P(AG \cap BG)/p(G) \\ &= (P(AG) + P(BG))/p(G) \\ &= P(A|G) + P(B|G) \end{aligned}$$

- Probability Chain Rule

$$P(A_1...A_n) = P(A_1)P(A_2|A_1)...P(A_n|A_{n-1}, ..., A_1) = P(A_1)\prod_{i=2}^{n} P(A_i|A_{i-1}, ..., A_1) = \prod_{i=1}^{n} P(A_i|A_{i-1}, ..., A_1)$$

- Independence: two events A and B are independent if

$$P(AB) = P(A \cup B) = P(A)P(B)$$

  - if A and B are independent then $P(A|B) = P(A)$

- *Bayes' Theorem*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  - Example: A is the event the email is spam and B is the event the email contains "free". We know $P(B|A) = 0.8$ and $P(B|not\ A) = 0.1$ and $P(A) = 0.25$ What is the probability the email is spam given the email contains "Free"?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.8 * 0.25}{0.8 * 0.25 + 0.1 * 0.75} = 0.727$$

  - This is an example of an *expert* system.

1.5 Random Variables

- Definition: Given a probability space(Ω, P), a random variable is a function $X(\omega)$ which maps each element ω of the sample space Ω onto a point on the real line.

  - Example: Flipping a coin twice. Sample Space: $\Omega = \{HH, HT, TH, TT\}$ Define $X(\omega)$ be the number of heads.

ω	$P(\{\omega\})$	$X(\omega)$
TT	0.25	0
TH	0.25	1
HT	0.25	1
HH	0.25	2

x	$\Pr(X = x)$
0	0.25
1	0.5
2	0.25

  - The second table does not mention the sample space. The range of X is listed along with the probability associated.
  - However, there is a sample space lurking behind every definition of a rv.
  - The Probability that $X = x$ is inherited from the definition of (Ω, P) and the mapping $X(\omega)$
  - For any set $A \subset (-\infty, \infty)$, we define

$$Pr(X \in A) = P(\{\omega : X(\omega) \in A\})$$

- Discrete random variable: range is a finite set, say $\{x_1, ..., x_i, ..., x_M\}$ or a countable set, say$\{x_1, x_2, ...\}$.
  - A set E is countable if you can define a one-to-one mapping from E to the set of integers .
  - Examples: all rational number, all even number. The interval$[0, 1]$ is not countable.
  - Definition: Discrete rv X with range $\{x_1, x_2, ...\}$, the pmf is the function $p_x : \{x_1, x_2, ...\} \to [0, 1]$ where

$$p_X(x_i) = Pr(X = x_i) \text{and } \sum_{i=1}^{\infty} p_X(x_i) = 1$$

    The pmf is a complete description: for any set A,

$$Pr(X \in A) = \sum_{i=1}^{\infty} \mathbb{I}_A(x_i) p_X(x_i)$$

- Continuous random variable: defined as having a probability density function(pdf)
  - Definition: A random variable is continuous if there exists a non-negative function $f_X(x) \geq 0$ such that $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$ and for any set A

$$Pr(X \in A) = \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_X(x)\, dx$$

  - Example: $A = [a, b]$ then

$$Pr(X \in A) = Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

    * pdf f_X assigns 0 probability to any particular point $x \in \mathbb{R}$ Thus $Pr(X = x) = 0$ for all x.

$$Pr(X \in [a, b]) = Pr(X \in (a, b]) = Pr(X \in (a, b))$$

    * This means a continuous rv has no concentration of probability at points like a discrete rv does

- Cumulative distribution function: Describe both discrete and continuous random variables and is defined to be

$$F_X(x) = Pr(X \leq x)$$

Properties:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(x)$ is non-decreasing as x increases
3. $\Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$
4. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$
5. If X is a continuous r.v. then $F_X(x)$ is continuous
6. If X is discrete then $F_X(x)$ is right-continuous: $F_X(x) = \lim_{t\downarrow x} F(t)$ for all x

For Property 6

- For a discrete rv with range $x_1, ..., x_i, ..., x_M$

$$F_X(x) = \sum_{j=1}^{M} P(x_j)\mathbb{I}_{[x_j,\infty)}(x) \qquad (\ [\ \text{touch}\ (\ \text{not touch})$$

  is a step function

- CDF and PDF for continuous rv

$$F_X(x) = Pr(X \leq x) = \int_{-\infty}^{x} f_x(t)\, dt$$

$$f_X(t) = \frac{dF_X(t)}{dx}$$

- CDF is useful when transformation of a random variable

$$\begin{aligned} Y &= r(X) \qquad \text{r is a strictly increasing function} \\ F_Y(y) &= Pr(Y \leq y) \\ &= Pr(r(X) \leq y) \\ &= Pr(X \leq r^{-1}(y)) \\ &= F_X(r^{-1}(y)) \\ f_Y(y) &= f_X(r^{-1}(y)) * \frac{dr^{-1}(y)}{dy} \end{aligned}$$

2 Multivariates

2.1 Bivariates

2.1.1 Discrete bivariates

- joint pmf: $p_{X,Y}(x_i, y_j) = Pr(X = x_i, Y = y_j)$
- marginal pmf:

$$P_X(x_k) = \sum_{j=1}^{n} P_{X,Y}(x_k, y_j), \qquad P_Y(y_k) = \sum_{i=1}^{m} P_{X,Y}(x_i, y_k)$$

- Independent if:
$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \text{for all (x,y)}$$
- Conditional Probability
$$p_{X|Y}(x|y) = \frac{p(X,Y)(x,y)}{P_Y(y)}$$

2.1.2 Continuous bivariates

- For continuous random variables X and Y, we call $f(x,y)$ their **Joint probability density function**:
  - $\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(x,y)\,dx\right)\,dy = 1$ and
  - for any sets(events) $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$
$$Pr(X \in A, Y \in B) = \int_{-\infty}^{\infty} \mathbb{I}_B(y)\left(\int_{-\infty}^{\infty} \mathbb{I}_A(x) f(x,y)\,dx\right)\,dy$$
- Independent
$$\text{If and only if: } f_{X,Y}(x,y) = f_X(x)f_Y(y)$$
- Conditional probability density function:
$$f_{X|Y}(x|y) = \frac{f(X,Y)(x,y)}{f_Y(y)}$$
Moreover, for all sets A
$$Pr(X \in A|Y = y) = \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_{X|Y}(x|y)\,dx$$
- Example: Let X_1, X_2 be two independent rvs with $f_1(x_1), f_2(x_2)$ and let $Y = X_1 + X_2$. Find the pdf $f_{X_1,y}$ and f_Y.

  Write the joint pdf using conditional pdf formula:
$$f_{X_1,y}(x_1,y) = f_1(x_1)fY|X_1(y|x_1).$$
  Since $Y = X_2 + x_1$, $fY|X_1(y|x_1) = f_2(y - x_1)$
$$f_Y(y) = \int_{-\infty}^{\infty} f_2(y - x_1)f_1(x_1)\,dx_1$$
  which is the convolution of f_1 and f_2

2.1.3 Expected Value Operations

- Expectation
  - Definition: The *Expected value* or *mean value* or *first moment* of X is
$$\mathbb{E}\{X\} = \begin{cases} \sum_x & xp_X(x) & \text{Discrete} \\ \int_{-\infty}^{\infty} & xf_X(x)\,dx & \text{Continuous} \end{cases}$$

- Expectation of a function of rv
  - Definition: For any function $r(\cdot)$ compute $\mathbb{E}\{r(X)\}$ by replacing x in the above formulae with $r(x)$ For example, the higher moments are $\mathbb{E}(X^n)$ set $r(X) = X^n$
  - Example: For an event A:

$$\mathbb{E}\{\mathbb{I}_A(X)\} = \begin{cases} \sum_x & \mathbb{I}_A(X)p_X(x) & \text{Discrete} \\ \int_{-\infty}^{\infty} & \mathbb{I}_A(X)f_X(x)\,dx & \text{Continuous} \end{cases}$$

    Then $\mathbb{E}\{\mathbb{I}_A(X)\} = \Pr\{X \in A\}$
  - Example: Take a unit length stick and break it at random. Find the mean of the long piece. Call the longer piece Y and the break point X. Then X is a uniform rv in $[0, 1]$, $Y = \max\{X, 1 - X\}$ and,

$$\begin{aligned} \mathbb{E}Y &= \mathbb{E}(max\{X, 1 - X\}) \\ &= \int_{-\infty}^{\infty} max\{x, 1 - x\} f_X(x)\,dx \\ &= \int_0^1 max\{x, 1 - x\}\,dx \\ &= \int_0^0 .5(1 - x)\,dx + \int_0 .5^1 x\,dx = 0.75 \end{aligned}$$

- Expectation of a function of bivariates
  - Definition: The mean of a function $r(X, Y)$ of the bivariate (X, Y) is

$$\mathbb{E}\{r(X,Y)\} = \begin{cases} \sum_y \sum_x & r(x,y)p_{X,Y}(x,y) & \text{Discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} & r(x,y)f_{X,Y}(x,y)\,dxdy & \text{Continuous} \end{cases}$$

  - The conditional expectation is

$$\mathbb{E}\{r(X,Y)|Y = y\} = \begin{cases} \sum_x & r(x,y)p_{X|Y}(x|y) & \text{Discrete} \\ \int_{-\infty}^{\infty} & r(x,y)f_{X|Y}(x|y)\,dx & \text{Continuous} \end{cases}$$

  - By using conditional probability we can calculate $\mathbb{E}\{r(X, Y)\}$:

$$\begin{aligned} \mathbb{E}\{r(X,Y)\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x,y) f_{X,Y}(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x|y)\,dx \right) dy \\ &= \int_{-\infty}^{\infty} \mathbb{E}\{r(X,Y)|Y = y\} f_Y(y)\,dy \end{aligned}$$

- Rule of iterated expectation
  Discrete:

$$\mathbb{E}\{r(X,Y)\} = \mathbb{E}(\mathbb{E}\{r(X,Y)|Y\})$$

  Continuous:

$$\begin{aligned} \mathbb{E}\{r(X,Y)|Y = y\} &= \int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x|y)\,dx \\ \mathbb{E}\{r(X,Y)\} &= \int_{-\infty}^{\infty} \mathbb{E}\{r(X,Y)|Y = y\} f_Y(y)\,dy \end{aligned}$$

2.2 Multivariates

2.2.1 Definition

- Let $X_1, X_2, ..., X_n$ be n continuous/discrete random variables. We call $X = (X_1, ..., X_n \in \mathbb{R}^n$ a continuous/discrete random vector.
- Let $X = (X_1, ..., X_n \in \mathbb{R}^n$ be a continuous random vector. Let $f(x_1, ...x_n)$ be a non-negative function that integrates to 1. Then f is called the pdf of the random vector X if

$$Pr(X_1 \in A_1, ..., X_n \in A_n) = \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) ... \int_{-\infty}^{\infty} \mathbb{I}_{A1}(x_1) f(x_1, ...x_n) d_{x_1} ... d_{x_n}$$

- pdf of X_i is obtained by integrating $f(x_1, .., x_n)$ over the full range except x_i:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, .., x_n) d_{x_1} ... d_{x_{i-1}} d_{x_{i+1}} d_{x_n}$$

  This is called ith marginal of $f(x_1, .., x_n)$

2.2.2 Independence

- Definition: The n random variables $X_1, ...X_n$ are independent if and only if for every $A_1, ...A_n$

$$Pr(X_1 \in A_1, ..., X_n \in A_n) = Pr(X_1 \in A_1) ... Pr(X_n \in A_n)$$

- joint pdf = product of marginals:

$$f(x_1, ..., x_n) = f_{X_1}(x_1) ... f_{X_n}(x_n)$$

  - Example: The pdf $f(x_1, ..., x_n)$ of a Gaussian random vector $X = (X_1, ..., X_n)$ is

$$\frac{1}{(2\pi)^{n/2} (\det C)^{1/2}} \exp\left\{ -\frac{1}{2} (x - m) C^{-1} (x - m)^T \right\}$$

    Where $m = (m_1, ..., m_n)$ is the row vector of means and C is the covariance matrix

$$m_i = \mathbb{E}\{X_i\} \qquad \text{and} \qquad [C]_{i,j} = \mathbb{E}\{(X_i - m_i)(X_j - m_j)\}$$

    Show that if independent, $C_{i,j} = 0$ for $i \neq j$ then

$$f(x_1, ..., x_n) = f_{X_1}(x_1) ... f_{X_n}(x_n)$$

    Proof: Call $C_{i,i} = \sigma_i^2$

$$(x - m) C^{-1} (x - m)^T = \sum_{i=1}^{n} \frac{(x_i - m_i)^2}{\sigma_i^2}$$

    Hence $f(x_1, ..., x_n)$ is

$$\frac{1}{(2\pi)^{n/2} (\det C)^{1/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - m_i)^2}{\sigma_i^2} \right\}$$

$$= \frac{1}{\sqrt{(2\pi)}\sigma_1 ... \sqrt{(2\pi)}\sigma_n} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} \frac{(x_i - m_i)^2}{\sigma_i^2} \right\}$$

$$= f_{X_1}(x_1) ... f_{X_n}(x_n)$$

- If $X_1, ..., X_n$ are independent then

$$\mathbb{E}\{\prod_{i=1}^{n} X_i\} = \prod_{i=1}^{n} \mathbb{E}\{X_i\}$$

That is the expectation of the product is the product of expectation

2.2.3 Change of variables

- The change of variable formula can be applied to random vectors. Let

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g_1(X_1, ..., X_n) \\ \vdots \\ g_n(X_1, ..., X_n) \end{bmatrix}$$

or

$$Y = G(X)$$

- If G is invertible then $X = G^{-1}(Y)$. Let $H(Y) = G^{-1}(Y)$. So

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} h_1(Y_1, ..., Y_n) \\ \vdots \\ h_n(Y_1, ..., Y_n) \end{bmatrix}$$

- The *Jacobian* matrix of partial derivatives of $H(y)$ is formed:

$$J(y) = \begin{bmatrix} \frac{\partial}{\partial y_1} h_1 & \dots & \frac{\partial}{\partial y_n} h_1 \\ & \vdots & \\ \frac{\partial}{\partial y_1} h_n & \dots & \frac{\partial}{\partial y_n} h_n \end{bmatrix}$$

Then

$$f_Y(y) = f_X(H(y)) |\det J(y)|$$

  - Example: Let $X_1, ..., X_n$ be independent Gaussian rv where X_i is $\mathcal{N}(0, 1)$ Let S be an invertible matrix and m a column vector. Let $Y = m + SX$ where $X = (X_1, .., X_n)^T$. Show Y is also a Gaussian random vector.

    Use the Change of variable result:

$$H(Y) = S^{-1}(Y - m)$$

    The Jacobian Matrix$J(y)$:

$$J(y) = S^{-1}$$

    Applying change of variable formula gives

$$f_Y(y) = f_X(S^{-1}(y - m)) |\det S^{-1}|$$

    where $f_X(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2} x^T x\right\}$

$$f_Y(y) = \frac{|\det S^{-1}|}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(y - m)^T (S^{-1})^T S^{-1}(y - m)\right\}$$

    is the density of a Gaussian vector with mean m and covariance matrix SS^T . Note that $\det S^{-1} = 1/detS$, $\det(SS^T) = \det S \det S^T = (\det S)^2$

- An affine transformation of a Gaussian vector is still a Gaussian vector. This gives a method for generating any Gaussian vector from iid Gaussian random variables.
- To Generate a $\mathcal{N}(m, \Sigma)$ vector:
  * Decompose the symmetric matrix $\Sigma = SS^T$.
  * Output $m + SX$ where $X = (X_1, ..., X_n)^T$ where $X_1, ..., X_n$ are independent $\mathcal{N}(0, 1)$

2.2.4 Characteristic function

- Definition: The characteristic function of a discrete or continuous random variable X is:

$$\varphi_X(t) = \mathbb{E}\{\exp(itX)\}, \qquad t \in \mathbb{R}$$

For a random vector $X = (X_1, X_2, ..., X_n)$,

$$\varphi_X(t) = \mathbb{E}\{\exp(it^T X)\}, \qquad t \in \mathbb{R}^n$$

Similarly to Fourier Transform, the characteristic function uniquely describes a pdf.

- Example: Show $\varphi_X(t) = \exp(itX)\exp(-\frac{1}{2}\sigma^2 t^2)$ when X is a Gaussian random variable with mean μ and variance σ^2.

$$\begin{aligned}
&\mathbb{E}\{\exp(itX)\} \\
&= \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\
&= e^{it\mu} \int_{-\infty}^{\infty} e^{its} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}s^2\right) ds, \qquad \text{let } s = x - \mu \\
&= e^{it\mu} e^{-\frac{1}{2}\sigma^2 t^2} \qquad \text{Fourier transform table}
\end{aligned}$$

- Example: Compute the characteristic function $\varphi_Y(t)$ of $Y = \sum_{i=1}^n X_i$ where X_i are independent random variables.

$$\begin{aligned}
&\mathbb{E}\{\exp(itY)\} \\
&= \mathbb{E}\{\exp(itX_1)\exp(itX_2)...\exp(itX_n)\} \\
&= \mathbb{E}\{\exp(itX_1)\}\mathbb{E}\{\exp(itX_2)\}...\mathbb{E}\{\exp(itX_n)\} \\
&= \varphi_{X_1}(t)...\varphi_{X_n}(t)
\end{aligned}$$

- The characteristic function of the sum of independent random variables is the product of their individual characteristic functions.
- Example: (Moments) Using $\varphi_X(t)$, compute $\mathbb{E}\{X^n\}$

$$\frac{d^n}{dt^n}\varphi_X(t) = \mathbb{E}\left\{\frac{d^n}{dt^n}\exp(itX)\right\} = \mathbb{E}\{i^n X^n \exp(itX)\}$$

Thus $i^n\mathbb{E}\{X^n\} = \frac{d^n}{dt^n}\varphi_X(t=0)$ (Putting t=0 for the above equation and make the exponential go to 1)

- Equality of characteristic functions: Suppose that X and Y are random vectors with same characteristic functions: $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}^n$. Then X and Y have the same probability distribution

- Example using characteristic function: Let $X_1, X_2, ..., X_n$ be independent Gaussian random variables where X_i is $\mathcal{N}(0,1)$. Then $Y = m + SX$, where $m \in \mathbb{R}^d$ where $d < n$, is the multivariate Gaussian with mean m and covariance SS^T.

  Verify the result using characteristic function, that is let $t \in \mathbb{R}^d$ and compute $\mathbb{E}\{\exp(it^T Y)\}$

$$\begin{aligned}\exp(it^T Y) &= \exp(it^T m)\exp(it^T SX) \\ &= \exp(it^T m)\exp(ir_1 X_1)...\exp(ir_n X_n)\end{aligned}$$

  Where vector $r = t^T S$

$$\begin{aligned}&\mathbb{E}\{\exp(it^T Y)\} \\ &= \exp(it^T m)\mathbb{E}\{\exp(ir_1 X_1)...\exp(ir_n X_n)\} \\ &= \exp(it^T m)\exp(-\frac{1}{2}r_1^2)...\exp(-\frac{1}{2}r_n^2) \\ &= \exp(it^T m)\exp(-\frac{1}{2}t^T SS^T t)\end{aligned}$$

3 Random process

3.1 Definition of random process

- Definition:A discrete random (or stochastic) process is one of the following infinite collection of random variables

$$\{..., X_{-1}, X_0, X_1, ...\} \quad or \quad \{X_0, X_1, ...\} \quad or \quad \{X_1, X_2, ...\}$$

  Notation: $\{X_n\}_{n_i}^{j} = \{X_i, X_{i+1}, ..., X_j\}$

  - Example: Random phase cosine. Let $X_n = \cos(2\pi fn + \phi)$ where ϕ is a Uniform random variable drawn from $[0, 2\pi)$ To generate

$$\{X_n\}_{n=0}^{\infty} = \{X_0, X_1, ...\}$$

    first sample ϕ and then set

$$X_n = \cos(2\pi fn + \phi)$$

    for $n = 0, 1, ...$

  - Example: infinite collection of independent random variables
    Let $0 < q < 1$ and $U_1, U_2, ...$ be iid discrete random variables such that

$$Pr(U_n = 1) = q, \qquad Pr(U_n = -1) = 1 - q$$

  - Example: Random walk
    Generate the sequence $U_1, U_2, ...$ as in the previous example and define a new random process $X_0, X_1...$ as follows: set $X_0 = 0$ and

$$X_n = X_{n-1} + U_n$$

for $n > 0$
We could equivalently write

$$X_n = \begin{cases} X_{n-1} + 1 & w.p.q \\ X_{n-1} - 1 & w.p.1 - q \end{cases}$$

and $X_0 = 0$.

- Definition (Finite dimensional distributions)
  - To completely specify a discrete time random process $X_0, X_1, ...$, we must specify their joint probability density function
    $$f_{X_0,X_1,...,X_n}(x_0, x_1, ..., x_n)$$
    for all integers $n \geq 0$ when $X_0, X_1, ...$ is a collection of continuous random variables
  - For discrete time random process $X_0, X_1, ...$, we must specify their joint probability mass function
    $$p_{X_0,X_1,...,X_n}(x_0, x_1, ..., x_n)$$
    for all integers $n \geq 0$
  - For any fixed n, you can treat $(X_0, X_1, ..., X_n)$ as a random vector and just as in the case of random vectors, we use their joint pdf or joint pmf to describe how the random vector should be generated.
  - For many interesting random processes, specifying $p_{X_0,X_1,...,X_n}(x_0, x_1, ..., x_n$ is not too arduous. One such process which underpins many real world statistical models is a **Markov chain**.

3.2 Markov Chain

3.2.1 Introduction and Properties

- Example
  A gambler has initial wealth r bets and keep playing until wealth is R or zero. Amount bet is b at every bet. The random process now is:

$$X_{n+1} = \begin{cases} X_n & \text{if} X_n \in \{0, R\} = 0 \\ X_n + b & w.p.q \\ X_n - b & w.p.1 - q \end{cases}$$

The generate X_{n+1}, only the value of X_n is needed and not its past values. Any discrete time random process with this property is called a Markov process.
It can be shown that the probability of wealth doubling when $q \leq 0.5$is

$$\left[1 = \left(\frac{1-q}{q}\right)^{r/b}\right]^{-1}$$

	b=1pound	b=10pence	b=1pence
$q = 0.5$	0.5	0.5	0.5
$q = 0.49$	0.40	0.02	4.3×10^{-18}

Table 2: Probability of doubling wealth with an initial fortune of 10 pounds.

- Definition of Markov chain Let $\{X_n\}_{n\geq 0}$ be discrete random variables taking values in $S = \{1, ..., L\}$.
  - The transition probability matrix Q is a non-negative matrix

$$\begin{bmatrix} Q_{1,1} & Q_{1,2} & \cdots & Q_{1,L} \\ Q_{1,1} & Q_{1,2} & \cdots & Q_{1,L} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{L,1} & Q_{L,2} & \cdots & Q_{L,L} \end{bmatrix}$$

  and each row sums to one.

$$Q_{1,L} = Pr(X_{n+1} = L | X_n = 1)$$

  from state 1 jump to state L is the probability of L given current state is L
  - The conditional pmf of X_n given $X_0 = i_0, ..., X_{n-1} = i_{n-1}$

$$Pr(X_n = i_n | X_0 = i_0, ..., X_{n-1} = i_{n-1}) = Q_{i_{n-1}, i_n} = Pr(X_n = i_n | X_{n-1} = i_{n-1})$$

- Example: Two state Markov chain
  - For a two state Markov Chain, $S = 1, 2$, let.

$$Q = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = \begin{bmatrix} Pr(X_{n+1} = 1 | X_n = 1) & Pr(X_{n+1} = 2 | X_n = 1) \\ Pr(X_{n+1} = 1 | X_n = 2) & Pr(X_{n+1} = 2 | X_n = 2) \end{bmatrix}$$

  - Think of MC as an evolving sequence of random variables, generate X_0, then X_1, then X_2 etc.
  - The chain jumps between values 1 and 2 according to Q of pictorially as above.
  - Assume the pmf of X_0 is

$$p_{X_0} = i = \lambda_i, \quad i = 1, ..., L$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_L \end{bmatrix}$$

    Note that $\lambda_i > 0$ and $\sum \lambda_i = 1$
  - The definition states how X_0 should be generated and then how every other X_k should.
  - The process just defined is called a Markov chain. The pair (λ, Q) completely defines the Markov Chain.
  - We call Q the *transition probability matrix* of the MC and λ the *initial distribution* of the chain.
- **Limited memory property** is known as Markov property: Only the most recent value $X_{n-1} = i_{n-1}$ is needed to generate X_n, that is we don't need to know the values before.

3.2.2 Further properties

1. **Marginals of a Markov Chain**.

