Probability and random variables

Probability

Sample space, probability and probability space.

- A random expt is any situation which has a set of possible outcomes, each of which occurs w/ a particular probability
- To mathematically describe a random expt, we must specify:
  - ↳ 1) the sample space Ω : the set of all possible outcomes of the random expt. We can call any subset $A \subseteq \Omega$ an event
  - ↳ 2) the probability P : A mapping/function from events to a no. in the interval $[0,1]$
- We call (Ω, P) the probability space.

Axioms of probability

- Probability theory is based on the Kolmogorov axioms:
  - ↳ Axiom I: The probability of an event is a non-negative real no.

$$P(A) \in \mathbb{R}, \quad P(A) \geq 0 \quad \forall A \subseteq \Omega$$

  - ↳ Axiom II: The sample space / certain event has unit probability

$$P(\Omega) = 1.$$

  - ↳ Axiom III: Additivity for incompatible events (disjoint sets)

$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \cap B = \phi.$$

- These three axioms are sufficient to prove the following:
  - ↳ Monotonicity: if $A \subseteq B$, then $P(A) \leq P(B)$
  - ↳ Probability of the empty set: $P(\phi) = 0$
  - ↳ Complement rule: $P(A') = 1 - P(A)$
  - ↳ Numeric bound: $0 \leq P(A) \leq 1 \quad \forall A \subseteq \Omega$
  - ↳ Addition law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - ↳ Sum rule: $P(A) = P(A \cap B) + P(A \cap B')$
  - ↳ Additivity extended: $P(A) = \sum_{w \in A} P(w)$ where $\{w \in \Omega\}$ are mutually-disjoint individual outcomes

Defining probability

- Define the indicator function for a set or event E

$$I_E(t) = \begin{cases} 0 & t \in E \\ 1 & t \notin E \end{cases}$$

- For a finite discrete set $\Omega = \{\omega_1, \omega_2 \dots \omega_n\}$, let $p_1, p_2 \dots p_n$ be non-negative no. that sum to 1. For any event E, set

$$P(E) = \sum_{i=1}^{n} I_E(\omega_i) p_i$$

then P satisfies the axioms.

- For an infinite discrete set $\Omega = \{\omega_1, \omega_2 \dots\}$, let $p_1, p_2 \dots$ be a non-negative sequence that sum to 1, For any event E, set

$$P(E) = \sum_{i=1}^{\infty} I_E(\omega_i) p_i$$

then P satisfies the axioms

- For the set of real no $\Omega = \mathbb{R}$, we define a pdf $f(t)$ that satisfies

$$f(t) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} f(t)\,dt = 1$$

For any event E, set

$$P(E) = \int_{-\infty}^{\infty} I_E(t) f(t)\,dt$$

then P satisfies the axioms.

Conditional probability

- The conditional probability of event A occuring given that event B has occured is defined to be.

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

- From this, we can get the multiplication rule

$$P(A \cap B) = P(A|B) \cdot P(B)$$

In general, we can define the probability chain rule

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

- Applying the sum rule, we get the law of total probability

$$P(A) = P(A \cap B) + P(A \cap B')$$

$$= P(A|B)P(B) + P(A|B')P(B')$$

In general, for a set of pairwise incompatible events B_i, s.t. $\bigcup_i B_i = \Omega$.

$$P(A) = \sum_i P(A|B_i) P(B_i)$$

Bayes' theorem

- We can write $P(A \cap B)$ in two ways using the multiplication rule.

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Rearranging, we get Bayes's thm. (relates $P(A|B)$ w/ $P(B|A)$)

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

- Bayes thm. is often used w/ the law of total probability.

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

Independence

- Two events are independent if the knowledge that one occured does not change the probability that the other one occurs.
- Mathematically, A and B are independent if.

$$P(A) = P(A|B) = P(A|B').$$

Using the multiplication rule, we get the formal defn of independence

$$P(A \cap B) = P(A)P(B).$$

Random variables

Random variables (RVs)

- Given a probability space (Ω, P), a RV is a function $X(\omega)$ which maps each element ω of the sample space Ω onto a pt. on the real line.
- We can define a RV by listing the range of $X(\omega)$ along w/ the probability the RV takes those values.
- For any set $A \subset (-\infty, \infty)$, we define

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

- For a discrete RV X, we define the probability mass function (pmf) p_X to be.

$$p_X(x_i) = P(X = x_i), \quad \text{where } \sum_i p_X(x_i) = 1.$$

For any set A, $P(X \in A) = \sum_i I_A(x_i) p_X(x_i)$

- For a continuous RV X, we define the probability density function (pdf) f_X to be

$$f_X(x) = P(X | \in \varepsilon), \quad \text{where } \int_{-\infty}^{\infty} f_X(x)dx = 1.$$

For any set A, $P(X \in A) = \int_{-\infty}^{\infty} I_A(x) f_X(x) dx$.

cumulative distribution function (cdf)

- For discrete/continuous RV, the cdf F_X is defined as (note $P(X>x) = 1 - F_X(x)$)

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

- the cdf $F_X(x)$ has the following properties:
  - ↳ $0 \leq F_X(x) \leq 1$
  - ↳ $P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$
  - ↳ $F_X(x)$ is non-decreasing as x increases
  - ↳ $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
- For continuous RV X, $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$ → area under f_X is a continuous function → F_X is continuous.
- For continuous RV X, the cdf F_X and pdf f_X are related by.

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx \xrightarrow{FTC} f_X(x) = \frac{dF_X(x)}{dx}$$ ← useful for finding pdf of transformed variables!

- For discrete RV X w/ range $\{x_1, \ldots x_n\}$, $F_X(x) = \sum_{i=1}^{n} p(x_i) I_{[x_i, \infty)}(x)$ → $F_X(x)$ is a step function that jumps w/ each x_i → F_X is right-continuous, i.e. $F_X(x) = \lim_{x \to x_+} F_X(x)$ $\forall x$.

Expectation

- The expected value of X, $E[X]$ is defined as.

$$E[X] = \begin{cases} \sum_{x \in \mathbb{R}} x P_X(x) & \text{[discrete]} \\ \int_{-\infty}^{\infty} x f_X(x)\,dx & \text{[continuous]} \end{cases}.$$

- The law of large nos (LLN) states that.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X] \quad \text{as} \quad n \to \infty$$

so we can think of $E[X]$ as the empirical average of independent samples of X, X_i.

- For any function $g(x)$, its expectation $E[g(X)]$ is computed using.

$$E[g(X)] = \begin{cases} \sum_{x} g(x) P_X(x) & \text{[discrete]} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\,dx & \text{[continuous]} \end{cases}.$$

- For $g(x) = I_A(x)$, its expectation is the probability of event A.

$$E[I_A(x)] = \begin{cases} \sum_{x} I_A(x) P_X(x) \\ \int_{-\infty}^{\infty} I_A(x) f_X(x)\,dx \end{cases} = P(X \in A)$$

The LLN formalises the freq. interpretation of probability.

$$\frac{1}{n}\sum_{i=1}^{n} I_A(x) \to E[I_A(x)] = P(A(x)) \quad \text{as} \quad n \to \infty.$$

Transformation of random variable.

- Consider a continuous RV X w/ pdf f_X and cdf F_X. Define a new RV $Y = g(X)$. For strictly increasing/decreasing g, there exist an inverse g^{-1}, and the pdf of the transformed RV, f_Y, is given by.

$$\boxed{f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d(g^{-1}(y))}{dy} \right|}.$$

↳ increasing g [$g(x) \leq y$ ⇔ $x \leq g^{-1}(y)$]: $F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$

$$\therefore f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy} F_X(g^{-1}(y)) = \frac{dF_X(g^{-1}(y))}{dx}\frac{dx}{dy} = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

↳ decreasing g [$g(x) \leq y$ ⇔ $x \geq g^{-1}(y)$]: $F_Y(y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$

$$\therefore f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left(1 - F_X(g^{-1}(y))\right) = -\frac{dF_X(g^{-1}(y))}{dx}\frac{dx}{dy} = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

* In general, express $F_Y(y)$ in terms of F_X, then differentiate wrt y to get $f_Y(y)$.

Bivariate random variables.

Bivariate discrete random variables.

- consider two discrete RVs X, Y where $X \in \{x_1 \dots x_m\}$ and $Y \in \{y_1 \dots y_n\}$. Define the joint pmf P_{XY}

$$P_{XY}(x_i, y_j) = P(X = x_i, Y = y_j)$$

- We can derive the marginal pmfs P_X, P_Y given the joint pmf P_{XY}.

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

- Two discrete RVs X and Y are independent if

$$P_{XY}(x,y) = P_X(x) P_Y(y) \quad \forall (x,y)$$

- For the discrete RVs X, Y, the conditional pmf of X given $Y = y$, $P_{X|Y}$ is defined as

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)}$$

Bivariate continuous random variables.