   - Show that $p_{X_n}(i_n) = (\lambda Q^n)_{i_n}$ where λ is the row vector $\lambda = (\lambda_1, .., \lambda_L)$
   - The joint pmf can be expressed as

   $$p(i_0, ..., i_n) = \frac{p(i_0, ..., i_n)}{p(i_0, ..., i_{n-1})} \frac{p(i_0, ..., i_{n-1})}{p(i_0, ..., i_{n-2})} \cdots \frac{p(i_0, i_1)}{p(i_0)} p(i_0)$$

   or

   $$[p(i_0, ..., i_n) = p(i_n|i_0, ..., i_{n-1})p(i_{n-1}|i_0, ..., i_{n-2})...p(i_1|i_0)p(i_0).$$

   - Thus using the Markov property,

   $$p(i_0, ..., i_n) = \lambda_{i_0} Q_{i_0,i_1} ... Q_{i_{n-2},i_{n-1}} Q_{i_{n-1},i_n}$$

   - Summing over all possible values for $i_0, ..., i_{n-1}$ gives the result $p(i_n) = (\lambda Q^n)_{i_n}$

2. **Strict stationarity**

   - Definition
     - A discrete time random process $X_0, X_1, ...$ is strictly stationary if

     $$f_{X_0,...,X_k}(x_0, ..., x_k) = f_{X_m,...,X_{k+m}}(x_m, ..., x_{k+m})$$

     for all k and displacement $m > 0$. That is the joint pdf of $(X_0, ..., X_k)$ and $(X_m, ..., X_{m+k})$ are the same.
     - This definition apply to any discrete time process Markov or otherwise.
     - Strict stationarity means any two "sections" of the process

     $$(X_0, ..., X_k) \text{ and } (X_m, ..., X_{m+k})$$

     are statistically indistinguishable for any displacement m.
     - If $X_0, X_1, ...$ were *discrete* random variables then the *joint pmf* instead of joint pdf is used to define strict stationarity.
   - Example:
     - Let X_i be the average temperature of day i. Clearly $Pr(X_i \in A) \neq Pr(X_j \in A)$ for some A, i.e. X_i and X_j do not have the same probability distributions, if i is a day in summer and j a day in winter.
     - But if we remove the seasonal effects to obtain $\{Y_k = X_k - S_k\}_{k \in I}$. Y_k is the deviation of the measured temperature from the anticipated seasonal value S_k. It is sensible to assume $\{Y_k\}_{k \in I}$ is stationary.
     - Thus stationarity is a useful model for random deviations which is sometimes called noise.

3. Invariant distribution of a Markov chain.

   - Definition:
     - When the Markov chain is initialised in a very specific way.

– Consider the transition probability matrix Q with state-space S. The pmf $\pi = (\pi_i : i \in S)$ is invariant for Q if for all $j \in S$,

$$\sum_{i \in S} \pi_i Q_{i,j} = \pi_j.$$

or

$$\pi Q = \pi$$

for row vector pi

- Example:
  - For a two state Markov chain, $S = \{1, 2\}$, let

$$Q = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

  - Check that $\pi = [\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}]$ is invariant for Q or show that $\pi Q = \pi$,

$$\left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right] \times \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = \left[\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right]$$

  - there can be many pmf π satisfies the invariant properties check that any multiple of $\pi = [\beta, \alpha]$ is invariant for the Q.

4. **Fact**: *The Markov chain (π, Q) is strictly stationary*

- Verification: apply the pmf of X_n calculated before

$$\begin{aligned} p_{X_n}(i_n) =& (\pi Q^n)_{i_n} \\ =& (\pi Q\, Q^{n-1})_{i_n} \\ =& (\pi Q^{n-1})_{i_n} \quad (\pi Q = \pi) \\ & \vdots \\ =& \pi_{i_n} \end{aligned}$$

- Also, the pmf of $(X_m, ..., X_{m+k})$, for any $m \in \{0, 1, ...\}$, can be written as

$$\begin{aligned} p(i_m, ..., i_{m+k}) = & p(i_{m+k} | i_m, ..., i_{m+k-1}) \\ & \times P(i_{m+k-1} | i_m, ..., i_{m+k-2}) \\ & \vdots \\ & \times p(i_{m+1} | i_m) \\ & \times p(i_m) \\ = & \pi_{i_m} Q_{i_m, i_{m+1}} \cdots Q_{i_{m+k-1}, i_{m+k}}, \end{aligned}$$

  which follows from the Markov property and invariance, i.e. $p(i_m) = \pi_{i_m}$

- Thus the pmf of $(X_0, ..., X_k)$, setting $m = 0$, is

$$p(i_0, ..., i_k) = \pi_{i_0} Q_{i_0, i_1} ... Q_{i_{k-1}, i_k}.$$

- These two joint pmfs are equal, which implies strict stationarity.

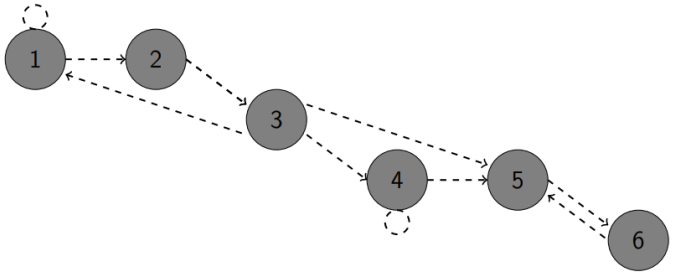
5. **Fact** *Ergodic theorem.* When the MC is irreducible then for any initial distribution λ, the sample (or empirical) average converges to $\sum_{i\in S}\pi_i r(i)$.

$$\underbrace{\frac{1}{n+1}\sum_{k=0}^{n} r(X_k)}_{\text{Empirical Mean}} \to \sum_{i\in S}\pi_i r(i).$$

- Definition:
  - An irreducible Markov chain refers to a chain where all state values in S communicate with each other.
  - This means for any pair of states (i, j), the Markov chain starting in i will eventually visit j and vice versa.
  - We can identify communicating states by inspecting the elements of the transition probability matrix
- Example:
  - The communicating sets of states for a MC with transition probability matrix

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

    is $\{1, 2, 3\}$, $\{4\}$ and $\{5, 6\}$. So this chain is not irreducible and the Ergodic theorem does not hold.



  - The dashed lines (arrows) indicate possible transition between states.
  - The dashed loops indicate self-transitions.
  - Each arrow should carry a weight equal to the probability of moving between the corresponding states.

4 Time-series Analysis

4.1 Definition of a time series

A **time series** is a set of observations y_n, $n = 0, 1, ...$, arranged in increasing time.

- Typically observations are recorded at regular real-time intervals, e.g. $t_0 + n\Delta$ where t_0 is start time for recording observations and Δ is the time-increment between observations.
- The nth observation is recorded at real-time $t_0 + n\Delta$

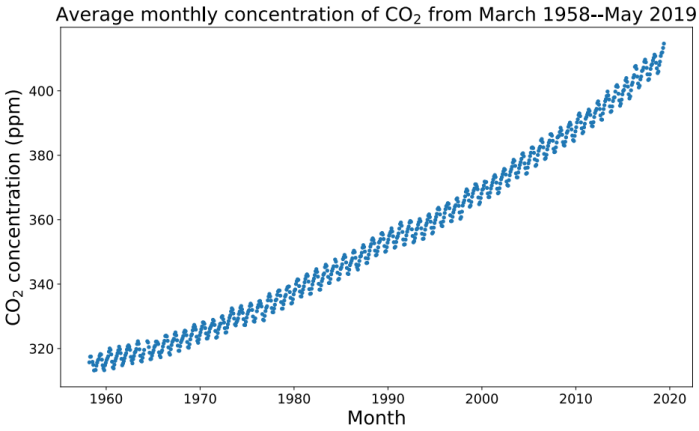


Figure 1: A time-series $\{y_0, y_1, \dots\}$, the CO_2 data. Time interval between observations $\Delta = 1$ month.

4.2 Workflow for time-series analysis

When faced with a novel time-series data $y_0, y_1, \dots$

- **Step 1**: Set up a probability model to represent the data.

$$Y_n = \text{trend} + \text{seasonal component} + \text{residual}$$

where

Trend m_n: evolution of mean over time

Seasonal S_n: sinusoid(s) with periodicity related to the calendar

Residual X_n: zero-mean random variable (not just independent over time)

- **Step 2**: Estimate model parameters and check goodness of fit
- **Step 3**: Deploy: Simulate, forecasting,...

4.3 Estimating, or fitting, the trend

- Estimating the trend
  - Assume $Y_n = m_n + S_n + X_n$ where the trend is assumed to be polynomial

$$m_n = a + bn + cn^2$$

  - Using the data $y_0, \dots, y_n$ estimate (a, b, c) via least squares:

$$\left(\frac{\partial}{\partial a}, \frac{\partial}{\partial b}, \frac{\partial}{\partial c}\right) \sum_{n=0}^{N} (Y_n - a - bn - cn^2)^2 = (0, 0, 0)$$

  - De-trended data is

$$Y_n - \hat{m}_n = Y_n - \hat{a} - \hat{b}_n - \hat{c}n^2$$

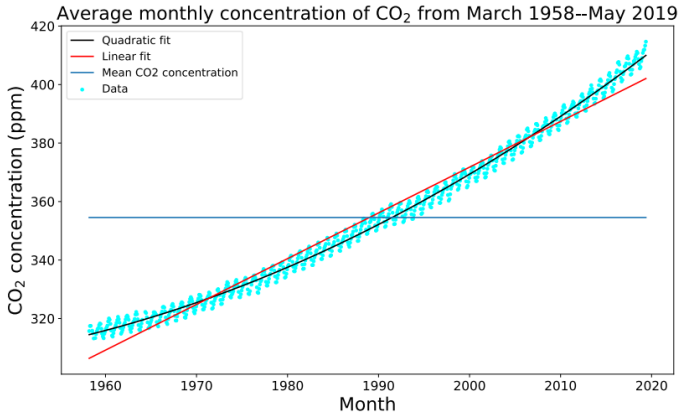


Figure 2: Different fits for the data

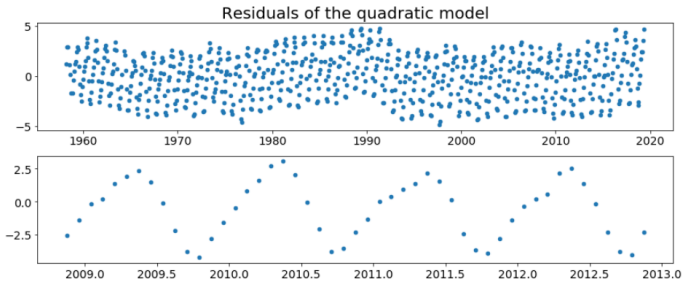


Figure 3: De-trended data $Y_n - \hat{m}_n$

- Biased
  - Getting the trend model wrong will result in a biased estimate. Assume $Y_n = a + bn + X_n$ where $\mathbb{E}(X_n) = 0$ but we fit a constant trend $Y_n = m + X_n$. Then

$$\hat{m} = \frac{1}{N+1}\sum_{n=0}^{N} Y_n = a + \frac{N}{2}b + \frac{1}{N+1}\sum_{n=0}^{N} X_n$$

    which is biased since $\mathbb{E}(\hat{m}) = \text{a} + \text{b N}/2$
  - Example: Assume $Y_n = a + b_n + X_n$ where $\mathbb{E}(X_n) = 0$. Estimate the trend with the moving average where $\hat{m}_n = (Y_{n-1} + Y_n + Y_{n+1})/3$ and find $\mathbb{E}(\hat{m}_n)$.

$$\begin{aligned}\hat{m}_n &= a + b(n-1+n+n+1)/3 + (X_{n-1} + X_n + X_{n+1})/3 \\ &= a + bn + (X_{n-1} + X_n + X_{n+1})/3\end{aligned}$$

    which is unbiased. But verify this estimate is biased if the true data was $Y_n = a + bn + cn^2 + X_n$.

- Fitting the seasonal term:
  - Assume $S_n = A\cos(2\pi f n + \phi)$ where A and ϕ are independent random variable, $0 \le \phi < 2\pi$ is uniformly distributed. Note

$$S_n = \alpha_1 \cos(2\pi f n) + \alpha_2 \sin(2\pi f n)$$

    where $\alpha_1 = A\cos(\phi)$ and $\alpha_2 = -A\sin(\phi)$

– Solve with least squares similarly

– Then, left with residuals:

$$Y_n - \hat{m}_n - \hat{S}_n = Y_n - \hat{m}_n - \hat{\alpha}_1 \cos(2\pi f n) - \hat{\alpha}_2 \sin(2\pi f n)$$

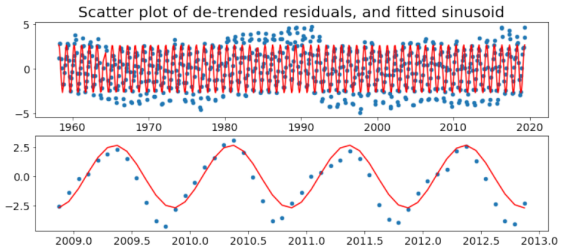


Figure 4: Fitted sinusoid

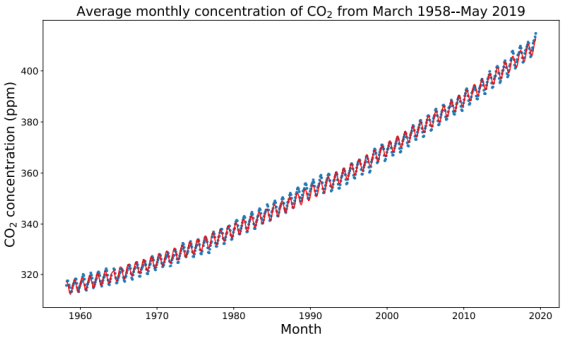


Figure 5: Combined Fit

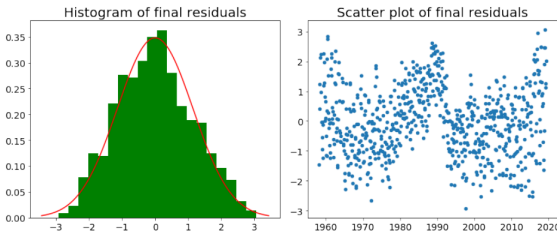
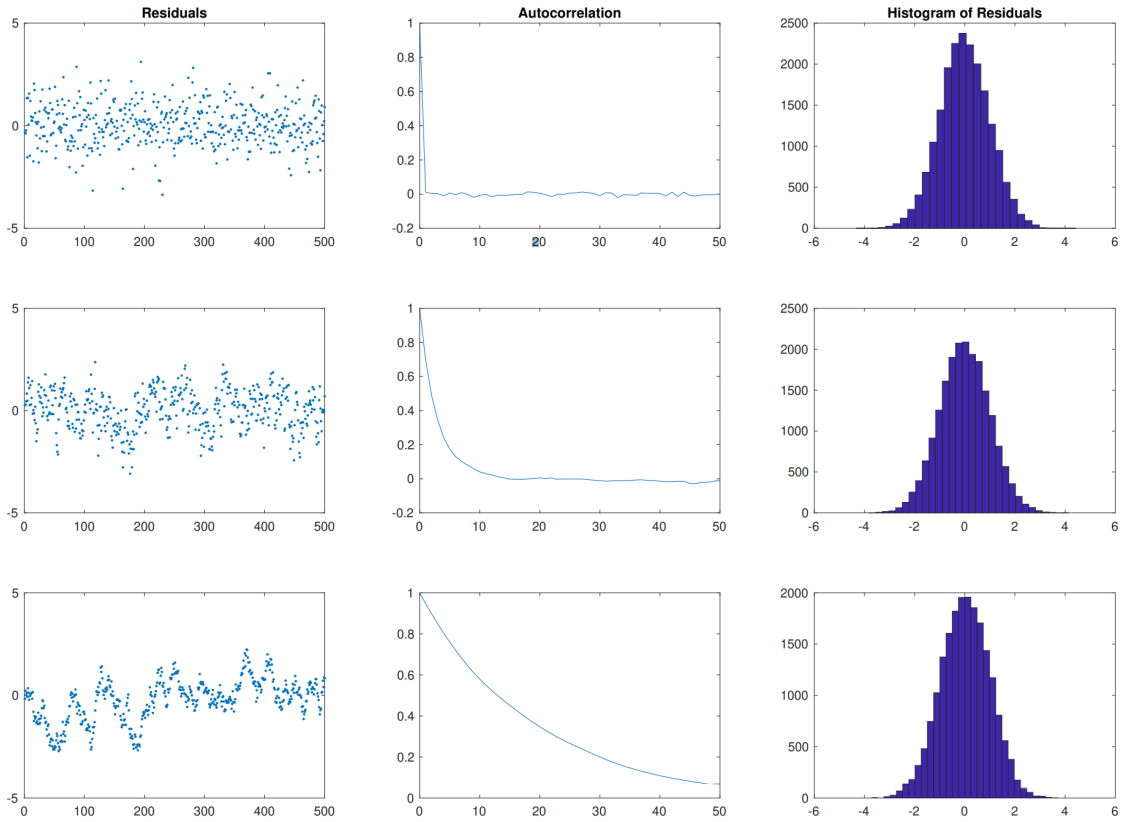


Figure 6: Final Residual

- Fitting the residual
  - Plotting residuals $Y_n - \hat{m}_n - \hat{S}_n$, its autocorrelation, the histograms of residuals.
  - Each row is a scenario you could face:
    - Top row: independent zero mean residuals
    - Middle row: correlated zero mean residuals
    - Bottom row: strongly correlated zero mean residuals
  - Note that all residuals have the same histograms ($\mathcal{N}(0,1)$)



5 AR and MA process

5.1 Auto-regressive (AR) process

5.1.1 Definition

- Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of random variables such that $\mathbb{E}(W_n) = 0$ for all n,

$$\mathbb{E}(W_i W_j) = \begin{cases} \sigma^2 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

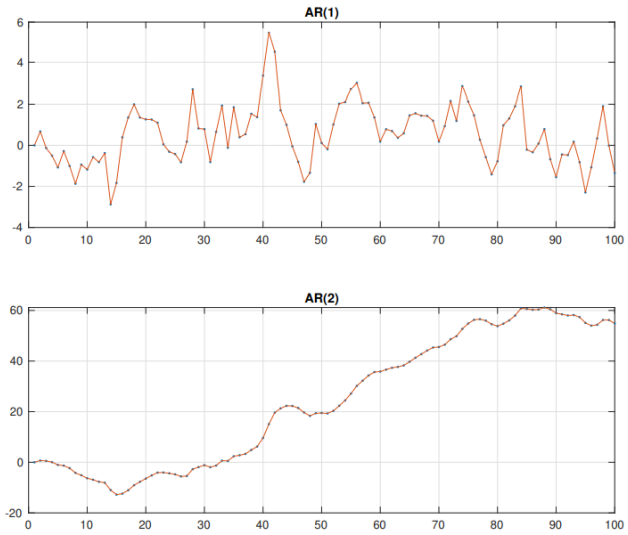
- The AR(p) process $\{X_n\}_{n=-\infty}^{\infty}$ is

$$X_n = \left(\sum_{i=1}^{p} a_i X_{n-1} \right) + W_n$$

  where $a_1, ..., a_p$ are constants and p is the order of the process.

- A real time-series (data set) X_n is modelled as a function of its previous values and random part:

$$X_n = \underbrace{a_1 X_{n-1} + ... + a_p X_{n-p}}_{\text{predictable part}} + \underbrace{W_n}_{\text{random part}}$$



- Comparing AR(1) and AR(2), AR(2) is more "sluggish" or is more influenced by its past values.

5.1.2 AR(1) process

- For the AR(1), we write
$$X_n = aX_{n-1} + W_n$$
i.e. Only take into account one term before

- Expand X_n in terms of W_k, $k \leq n$ and compute the mean and variance:

$$\begin{aligned} X_n &= aX_{n-1} + W_n \\ &= a^2 X_{n-2} + aW_{n-1} + W_n \\ &\vdots \\ &= \sum_{k=0}^{\infty} W_{n-k} a^k \\ X_n &= \sum_{k=0}^{\infty} W_{n-k} h_k \end{aligned}$$

- AR(1) is causal with impulse response $\{h_k\}_{k\geq 0}$.
- No future W_j, $j > n$, terms in X_n.
- Using linearity of $\mathbb{E}\{\cdot\}$ to get mean,
$$\mathbb{E}\{X_n\} E\left\{\sum_{k=0}^{\infty} W_{n-k} a^k\right\} = \sum_{k=0}^{\infty} \{W_{n-k} a^k\} = 0$$
since each W_n has mean zero.

- For the variance

$$\begin{aligned}\mathbb{E}\{X_n^2\} &= \mathbb{E}\left\{\left(\sum_{k=0}^{\infty} W_{n-k}a^k\right)^2\right\} \\ &= \mathbb{E}\left\{\left(\sum_{k=0}^{\infty} W_{n-k}^2 a^{2k}\right) + \text{cross terms}\right\} \\ &=_{(a)} \sum_{k=0}^{\infty} \mathbb{E}\{W_{n-k}^2\}a^{2k} \\ &= \sum_{k=0}^{\infty} \sigma^2 a^{2k} =_{(b)} \frac{\sigma^2}{1-a^2}\end{aligned}$$

  (a) Since $\mathbb{E}(W_i W_j) = 0$ for $i \neq j$ so cross terms have zero mean.

  (b) Geometric sum converges provided $|a| < 1$.

5.2 Wide sense stationary (WSS)

5.2.1 Definition

$\{X_n\}_{n=-\infty}^{\infty}$ is wide-sense stationary if

- $\mathbb{E}\{X_n\} = \mu$ for all n i.e.,X_n has constant mean;
- has finite variance, i.e., $\mathbb{E}\{X_n^2\} < \infty$ for all n;
- $\mathbb{E}\{X_{n1}X_{n2}\} = \mathbb{E}\{X_{n1+k}X_{n2+k}\}$ for any n_1, n_2, k.

5.2.2 Correlation function

$$R_X(k) = \mathbb{E}\{X_0 X_k\}, \quad k \in \mathbb{Z}$$

- $R_X(k)$ is the **correlation function** of a WSS process.
- For any $n_2 > n_1$,
$$\mathbb{E}\{X_{n_1}X_{n_2}\} = \mathbb{E}\{X_{n_1-n_1}X_{n_2-n_1}\} = R_X(n_2 - n_1)$$
- We see that $\mathbb{E}\{X_{n_1}X_{n_2}\}$ only depends on $|n_2 - n_1|$.
- $R_X(k)$ is an even function.
- Strict stationarity is too strong a requirement to be relevant for many application. Recall it implies any two sections of the process,

  $$(X_0, ..., X_k) and (X_m, ..., X_{k+m})$$

  are statistically indistinguishable for any section size k and displacement m. This is an inappropriate model for many real-world processes.
- Wide sense stationarity is a weaker modelling restriction and useful for the analysis of time-series.

5.2.3 Example: Random phase cosine

Let $X_n = \cos(2\pi f n + \phi)$,we showed earlier $\mathbb{E}\{X_n\} = 0$. Show $\mathbb{E}\{X_{n_1}X_{n_2}\}$ only depends on the time difference.

$$\begin{aligned}\mathbb{E}\{X_{n_1}X_{n_2}\} &= \int_0^{2\pi} \cos(2\pi f n_1 + \phi)\cos(2\pi f n_2 + \phi)\frac{1}{2\pi}\,d\phi \\ &= \int_0^{2\pi} \frac{1}{2}\cos(2\phi + 2\pi f(n_1+n_2))\frac{1}{2\pi}\,d\phi + \int_0^{2\pi}\frac{1}{2}\cos(2\pi f(n_1-n_2))\frac{1}{2\pi}\,d\phi \\ &= \frac{1}{2}\cos(2\pi f \underbrace{(n_2-n_1)}_{\text{only difference}})\end{aligned}$$

- For random phase example, $\mathbb{E}\{X_{n_1}X_{n_2}\}$ only depends on the time difference.
- Note that $R_X(k) = \mathbb{E}\{X_0X_k\}$ is periodic and does not decay. (The period is the same as the data.)
- So X_n and X_{n+k} remain strongly correlated even as k increases.

5.2.4 Example: AR(1) process.

We showed earlier that $\mathbb{E}\{X_n\} = 0$ and $\mathbb{E}\{X_n^2\} = \frac{\sigma^2}{(1-a^2)}$ for all n.

Calculate $\mathbb{E}\{X_{n-k}X_n\}$ for $k \geq 1$ for AR(1): $X_n = aX_{n-1} + W_n, |a| < 1$.

- $$\begin{aligned}\mathbb{E}\{X_{n-1}X_n\} &= \mathbb{E}\{X_{n-1}(aX_{n-1}+W_n)\} \\ &= a\mathbb{E}\{X_{n-1}^2\} + \underbrace{\mathbb{E}\{X_{n-1}W_n\}}_{=0} \\ &= a\sigma_X^2\end{aligned}$$
  - Note $\mathbb{E}\{X_{n-1}W_n\} = 0$ as no W_n term in X_{n-1}.
- $$\mathbb{E}\{X_{n-2}X_n\} = a\mathbb{E}\{X_{n-2}X_{n-1}\} + \underbrace{\mathbb{E}\{X_{n-2}W_n\}}_{=0} = a^2\sigma_X^2$$
- $$\mathbb{E}\{X_{n-3}X_n\} = a\mathbb{E}\{X_{n-3}X_{n-1}\} + \underbrace{\mathbb{E}\{X_{n-3}W_n\}}_{=0} = a^3\sigma_X^2$$
- Therefore, $\mathbb{E}\{X_{n-k}X_n\}$ does not depend on n. Thus the AR(1) process is WSS and
$$R_X(k) = a^k\sigma_X^2$$

5.2.5 More on AR process

- The AR(p) process is a useful model for time-series data:
$$X_n = \underbrace{a_1X_{n-1} + ... + a_pX_{n-p}}_{\text{predictable part}} + \underbrace{W_n}_{\text{random part}}$$
- The AR(p) process is wide sense stationary.
- It is easy to estimate the best fitting AR(p) model for a real time-series data set using its WSS property.

5.3 Moving average MA(q) process.

5.3.1 Definition

- Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of random variables such that $\mathbb{E}(W_n) = 0$ for all n,

$$\mathbb{E}(W_i W_j) = \begin{cases} \sigma^2 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

- The MA(q) process $\{X_n\}_{n=-\infty}^{\infty}$ is

$$X_n = \left(\sum_{i=1}^{q} b_i W_{n-i} + W_n \right)$$

  where $b_1, ..., b_q$ are constants and q is the *order* of the process

5.3.2 Correlation of the MA(q) process

- Since $\{W_n\}_{n=-\infty}^{\infty}$ satisfies $\mathbb{E}\{W_n W_k\} = 0$ when $n \neq k$

$$\begin{aligned} \mathbb{E}\{X_n^2\} &= \mathbb{E}\left\{ \left(W_n + \sum_{i=1}^{q} b_i W_{n-i} \right)^2 \right\} \\ &= \sum_{i=1}^{q} b_i^2 \mathbb{E}\{W_{n-1}^2\} + \mathbb{E}\{W_n^2\} + \mathbb{E}\{\text{cross term}\} \\ &= \sigma^2(1 + b_1^2 + ... + b_q^2). \end{aligned}$$

- We may write

$$X_n = \sum_{i=0}^{\infty} h_i W_{n-i}$$

  $h_i = b_i$ for $0 < i \leq q$, $h_0 = 1$ and $h_i = 0$ for all other i

$$h_i = \begin{cases} b_i & \text{if } 0 < i \leq q \\ 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases}$$

- Thus to compute $\mathbb{E}\{X_{n_1} X_{n_2}\}$, we need to compute the correlation of the output of an LTI system with impulse response $\{h_k\}_{k=-\infty}^{\infty}$

5.3.3 Wide sense stationarity

- **Fact:** If the input $\{W_n\}_{n=-\infty}^{\infty}$ of a discrete time LTI system with impulse response $\{h_n\}_{n=-\infty}^{\infty}$ is WSS then its output $\{y_n\}_{n=-\infty}^{\infty}$ is also WSS.

- **Verification:**

  - Output is found by convolution: $Y_n = \sum_{k=-\infty}^{\infty} h_{n-k} W_k$.

– The mean of Y_n is then:

$$\begin{aligned}\mathbb{E}\{Y_n\} &= \sum_{k=-\infty}^{\infty} h_{n-k}\mathbb{E}\{W_k\} && \text{(linearity of expectation)}\\ &= \mathbb{E}\{W_0\}\sum_{k=-\infty}^{\infty} h_{n-k} && \text{(constant mean property)}\end{aligned}$$

– and thus $\mathbb{E}\{Y_n\}$ is time independent.

– The correlation is

$$\begin{aligned}\mathbb{E}\{Y_{n_1}Y_{n_2}\} &= \mathbb{E}\left\{\left(\sum_{k=-\infty}^{\infty} h_k W_{n_1-k}\right)\left(\sum_{l=-\infty}^{\infty} h_l W_{n_1-l}\right)\right\}\\ &= \mathbb{E}\left\{\sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} h_k h_l W_{n_1-k}W_{n_1-l}\right\}\\ &= \sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} h_k h_l \mathbb{E}\{W_{n_1-k}W_{n_1-l}\}\\ &= \sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} h_k h_l R_W(n_2-n_1+k-l)\end{aligned}$$

– Which is indeed only a function of $n_2 - n_1$ Thus $\{Y_n\}_{n\in\mathbb{Z}}$ is also WSS.

– Back to the MA example, $\{W_n\}_{n\in\mathbb{Z}}$ is WSS since $\mathbb{E}\{W_n\} = 0, \mathbb{E}\{W_nW_k\} = 0$ for $n\perp = k$. Thus the MA process is WSS.

6 Power Spectrum

Overview:

- The *Fourier transform* is an important tool for analysing deterministic signals and is equally important for random process as well.
- A *frequency domain* representation is obtained by computing the Fourier transform of the correlation function to yield the *Power spectrum*.
- As we will see, it is called the power spectrum because it gives the power of the random process at each frequency of the spectrum.

6.1 Definition

Let $R_X(k)$ be the correlation function of a **discrete time** WSS process.

- The power spectrum density $S_X(f)$ is

$$S_X(f) = \sum_{k=-\infty}^{\infty} R_X(k)e^{-j2\pi fk}$$

- The inversion formula:

$$R_X(n) = \int_{-1/2}^{1/2} S_X(f) e^{j2\pi fn}\, df.$$

- Note that $S_X(f) = S_X(f+n)$, n is integer. S_X has period 1.
- $S_X(f)$ is an even function

6.2 Using the inversion as an interpretation

- We will show (by symmetry)

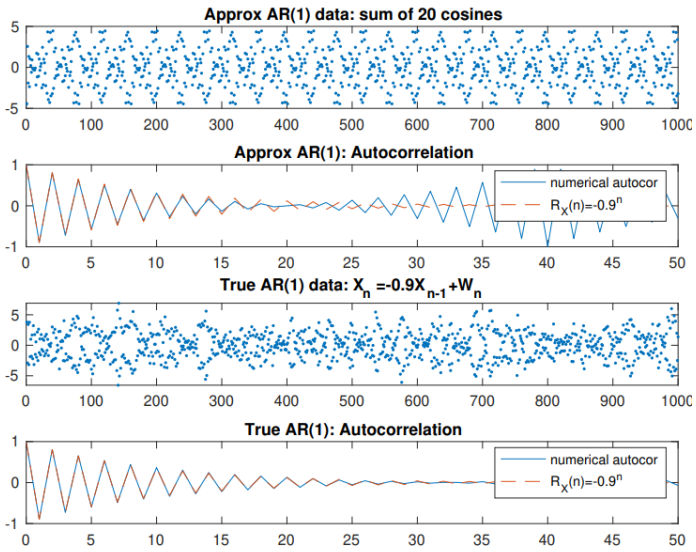
$$\begin{aligned} R_X(n) &= 2\int_{-1/2}^{1/2} S_X(f)\cos(j2\pi fn)\, df \\ &\approx \sum_{i=1}^{k} \frac{1}{K} S_X(f_i)\cos(j2\pi f_i n) \end{aligned}$$

  By dividing $[0, \frac{1}{2}]$ into K equal size segments, f_i is the mid-point of segment i.

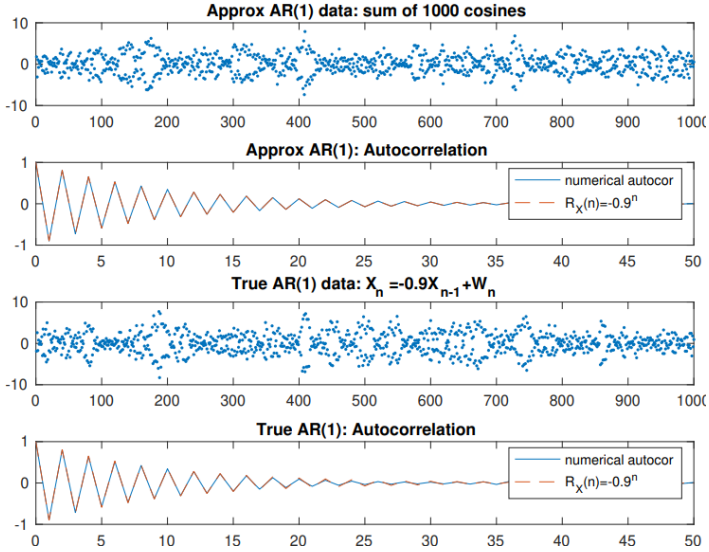
- We can use a sum of cosines with random phases to find a WSS random process with the same $R_X(n)$:

$$X_n = \sum_{i=1}^{k} \sqrt{2S_X(f_i)/K}\cos(2\pi f_i n + \phi_i)$$

  where $\phi_i \sim \mathcal{U}[0, 2\pi]$, each ϕ_i independent.



(a) $K = 20$ Cosine approximation



(b) $K = 1000$ Cosine approximation

6.3 Example: PSD of AR(1) process.

- The AR(1) process:

$$X_n = aX_{n-1} + w_n, \quad |a| < 1$$

$$\mathbb{E}\{X_n\} = 0, \quad \mathbb{E}\{X_n^2\} = \sigma_X^2 = \frac{\sigma^2}{1-a^2}, \quad R_X(k) = a^{|k|}\sigma_X^2$$

- PSD :

$$\begin{aligned} S_X(f) &= \sigma_X^2 \sum_{k=-\infty}^{\infty} a^{|k|} e^{-j2\pi fk} \\ &= -\sigma_X^2 + \sigma_X^2 \sum_{k=-\infty}^{0} a^{-k} e^{-j2\pi fk} + \sigma_X^2 \sum_{k=0}^{\infty} a^k e^{-j2\pi fk} \\ &= \sigma_X^2 \left(-1 + \frac{1}{1 - ae^{j2\pi f}} + \frac{1}{1 - ae^{-j2\pi f}} \right) \\ &= \frac{\sigma^2}{(1 - ae^{j2\pi f})(1 - ae^{-j2\pi f})} \\ &= \frac{\sigma^2}{1 + a^2 - 2a\cos(2\pi f)} \end{aligned}$$

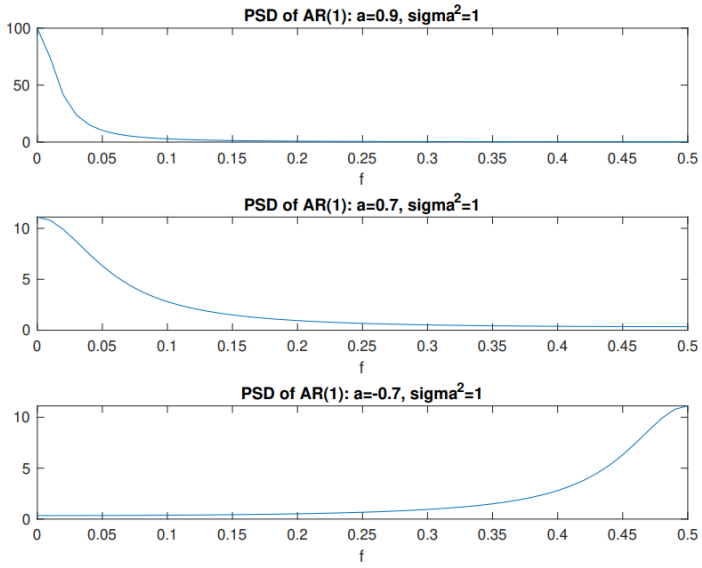


Figure 7: Power spectrum of the AR(1) process

.

6.4 Property of WSS PSD

- **Fact**: If $R_X(k)$ is the correlation function of a discrete time WSS process then the power spectrum density $S_X(f)$ is an even, real valued and nonnegative function of f. Moreover, $S_X(f)$ is a continuous function if $\sum_{k=-\infty}^{\infty} |R_X(k)| < \infty$.

- *Verification*: Use the definition of the PSD

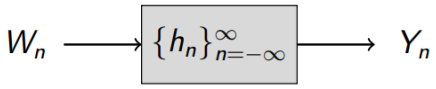
$$\begin{aligned} S_X(f) &= \sum_{k=-\infty}^{\infty} R_X(k) e^{-j2\pi fk} \\ &= \sum_{k=-\infty}^{\infty} R_X(k)\cos(2\pi fk) - j \sum_{k=-\infty}^{\infty} R_X(k)\sin(2\pi fk) \\ &= \sum_{k=-\infty}^{\infty} R_X(k)\cos(2\pi kf) \end{aligned}$$

- NO sine terms as $R_X(k)$ is even and $\sin(2\pi fk)$ is odd, and summing from negative to positive cancel out this whole odd function. Thus $S_X(f)$ is even.

$$S_X(f) = S_X(-f) \quad \text{and} \quad S_X(f) = S_X^*(f).$$

- Continuity follows as the weighted sum of continuous functions (note $\cos(2\pi ft)$ is continuous) is continuous.

6.5 PSD of an LTI system



- If the input $\{W_n\}_{n=-\infty}^{\infty}$ if a LTI system with impulse response $\{h_n\}_{n=-\infty}^{\infty}$ is WSS then its output $\{Y_n\}_{n=-\infty}^{\infty}$ is also WSS.
- Relating PSD of input and output:

$$S_Y(f) = S_W(f)|H(f)|^2.$$

- *Verification*:
  - $$R_Y(n) = \mathbb{E}\{Y_0 Y_n\} = \sum_{k=-\infty}^{\infty} h_k \left(\sum_{l=-\infty}^{\infty} h_l R_W(n+k-l) \right)$$
  - Inner sum is the convolution of R_W and the impulse response:

$$\sum_{l=-\infty}^{\infty} h_l R_W(n+k-l) = g(n+k)$$

  - Thus:

$$R_Y(n) = \sum_{k=-\infty}^{\infty} h_k g(n+k) = \sum_{k=-\infty}^{\infty} h_{-k} g(n-k)$$

  - The Fourier transform is

$$S_Y(f) = H^*(f)G(f)$$

    where

$$S_Y(f) = \sum_{n=-\infty}^{\infty} R_Y(n)e^{-j2\pi fn}, \quad H(f) = \sum_{n=-\infty}^{\infty} h(n)e^{-j2\pi fn}, \quad G(f) = \sum_{n=-\infty}^{\infty} g(n)e^{-j2\pi fn}$$

  - Thus

$$S_Y(f) = H^*(f)G(f) = H^*(f)S_W(f)H(f) = S_W(f)|H(f)|^2$$

- Example:For the AR model $X_n = aX_{n-1} + W_n$, compute the PSD using the Fourier method.
  - AR(1) model can be written as :

$$X_n = \sum_{k=0}^{\infty} W_{n-k} a^k = \sum_{k=0}^{\infty} W_{n-k} h_k.$$

– $S_X(f) = S_W(f)|H(f)|^2$ where

$$S_W(f) = \sum_{k=-\infty}^{\infty} R_W(k) e^{-j2\pi fk} = \sigma^2$$

Since $R_W(k) = 0$ for $k \neq 0$ and $R_W(0) = \mathbb{E}\{W_n^2\} = \sigma^2$

– Then

$$\begin{aligned} H_X(f) &= \sum_{k=-\infty}^{\infty} h_k e^{-j2\pi fk} \\ &= \sum_{k=0}^{\infty} a^k e^{-j2\pi fk} \\ &= \left(\frac{1}{1 - ae^{-j2\pi f}} \right) \\ |H_X(f)|^2 &= \frac{1}{(1 - ae^{-j2\pi f})(1 - ae^{j2\pi f})} \end{aligned}$$

– Finally, using $S_X(f) = S_W(f)|H(f)|^2 = \sigma^2|H(f)|^2$ we get the same answer as before.

6.6 The ARMA model

6.6.1 Definition

- Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of random variables with common mean 0 common variance σ^2 and $\mathbb{E}\{W_n W_k\} = 0$ for $n \neq k$.
- The ARMA(p, q) process $\{X_n\}_{n=-\infty}^{\infty}$ is the discrete time process satisfying

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + W_n + \sum_{i=1}^{q} b_i W_{n-i}$$

where a_i and b_j are constants.

- It can be expressed as a causal filter applied to $\{W_n\}_{n=-\infty}^{\infty}$,

$$X_n = \sum_{k=0}^{\infty} h_k W_{n-k}.$$

- **Fact**: The ARMA(p, q) is WSS.
- Verification of WSS:
  - The ARMA(p, q)model can be interpreted as an LTI system with input $\{W_n\}_{n=-\infty}^{\infty}$.
  - Use the previous result that showed an LTI system preserves the WSS property.
  - Since $\mathbb{E}\{W_n\} = 0$ and $\mathbb{E}\{W_n W_m\} = 0$ for $m \neq n$ which implies WSS of $\{W_n\}_{n=-\infty}^{\infty}$, therefore ARMA$(p, q)$is also WSS.

7 Random Process revisit

7.1 Discrete-time Random Process

7.1.1 Definition

- A discrete random process is defined as an ensemble of functions

$$\{X_n(w)\}, \quad n = -\infty, ..., -1, 0, 1, ..., \infty$$

- w is a random variable having a probability density function $f(w)$.
- Think of a generative model for the waveforms you might observe in practice:

  1. First draw a random value $\tilde{w}$ from the density $f(w)$
  2. The observed waveform for this value $w = \tilde{w}$ is given by

$$X_n(\tilde{w}), \quad n = -\infty, ..., -1, 0, 1, ..., \infty$$

  3. the 'ensemble' is built up by considering all possible values $\tilde{w}$ ('the sample space') and their corresponding time waveforms $X_n(\tilde{w})$
  4. $f(w)$ determines the relative frequency (or probability) with which each waveform $X_n(w)$ can occur.
  5. Where no ambiguity can occur, w is left out for notational simplicity, i.e. we refer to 'random process $\{X_n\}$

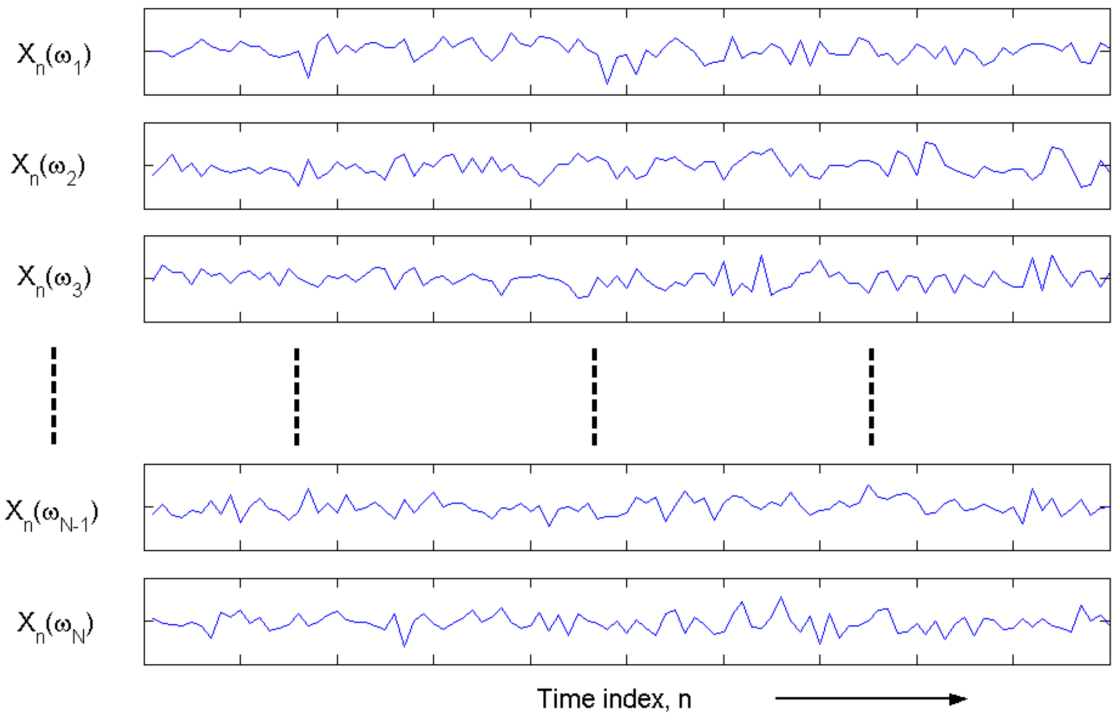


Figure 8: Ensemble representation of a discrete-time random process

7.1.2 Example: the harmonic process

The harmonic process is important in a number of applications including radar, sonar, speech and audio modelling. An example of a real-valued harmonic process is the random phase sinusoid.

- Sine-waves of known amplitude a and frequency w_0.
- Phase, unknown and random
- Random phase correspond to an unknown delay in a system for example.
- This random process could be expressed in:

$$X_n = a\sin(nw_0 + \Phi) \qquad \text{with fixed constant } a \text{ and } w_0 \text{ and random variable } \Phi$$

- Random variable Φ has a uniform probability distribution over the range $-\pi$ to $+\pi$:

$$f(\phi) = \begin{cases} 1/(2\pi) & -\pi < \phi < +\pi \\ 0, & \text{otherwise} \end{cases}$$

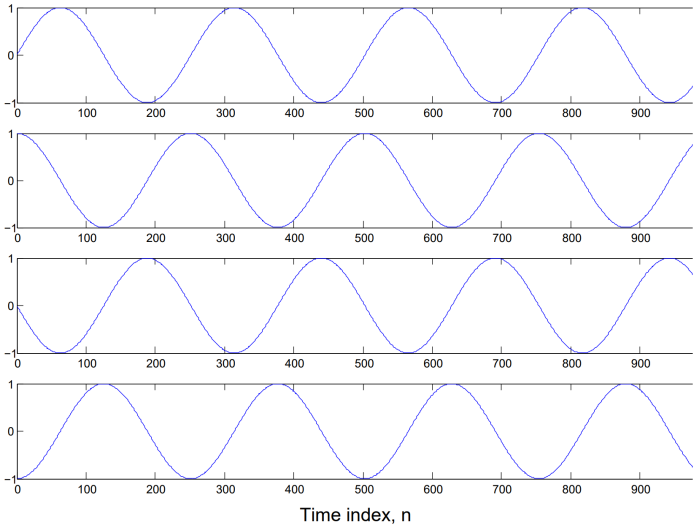


Figure 9: Some members of the random phase sine ensemble

7.2 Correlation functions

7.2.1 Autocorrelation function

- The mean of a random process X_n is - $\mathbb{E}\{X_n\}$
- The **Autocorrelation function** of process $\{X_n\}$ is

$$r_{xx}[n, m] = \mathbb{E}[X_n X_m]$$

7.2.2 Cross-Correlation function

- The **Cross-correlation** function between two processes $\{X_n\}$ and $\{Y_n\}$ is:

$$r_{xy}[n, m] = \mathbb{E}[X_n Y_m]$$

7.3 Stationarity

7.3.1 Strict-Sense Stationary

- A stationary process has the *same statistical characteristics* irrespective of shifts along the time axis.
  - An observer looking at the process from sampling time n_1 would not be able to tell the difference in the *statistical* characteristics of the process if the time moved to n_2
  - Could be formalised by considering the Nth order density function for the process:

    $$f_{X_{n_1},X_{n_2},...,X_{n_N}}(x_{n_1}, x_{n_2}, ..., x_{n_N})$$

    i.e., the joint probability density function for N arbitrarily chosen time indices $\{n_1, n_2, ..., n_N\}$.
  - Since the probability distribution of a random vector contains all the statistical information about that random vector, we should expect the probability distribution to be unchanged if we shifted the time axis any amount to the left or the right, for a stationary signal
- A random process is strict-sense stationary if, for any finite c, N and $\{n_1, n_2, ..., n_N\}$:

  $$f_{X_{n_1},X_{n_2},...,X_{n_N}}(\alpha_1, ..., \alpha_N) = f_{X_{n_1+c},X_{n_2+c},...,X_{n_N+c}}(\alpha_1, ..., \alpha_N)$$

- Strict-sense stationarity is hard to prove for most systems. In this course, a less stringent condition is used, wide-sense stationarity, which only requires **first and second moments**(i.e. mean and autocorrelation function) to be invariant to time shifts.

7.3.2 Wide-sense Stationary

A random process is *wide-sense stationary(WSS)* if:

1. $\mu_n = \mathbb{E}[X_n] = \mu$ (Constant mean)
2. $r_{xx}[n, m] = r_{xx}[n + k, m + k]$ (Autocorrelation only depends on the difference)
3. Finite variance: $\mathbb{E}[(X_n - \mu)^2] < \infty$

Note that strict-sense stationarity plus finite variance implies wide-sense stationarity, but not vice versa.