- consider two continuous RVs X, Y, we define the joint cdf f_{XY} to be a non-negative function st.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_A(x) I_B(y) f_{XY}(x,y)\,dx\,dy = P(X \in A, Y \in B), \text{ where } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1.$$

- We can derive the marginal pdfs f_X, f_Y given the joint pdf f_{XY}.

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$$

- Two continuous RVs X and Y are independent iff

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$

- For the continuous RVs X, Y, the conditional pdf of x given $Y = y$, $f_{X|Y}$ is defined as

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Sum of random variables.

- Let X, Y be two independent RVs and let $Z = X + Y$.

For discrete X, Y, $P_Z(z) = \sum_{x \in \mathcal{X}} P_X(x) P_Y(z-x) = P_X * P_Y$

↳ $P_Z(z) = \sum_{x \in \mathcal{X}} P_{XZ}(x,z) = \sum_{x \in \mathcal{X}} P_{Z|X}(z|x) P_X(x) = \sum_{x \in \mathcal{X}} P(X+Y=z \mid X=x) P_X(x) = \sum_{x \in \mathcal{X}} P(Y=z-x \mid X=x) P_X(x)$

$= \sum_{x \in \mathcal{X}} P_{Y|X}(z-x) P_X(x) = \sum_{x \in \mathcal{X}} P_Y(z-x) P_X(x) = P_X * P_Y.$

For continuous X, Y, $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx = f_X * f_Y$

↳ $F_Z(z) = P(Z \le z) = P(X+Y \le z) = \iint_{x+y \le z} f_{XY}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{XY}(x,y)\,dx\,dy$

Sub $u = x+y$: $F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_{XY}(u-y, y)\,du\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_X(u-y) f_Y(y)\,du\,dy$

Diff. wrt z: $f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy = f_X * f_Y.$

Expectation of bivariates

- The expectation of a function $g(X,Y)$ of the bivariate (X,Y) is given by.

$$E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y) p_{X,Y}(x,y) & \text{[discrete]} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy & \text{[continuous]} \end{cases}$$

The conditional expectation of $g(X,Y)$ given $Y=y$ is then

$$E[g(X,Y) \mid Y=y] = \begin{cases} \sum_x g(x,y) p_{X|Y}(x|y) & \text{[discrete]} \\ \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y)\,dx & \text{[continuous]} \end{cases}$$

- We can compute $E[g(X,Y)]$ by considering $E[g(X,Y)|Y=y]$. For the continuous case,

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y)\,dx\right) f_Y(y)\,dy$$

$$E[g(X,Y)] = \int_{-\infty}^{\infty} E[g(X,Y)\mid Y=y] f_Y(y)\,dy = E_Y\left[E_X[g(X,Y)|Y]\right]$$

Note fact for a function $h(y)$, $E_Y[h(Y)] = \int_{-\infty}^{\infty} h(Y=y) f_Y(y)\,dy$

* The above result also holds for the discrete case

Law of iterated expectation (Adam's law)

- The law of iterated expectation states that

$$E[X] = E_Y[E_X[X|Y]]$$

- Consider the continuous case (result also holds for discrete case)

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right)dx = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx\right) f_Y(y)\,dy = \int_{-\infty}^{\infty} E[X|Y=y] f_Y(y)\,dy = E_Y[E_X[X|Y]]$$

Multivariate random variables

Random vectors

- Let $X_1, X_2 \ldots X_n$ be n continuous RVs. We call $X = (X_1 \ldots X_n) \in \mathbb{R}^n$ a continuous random vector

(similarly we can define a discrete random vector, but we will consider the continuous case in this section)

- For a continuous random vector $X = (X_1 \ldots X_n) \in \mathbb{R}^n$, we define the pdf f to be a non-negative function that integrates to unity which satisfies the following for all events $A_1 \ldots A_n$.

$$P(X_1 \in A_1, \ldots X_n \in A_n) = \int_{-\infty}^{\infty} I_{A_1}(x_n) \cdots \int_{-\infty}^{\infty} I_{A_1}(x_1) f(x_1 \ldots x_n)\,dx_1 \ldots dx_n$$

If event $A_i = [a_i, b_i]$, then we can replace the range w/.

$$P(X_1 \in A_1, \ldots X_n \in A_n) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(x_1 \ldots x_n)\,dx_1 \ldots dx_n$$

- The marginal pdf of X_i can be obtained by integrating $f(x_1 \ldots x_n)$ over the full range of all variables except X_i

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1 \ldots x_n)\,dx_1 \ldots dx_{i-1}\,dx_{i+1} \ldots dx_n$$

- The n RVs $X_1 \ldots X_n$ are independent iff for every $A_1, \ldots A_n$

$$P(X_1 \in A_1 \ldots X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n)$$