7.3.3 Example: the harmonic process continued

1. Mean:

   $$\begin{aligned}\mathbb{E}[X_n] =& \mathbb{E}[a\sin(nw_0 + \Phi)] \\ =& a\{\mathbb{E}[\sin(nw_0)\cos(\Phi) + \cos(nw_0)\sin(\Phi)]\} \\ =& a\{\sin(nw_0)\mathbb{E}[cos(\Phi)] + \cos(nw_0)\mathbb{E}[\sin(\Phi)]\} \\ =& 0\end{aligned}$$

   since $\mathbb{E}[\cos(\Phi)] = \mathbb{E}[\sin(\Phi)] = 0$ under the assumed uniform pdf $f(\phi)$.

2. Autocorrelation:

$$\begin{aligned} r_{XX}[n,m] =& E[X_n X_m] \\ =& \mathbb{E}[a\sin(nw_0+\Phi)\cdot a\sin(mw_0+\Phi)] \\ =& 0.5a^2\{\mathbb{E}[\cos[(n-m)w_0]-\cos[(n+m)w_0+2\Phi]]\} \\ =& 0.5a^2\{\cos[(n-m)w_0]-\mathbb{E}[\cos[n+m]w_0+2\Phi]\} \\ =& 0.5a^2\cos[(n-m)w_0] \end{aligned}$$

3. To verify finite variance, just set $m = n$ in the autocorrelation function. Therefore, the harmonic process satisfy all three conditions and is WSS.

7.4 Power Spectra

7.4.1 Definition

- For a wide-sense stationary random process $\{X_n\}$, the power spectrum is defined as the discrete-time Fourier transform (DTFT) of the autocorrelation function:

$$S_X(e^{j\Omega}) = \sum_{m=-\infty}^{\infty} r_{xx}[m]e^{-jm\Omega}$$

  where $\Omega = \omega T$, the normalised frequency, in radians per sample is used for convenience.

- T is the sampling interval of the discrete time process and therefore $\Omega = 2\pi$ corresponds to the sampling frequency $\omega = 2\pi/T$ rad/s

- The definition here is slightly different compared to the before. When we write the PSD in a function of $e^{j\Omega}$. The period becomes 2π

7.4.2 Inverse formula for PSD

- The autocorrelation function can thus be found from the power spectrum by inverting the transform using the inverse DTFT:

$$r_{xx}[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(e^{j\Omega})e^{jm\Omega}\,d\Omega$$

7.4.3 Properties

- The *power spectrum* is a **real, positive, even** and **periodic** function of frequency.

- The power spectrum can be interpreted as a density spectrum in the sense that the mean-squared signal value at the output of an ideal band-pass filter with lower and upper cut-off frequencies ω_l and ω_u is given by

$$\frac{1}{\pi}\int_{\omega_l T}^{\omega_u T} S_X(e^{j\Omega})\,d\Omega$$

7.4.4 Example: random phase sine-wave

- The autocorrelation function is obtained as:

$$r_{xx}[m] = 0.5a^2 \cos[m\omega_o]$$

- Hence the power spectrum is obtained as:

$$\begin{aligned} S_X(e^{j\Omega}) &= \sum_{m=-\infty}^{\infty} r_{xx}[m]e^{-jm\Omega} \\ &= \sum_{m=-\infty}^{\infty} 0.5a^2 \cos[m\omega_0]e^{-jm\Omega} \\ &= 0.25a^2 \times \sum_{m=-\infty}^{\infty} (e^{jm\omega_0} + e^{-jm\omega_0})e^{-jm\Omega} \\ &= 0.5\pi a^2 \times \sum_{m=-\infty}^{\infty} \delta(\Omega - \omega_0 - 2m\pi) + \delta(\Omega + \omega_0 - 2m\pi) \end{aligned}$$

  where Normalised frequency $\Omega = \omega T$ is used for shorthand

- The last line above is obtained from the Fourier series of a periodic train of δ functions

$$\sum_{m=-\infty}^{\infty} \delta(t + t_0 - 2m\pi) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \exp(-jmt_0) \exp(+jmt)$$

- Alternatively (and equivalently) just take the inverse DTFT of the delta function to check the result:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \delta(\Omega + \omega_0 - 2m\pi)e^{jn\Omega}\, d\Omega = \frac{1}{2\pi} e^{jn\omega_0}$$

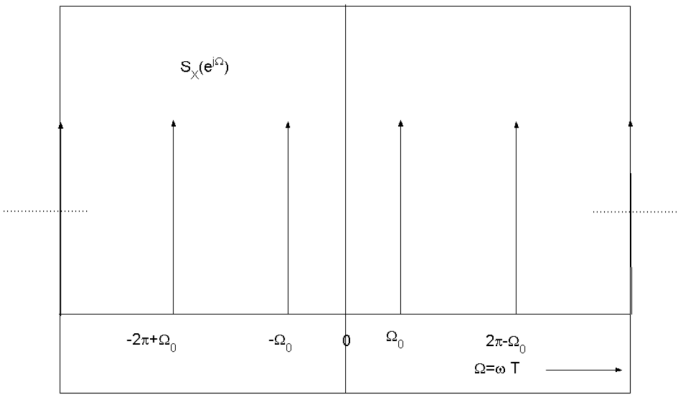


Figure 10: Power spectrum of harmonic process - $\Omega_0 = \omega_0$ in this plot

7.5 White Noise

7.5.1 Definition

- White noise is defined in terms of its auto-covariance function. A wide sense stationary process is termed white noise if:

$$r_{xx}[m] = \mathbb{E}[(X_n - \mu)(X_{n+m} - \mu)] = \sigma_X^2 \delta[m]$$

where $\delta[m]$ is the discrete impulse function:

$$\delta[m] = \begin{cases} 1, & m = 0 \\ 0, & \text{otherwise} \end{cases}$$

$\sigma_X^2 = \mathbb{E}[(X_n - \mu)^2]$ is the variance of the process.

- If $\mu = 0$ then σ_X^2 is the *mean-squared* value of the process, which will refer to as the 'power'.
- The power spectrum of zero mean white noise is:

$$S_X(e^{j\omega T}) = \sum_{m=-\infty}^{\infty} r_{xx}[m]e^{-jm\Omega} = \sigma_X^2$$

i.e., flat across all frequencies.

7.5.2 Example: White Gaussian noise (WGN)

- The values X_n are drawn *independently* from a Gaussian distribution with mean 0 and variance σ_X^2.
- The N$^{\text{th}}$ order pdf for the Gaussian white noise process is:

$$f_{Xn_1,Xn_2,...,Xn_N}(\alpha_1, \alpha_2, ..., \alpha_N) = \prod_{i=1}^{N} \mathcal{N}(\alpha_i|0, \sigma_X^2)$$

where

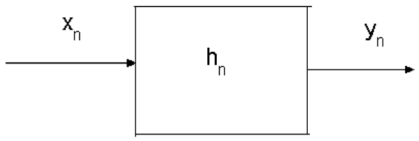
$$\mathcal{N}(\alpha|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\alpha - \mu)^2\right)$$

is the univariate normal pdf.

- The Gaussian white noise process is *Strict sense stationary*, since:

$$f_{Xn_1,Xn_2,...,Xn_N}(\alpha_1, \alpha_2, ..., \alpha_N) = \prod_{i=1}^{N} \mathcal{N}(\alpha_i|0, \sigma_X^2) = f_{Xn_1+c,Xn_2+c,...,Xn_N+c}(\alpha_1, \alpha_2, ..., \alpha_N)$$

7.6 Linear systems and random process



7.7 Properties

7.7.1 Wide-sense Stationarity

When a wide-sense stationary discrete random process $\{X_n\}$ is passed through a stable, linear time invariant (LTI) system with digital impulse response $\{h_n\}$, the output process $\{Y_n\}$, i.e.

$$y_n = \sum_{k=-\infty}^{\infty} h_k x_{n-k} = X_n * h_n$$

is also Wide-sense stationary

7.7.2 Correlation function

We can express the output correlation functions and power spectra in terms of the input statistics and the LTI system:

$$r_{xy}[k] = \mathbb{E}[X_n Y_{n+k}] = \sum_{l=-\infty}^{\infty} h_l r_{XX}[k-l] = h_k * r_{XX}[k]$$

$$r_{yy}[l] = \mathbb{E}[Y_n Y_{n+l}] = \sum_{k=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} h_k h_i r_{xx}[l+i-k] = h_l * h_{-l} * r_{XX}[l] \tag{1}$$

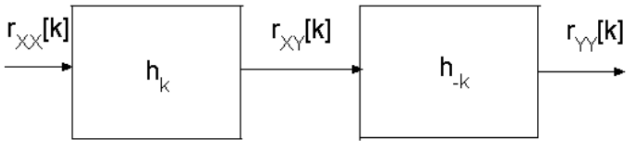


Figure 11: Correlation functions with convolution

7.7.3 Power spectrum

Power spectrum is found by converting this result to the frequency domain by taking DTFT of both sides of Eq. 1:

$$S_Y(e^{j\omega T}) = |H(e^{j\omega T})|^2 S_X(e^{j\omega T}) \tag{2}$$

Here $H(e^{j\omega T} = \sum_{n=-\infty}^{\infty} h_n \exp -jn\omega T$ is the frequency response of the system.

7.7.4 Example: Filtering white noise

- Suppose we filter a zero mean white noise process $\{X_n\}$ with a first order *finite impulse response (FIR) filter*:

$$y_n = \sum_{m=0}^{1} b_m X_{n-m}, \quad \text{or} \quad Y(z) = (b_0 + b_1 z^{-1}) X(z)$$

  with $b_0 = 1, b_1 = 0.9$. This is a *moving average* process

1. The impulse response of this causal filter is:

$$\{h_n\} = \{b_0, b_1, 0, 0...\}$$

2. The autocorrelation function of $\{Y_n\}$ is obtained as:

$$r_{YY}[l] = \mathbb{E}[Y_n Y_{n+l}] = h_l * h_{-l} * r_{XX}[l]$$

3. This convolution can be performed directly. However, it its more straightforward in the frequency domain.

4. The frequency response of the filter is:

$$H(e^{j\Omega}) = b_0 + b_1 e^{-j\Omega}$$

5. The power spectrum of $\{X_n\}$ (white noise) is:

$$S_X(e^{j\Omega}) = \sigma_X^2$$

6. Hence the power spectrum of $\{Y_n\}$ is:

$$\begin{aligned} S_Y(e^{j\Omega}) &= |H(e^{j\Omega})|^2 S_X(e^{j\Omega}) \\ &= |b_0 + b_1 e^{-j\Omega}|^2 \sigma_X^2 \\ &= (b_0 b_1 e^{j\Omega} + (b_0^2 + b_1^2) + b_0 b_1 e^{-j\Omega})\sigma_X^2 \end{aligned}$$

7. Comparing this expression with the DTFT of $r_{yy}[m]$:

$$S_Y(e^{j\Omega}) = \sum_{m=-\infty}^{\infty} r_{yy}[m] e^{-jm\Omega}$$

we can identify non-zero terms in the summations only when $m = -1, 0, +1$ as follows:

$$r_{YY}[-1] = \sigma_X^2 b_0 b_1, \quad r_{YY}[0] = \sigma_X^2 (b_0^2 + b_1^2), \quad r_{YY}[1] = \sigma_X^2 b_0 b_1$$

7.8 Ergodic Random process

7.8.1 Why ergodic

In practical signal processing systems, correlation function or power spectra may not be known in advance. How could we estimate these for a WSS process? If the process is **ergodic**, then easy methods exists.

7.8.2 Definition of Mean and Correlation ergodic

If we measure a realisation $\{x_n\}$ (i.e. one waveform from the ensemble of possible waveforms) of the process $\{X_n\}$ for some long (ideally infinite) period of time, we can use this to estimate means and correlation functions as follows:

- For an **Ergodic** random process we can estimate expectations by performing time-averaging on a single sample function, e.g.

$$\mu = \mathbb{E}[X_n] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n \qquad \text{Mean ergodic}$$

$$r_{xx}[k] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n x_{n+k} \qquad \text{Correlation ergodic}$$

- These formulae allow us to make estimation for 'sufficiently' large N:

$$\mu = \mathbb{E}[X_n] \approx \frac{1}{N} \sum_{n=0}^{N-1} x_n$$

$$r_{xx}[k] \approx \frac{1}{N} \sum_{n=0}^{N-1} x_n x_{n+k}$$

Clearly, larger N better estimation

7.8.3 When ergodic?

- It is hard in general to determine whether a given process is ergodic
- However, a *necessary* and *sufficient* condition for *mean* ergodicity is given by:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c_{XX}[k] = 0$$

  where $c_X X$ is the *autocovariance* function:

$$c_{XX}[k] = \mathbb{E}[(X_n - \mu)(X_{n+k} - \mu)]$$

  and $\mu = \mathbb{E}[X_n]$

- A simpler sufficient condition for mean ergodicity is that $C_{XX}[0] < \infty$ and

$$\lim_{N\to\infty} c_{XX}[N] = 0$$

- The logic of sufficient and necessary is that if we can prove the sufficient condition, that guarantees mean ergodicity.
- However, if we can not prove it (the sufficient), or it turns out to be false, we don't know for sure whether the process is ergodic.
- For necessary and sufficient condition, if we can prove that, then definitely ergodic. If we prove the condition is false, then the process is definitely not ergodic.
- Correlation ergodicity can be studied by extensions of the above theorems. Not examinable.
- Unless otherwise stated, we will always assume that the signals we encounter are both wide-sense stationary and ergodic.

7.8.4 Example 1

For measurement of the voltages of different battery cells, the population of batteries has a mean value 9V and standard deviation 0.1V and the population is believed to be Gaussian.The voltage across one randomly selected battery cell is measured for a time interval of N samples. It is intuitively obvious that we can't estimate the mean value of the population from this one set of N time measurements, but let's prove it:

1. Consider the 'd.c. level' random process:

$$X_n = A$$

   where A is a random variable having the standard Gaussian distribution

$$f(A = a) = \mathcal{N}(a|9, 0.01)$$

2. The mean of the random is:

$$\mathbb{E}[X_n] = \int_{-\infty}^{\infty} x_n(a) f(a)\, da = \int_{-\infty}^{\infty} a f(a)\, da = 9$$

3. Now consider a random sample function measured from the ranodm process, say

$$x_n = a_0$$

4. The 'ergodic' average value of this particular sample function is

$$1/N \sum_{n=0}^{N-1} a_0 = a_0$$

whatever signal duration N we have available.

5. Since in general $a_0 \neq E[A] = 9$, the process is clearly not mean ergodic, but could we have checked this using the mean ergodicity theorems?

6. Check this using the mean ergodic theorem. The autocovariance function is:

$$\begin{aligned} c_{XX}[k] &= \mathbb{E}[(X_n - \mu)(X_{n+k} - \mu)] \\ &= \mathbb{E}[(X_n - 9)(X_{n+k} - 9)] = \mathbb{E}[A^2 - 9^2] = \text{var}(A) = 0.01 \end{aligned}$$

7. Now

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c_{XX}[k] = 0.01 \times N/N = 0.01 \neq 0$$

8. Hence the theorem confirms our finding that the process is not ergodic in mean

9. While this example highlights a possible pitfall in assuming ergodicity, most of the processes we deal with will, however, be ergodic.

7.8.5 Ergodic example: MA process

- The autocorrelation function for the zero mean MA process as

$$r_{YY}[-1] = \sigma_X^2 b_0 b_1, \quad r_{YY}[0] = \sigma_X^2 (b_0^2 + b_1^2), \quad r_{YY}[1] = \sigma_X^2 b_0 b_1$$

and zero for all other lags.

- Since it is zero mean, the autocovariance function equals autocorrelation function r_{YY}. Hence we can use either the necessary or sufficient condition to show mean ergodicity.
  - Sufficient condition:

$$\lim_{N\to\infty} C_{YY}[N] = 0$$

  - Necessary condition:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c_{YY}[k] = \lim_{N\to\infty} \frac{1}{N} (\sigma_X^2 b_0 b_1 + \sigma_X^2 (b_0^2 + b_1^2) + \sigma_X^2 b_0 b_1) = 0$$

8 Optimal Filtering Theory - Wiener Filters

8.1 Overview

- Optimal filtering is an area in which we design filters that are optimally adapted to the statistical characteristics of a random process.
- Combination of standard filter design for deterministic signals with the random process theory.
- Pioneered in 1940's by Norbert Wiener who designed methods for optimal estimation of a signal measured in noise
- A desired signal d_n is observed in noise v_n:

$$x_n = d_n + v_n$$

- Wiener showed how to design a linear filter which would optimally estimate d_n given just the noisy observations x_n and some assumptions about the statistics of the random signal and noise processes. This class of filters the *wiener filter*, forms the basis of many fundamental signal processing applications.
- Typical applications include:
  - Noise reduction e.g. speech and music signals
  - Predication of future values of a signal, e.g. in finance
  - Noise cancellation, e.g. for aircraft cockpit noise
  - Deconvolution, e.g. removal of room acoustics (dereverberation) or echo cancellation in telephony
- The Wiener filter is a very powerful tool. However, it is only the optimal linear estimator for stationary signals. The *Kalman filter* offers an extension for non-stationary signals via *state space models*. In cases where a linear filter is still not good enough, non-linear filtering techniques can be adopted. (Non-examinable)

8.2 General filters

- The observed signal x_n can be filtered with an infinite dimensional filter, having a non-causal impulse response h_p:

$$\{h_p; p = -\infty, ..., -1, 0, 1, ..., \infty\}$$

- An estimate $\hat{(d)}_n$ of the desired signal can be obtained by filtering the noisy signal using the filter $\{h_p\}$:

$$\hat{d}_n = \sum_{p=-\infty}^{\infty} h_p X_{n-p}$$

- Since both d_n and x_n are drawn from random processes $\{d_n\}$ and $\{x_n\}$, performance can only be measured in terms of expectations. The criterion used for Wiener filtering is the mean-squared error (MSE). First, form the error signal ϵ_n:

$$\epsilon_n = d_n - \hat{d}_n = d_n - \sum_{p=-\infty}^{\infty} h_p X_{n-p}$$

The *mean-squared error (MSE)* is then defined as:

$$J = \mathbb{E}[\epsilon_n^2] \tag{3}$$

where the expectation is w.r.t the random signal d and the random noise v.

- The wiener filter minimises J w.r.t the filter coefficient $\{h_p\}$

8.3 Derivation of Wiener filter

8.3.1 Assumption

- The Wiener filter assumes that $\{x_n\}$ and $\{d_n\}$ are *jointly wide-sense stationary*.
  - Both means are constant
  - All autocorrelation and cross-correlation functions depend only on the time difference between data points.
- Assume $\{d_n\}$ and $\{v_n\}$ have zero mean

$$\mathbb{E}[d_n] = 0 \quad \mathbb{E}[v_n] = 0$$

Non-zero mean processes can be dealt with by first subtracting the mean before filtering.

8.3.2 Derivation

1. The expected error in Eq. 3 may be minimised w.r.t the impulse response values $\{h_q\}$. A sufficient condition for a minimum is:

$$\frac{\partial J}{\partial h_q} = \frac{\partial \mathbb{E}[\epsilon_n^2]}{\partial h_q} = \mathbb{E}\left[\frac{\partial \epsilon_n^2}{\partial h_q}\right] = \mathbb{E}\left[2\epsilon_n \frac{\partial \epsilon_n}{\partial h_q}\right] = 0$$

   simultaneously for all $q \in \{-\infty, ..., -1, 0, 1, ..., \infty\}$.

2. The term $\frac{\partial \epsilon_n}{\partial h_q}$ is then calculated as:

$$\frac{\partial \epsilon_n}{\partial h_q} = \frac{\partial}{\partial h_q}\left\{d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right\} = -x_{n-q}$$

   since d_n and X_{n-p} do not depend on h_p and are treated as constant in the partial derivative.

3. Hence the coefficient must satisfy, for all q:

$$\mathbb{E}\left[\epsilon_n \frac{\partial \epsilon_n}{\partial h_q}\right] = -\mathbb{E}[\epsilon_n x_{n-q}] = 0$$

   i.e.

$$\mathbb{E}[\epsilon_n x_{n-q}] = 0; \qquad -\infty < q < \infty \tag{4}$$

   This is known as the *orthogonality principle*, two random variables X and Y are termed *orthogonal* if

$$\mathbb{E}[XY] = 0$$

4. Substituting for ϵ_n in E.q. 4:

$$\begin{aligned}\mathbb{E}[\epsilon_n x_{n-q}] &= \mathbb{E}\left[\left(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right) x_{n-q}\right] \\ &= \mathbb{E}[d_n x_{n-q}] - \sum_{p=-\infty}^{\infty} h_p \mathbb{E}[x_{n-q} x_{n-q}] \\ &=^{(a)} r_{xd}[q] - \sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] \qquad\qquad = 0\end{aligned}$$

(a) since $\mathbb{E}[d_n x_{n-q}] = r_{dx}[-q] = r_{xd}[q]$

5. Rearranging, the solution must satisfy

$$\sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = r_{xd}[q], \qquad -\infty < q + \infty \tag{5}$$

Eq. 5 is the *Wiener-Hopf* equations, which involve an infinite number of unknowns h_q. The simplest way to solve this is in the frequency domain.

6. Rewrite the *Wiener-Hopf* equations as a discrete-time convolution:

$$h_q * r_{xx}[q] = r_{xd}[q], \qquad -\infty < q < +\infty$$

7. Taking DTFT of both sides:

$$H(e^{j\Omega}) S_x(e^{j\Omega}) = S_{xd}(e^{j\Omega})$$

where $S_{xd}(e^{j\Omega})$ is defined as the DTFT of $r_{xd}[q]$ and is termed the *cross-power spectrum* of d and x.

- The cross-power spectrum is in general complex valued and measures the coherence between two process at particular frequency.
- It has the property:

$$S_{xd}(e^{j\Omega}) = S_{dx}(e^{j|\Omega})^*$$

8. Rearranging:

$$H(e^{j\Omega}) = \frac{S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})} \tag{6}$$

8.3.3 Results

- Result E.q 6 tells us that the frequency response of the optimal filter for estimating d can be computed given knowledge of just the power spectrum (or equivalently the autocorrelation function) of the noisy signal x, and the cross power spectrum (or equivalently the cross-correlation function) between the noisy signal x and the desired signal d.
- Depending on scenario, it may be possible to estimate these quantities directly from data, and/or from physical modelling considerations about the system.
- Estimation from data will typically require the process to be ergodic, hence time averages converge to ensemble averages of correlation functions.
- This result in general yields a **non-causal** filter that is not implementable in practice. The practical approach is either to approximate the filtering in the frequency domain using DFTs, or derive **sub-classes of Wiener filter** that are *casual* and *implementable*, as considered shortly.

8.4 Mean-squared error for the Optimal filter

- E.q 6 show how to calculate the optimal filter for a given problem.
- However, have not given the performance of that optimal filter.
- This can be assessed from the mean-squared error value of the optimal filter:

$$\begin{aligned} J = \mathbb{E}[\epsilon_n^2] &= \mathbb{E}[\epsilon_n(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}) \\ &= \mathbb{E}[\epsilon_n d_n] - \sum_{p=-\infty}^{\infty} h_p \mathbb{E}\underbrace{[\epsilon_n x_{n-p}]}_{(a)} \end{aligned}$$

  (a): =0 at optimal from E.q 4

- Thus, the minimum error is:

$$\begin{aligned} J_{\min} &= \mathbb{E}[\epsilon_n d_n] \\ &= \mathbb{E}[(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}) d_n] \\ &= r_{dd}[0] - \sum_{p=-\infty}^{\infty} h_p r_{xd}[p] \end{aligned}$$

Steps to find the minimum error

1. Compute the autocorrelation function for the error signal

$$\begin{aligned} r_{\epsilon\epsilon}[k] = \mathbb{E}[\epsilon_n \epsilon_{n+k}] &= \mathbb{E}\left[(d_n - \sum_{p_1=-\infty}^{\infty} h_{p_1} x_{n-p_1})(d_{n+k} - \sum_{p_2=-\infty}^{\infty} h_{p_2} x_{n+k-p_2})\right] \\ &= \mathbb{E}\left\{d_n d_{n+k} - d_n \sum h_{p_2} x_{n+k-p_2} - d_{n+k} \sum h_{p_1} x_{n-p_1} + \sum\sum h_{p_1} h_{p_2} x_{n-p_1} x_{n+k-p_2}\right\} \\ &= r_{dd}[k] - \mathbb{E}\left\{\sum h_p r_{dx}[n+k-p-n] - \sum h_p r_{dx}[n-p-n-k] - \sum\sum r_{xx}[k-p_2+p_1]\right\} \\ &= r_{dd}[k] - \mathbb{E}\left\{\sum h_p r_{dx}[k-p] - \sum h_p r_{xd} k + p] - \sum\sum r_{xx}[k-p_2+p_1]\right\} \\ &= r_{dd}[k] - h_p * r_{dx}[k] - h_p^* * r_{xd}[k] + h_p * h_p^* * r_{xx}[k] \end{aligned}$$

2. Take its DTFT to get the power spectrum of the error signal

$$S_\epsilon(e^{j\Omega}) = S_D - S_{DX} \cdot H - S_{XD} \cdot H^* + S_X \cdot |H|^2$$

3. By taking the Inverse DTFT and put in $H = H^{\text{opt}}$ to get the power of minimum of the error

$$\begin{aligned}
J_{\min} &= \mathbb{E}[\epsilon_n^2] = r_{\epsilon\epsilon}[m=0] \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_\epsilon(e^{j\Omega})e^{jm\Omega}\,d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_\epsilon(e^{j\Omega})\,d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D - S_{DX}\cdot H - S_{XD}\cdot H^* + S_X\cdot|H|^2\,d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D - S_{DX}\cdot\frac{S_{XD}}{S_X} - H\cdot S_X\cdot H^* + S_X\cdot|H|^2\,d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D - S_{DX}\cdot\frac{S_{XD}}{S_X}\,d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) - S_{XD}^*\cdot H(e^{j\Omega})\,d\Omega
\end{aligned}$$

Since $H^{\text{opt}} = \frac{S_{XD}}{S_X}$ and $S_{XD} = S_{DX}^*$

8.5 Special Case: Uncorrelated Signal and Noise Process

- An important sub-class of the Wiener filter can be gained by considering the case where the desired signal process $\{d_n\}$ is uncorrelated with the noise process $\{v_n\}$, i.e.

$$r_{dv}[k] = \mathbb{E}[d_v v_{n+k}] = 0, \quad -\infty < k\infty$$

- This is a typical scenario in which v might be environmental noise that is independent of the desired signal and hence uncorrelated.
- One side concept: random variables that are independent and zero mean are uncorrelated; however, uncorrelated random zero mean random variables are not assured to be independent. (Independent stronger than correlation)
- In the Wiener-Hopf equations:

$$\sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = r_{xd}[q], \qquad -\infty < q + \infty$$

1. r_{xd}.

$$\begin{aligned}
r_{xd}[q] &= \mathbb{E}[x_n d_{n+q}] \\
&= \mathbb{E}[(d_n + v_n)d_{n+q}] \\
&= \mathbb{E}[d_n d_{n+q}] + \underbrace{\mathbb{E}[v_n d_{n+q}]}_{=0} \\
&= \mathbb{E}[d_n d_{n+q}]
\end{aligned}$$

Taking DTFT:

$$S_{xd}(e^{j\Omega}) = S_d(e^{j\Omega})$$

2. r_{XX}.

$$\begin{aligned} r_{XX}[q] &= \mathbb{E}[x_n x_{n+q}] \\ &= \mathbb{E}[(d_n + v_n)(d_{n+q} + v_{n+q})] \\ &= \mathbb{E}[d_n d_{n+q}] + \mathbb{E}[d_n v_{n+q}] + \mathbb{E}[v_n d_{n+q}] + \mathbb{E}[v_n v_{n+q}] \\ &= r_{dd}[q] + r_{vv}[q] \end{aligned}$$

Taking DTFT:

$$S_x(e^{j\Omega}) = S_d(e^{j\Omega}) + S_v(e^{j\Omega})$$

3. Thus the Wiener filter becomes

$$H(e^{j\Omega}) = \frac{S_{xd}}{S_x} = \frac{S_d(e^{j\Omega})}{S_d(e^{j\Omega}) + S_v(e^{j\Omega})} = \frac{1}{1 + 1/\rho(\Omega)} \tag{7}$$

where $\rho(\Omega) = S_d(e^{j\Omega}/S_v(e^{j\Omega}$ is the signal-to-noise (SNR) power ratio.

- From E.q 7 the behaviour of the filter is intuitively reasonable:
  - The gain is always non-negative, and ranges between 0 and 1. Hence the filter will never boost a particular frequency component; rather it acts as an optimal attenuation rule.
  - At those frequencies where the SNR is large, the gain of the filter tends to unity; whereas the gain tends to a small value at those frequencies where the SNR is small. Essentially, when the signal is very noisy, the best estimate the filter can make in a MSE sense is zero.
- The minimum expected error in this case:

$$\begin{aligned} J_{\min} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) - S^*_{XD} \cdot H(e^{j\Omega})\, d\Omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) - S_D \cdot H(e^{j\Omega})\, d\Omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) \left(1 - \frac{S_d(e^{j\Omega})}{S_d(e^{j\Omega}) + S_v(e^{j\Omega})}\right) d\Omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) \left(1 - \frac{1}{1 + 1/\rho(\Omega)}\right) d\Omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_D(e^{j\Omega}) \left(\frac{1}{1 + \rho(\Omega)}\right) d\Omega \end{aligned}$$

in which error tends to the power spectrum of $\{d_n\}$ when SNR is poor and tends to zero when SNR is good at a particular frequency.

8.6 Example:AR process

Problem Setting:

- An autoregressive process $\{d_n\}$ of order 1 is generated as:

$$d_n = a_1 d_{n-1} + e_n$$

with e_n as zero mean, variance σ_e^2 white noise

- Rewrite in z-transform domain and zero initial conditions:

$$D(z) = a_1 z^{-1} D(z) + E(z)$$

hence

$$D(Z) = \frac{E(z)}{1 - a_1 z^{-1}} = H(z)E(z)$$

where $H(z) = \frac{1}{1-a_1 z^{-1}}$ is a transfer function between e and d.

- The frequency response is thus:

$$H(e^{j\Omega}) = \frac{1}{1 - a_1 exp^{-j\Omega}}$$

- Power spectrum is then obtained from the linear systems result

$$S_d(e^{j\Omega}) = |H(e^{j\Omega})|^2 S_e(e^{j\Omega}) = \frac{\sigma_e^2}{(1 - a_1 exp^{-j\Omega})(1 - a_1 exp^{+j\Omega})}$$

- Suppose the process is observed in zero mean white noise with variance σ_v^2 and is uncorrelated with $\{d_n\}$:

$$x_n = d_n + v_n$$

- Design the Wiener filter for estimation of d_n.

Solution:

1. Use the uncorrelated frequency response formula:

$$\begin{aligned} H^{\text{opt}}(e^j\Omega) &= \frac{S_d}{S_d + S_v} \\ &= \frac{\frac{\sigma_e^2}{(1 - a_1 e^{-j\Omega})(1 - a_1 e^{+j\Omega})}}{\frac{\sigma_e^2}{(1 - a_1 e^{-j\Omega})(1 - a_1 e^{+j\Omega})} + \sigma_v^2} \\ &= \frac{\sigma_e^2}{\sigma_e^2 + \sigma_v^2 (1 - a_1 e^{-j\Omega})(1 - a_1 e^{+j\Omega})} \end{aligned}$$

2. Inverse DTFT is applied to find the impulse response of the filter

8.7 FIR Wiener filter

- In the previous part, the filter is non-causal. The impulse response h_p is defined for values of p less than 0.

- Here a practical alternative in which a causal Pth order Finite Impulse Response Wiener filter is developed.

- In the FIR case the signal estimate is:

$$\hat{d}_n = \sum_{p=0}^{P} h_p X_{n-p}$$

and we minimise, as before the MSE:

$$J = \mathbb{E}[(d_n - \hat{d}_n)^2]$$

8.7.1 Derivation of the FIR filter

- The filter derivation proceeds much as before, and we need to solve

$$\frac{\partial J}{\partial h_q} = \mathbb{E}\left[2\epsilon_n \frac{\partial \epsilon_n}{\partial h_q}\right] = 0, \qquad \text{for } q = 0, 1, ...P.$$

- Then, as before:

$$\frac{\partial \epsilon_n}{\partial h_q} = \frac{\partial}{\partial h_q}\left\{d_n - \sum_{p=0}^{P} h_p x_{n-p}\right\} = -x_{n-q}$$

  leading to the orthogonality principle:

$$\mathbb{E}[\epsilon_n x_{n-q}] = 0; \qquad q = 0, .., P$$

  and Wiener-Hopf equations as follows:

$$\sum_{p=0}^{P} h_p r_{xx}[q-p] = r_{xd}[q], \qquad q = 0, 1, ..., P$$

- This is a simple finite set of simultaneous equations that we can solve in the time domain. The equations may be written in matrix form as:

$$R_x h = r_{xd}$$

  where:

$$h = \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_P \end{bmatrix} \qquad r_{xd} = \begin{bmatrix} r_{xd}[0] \\ r_{xd}[1] \\ \vdots \\ r_{xd}[P] \end{bmatrix}$$

  and

$$R_X = \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \dots & r_{xx}[P] \\ r_{xx}[1] & r_{xx}[0] & \dots & r_{xx}[P-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[P] & r_{xx}[P-1] & \dots & r_{xx}[0] \end{bmatrix}$$

  R_x is known as the *correlation matrix*.

- Note that $r_{xx}[k] = r_{xx}[-k]$ so that the correlation matrix R_x is symmetric and has constant diagonals (a symmetric *Toeplitz* matrix)

- The coefficient vector can be found by matrix inversion:

$$h = R_X^{-1} r_{xd}$$

- For finding the impulse response of the FIR filter, we need to have *a-priori* knowledge of the autocorrelation matrix R_x and the cross-correlation r_{xd}

- The minimum mean-square error is given by:

$$\begin{aligned} J_{\min} =& \mathbb{E}[\epsilon_n d_n] \\ =& \mathbb{E}[(d_n \sum_{p=0}^{P} h_p x_{n-p}) d_n] \\ =& r_{dd}[0] - \sum_{p=0}^{P} h_p r_{xd}[p] \\ =& r_{dd}[0] - r_{xd}^T h \\ =& r_{dd}[0] - r_{xd}^T R_x^{-1} r_{xd} \end{aligned}$$

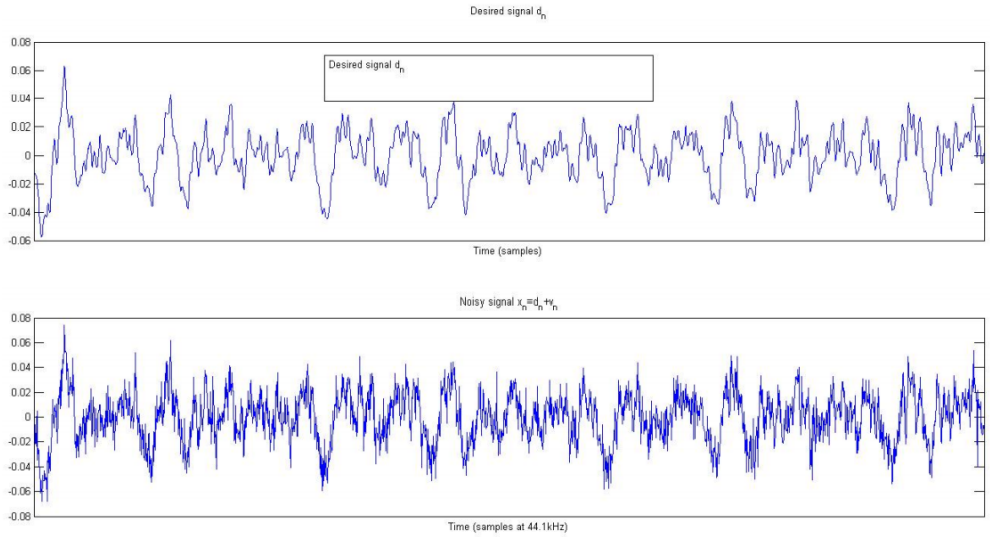
8.8 Case study: audio noise reduction

8.8.1 Overview and assumptions

- Consider a section of acoustic waveform d_n that is corrupted by additive noise v_n

$$x_n = d_n + v_n$$

d_n: Clean audio signal $\quad$ x_n: Noisy audio signal



- Implement FIR Wiener filter to reduce the noise in the signal.
- Assume that the section of data is wide-sense stationary and ergodic (approx. true for a short segment around 1/40 s).
- Assume also that the noise is white and uncorrelated with the audio signal and have variance σ_v^2 i.e.

$$r_{vv}[k] = \sigma_v^2 \delta[k]$$

8.8.2 Wiener filter

- The wiener filter in this case needs
  - $r_{xx}[k]$, Autocorrelation of noisy signal

- $r_{xd}[k] = r_{dd}[k]$, Autocorrelation of desired signal

- Since signal is assumed ergodic, these quantities can be estimated:

$$r_{xx}[k] \approx \frac{1}{N}\sum_{n=0}^{N-1} x_n X_{n+k}$$

$$r_d d[k] = r_{xx}[k] - r_{vv}[k] = \begin{cases} r_{xx}[k], & k \neq 0 \\ r_{xx}[0] - \sigma_v^2, & k = 0 \end{cases}$$

- Under what conditions would $r_{dd}[k]$ form a valid autocorrelation sequence for construction of R_x?
  - A necessary condition is that the resulting R_x matrix is non-negative definite.
  - Definition of non-negative definiteness :

$$a^T R_x a \geq 0$$

    for any length $P+1$ vector a.
  - Now, form the vector $x_n = [X_n, X_{n-1}...X_{n-P}]^T$.
  - Take now the non-negative quantity:

$$(a^T X_n)^2 = (a^T X_n)(X_n^T a) = a^T (x_n X_n^T) a$$

  - The (i,j)th element of $x_n X_n^T$ is $x_{n-i+1}x_{n-j+1}$ and $\mathbb{E}[x_{n-i+1}x_{n-j+1}] = r_{xx}[i-j]$. Hence $\mathbb{E}[x_n x_n^T] = R_x$
  - Now take the expectation of the non-negative quantity which itseld will therefore be non-neagtive:

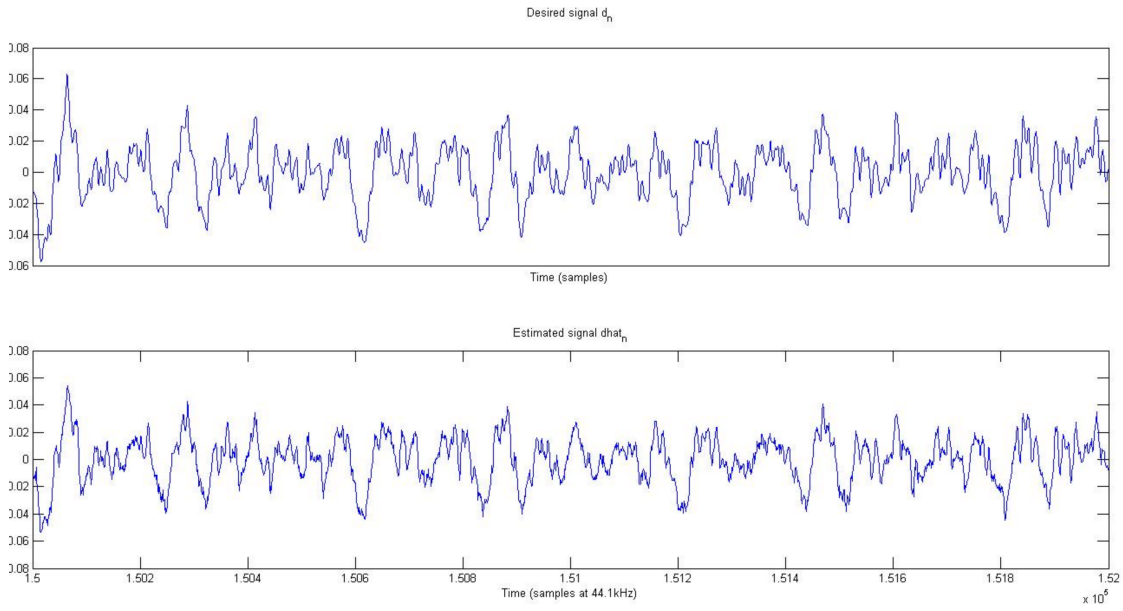
$$\mathbb{E}[(a^T X_n)^2] = a^T \mathbb{E}[(x_n X_n^T)] a = a^T R_x a \geq 0$$

    and hence the correlation matrix of any vector x_n must be non-negative definite.

- Choose the filter length P, form the autocorrelation matrix and cross-correlation vector and solve in Matlab:

$$h = R_x^{-1} r_{xd}$$

- The output looks like this with $P = 350$



- The theoretical MSE is calculated as:

$$J_{\min} = r_{dd}[0] - r_{xd}^T h$$

- This can be computed for various filter lengths, and in this artificial scenario we can compare the theoretical error performance with the actual MSE, since we have access to the true d_n itself:

$$J_{\text{true}} = \frac{1}{N}\sum_{n=0}^{N-1}(d_n - \hat{d}_n)^2$$

Not necessarily equal to the theoretical value since we estimated the autocorrelation functions from finite pieces of data and assumed stationarity of the processes.

8.9 Extending the Wiener filter

- The Wiener Filter can readily be extended to deal with cases outside the regular noise reduction case. You could replace the desired signal d with whatever one wants to predict or estimate and then rederive the new version of the filter.
- Generally, for the FIR case the formula

$$h = R_x^{-1} r_{xd}$$

applies to whatever the form of the desired signal and the 'noisy' signal, provided you can calculate the necessary correlation functions.

- The theory for non-standard cases will be examinable.
- Examples are:

1. Prediction of a noisy signal $\{u_n\}$. The model for the noisy signal is as before:

$$x_n = u_n + v_n$$

but now we define the desired signal to be the predicted value of the signal, $d_n = u_{n+p}$, where p is the desired prediction interval. The FIR Wiener estimate takes the same form as before:

$$\hat{d}_n = \sum_{p=0}^{P} h_p x_{n-p}$$

the error is $\epsilon_n = d_n - \hat{d}_n$ just as before

2. Smoothing of a noisy signal. In this case we use the current samples to get an even better estimate of the signal at some point in the past. The setup is exactly the same as for prediction, except that $d_n = u_{n-p}$ where p is the amount of 'lookahead' that we can allow. This would be appropriate in systems where a certain time-lag or or latency is allowable before the signal estimate needs to be obtained.
3. Deconvolution. To extract a signal u_n from a noisy convolved version of itself:

$$x_n \sum_{q=0}^{Q} h_q u_{n-q} + v_n$$

The setup is the same as for the regular filter, setting $d_n = u_n$, but we will have a more complex expression for the autocorrelation function of x_n and the cross-correlation between d and x. One example is the removal of room acoustics from a voice signal u_n.

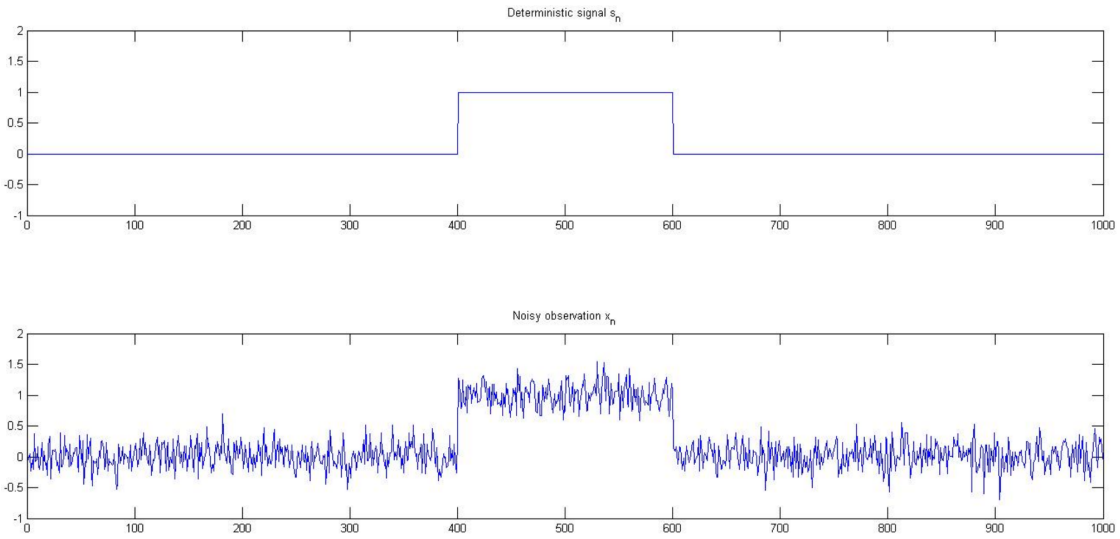
9 Optimal Detection - Matched Filter

9.1 What is Matched Filter

- The Wiener filter shows how to extract a random signal from a random noise environment.
- How about detecting a known deterministic signal s_n, $n = 0, ..., N-1$, buried in random noise v_n:

$$x_n = s_n + v_n$$

for example:



- The method is *Matched filter*
- It finds extensive application in detection of pulses in communications data, radar and sonar data

9.2 Formulate the problem

- First 'vectorise' the equation:

$$x = s + v$$

$$s = [s_0, s_1, ..., s_{N-1}]^T, \quad x = [x_0, x_1, ..., x_{N-1}]^T$$

- Once again, we will design an optimal FIR filter for performing the detection task. Suppose the filter has coefficients h_m, for $m = 0, 1, ..., N-1$, then the output of the filter at time $N-1$ is:

$$\begin{aligned} y_{N-1} &= \sum_{m=0}^{N-1} h_m x_{N-1-m} \\ &= h^T \tilde{x} \\ &= h^T(\tilde{s} + \tilde{v}) \\ &= h^T \tilde{s} + h^T \tilde{v} \\ &= y^s_{N-1} + y^n_{N-1} \end{aligned}$$

where $\tilde{x} = [x_{N-1}, X_{N-2}, ..., x_0]^T$ is the 'time-reversed' vector. y^s_{N-1} is defined as the output from the signal-only part and y^n_{N-1} is the output from just the noise going through the filter.

- Instead of minimising the MSE, we maximise the signal-to-noise ratio (SNR) at the output of the filter, hence giving best possible chance of detecting the signal s_n.
- Define output SNR as:
$$\frac{\mathbb{E}[|y^s_{N-1}|^2]}{\mathbb{E}[|y^n_{N-1}|^2]} = \frac{\mathbb{E}[|h^T\tilde{s}|^2]}{\mathbb{E}[|h^T\tilde{v}|^2]} = \frac{|h^T\tilde{s}|^2}{\mathbb{E}[|h^T\tilde{v}|^2]}$$
since numerator is not a random quantity and only the noise (v) part is random.

9.3 Signal output energy

- The signal component at the output is $y^s_{N-1} = h^T\tilde{s}$, with energy
$$|h^T\tilde{s}|^2 = (h^T\tilde{s})(\tilde{s}^T h) = h^T(\tilde{s}\tilde{s}^T)h$$
- To analyse this, consider the matrix $M = \tilde{s}\tilde{s}^T$. What are its eigenvectors/eigenvalues ?
  - Recall the definition of eigenvectors (e) and eigenvalues (λ):
$$Me = \lambda e$$
- Try $e = \tilde{s}$:
$$M\tilde{s} = (\tilde{s}\tilde{s}^T)\tilde{s} = \tilde{s}(\tilde{s}^T\tilde{s}) = (\tilde{s}^T\tilde{s})\tilde{s}$$
Hence the unit length vector $e_0 = \tilde{s}/|\tilde{s}|$ is an eigenvector and $\lambda = (\tilde{s}^T\tilde{s}) is the corresponding$
- Now consider any vector e' which is orthogonal to e_0 (i.e. $\tilde{s}^T e' = 0$):
$$Me' = \tilde{s}\tilde{s}^T = 0$$
  - Hence e' is also an eigenvector, but with eigenvalue $\lambda' = 0$.
  - Since we can construct a set of $N-1$ orthonormal (unit length and orthogonal to each other) vectors which are orthogonal to $\tilde{s}$, call these $e_1, e_2, ..., e_{N-1}$, we have now discovered all N eigenvectors/eigenvalues of M.
- Since the N eigenvectors form an orthonormal basis, we may represent any filter coefficient vector h as a linear combination of these:
$$h = \alpha e_0 + \beta e_1 + \gamma e_2 + ... + ...e_{N-1}$$
- Thus we can express
$$\begin{aligned} Mh &= M(\alpha e_0 + \beta e_1 + ... + ...e_{N-1}) \\ &= \alpha M e_0 + ... + ...M e_{N-1} \\ &= \alpha(\tilde{s}^T\tilde{s})e_0 + 0 + 0 + 0 + ... + 0 \\ &= \alpha(\tilde{s}^T\tilde{s})e_0 \end{aligned}$$
since all but the first eigenvalue is zero.

- Now, consider the signal output energy again:

$$\begin{aligned} h^T\tilde{s}\tilde{s}^T h &= h^T M h \\ &= h^T \alpha(\tilde{s}^T\tilde{s})e_0 \\ &= \alpha(\tilde{s}^T\tilde{s})h^T e_0 \\ &= (\alpha\tilde{s}^T\tilde{s})(\alpha e_0 + \beta e_1 + ... + ...e_{N-1})^T e_0 \\ &= \alpha\tilde{s}^T\tilde{s} \end{aligned}$$

since $e_0^T e_j = \delta[j]$ (eigenvectors are othonormal).