This is equivalent to checking that the joint pdf reduces to the product of marginals.

$$f(x_1 \ldots x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$$

Transformation of random vectors

- Consider the transformation $\underline{Y} = G(\underline{X})$, i.e.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g_1(x_1 \dots x_n) \\ \vdots \\ g_n(x_1 \dots x_n) \end{bmatrix}$$

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

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} h_1(x_1 \dots x_n) \\ \vdots \\ h_n(x_1 \dots x_n) \end{bmatrix}$$

Performing a change of variable during integration, we get

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

where the Jacobian $|\det J(y)|$ is the magnitude of the determinant of the matrix of partial derivatives of $H(y)$

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

Characteristic function.

- The characteristic function of a RV X is defined as

$$\phi_X(t) = E[\exp(itX)], \quad t \in \mathbb{R}.$$

The characteristic function of a random vector $\underline{X} = (X_1 \dots X_n)$ is defined as

$$\phi_{\underline{X}}(\underline{t}) = E[\exp(i \underline{t}^T \underline{X})], \quad \underline{t} \in \mathbb{R}^n$$

- Suppose those X and Y are RVs/random vectors w/ $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$ / $\underline{t} \in \mathbb{R}^n$, then X and Y have the same probability distribution. (since the characteristic function uniquely describes a pdf).

- Consider the RV $Y = \sum_{i=1}^n X_i$, where X_i are independent RVs. Its characteristic function $\phi_Y(t)$ is

$$\phi_Y(t) = E[\exp(itY)] = E[\exp(itX_1) \dots \exp(itX_n)] = E[\exp(itX_1)] \dots E[\exp(itX_n)] = \phi_{X_1}(t) \dots \phi_{X_n}(t)$$

i.e. the characteristic function of the sum of independent RVs is the product of the individual characteristic functions.

- We can use the characteristic function to compute moments, $E[X^n]$. Consider $\frac{d^n}{dt^n} \phi_X(t)$.

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

If we set $t = 0$, we get $\frac{d^n}{dt^n} \phi_X(t)\big|_{t=0} = i^n E[X^n]$. $\rightarrow$ $E[X^n] = \frac{1}{i^n} \frac{d^n}{dt^n} \phi_X(t)\big|_{t=0}$

The Gaussian vector.

- For a Gaussian vector $\underline{X} = (X_1 \dots X_n) \in \mathbb{R}^n$, the pdf $f(x_1 \dots x_n)$ is given by

$$f(x_1 \dots x_n) = \frac{1}{(2\pi)^{n/2} |\underline{\underline{\Sigma}}|^{1/2}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{m})^T \underline{\underline{\Sigma}}^{-1} (\underline{x} - \underline{m})\right)$$

where $\underline{m}$ is the mean vector and $\underline{\underline{\Sigma}}$ is the covariance matrix.

$$m_i = E[X_i] \qquad [\underline{\underline{\Sigma}}]_{ij} = E[(X_i - m_i)(Y_j - m_j)]$$

* For independent X_i, the covariance matrix $\underline{\underline{\Sigma}}$ is diagonal, i.e. $[\underline{\underline{\Sigma}}]_{ij} = 0$ for $i \neq j$.

Transformation of the Gaussian vector

- consider a gaussian vector $X=(X_1 \dots X_n)$, where each X_i is independent and $X_i \sim N(0,1)$.

Let $\underline{\underline{S}}$ be an invertible matrix and $\underline{m}$ be a column vector, and define $\underline{Y} = \underline{m} + \underline{\underline{S}}\underline{X}$.

- using the change of variable result, where $\underline{Y} = g(\underline{X}) = \underline{m} + \underline{\underline{S}}\underline{X} \rightarrow H(\underline{Y}) = g^{-1}(\underline{Y}) = \underline{\underline{S}}^{-1}(\underline{Y} - \underline{m}) \rightarrow J(\underline{Y}) = \underline{\underline{S}}^{-1}$.

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{\underline{S}}^{-1}(\underline{y} - \underline{m})) \left| \det \underline{\underline{S}}^{-1} \right|$$

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

- we can therefore rewrite $f_{\underline{Y}}(\underline{y})$ as:

$$\boxed{f_{\underline{Y}}(\underline{y}) = \frac{|\det \underline{\underline{S}}^{-1}|}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\underline{y} - \underline{m})^T (\underline{\underline{S}}^{-1})^T \underline{\underline{S}}^{-1} (\underline{y} - \underline{m})\right]}$$

i.e. $\underline{Y}$ is a Gaussian vector w/ mean $\underline{m}$ and covariance matrix $\underline{\underline{S}}\underline{\underline{S}}^T$

(Note that $|\det \underline{\underline{S}}^{-1}| = \frac{1}{|\det \underline{\underline{S}}|}$ and $\frac{1}{|\det \underline{\underline{S}}\underline{\underline{S}}^T|^{1/2}} = \frac{1}{|\det \underline{\underline{S}} \det \underline{\underline{S}}^T|^{1/2}} = \frac{1}{|\det \underline{\underline{S}}|}$).