9.4 Noise output energy

- Now consider the expected noise output energy, which may be simplified as follows:

$$\mathbb{E}[|h^T\tilde{v}|^2] = \mathbb{E}[h^T\tilde{v}\tilde{v}^T h] = h^T\mathbb{E}[\tilde{v}\tilde{v}^T]h$$

- Consider the case where the noise is white and zero mean with variance σ_v^2. Then, for any time indexes $i = 0, .., N-1$ and $j = 0, ..., N-1$:

$$\mathbb{E}[v_i v_j] = \begin{cases} \sigma_v^2, & i = j \\ 0, & i \neq j \end{cases}$$

and hence

$$\mathbb{E}[\tilde{v}\tilde{v}^T] = \sigma_v^2 I$$

where I is the $N \times N$ identity matrix, as the diagonal elements are those '$i = j$' terms and the off-diagonal are '$i \neq j$' terms.

- So we have the expression for the noise output

$$\mathbb{E}[|h^T\tilde{v}|^2] = \sigma_v^2 I h^T h$$

and expending h in terms of the eigenvectors of M:

$$\sigma_v^2 I h^T h = \sigma_v^2(\alpha^2 + \beta^2 + \gamma^2 + ...)$$

once again, all the other crossover terms are zero.

9.5 SNR Maximisation

- The SNR may now be expressed as:

$$\frac{|h^T\tilde{s}|^2}{\mathbb{E}[|h^T\tilde{v}|^2]} = \frac{\alpha^2\tilde{s}^T\tilde{s}}{\sigma_v^2(\alpha^2 + \beta^2 + \gamma^2 + ...)}$$

- Scaling h by some factor ρ will not change the SNR since numerator and denominator will both scale equally be ρ^2. So, we can arbitrarily fix $|h| = 1$ and then maximise.

- With $|h| = 1$ we have $(\alpha^2 + \beta^2 + ...) = 1$ and the SNR becomes just equal to

$$\frac{\alpha^2\tilde{s}^T\tilde{s}}{\sigma_v^2}$$

- The largest possible value of α given that $|h| = 1$ is 1 and that implies $\beta = \gamma = 0$ and the solution becomes:

$$h^{\text{opt}} = 1 \times e_0 = \frac{\tilde{s}}{|\tilde{s}|}$$

  i.e. the optimal filter coefficients are just the normalised time-reversed signal

- The SNR at the optimal filter setting is given by

$$\text{SNR}^{\text{opt}} = \frac{\tilde{s}^T \tilde{s}}{\sigma_v^2}$$

  and clearly the performance depends very much on the energy of the signal s and the noise v.

9.6 Practical Implementation of the matched filter

- A batch of data of same length as the signal s and optimised a filter h of the same length are chosen.

- In practice we would now run this filter over a much longer length of data x which maximum energy occurs. This is the point at which s can be detected and optimal thresholds can be devised to make the decision on whether a detection of s should be declared at that time.

- In fact, we have only proved that the signal to noise ratio is maximised at one single filter output time, $n = N - 1$.However, given a stationary noise process, it is straightforward to show that the signal output power term for the optimal filter is always less than that computed for $n = N - 1$ since the deterministic correlation function of s_n always has its maximum at lag zero.

- Example: a square pulse radar detection problem

$$s_n = \text{Rectangle pulse} = \begin{cases} 1, & n = 0, 1, ..., T-1 \\ 0, & \text{otherwise} \end{cases}$$

  - Optimal filter is the normalised time reversed version of s_n:

$$h_n^{\text{opt}} = \begin{cases} 1/\sqrt{T}, & n = 0, 1, ..., T-1 \\ 0 & \text{otherwise} \end{cases}$$

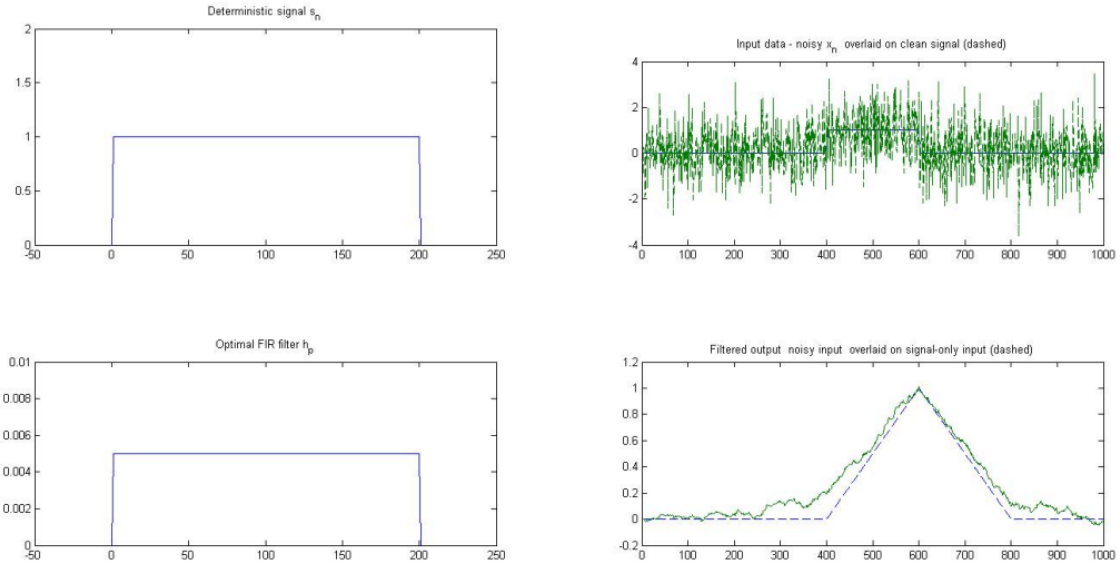
  - SNR achievable at detection point:

$$\text{SNR}^{\text{opt}} = \frac{\tilde{s}^T \tilde{s}}{\sigma_v^2} = \frac{T}{\sigma_v^2}$$

  - Compare with the best SNR attainable before matched filtering:

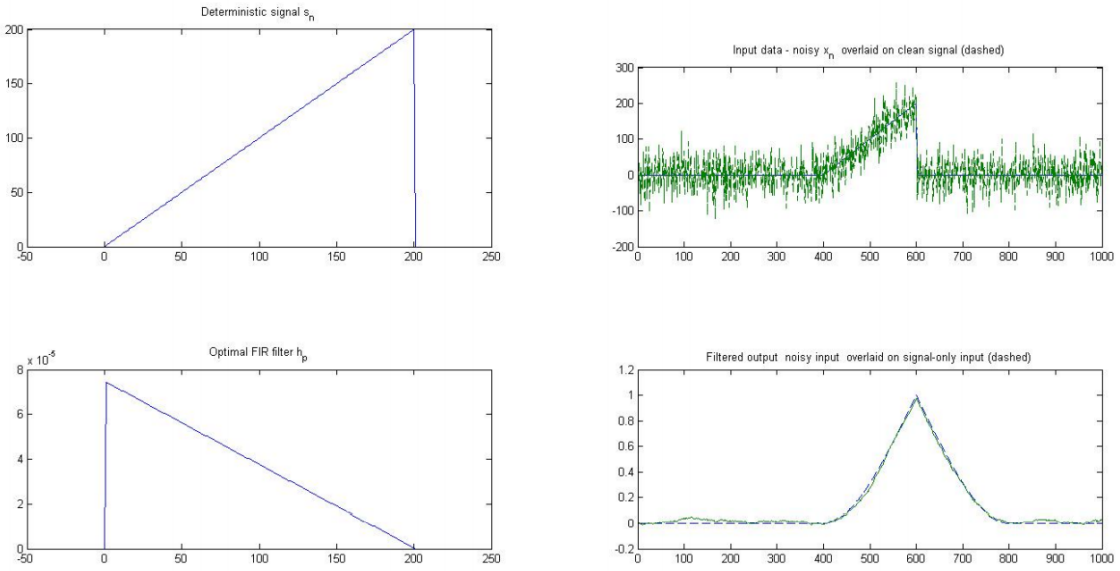
$$\text{SNR} = \frac{\text{Max signal value}^2}{\text{Average noise energy}} = \frac{1}{\sigma_v^2}$$

    i.e. a factor of T improvement, which could be substantial for long pulses $T >> 1$



- Example: saw-tooth pulse:

$$s_n = \text{Sawtooth pulse} = \begin{cases} n+1, & n = 0, 1, ..., T-1 \\ 0, & \text{otherwise} \end{cases}$$



10 Estimation theory and Inference

- Wiener Filter is a special example of estimation theory
- In general, statistical estimation involves the analysis of random data in order to determine quantities of interest. Wherever models need to be fitted to signal data, or quantitative hypothesis tested about datasets, estimation and inference will be required.
- Simple examples include the estimation of mean and variance for a collection of random mesurements, while more sophisticated examples might involve the estimation of parameters for some complex probability model of a signal such as multiple sinusoid model or an autoregressive model.

10.1 Estimation and Inference

1. In estimation theory, we start off with a vector of signal measurements x:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{bmatrix}$$

and some unknown quantities or parameters that we wish to infer:

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{P-1} \end{bmatrix}$$

Usually $P << N$ i.e., we have a lot more data than parameters, but in some of the modern modelling scenarios used in genomics, could be conversely.

2. Suppose the probability distribution of the data x can be expressed in terms of a joint probability density function (or probability mass function if discrete), or likelihood function:

$$p(x|\theta)$$

3. The likelihood function forms a generative model for the data, expressing how 'likely' different realisations of the observed data would be if we knew the underlying model parameters θ. This is conceptually the same thing as the random ensemble in random process theory.

   - E.g. A measured battery voltage is

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

4. In many problems we may treat θ as a random vector, and from physical or other modelling considerations, a prior probability density function can be formulated for θ:

$$p(\theta)$$

5. Such a probability distribution will represent prior belief about likely parameter configurations before any data have been seen. The prior distribution could be used to regularise the inference problem by constraining the paramter search to reasonable parts of the domain of θ.

   - For example, for the battery example, the population of batteries is known to have a mean around 9V, but we don't know it exactly. The prior could be assumed Gaussian and centered on 9V and the prior variance, say 0.1:

$$p(\mu) = \mathcal{N}(9, 0.1)$$

6. A strong prior regulariser would be required for cases where $P >> N$, for example the notion of 'sparsity' in which the prior encodes the notion that only a few of the θ elements are non-zero. Such framework fit broadly in to the category of Bayesian estimation, since Bayes' Theorem will be used to carry out the inference.

7. In other problems, we may not wish to consider θ to be a random quantity at all. In theses cases we can rely only on the likelihood function, and this is known as *Classical* or *Likelihood-based* inference. In such cases, it will be hard to regularise the solutions in the way that is possible within a Bayesian framework, and artificial regularisations may need to be introduced for analysis of complex models.

8. In *Estimation* problems, the task will be ti formulate an estimate $\hat{\theta}$ which is close in some sense to the true θ

9. *Inference* on the other hand, has estimation as a special case, but is more general in that we attempt to study the whole probability distribution of the unknown, including the uncertainties that remain about the value of θ once the data x have been observed

10.2 General Linear Model

10.2.1 Definition

- Models and likelihoods can take all sorts of forms. This class of models includes autoregressive model, the random sinusoid and various ad hoc/physically based structures that find use in applications.

- The general Linear Model is also known as the *Linear Regression* model in Machine Learning and Statistics.

10.2.2 Formulation

1. In the Linear Model it is assumed that the data x are generated as a linear function of the parameters θ with an additive random modelling error term e_n:

$$x_n = g_n^T \theta + e_n$$

   where g_n is a P-dimensional column vector

2. The expression may be written for the whole vector x as

$$\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$$

   where

$$G = \begin{bmatrix} g_0^T \\ g_1^T \\ \vdots \\ g_{N-1}^T \end{bmatrix}$$

3. Choice of the matrix G, the Design Matrix, will lead to a wide range of possible models, G may contain regression variables from another observed process, structural terms from the model, or even regressed values of x, as will be seen now

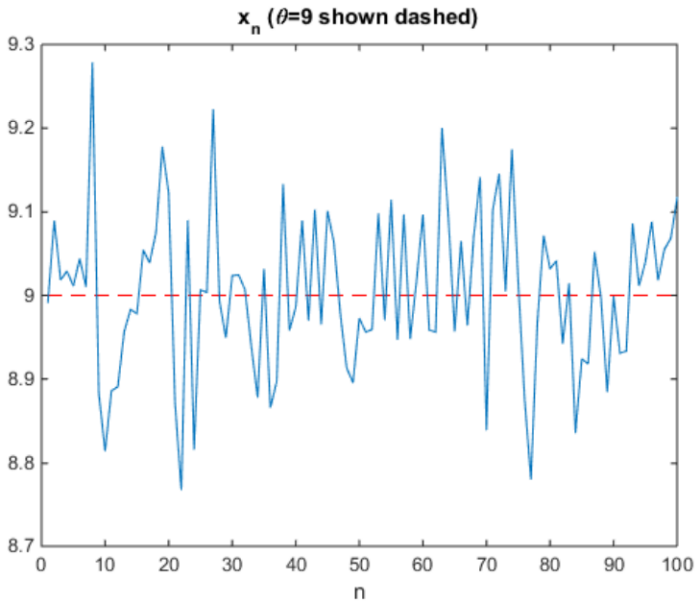
10.2.3 Example 1: Constant level in noise

Here we have a single parameter θ to model the unknown constant level;

$$x_n = \theta = e_n$$

and hence a very simple form for the design matrix:

$$G = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$



10.2.4 Example 2: the Sinusoidal model

- This model forms the basic building block for many frequency and spectral estimation algorithms. It is also used in sinusoidal speech coders.
- If we write a single sinusoid as the sum of sine and cosine components at a particular (known for now) frequency ω, we have:

  $$x_n = a\cos(\omega n) + b\sin(\omega n) + e_n$$

  where a and b are unknown parameters. This is equivalent to a sinusoid with uniformly random phase and random amplitude.
- Thus we can form a second order ($P = 2$) linear model from this if we take:

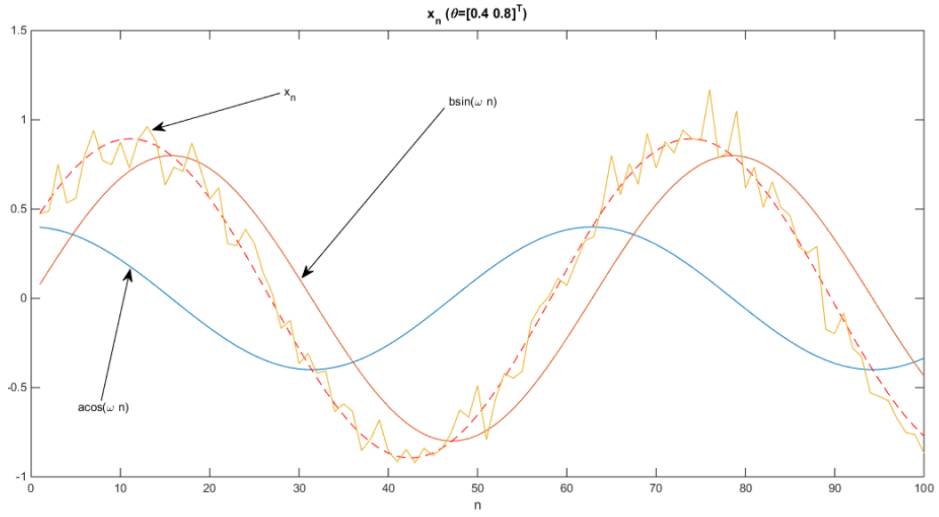
  $$G = \begin{bmatrix} c(\omega) & s(\omega) \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix}$$

  where

  $$c(\omega) = [\cos(0), \cos(\omega), \cos(2\omega), ..., \cos((N-1)\omega)]^T$$

  and

  $$s(\omega) = [\sin(0), \sin(\omega), \sin(2\omega), ..., \cos((N-1)\omega)]^T$$



- And similarly, a more complex model can be built with J sinusoids at different frequencies ω_j:

$$x_n = \sum_{j=1}^{J} a_j \cos(w_j n) + b_j \sin(\omega n) + e_n$$

and the linear model expression is

$$\mathbf{G} = \begin{bmatrix} c(\omega_1) & s(\omega_1) & c(\omega_2) & s(\omega_2) & ...c(\omega_J) & s(\omega_J) \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 & b_1 & a_2 & b_2 & ... & a_J & b_J \end{bmatrix}^T$$

- Thus we can make up a very complicated signal composed of lots of 'sinusoids' all added together. If we estimate the parameters θ from some data then we will be doing a kind of probabilistic 'spectrum estimation'.

10.2.5 Example 3: AR model

- The AR model is a standard time series model based on an all-pole filtered version of the noise residual:

$$x_n = \sum_{i=1}^{P} a_i x_{n-i} + e_n$$

where e_n is the zero mean white noise with variance σ_e^2

- The coefficients $\{a_i; i = 1...P\}$ are the filter coefficients of the all-pole filter, the AR parameters, and P, the number of coefficients, is the order of the AR process.

- $\{e_n\}$ can be interpreted as a 'prediction error' when predicting the next data point from the previous P.

- The transfer function for the filter is:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-1}}$$

- And hence the power spectrum for the model is :

$$S_x(e^{j\Omega}) = |H(\exp(j\Omega))|^2\sigma_e^2 = \frac{\sigma_e^2}{|\sum_{i=1}^{P} a_i e^{-j\Omega i}|^2}$$

- The shape of the power spectrum may readily be sketched by first sketching the magnitude frequency response of $H(z)$ and then squaring

- The model is used extensively in linear prediction of speech, speech synthesis and coding (especially in its adaptation to low bitrate CELP encoders)

- The AR modelling equation of (1) is now rewritten for the block of N data samples as

$$\mathbf{x} = \mathbf{G}\mathbf{a} + \mathbf{e}$$

  where $\mathbf{e}$ is the vector of N error values and the $(N \times P)$ matrix G is given by

$$\mathbf{G} = \begin{bmatrix} x_{-1} & x_{-2} & \cdots & x_{-(p-1)} & x_{-P} \\ x_0 & x_{-1} & \cdots & x_{-(p-2)} & x_{-p-1} \\ \vdots & & \ddots & & \vdots \\ x_{N-2} & X_{N-3} & \cdots & x_{N-P} & x_{N-P-1} \end{bmatrix}$$

- The $\mathbf{G}$ matrix has to contain data values prior to time $n = 0$ in order to calculate all the error terms e_n, $n = 0, ..., N-1$.

- This is fa special version of the linear model in which the matrix $\mathbf{G}$ is not fixed before making the measurements, but is actually made up of observed data points, owing to the feedback all-pole structure of the filter involved.

- This means that we can't directly generate a vector of data from an AR model using the formula $\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$, where $\mathbf{e}$ is generated from some suitable noise process, such as white Gaussian noise, since $\mathbf{G}$ now depends on the data we are trying to generate.

- However, we can generate the data sequentially by applying the all-pole IIR filter with coefficient $a_1, ..., a_P$ to a white noise input signal.

10.3 AR power spectrum

Follow from the last example, here we can find the power spectrum of the AR model:

- To sketch the power spectrum from the poles, for the general IIR filter:

$$H(e^{j\Omega}) = \frac{\sum_{k=0}^{M} b_k e^{-jk\Omega}}{1 - \sum_{k=1}^{N} a_k e^{-jk\Omega}}$$

  and in factorised form (where c_q are the zeros and d_q the poles):

$$H(e^{j\Omega}) = b_0 \frac{\prod_{q=1}^{M}(1 - c_q e^{-j\Omega})}{\prod_{q=1}^{M}(1 - d_q e^{-j\Omega})} = b_0 \frac{e^{-jM\Omega}\prod_{q=1}^{N}(e^{j\Omega} - c_q)}{e^{-jN\Omega}\prod_{q=1}^{N}(e^{j\Omega} - d_q)}$$

- For the power spectrum, taking square for the complex modulus:

$$|H(e^{j\Omega})|^2 = b_0^2 \frac{\prod_{q=1}^{M} |e^{j\Omega} - c_q|^2}{\prod_{q=1}^{N} |e^{j\Omega} - d_q|^2}$$
$$= b_0^2 \frac{\prod_{q=1}^{M} \text{ Squared distance from } e^{j\Omega} toc_q}{\prod_{q=1}^{N} \text{ Squared distance from } e^{j\Omega} tod_q}$$

- Hence for the AR model look at 1 over the product of squared distances from the unit circle to each pole, with each pole contributing a peak to the magnitude spectrum (high-Q peaks for poles close to the unit circle, low - Q peaks for poles distant from the unit circle). Note that low-Q peaks are easily hidden by neighbouring high-Q peaks:

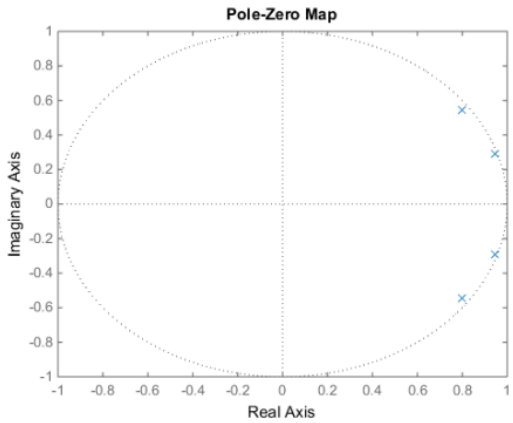


Figure 12: AR model poles with $P = 4$, poles at $(r, \theta) = 0.99 \exp(\pm j0.1\pi)$ and $(r, \theta) = 0.97 \exp(\pm j0.4\pi)$

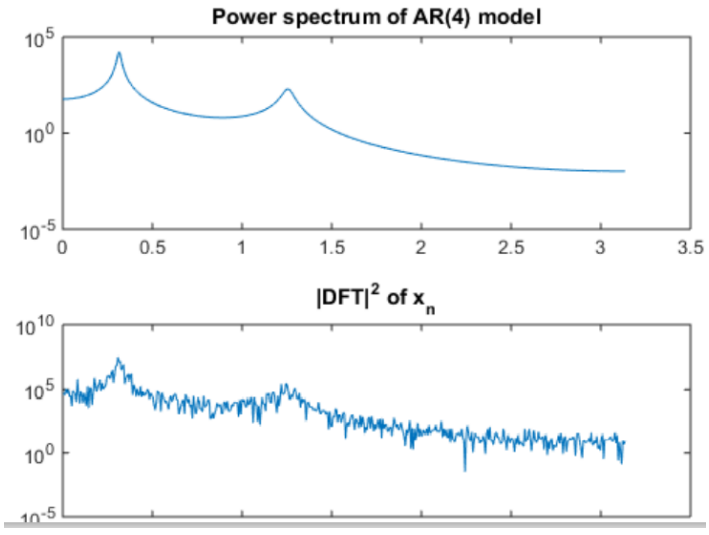


Figure 13: AR model PSD and DFT showing the similar shape but noisy version. Also showing the poles

- Next, Einstein-Wiener-Khinchin Theorem is introduced to prove this similar in shape behaviour.

10.4 Einstein-Wiener-Khinchin Theorem

1. Take a time-windowed version of the signal x_n, having duration $2N+1$ samples and zero elsewhere;

$$x_n^N = w_n^N x_n$$

where

$$w_n^N = \begin{cases} 1, & -N \leq n \leq N \\ 0, & \text{otherwise} \end{cases}$$

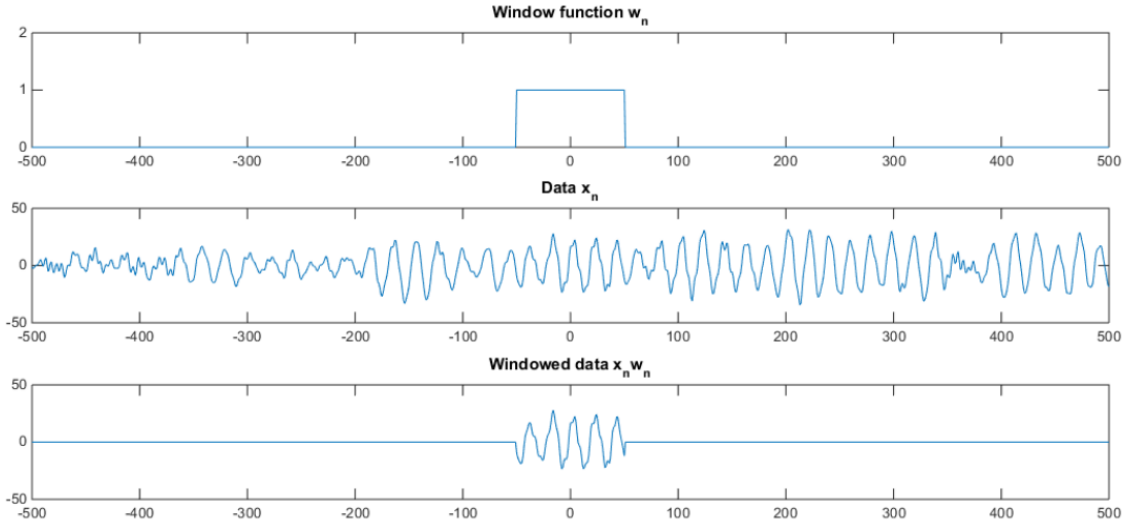


Figure 14: Window function applied to a data

2. Look at the DTFT:

$$X^N(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} w_n^N x_n e^{-jn\Omega}$$

- $X^N(e^{j\Omega})$ is the DFT of a frame of $2N+1$ data points. The shift of the origin back to $n = -N$ will not affect the magnitude of the DFT calculated, but will introduce a linear phase shift.

3. Now, its modulus squared can be expanded:

$$|X^N(e^{j\Omega})|^2 = X^N(e^{j\Omega})X^N(e^{j\Omega})^*$$

and hence we can write the following DTFT pair:

$$\text{DTFT}\{x_n^N * x_{-n}^N\} = X^N(e^{j\Omega})X^N(e^{j\Omega})^* = |X^N(e^{j\Omega})|^2$$

where x_{-n}^N is the time-reversed version of x_n^N. Notice that we are here using the result:

$$\text{DTFT}\{x_{-n}^N\} = \sum_{n=-\infty}^{\infty} x_{-n}^N e^{-jn\Omega} = \sum_{n'=-\infty}^{\infty} x_{n'}^N e^{jn'\Omega} = X^N(e^{j\Omega})^*$$

4. Expand the time-domain convolution:

$$x_n^N * x_{-n}^N = \sum_{n=-\infty}^{\infty} x_n^N x_{n-m}^N = \sum_{n=-\infty}^{\infty} x_n w_n^N x_{n-m} w_{n-m}^N$$

5. The DTFT results therefore becomes:

$$\text{DTFT}\{\sum_{n=-\infty}^{\infty} x_n w_n^N x_{n-m} w_{n-m}^N\} = |X^N(e^{j\Omega})|^2$$

6. Then divide both side by the window duration $(2N+1)$ and take expectations on both sides:

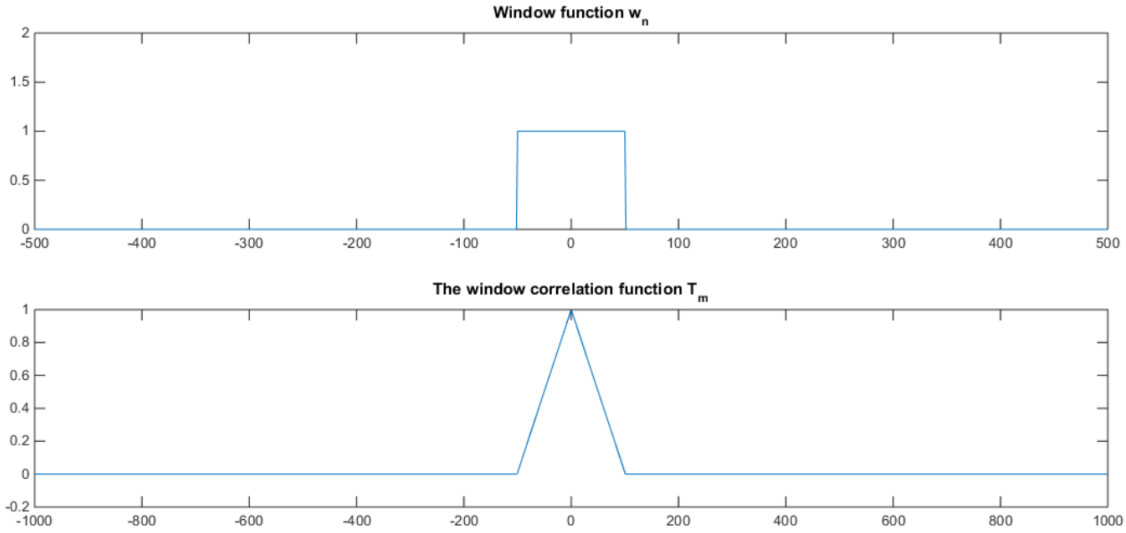
$$\frac{1}{2N+1}\mathbb{E}\left\{\text{DTFT}\{\sum_{n=-\infty}^{\infty} x_n w_n^N x_{n-m} w_{n-m}^N\}\right\} = \frac{1}{2N+1}\mathbb{E}\left\{|X^N(e^{j\Omega})|^2\right\}$$

$$\text{DTFT}\{\mathbb{E}[\frac{1}{2N+1}\sum_{n=-\infty}^{\infty} x_n w_n^N x_{n-m} w_{n-m}^N]\} = \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2]$$

7. And rewrite the term with autocorrelation function:

$$\begin{aligned}\mathbb{E}[\frac{1}{2N+1}\sum_{n=-\infty}^{\infty} x_n w_n^N x_{n-m} w_{n-m}^N] &= \frac{1}{2N+1}\sum_{n=-\infty}^{\infty} r_{xx}[m] w_n^N w_{n-m}^N \\ &= r_{xx}[m]\frac{1}{2N+1}\sum_{n=-\infty}^{\infty} w_n^N w_{n-m}^N \\ &= r_{xx}[m]t[m]\end{aligned}$$

where $t[m]$ is the deterministic autocorrelation function of the window-function w_n, as shown in the graph below:



- To summarise, we have the following DTFT relationship:

$$\text{DTFT}\{r_{xx}[m]t[m]\} = \mathbb{E}[\frac{1}{2N+1}|X^N(e^j\Omega)|^2]$$

- And the DTFT of r_{xx} is the power spectrum, $S_x(e^{j\Omega})$:

$$\text{DTFT}\{r_{xx}[m]t[m]\} = S_x(e^{j\Omega}) * T(e^{j\Omega})$$

where $T(e^{j\Omega})$ is the DTFT of $t[m]$.

- The plot below shows $T(e^{j\Omega})$ as N increases.

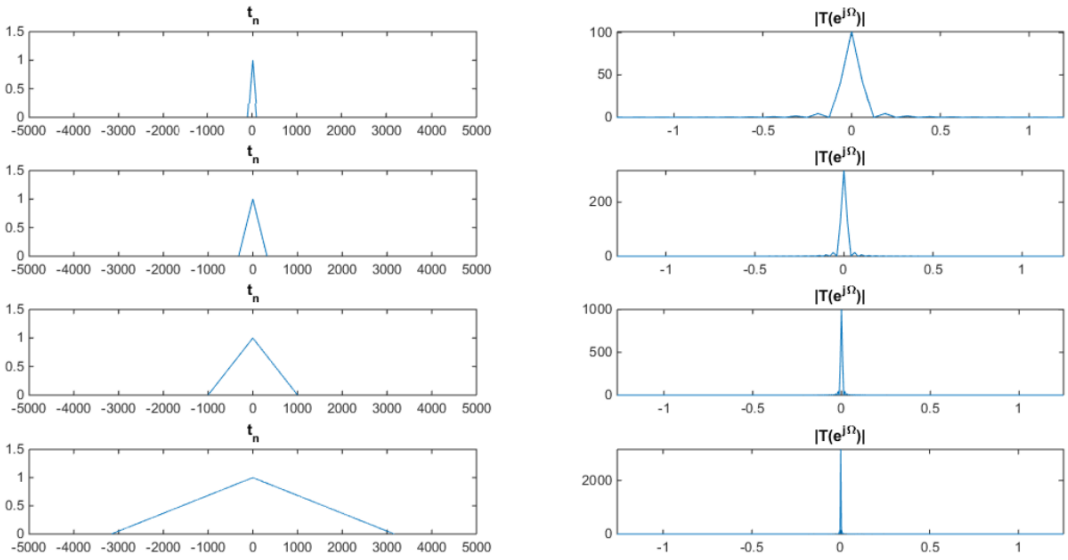


Figure 15: Spectrum of t_n as N increases

- As $t[m]$ gets wider and 'flatter', so $T(e^{j\Omega})$ tends to a delta function.
- Thus the final limiting expression becomes:

$$\lim_{N\to\infty} \text{DTFT}\{r_{xx}[m]t[m]\} = \text{DTFT}\{r_{xx}[m]\} = S_x(e^{j\Omega}) = \lim_{N\to\infty} \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2]$$

- In the limit, we have proved that the power spectrum is proportional to the expected value of the DTFT-squared of the data.
- The reason that there is a $1/(2N+1)$ factor is to make sure $t[m]$ remains finite-energy as N goes to infinity.
- To summarise, we have:

$$\lim_{N\to\infty} \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2] = S_x(e^{j\Omega})$$

- The power spectrum is the expected value of the time normalised DTFT-squared of the signal values. Hence, in our previous AR model example, with finite N, the DFT was random but had a similar shape to the underlying power spectrum of the process:

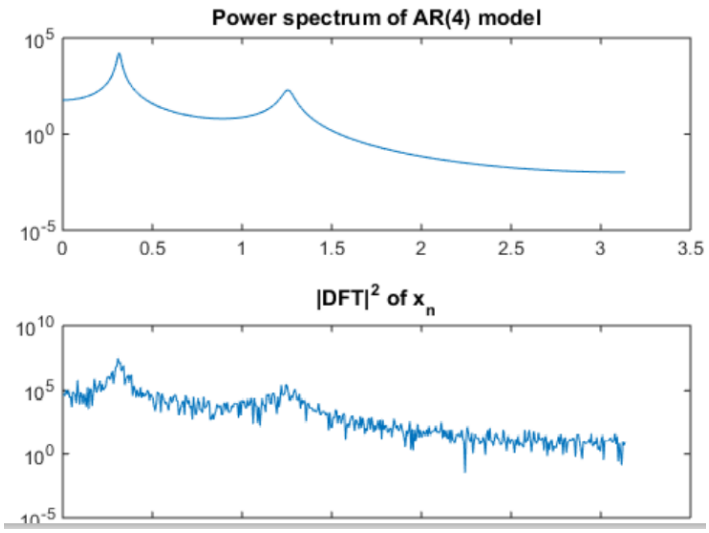


Figure 16: AR model PSD and DFT showing the similar shape but noisy version. Also showing the poles

10.5 Simple Estimators

We consider the mean and variance for estimators. For example, consider a task to estimate the mean of a random variable

- Suppose we have a set of N independent samples of a random variable X which has mean μ and standard deviation σ
  - Intuitively we can estimate the mean by taking the average of the samples:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - Intuitively this would be the best way to estimate the mean, but how do we know if it was a good estimator?
  - One way is to say that on average we should expect $\hat{\mu}$ to equal the true mean of X, $\mu = \mathbb{E}[X]$.
- This can be expressed in terms of expectations. An estimator is termed unbiased if

$$E[\hat{\mu}] = \mu$$

  where the expectations is taken with respect to the distribution of all of the random variables X_i;

$$E[\hat{\mu}] = \int_{x_1}\int_{x_2}\cdots\int_{x_N}\frac{1}{N}\sum_{i=1}^{N} x_i p(x_1)p(x_2)...p(x_N)d_{x_1}d_{x_2}...d_{x_N}$$

  and $p(x)$ is the common pdf of all of the random variables X_i. The mean of X is μ, i.e.

$$\mu = \int_x xp(x)d_x$$

- Such an estimator is described as unbiased, clearly a desirable property. Check that for the proposed estimator:

$$\begin{aligned}\mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i] \\ &= \mu\end{aligned}$$

Variance:

- However this is not the end of the whole story. Just because an estimator is correct on average doesn't stop it being inaccurate for much of the time. One way of measuring this property for an estimator is to measure the **variance** of the estimator:

$$\text{var}(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \mathbb{E}[\hat{\mu}^2] - \mathbb{E}[\hat{\mu}]^2$$

- Now test the variance of the proposed mean estimator:

$$\text{var}(\hat{\mu}) = \mathbb{E}[\hat{\mu}^2] - \mathbb{E}[\hat{\mu}]^2 = \mathbb{E}[\hat{\mu}^2] - \mu^2$$

- $\mathbb{E}[\hat{\mu}^2]$ simplifies as:

$$\begin{aligned}\mathbb{E}[\hat{\mu}^2] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2\right] \\ &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\mathbb{E}[X_i X_j]\end{aligned}$$

  Since the variable X_i are independent,

$$\mathbb{E}[X_i X_j] = \begin{cases} \mathbb{E}[X_i^2] = \mu^2 + \sigma^2, & i = j \\ \mu^2, & i \neq j \end{cases}$$

- Therefore,

$$\begin{aligned}\mathbb{E}[\hat{\mu}^2] &= \frac{1}{N^2}(N^2\mu^2 + N\sigma^2) \\ &= \mu^2 + \sigma^2/N\end{aligned}$$

- Hence:

$$\text{var}(\hat{\mu}) = \mathbb{E}[\hat{\mu}^2] - \mathbb{E}[\hat{\mu}]^2 = \mathbb{E}[\hat{\mu}^2] - \mu^2 = \mu^2 + \sigma^2/N - \mu^2 = \sigma^2/N$$

  therefore, as $N \to \infty$ the variance tends to zero.

Summary:

- An estimator such as this, which is unbiased and whose variance tends to zero as $N \to \infty$ is termed *consistent*. Such an estimator will 'definitely' get the correct answer with enough data.
- These definitions of *unbiased* estimators and their variance can in principle be used to measure the performance of any proposed estimation scheme.
- In fact, estimators can be designed for a given problem specifically to lead to no bias and have minimum variance. Such estimators are *minimum variance unbiased (MVU)* estimators.

10.6 Linear estimator

10.6.1 General Linear Model

General linear model has the form:

$$\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$$

First derive the *Ordinary Least Squares* estimator for the general linear model. Here we will carry out using matrix-vector derivatives.

1. Try to find the 'best fit' model that minimise the following error:

$$J = \sum_{n=0}^{N-1} e_n^2 = \mathbf{e}^T\mathbf{e}$$

2. Expand using $\mathbf{e} = \mathbf{x} - \mathbf{G}\theta$

$$J = \mathbf{e}^T\mathbf{e} = (\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{G}^T\mathbf{G}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{G}^T\mathbf{x}$$

3. By defining the vector gradient in the usual way:

$$\nabla\phi = \frac{d\phi(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial\phi(\boldsymbol{\theta})}{\partial\theta_0} \\ \frac{\partial\phi(\boldsymbol{\theta})}{\partial\theta_1} \\ \vdots \\ \frac{\partial\phi(\boldsymbol{\theta})}{\partial\theta_{P-1}} \end{bmatrix}$$

   we obtain:

$$\frac{dJ}{d\boldsymbol{\theta}} = 2\mathbf{G}^T\mathbf{G}\boldsymbol{\theta} - 2\mathbf{G}^T x$$

4. For a stationary point, setting $\frac{dJ}{d\boldsymbol{\theta}} = 0$

$$\mathbf{G}^T\mathbf{G}\boldsymbol{\theta} = \mathbf{G}^T x$$

   for invertible $\mathbf{G}^T\mathbf{G}$,

$$\boldsymbol{\theta}^{\text{OLS}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}$$

   i.e. the classical *Ordinary Least Squares* estimate of $\boldsymbol{\theta}$

   - Another useful way to think about the expansion of J is by 'completing the square':

$$\mathbf{x}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{G}^T\mathbf{G}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{G}^T\mathbf{x} = (\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{OLS}})^T\mathbf{G}^T\mathbf{G}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{OLS}}) - \boldsymbol{\theta}^{\text{OLS}^T}\mathbf{G}^T\mathbf{x} + \mathbf{x}^T\mathbf{x}$$

   - This serves to show that the OLS estimator is globally optimal, and will also come in handy shortly under likelihood and Bayesian inference schemes.
   - We will come back to this under Maximum Likelihood estimation, but for now consider the properties of the OLS estimator for the General Linear Model.

10.6.2 Properties of the Linear Estimator

What are the properties and could we ever do better than OLS?

- First consider the bias:

$$\mathbb{E}[\boldsymbol{\theta}^{\text{OLS}}] = \mathbb{E}[(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}] = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbb{E}[\mathbf{x}]$$

- For the linear model: $\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$:

$$\mathbb{E}[\mathbf{x}] = \mathbf{G}\boldsymbol{\theta} + \mathbf{0} = \mathbf{G}\boldsymbol{\theta}$$

  since the noise process $\{e_n\}$ has zero mean

- Substituting back into the first expectation gives

$$\mathbb{E}[\boldsymbol{\theta}^{\text{OLS}}] = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{G}\boldsymbol{\theta} = (\mathbf{G}^T\mathbf{G})^{-1}(\mathbf{G}^T\mathbf{G})\boldsymbol{\theta} = \boldsymbol{\theta}$$

  hence proving that OLS is unbiased which sounds like good news

10.6.3 Covariance of OLS

Compare the variance to other linear estimators:

- Define the OLS matrix term as

$$\mathbf{C} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$$

- Then examine the variance of any other *unbiased* estimator, which we can write:

$$\hat{\boldsymbol{\theta}} = \mathbf{D}\mathbf{x}$$

  where

$$\mathbf{D} = \mathbf{C} + \boldsymbol{\Delta}$$

  AND $\boldsymbol{\Delta}$ is some matrix perturbation away from the OLS solution.

- For $\mathbf{D}\mathbf{x}$ to be unbiased we require that:

$$\mathbb{E}[\mathbf{D}\mathbf{x}] = \boldsymbol{\theta},$$

  i.e.

$$\mathbb{E}[(\mathbf{C} + \boldsymbol{\Delta})\mathbf{x}] = (\mathbf{C} + \boldsymbol{\Delta})\mathbb{E}[\mathbf{x}] = (\mathbf{C} + \boldsymbol{\Delta})\mathbf{G}\boldsymbol{\theta} = \boldsymbol{\theta} + \Delta G\boldsymbol{\theta} = \boldsymbol{\theta},$$

  therefore we require

$$\boldsymbol{\Delta}\mathbf{G} = \mathbf{0}$$

- Now, the covariance matrix of the estimator of $\theta$is

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}])^T]$$

  This is the matrix-vector version of the scalar variance of a random variable. Note in particular that the (i, i)th element of the covariance matrix is the variance of $\hat{\theta_{i-1}}$

- But since we are dealing only with unbiased estimators we have $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$ and the calculation reduces to the covariance matrix of the estimation error:

$$\text{cov}(\boldsymbol{\theta}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] = \mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] - \boldsymbol{\theta}\boldsymbol{\theta}^T$$

- Now, for the linear estimator $\hat{\boldsymbol{\theta}} = \mathbf{Dx}$, we have

$$\mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] = \mathbb{E}[\mathbf{Dxx^T D^T}] = \mathbf{D}\mathbb{E}[\mathbf{xx^T}]\mathbf{D^T}$$

- And

$$\mathbf{xx^T} = (\mathbf{G}\boldsymbol{\theta} + \mathbf{e})(\mathbf{G}\boldsymbol{\theta} + \mathbf{e})^\mathbf{T} = \mathbf{G}\boldsymbol{\theta}\boldsymbol{\theta}^\mathbf{T}\mathbf{G^T} + \mathbf{ee^T} + \mathbf{e}\boldsymbol{\theta}^\mathbf{T}\mathbf{G^T} + \mathbf{G}\boldsymbol{\theta}\mathbf{e^T}$$

  hence

$$\mathbb{E}[\mathbf{xx^T}] = (\mathbf{G}\boldsymbol{\theta} + \mathbf{e})(\mathbf{G}\boldsymbol{\theta} + \mathbf{e})^\mathbf{T} + \sigma_e^2\mathbf{I}$$

  since $\{e_n\}$ is zero mean white noise with variance σ_e^2.s

- Now the expectation is obtained as

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] =& \mathbf{D}\mathbb{E}[\mathbf{xx^T}]\mathbf{D^T} \\ &= \mathbf{D}(\mathbf{G}\boldsymbol{\theta}\boldsymbol{\theta}^\mathbf{T}\mathbf{G^T} + \sigma_\mathbf{e}^\mathbf{2}\mathbf{I})\mathbf{D^T} \\ &= \boldsymbol{\theta}\boldsymbol{\theta}^T + \sigma_e^2\mathbf{D}\mathbf{D}^T\end{aligned}$$

  where the last line is obtained by:

$$\mathbf{DG} = (\mathbf{C} + \boldsymbol{\Delta})\mathbf{G} = \mathbf{CG} + \boldsymbol{\Delta}\mathbf{G} = \mathbf{CG} = \mathbf{I}$$

- Then we have,

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] - \boldsymbol{\theta}\boldsymbol{\theta}^T = \sigma_e^2\mathbf{D}\mathbf{D}^T \\ &= \sigma_e^2(\mathbf{C} + \boldsymbol{\Delta})(\mathbf{C} + \boldsymbol{\Delta})^T \\ &= \sigma_e^2(\mathbf{C}\mathbf{C}^\mathbf{T} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\mathbf{T} + \boldsymbol{\Delta}\mathbf{C}^\mathbf{T} + \mathbf{C}\boldsymbol{\Delta}^\mathbf{T}) \\ &= \sigma_e^2(\mathbf{C}\mathbf{C}^\mathbf{T} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\mathbf{T}) \\ &= \sigma_e^2((\mathbf{G}^\mathbf{T}\mathbf{G})^{-\mathbf{1}} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\mathbf{T})\end{aligned}$$

- Clearly with $\Delta = 0$ we have the OLS estimator covariance , so

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \text{cov}(\boldsymbol{\theta}^{\text{OLS}}) + \sigma_e^2\Delta\Delta^T$$

  "OLS estimator has the lowest convariance" and the rest part is the positive valued perturbation term

Summary:

- Now the variance of each parameter estimate $\hat{\theta_{i-1}}$ is the ith diagonal element of $\text{cov}(\hat{\boldsymbol{\theta}})$. And the diagonal elements of $\Delta\Delta^T$ are also ≥ 0 by its construction. Hence we have that

$$\text{var}(\hat{\theta}_i) \geq \text{var}(\theta_i^{\text{OLS}})$$

  with equality when $\Delta = 0$

- We have thus proved that the OLs estimator is the minimum variance unbiased estimator of $\boldsymbol{\theta}$. Such an estimator is termed as **Best Linear Unbiased Estimator (BLUE)**
- Also it can showed that the OLS is the uniqure BLUE for the Generale Linear Model.
- If in addition we have prior probability information $p(\boldsymbol{\theta})$ about $\boldsymbol{\theta}$ then a Bayesian estimator can give better mean-squared error performance at the cost of some small bias in the estimates

10.7 Likelihood Estimation

Defining e_n error term:

- The error sequence $\mathbf{e}$ will be assumed i.i.d. , that is

$$p(\mathbf{e}) = p_e(e_0)p_e(e_1)...p_e(e_{N-1})$$

  where p_e denotes some identical noise distribution. All of the p terms here are pdf. Note that here not necessarily zero mean so far.

- $\{e_n\}$ can be viewed as a modelling error, 'innovation' or observation noise, depending upon the type of model.

- When $p_e()$ is the zero-mean normal distribution we have the Linear Gaussian model, sometimes known as the *Gauss-Markov* model.

Multivariate Gaussian density function:

- For a length N random column vector $\mathbf{X}$:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt[N]{2\pi}\det(\mathbf{C}_{\mathbf{x}})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T {\mathbf{C}_{\mathbf{X}}}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Here $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ is the mean vector

- And $\mathbf{C}_{\mathbf{X}} = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$ is the *Covariance Matrix*

- Elementwise we have that $[C_X]ij = c_{X_iX_j}$ the covariance between elements X_{i-1} and X_{j-1}.

10.7.1 Maximum Likelihood (ML) Estimator

- The observed data $\mathbf{x}$ is considered random and we often obtain the pdf for $\mathbf{x}$ when the value of $\boldsymbol{\theta}$ is known. This pdf is termed the *likelihood* $L(\mathbf{x};\boldsymbol{\theta})$:

$$L(\mathbf{x};\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$$

- The Maximum Likelihood (ML) estimate for $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that maximises the likelihood for given observations $\mathbf{x}$:

$$\boldsymbol{\theta}^{\mathrm{ML}} = \arg\max_{\theta}\{p(\mathbf{x}|\boldsymbol{\theta})\}$$

- The ML solution corresponds to the parameter vector which would have generated the observed data $\mathbf{x}$ with highest probability.

- The maximisation task required for ML estimation can be achieved using standard differential calculus for well-behaved and differentiable likelihood functions

- It is convenient analytically to maximises the log-likelihood function

$$l(\mathbf{x};\boldsymbol{\theta}) = \log(L(\mathbf{x};\boldsymbol{\theta}))$$

  Since log is a monotonically increasing function, the two solutions are identical.

- The likelihood function is arrived at through knowledge of the stochastic model for the data.

- For example, for Gauss-Markov model we have:

$$\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$$

where the noise terms are i.i.d. as zero-mean Gaussian:

$$p(\mathbf{e}) = \prod_{n=0}^{N-1} \mathcal{N}(e_n|0, \sigma_e^2) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{1}{2\sigma_e^2} e_n^2}$$

- Which can be rewritten in the vector form:

$$\prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{1}{2\sigma_e^2} e_n^2} = \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{1}{2\sigma_e^2} \sum_{n=0}^{N-1} e_n^2} = \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{1}{2\sigma_e^2} \mathbf{e^T e}}$$

which is the multivariate Gaussian distribution with mean zero and covariance matrix $\sigma_e^2\mathbf{I}$

$$p(\mathbf{e}) = \mathcal{N}(\mathbf{e}|\mathbf{0}, \boldsymbol{\sigma_e^2 I})$$

- To get the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$, notice that the linear model equation $\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e}$ is a vector change of variables,

$$\mathbf{e} \to \mathbf{x}$$

- We are conditioning on $\boldsymbol{\theta}$ and therefore we can treat $\mathbf{G}\boldsymbol{\theta}$ as a constant term in the change of variables.

- Hence the change of variables is a very simple one with unity Jacobian and we get

$$p(\mathbf{x}|\boldsymbol{\theta}) = p_e(\mathbf{e})|_{\mathbf{e}=\mathbf{x}-\mathbf{G}\boldsymbol{\theta}}$$

- Thus the likelihood is:

$$L(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}) = p_{\mathbf{e}}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})$$

- Expanding this out we get:

$$p_e(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})\right)$$

and taking logarithms:

$$\begin{aligned}
\log L(\mathbf{x}; \boldsymbol{\theta}) =& -(N/2)\log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) \\
=& -(N/2)\log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{n=0}^{N-1}(x_n - \mathbf{g}_n^T\boldsymbol{\theta})^2 \\
=& \frac{1}{2\sigma_e^2}\sum_{n=0}^{N-1}(x_n - \mathbf{g}_n^T\boldsymbol{\theta})^2 + \text{constant}
\end{aligned}$$

- Thus maximisation of this function w.r.t. $\boldsymbol{\theta}$ is equivalent to minimising the sum-squared of the error sequence. This is exactly the criterion which is applied in the familiar ordinary least squares (OLS) estimation method

- Hence we get that ML estimator is:

$$\boldsymbol{\theta}^{\mathrm{ML}} = \boldsymbol{\theta}^{\mathrm{OLS}} = (\mathbf{G}^{\mathbf{T}}\mathbf{G})^{-\mathbf{1}}\mathbf{G}^{\mathbf{T}}\mathbf{x}$$

Summary:

- In general, then, when the error process $\{e_n\}$ is zero-mean independent and Gaussian with fixed variance, the OLS and ML solutions are identical.
- We would get a different solution if the noise were non-white/Gaussian.
- Moreover, the Bayesian inference method will give a new solution to the estimation problem even in white Gaussian noise case.

10.7.2 Example: AR model for speech

- We can apply the ML method directly to the AR model.
- We know for the AR model the form of the general linear model is:

$$\mathbf{x} = \mathbf{G}\mathbf{a} + \mathbf{e}$$

where $\mathbf{e}$ is the vector of N error values and the $(N \times P)$ matrix G is given by

$$\mathbf{G} = \begin{bmatrix} x_{-1} & x_{-2} & \cdots & x_{-(p-1)} & x_{-P} \\ x_0 & x_{-1} & \cdots & x_{-(p-2)} & x_{-p-1} \\ \vdots & & \ddots & & \vdots \\ x_{N-2} & X_{N-3} & \cdots & x_{N-P} & x_{N-P-1} \end{bmatrix}$$

- Then, measure some data $\mathbf{x}$d construct $\mathbf{G}$ as above from teh data vector and estimate the parameters by ML:

$$\mathbf{a}^{\mathrm{ML}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}$$

- This method is often referred to as the 'covariance' method.

10.7.3 Estimating the variance

- The noise variance can also be estimated in the Linear Gaussian Model by ML
- To see this, look at the log-likelihood function at the optimal parameter estimate $\boldsymbol{\theta}^{\mathrm{ML}}$ aldo considered as a function of σ_e^2:

$$\begin{aligned} \log L(\mathbf{x}; \boldsymbol{\theta}^{\mathrm{ML}}, \sigma_e^2) &= -(N/2)\log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}^{\mathrm{ML}})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}^{\mathrm{ML}}) \\ &= -(N/2)\log(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}J^{\mathrm{ML}} \end{aligned}$$

where J^{ML} is the minimum squared error term corresponding to the ML optimisation

- Differentiate wrt σ_e^2 and set to zero to get:

$$\frac{\partial \log L(\mathbf{x}; \boldsymbol{\theta}^{\mathrm{ML}}, \sigma_e^2)}{\partial \sigma_e^2} = \frac{(N/2)}{\sigma_e^2} + \frac{J^{\mathrm{ML}}}{2(\sigma_e^2)^2} = 0$$

and hence

$$\sigma_e^{2^{\mathrm{ML}}} = J^{\mathrm{ML}}/N$$

- Then apply it to some real speech.
- Estimate from recorded speech vector, an AR model $\mathbf{a}$ plus variance σ_e^2 for various orders P
- Reasynthesize some new speech by generating random white Gaussian noise with the estimated variance:

$$e_n^{\text{synth}} \sim \mathcal{N}(0, \sigma_e^{2^{\text{ML}}})$$

- Then running the noise through the estimated AR (all-pole) filter:

$$x_n^{\text{synth}} = \sum_{i=1}^{P} a_i^{\text{ML}} x_{n-i}^{\text{synth}} + e_n^{\text{synth}}$$

10.8 Bayesian Methods

10.8.1 Definition

- The ML methods treat parameters as unknown constants. If we are prepared to treat parameters as random variables/vectors it is possible to assign prior pdf to the parameters.
- These pdf should ideally express some prior knowledge about the relative probability of different parameter values before the data are observed
- Of course if nothing is known *a priori* about the paramters then the prior distributions should in some sense express no initial preference for one set of parameters over any other.
- In many cases a prior density is chosen to express some highly qualitative prior knowledge about the parameters.
  - In such cases the prior chosen will be more a reflection of a degree of belief concerning parameter values than any true modelling of an underlying random process which might have generated those parameters.
- The willingness to assign priors which reflect subjective information is a powerful feature and also one of the most fundamental differences between Bayesian and classical (likelihood-based) inferential procedures.

10.8.2 Properties and Process

- The precise form of probability distributions assigned *a priori* to the parameters requires careful consideration since misleading results can be obtained from erroneous priors, but in principle at least we can apply the Bayesian approach to any problem where statistical uncertainty is present.
- Bayes' Theorem is now stated as applied to estimation of random parameters $\boldsymbol{\theta}$ from a random vector $\mathbf{x}$ of observations, known as the posterior or *a posteriori* probability for the paramter:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{\overbrace{p(\mathbf{x}|\boldsymbol{\theta})}^{\text{Likelihood}} \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{Marginal Likelihood / Evidence}}}$$

Note that all of the distributions in this expression are implicitly conditioned upon all prior modelling assumptions, as was the likelihood function.

  - The distribution $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood as used for ML estimation,
  - while $p(\boldsymbol{\theta})$ is the prior or *a priori* distribution for the parameters. This term is one of the critical differences between Bayesian and classical techniques. It expresses in an objective fashion the probability of various model parameters values before the data $\mathbf{x}$ has been observed.

- The prior density may be an expression of highly subjective information about parameter values. This transformation from the subjective domain to an objective form for the prior can be of great significance and should be considered carefully when setting up an inference problem.

- The term $p(\boldsymbol{\theta}|\mathbf{x})$ the posterior or *a posteriori* distribution, expresses the probability of $\boldsymbol{\theta}$ given the observed data $\mathbf{x}$. This is now a true measure of how 'probable' a particular value of $\boldsymbol{\theta}$ is, given the observations $\mathbf{x}$

- $p(\boldsymbol{\theta}|\mathbf{x})$ is in a more intuitive form for parameter estimation than the likelihood, which expresses how probable the observations are given the parameters.

- The generation of the posterior distribution form the prior distribution when data $\mathbf{x}$ is observed can be thought of as a refinement to any previous ('prior') knowledge about the parameters.

- Before $\mathbf{x}$ is observed $p(\boldsymbol{\theta})$ expresses any information previously obtained concerning $\boldsymbol{\theta}$

- Any new information concerning the parameters contained in $\mathbf{x}$ is then incorporated to give the posterior distribution.

- Clearly if we start off with little or no information about $\boldsymbol{\theta}$ then the posterior distribution is likely to obtain information almost solely from $\mathbf{x}$.

- Conversely, if $p(\boldsymbol{\theta})$ expresses a significant amount of information about $\boldsymbol{\theta}$ then $\mathbf{x}$ will contribute relatively less new information to the posterior distribution.

- The denominator $p(\mathbf{x})$, referred to as the marginal likelihood, or the 'evidence' in machine learning, is a fundamentally useful quantity in model selection problems, and is constant for any given observation $\mathbf{x}$; thus it may be ignored if we are only interested in the relative posterior probabilities of different parameters.

- As a result of this, Bayes' theorem is often stated in the form:

$$p(\boldsymbol{\theta}|\mathbf{x}) = P(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- $p(\mathbf{x})$ may be calculated in principle by integration:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

  and this effectively serves as the normalising constant for the posterior density. For discrete replace integration with summation.

- We are here implicitly conditioning in this framework on many pieces of additional prior information beyond just the prior parameters of the model $\boldsymbol{\theta}$. For example, we are assuming a precise form for the data generation process in the model

- if the linear Gaussian model is assumed, then the whole data generation process must follow the probability law of that model, otherwise we cannot guarantee the quality of our answers; the same argument applies to ML estimation, although in the Bayesian setting the distributional form of the Bayesian prior must be assumed in addition.

10.8.3 Posterior inference

- The posterior distribution gives the probability for any chosen $\boldsymbol{\theta}$ given observed data $\mathbf{x}$, and as such optimally combines our prior information about $\boldsymbol{\theta}$ and any additional information gained about $\boldsymbol{\theta}$ from observing $\mathbf{x}$.
- We may in principle manipulate the posterior density to infer any required statistic of $\boldsymbol{\theta}$ conditional upon $\boldsymbol{\theta}$
- This is a significant advantage over ML and least squares methods which strictly give us only a single estimate of $\boldsymbol{\theta}$, known as a 'point estimate'.
- However, by producing a posterior p.d.f. with values defined for all $\boldsymbol{\theta}$ the Bayesian approach gives a fully interpretable probability distribution.
- In principle this is as much as one could ever need to know about the inference problem.
- In signal processing problems, however, we usually require a single point estimate for $\boldsymbol{\theta}$, and a principled way of choosing this is via an expected *cost function* $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ which express objectively a measure of the cost associated with a particular parameter estimate $\hat{\boldsymbol{\theta}}$ when the true parameter is $\boldsymbol{\theta}$

MAP

- The simplest and most intuitive way to perform Bayesian estimation is the maximum *a posteriori* (MAP) estimate, the value of $\hat{\boldsymbol{\theta}}$ which maximises the posterior distribution:

$$\boldsymbol{\theta}^{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \{p(\boldsymbol{\theta}|\mathbf{x})\}$$

- In other words, just like ML for the likelihood, we solve to find the parameters $\boldsymbol{\theta}$ which maximise the posterior probability.
- Work through the MAP estimation scheme under the Linear Gaussian model. Suppose that the prior on parameter vector $\boldsymbol{\theta}$ is the multivariate Gaussian:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_{\boldsymbol{\theta}}, \mathbf{C}_{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{P/2}|\mathbf{C}_{\boldsymbol{\theta}}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})\right)$$

  where $\mathbf{m}_{\boldsymbol{\theta}}$ is the prior parameter mean vector, $\mathbf{C}_{\boldsymbol{\theta}}$ is the parameter covariance matrix and P is the number of parameters in $\boldsymbol{\theta}$

- The likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ takes the same form as before for the Ml estimator, so the posterior distribution is as follows:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) \propto & p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta}) \\ \propto & \frac{1}{(2\pi)^{P/2}|\mathbf{C}_{\boldsymbol{\theta}}|^{1/2}} \; exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})\right) \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})\right) \end{aligned}$$

- Then $-2 \times$ log probability is given by:

$$-2\log(p(\boldsymbol{\theta}|\mathbf{x})) = (\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})^T \mathbf{C}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}}) + \frac{1}{\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) + \text{constant}$$

  where the constant term does not depend on $\boldsymbol{\theta}$

- Hence the MAP estimate $\boldsymbol{\theta}^{\text{MAP}}$ is obtained by differentiation and setting to 0:

$$\boldsymbol{\theta}^{\text{MAP}} = (\mathbf{G}^T\mathbf{G} + \sigma_e^2\mathbf{C}^{-1})(\mathbf{G}^T\mathbf{x} + \sigma_e^2\mathbf{C}^{-1}\mathbf{m}_{\boldsymbol{\theta}})$$

Compare to the ML estimator:

$$\boldsymbol{\theta}^{\text{ML}} = \boldsymbol{\theta}^{\text{OLS}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}$$

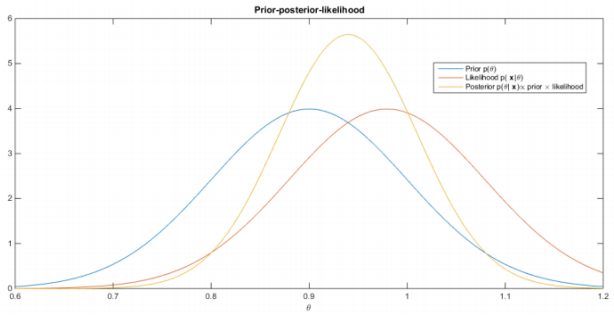


Figure 17: Prior, likelihood and posterior for 1-D Gaussians

- In the expression we can see the 'regularising' effect of the prior density on the ML estimate. As the prior becomes more 'diffuse', i.e. the diagonal elements of $\mathbf{C}_{\boldsymbol{\theta}}$ increases both in magnitude and relative to the off-diagonal elements, we imposes 'less' prior information on the estimate. In the limit the prior tends to a uniform (flat) prior with all $\boldsymbol{\theta}$ equally probable. In this limit $\mathbf{C}_{\boldsymbol{\theta}}^{-1} = 0$ and the estimate is identical to the ML estimate. This useful relationship demonstrates that the ML estimate ma be interpreted as the MAP estimate with uniform prior assigned to $\boldsymbol{\theta}$

- The MAP estimate will also tend towards the ML estimate when the likelihood is strongly 'peaked' around its maximum compared with the prior. Once again the prior will then have little influence on the shape of posterior density. It is in fact well known that as the sample size N tends to infinity the Bayes solution tends to the ML solution/

- This of course says nothing about small sample parameter estimates where the effect of the prior may be very significant.

- The choice of a multivariate Gaussian prior may well be motivated by physical considerations about the problem, or it may be motivated by subjective prior knowledge of about the value of $\boldsymbol{\theta}$ before the data $\mathbf{x}$ are seen in terms of a rough value $\mathbf{m}_{\boldsymbol{\theta}}$ and a confidence in that value through the covariance matrix $\mathbf{C}_{\boldsymbol{\theta}}$. In fact the choice of Gaussian also has the very special property that it makes the Bayesian calculations straightforward and available in closed form. Such a prior is known as a 'conjugate' prior.

- The required distribution can be obtained by rearranging the log probability function:

$$\frac{1}{\sigma_e^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}})^T\mathbf{C}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \mathbf{m}_{\boldsymbol{\theta}}) \tag{8}$$

$$= \frac{1}{\sigma_e^2}\left((\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{MAP}})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{MAP}}) + \mathbf{x}^T\mathbf{x} + \sigma_e^2\mathbf{m}_{\boldsymbol{\theta}}^T\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}} - \boldsymbol{\Theta}^T\boldsymbol{\theta}^{\text{MAP}}\right) \tag{9}$$

with terms defined as

$$\begin{aligned}\boldsymbol{\theta}^{\text{MAP}} &= \phi^{-1}\boldsymbol{\Theta} \\ \boldsymbol{\phi} &= \mathbf{G}^T\mathbf{G} + \sigma_e^2\mathbf{C}_{\boldsymbol{\theta}}^{-1} \\ \boldsymbol{\Theta} &= \mathbf{G}^T\mathbf{x} + \sigma_e^2\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}}\end{aligned}$$

- The first term in Eq.(9) is in exactly the correct form for the exponent of a multivariate Gaussian, with mean vector and covariance matrix as follows,

$$\mathbf{m}_{\boldsymbol{\theta}}^{\text{post}} = \boldsymbol{\theta}^{\text{MAP}} \qquad \mathbf{C}_{\boldsymbol{\theta}}^{\text{post}} = \sigma_e^2 \boldsymbol{\phi}^{-\mathbf{1}}$$

- Since then the remaining terms in Eq.(9) do not depend on $\boldsymbol{\theta}$, and we know that the multivariate density function must be proper (i.e. integrate to 1), we can conclude that the posterior distribution is itself a multivariate Gaussian,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\theta}^{\text{MAP}}, \sigma_e^2 \boldsymbol{\phi}^{-\mathbf{1}}).$$

  This formula would allow us to reinterpret the Bayesian estimator for the linear model in terms of its mean-squared error.

10.8.4 Example: Gaussian Model with one observation

- When $P = 1$ and $N = 1$, take $\mathbf{G} = 1$ so that:

$$x = \theta + e$$

  and prior :

$$p(\theta) = \mathcal{N}(\mu_\theta, \sigma_\theta^2)$$

- The likelihood is just:

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma_e^2) \propto e^{-\frac{1}{2\sigma_e^2}(x-\theta)^2}$$

  Note that this is Gaussian-shaped as a function of either x or θ.

- Using the above formulae, or re-deriving:

$$\theta^{\text{MAP}} = \frac{x + \sigma_e^2 \mu_\theta \sigma_\theta^2}{1 + \sigma_e^2/\sigma_\theta^2} = \frac{\sigma_\theta^2 x + \sigma_e^2 \mu_\theta}{\sigma_\theta^2 + \sigma_e^2} = \alpha x + (1-\alpha)\mu_\theta$$

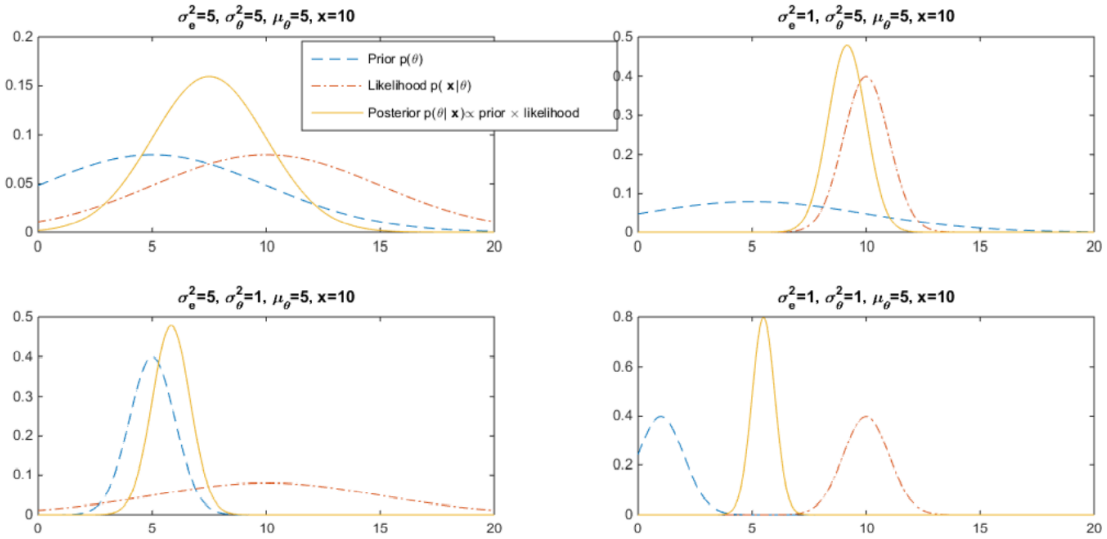
  with $\alpha = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2}$, and

$$\text{var}(\theta|x) = \frac{\sigma_e^2}{1 + \sigma_e^2/\sigma_\theta^2} = \frac{\sigma_e^2 \sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2}$$

  So,

$$p(\theta|x) = \mathcal{N}\left(\frac{\sigma_\theta^2 x + \sigma_e^2 \mu_\theta}{\sigma_\theta^2 + \sigma_e^2}, \frac{\sigma_e^2 \sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2}\right)$$

- See plots below for different prior-likelihood trade-offs:



10.8.5 MMSE Estimation

- Apart from intuition, we have not justified why the MAP estimator is a good thing to do.
- In fact, for the linear model we know that that OLS estimator has the best performance of any unbiased estimation scheme. (BLUE)
- Could it be, however, that the Bayesian estimator is actually introducing some bias in order to get better performance in some other sense?
- Now consider more carefully the notion of an expected cost function . $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ expresses the cost of estimating the parameter as $\hat{\boldsymbol{\theta}}$ when the true value is $\boldsymbol{\theta}$.
- A suitable cost function is non-negative and usually satisfies $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$.
- We can write the expected cost over all of the unknown paramters, conditional upon the observed data $\mathbf{x}$:

$$\mathbb{E}[C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})] = \int_{\boldsymbol{\theta}} C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

- The form of cost function will depend on the requirements of a particular problem.
- A cost of 0 indicates that the estimate is perfect for our requirements, while positive costs indicate poorer estimates.
- As usual, because of the random properties of the model, we can only estimate the expected cost of a particular estimator.
- A classic estimation technique related to the Wiener filtering objective function is the Minimum mean-squared error (MMSE) estimation method. Given some data $\mathbf{x}$ we attempt tot find an estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ which has minimum squared error, on average:

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2]$$

Any small bias introduced can be tolerated provided the MSE is low

- To derive the MMSE estimator we need the conditional distribution of θ given $\mathbf{x}$, $p(\theta|\mathbf{x})$, which will be obtained using Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}$$

- The MSE can now be expressed as:

$$J = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] = \int_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

- Differentiate w.r.t $\boldsymbol{\theta}$ gives:

$$\begin{aligned}\frac{dJ}{d\boldsymbol{\theta}} &= \frac{d}{d\boldsymbol{\theta}} \int_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \frac{d}{d\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}\end{aligned}$$

  and setting to zero:

$$\int_{\boldsymbol{\theta}} 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 0$$

  and

$$\begin{aligned}\hat{\boldsymbol{\theta}} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} &= \int_{\boldsymbol{\theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ \hat{\boldsymbol{\theta}} \times 1 &= \mathbb{E}[\boldsymbol{\theta}|\mathbf{x}]\end{aligned}$$

  - The LHS is simplified since we have:

$$\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 1$$

- Hence we have the result for the MMSE estimator as:

$$\hat{\boldsymbol{\theta}}^{\text{MMSE}} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{x}] = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

- The posterior distribution was obtained as:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\theta}^{\text{MAP}}, \sigma_e^2 \boldsymbol{\Phi}^{-1})$$

  The mean value of this distribution is $\boldsymbol{\theta}^{\text{MAP}}$

- Hence the MMSE estimator for the Linear Gaussian model is,

$$\boldsymbol{\theta}^{\text{MMSE}} = \boldsymbol{\theta}^{\text{MAP}}$$

  In other cases, the estimators do not necessarily coincide.

11 Summary of Estimators

We have examined three important estimation methods, each of increasing sophistication. It was seen that the Ordinary Least Squares (OLS) estimator was a special case of the Maximum Likelihood (ML) estimator when the noise was Gaussian with zero-mean and fixed variance. In turn the ML estimator was a special case of the Maximum a posteriori (MAP) estimator when the prior distribution on θ was uniform. However, ML requires specific knowledge of the likelihood function and MAP estimation requires in addition knowledge of a prior density $p(\theta)$. The choice of estimator will thus depend on the degree of knowledge available and the performance required.

The performance of the threee estimators:

- Least Squares
  - Requires no knowledge of probability distributions
  - Cannot incorporate prior knowledge about parameter probability distributions
  - Usually the simplest scheme to implement
  - Guarantee of performance as BLUE estimator
  - No guarantees of performance compared to nonlinear estimators
- Maximum Likelihood (ML)
  - Requires knowledge of noise probability distribution
  - Cannot incorporate prior knowledge about parameter probability distributions
  - Can be more complicated to implement than LS in the non-Gaussian case
  - Performance guaranteed to be optimal when the amount of data is large.
- Bayesian (MAP and MMSE)
  - Requires knowledge of noise (model) probability distributions
  - Requires knowledge of parameter prior probability distribution (might require subjective input)
  - Incorporates prior knowledge about parameter probability distribution
  - Can be more complicated to implement than LS or ML, depending on form of likelihood and prior
  - Performance guaranteed to be optimal for any amount of data (provided prior distribution is correct)