- In general, an affine transformation (shift + linear stretch) of a gaussian vector is a Gaussian vector.

We can use this result to generate any Gaussian vector $N(\underline{m}, \underline{\underline{\Sigma}})$.

↳ Decompose the symmetric matrix $\underline{\underline{\Sigma}} = \underline{\underline{S}}\underline{\underline{S}}^T$ (use Cholesky decomposition)

↳ Output $\underline{m} + \underline{\underline{S}}\underline{X}$, where $\underline{X} = (X_1 \dots X_n)$ and each X_i is independent w/ $X_i \sim N(0,1)$.

Characteristic function of a Gaussian

- consider a Gaussian RV X, where $X \sim N(M, \sigma^2)$. Its characteristic function $\phi_X(t)$ is

$$\boxed{E[\exp(itX)] = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-M)^2\right) dx \overset{s=x-M}{=} e^{itM} \int_{-\infty}^{\infty} e^{its} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{s^2}{2\sigma^2}\right) ds = e^{itM} e^{-\frac{1}{2}\sigma^2 t^2}}$$

since $FT\left[e^{-\frac{1}{2}\sigma^2 t^2}\right] = \frac{\sqrt{2\pi}}{\sigma\sqrt{2}} e^{-s^2/4(\sigma^2/2)} = \frac{\sqrt{2\pi}}{\sigma} e^{-\frac{s^2}{2\sigma^2}} \rightarrow FT\left[\frac{\sqrt{2\pi}}{\sigma} e^{\frac{-\sigma^2}{2}}\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\sqrt{2\pi}}{\sigma} e^{-\frac{s^2}{2\sigma^2}} e^{its} ds = e^{-\frac{1}{2}\sigma^2 t^2}$

- consider the transformation of a gaussian vector $\underline{X} = (X_1 \dots X_n)$, X_i iid $\sim N(0,1)$ and $\underline{Y} = \underline{m} + \underline{\underline{S}}\underline{X}$.

Noting that $i\underline{t}^T\underline{Y} = i\underline{t}^T\underline{m} + i\underline{t}^T\underline{\underline{S}}\underline{X} = i\underline{t}^T\underline{m} + i\underline{r}^T\underline{X}$, where we define $\underline{r} = \underline{\underline{S}}^T\underline{t}$

$$\phi_{\underline{Y}}(\underline{t}) = E[\exp(i\underline{t}^T\underline{Y})] = E[\exp(i\underline{t}^T\underline{X})\exp(i\underline{r}^T\underline{X})] = E[\exp(i\underline{t}^T\underline{X})\exp(ir_1X_1)\dots\exp(ir_nX_n)]$$

$$= \exp(i\underline{t}^T\underline{m})\, E[\exp(ir_1X_1)]\dots E[\exp(ir_nX_n)] = \exp(i\underline{t}^T\underline{m})\exp(-\tfrac{1}{2}r_1^2)\dots\exp(-\tfrac{1}{2}r_n^2)$$

$$= \exp(i\underline{t}^T\underline{m})\exp\left(-\frac{1}{2}\sum_{i=1}^{n} r_i^2\right) = \exp(i\underline{t}^T\underline{m})\exp\left(-\frac{1}{2}|\underline{t}^T\underline{\underline{S}}|^2\right) = \exp(i\underline{t}^T\underline{m})\exp\left(-\frac{1}{2}\underline{t}^T\underline{\underline{S}}\underline{\underline{S}}^T\underline{t}\right)$$

i.e. the characteristic function of a multivariate Gaussian w/ mean $\underline{m}$ and covariance $\underline{\underline{\Sigma}} = \underline{\underline{S}}\underline{\underline{S}}^T$.

Random process

Random process

Random process

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

$$\{X_n(\omega)\}, \quad n \in (-\infty, \infty)$$

where ω is a RV w/ pdf $f_\omega(\omega)$.

- We can consider a generative model for the waveforms. We first draw a random value $\tilde{\omega}$ from the density $f_\omega(\omega)$. Then, the waveform for this value $\omega = \tilde{\omega}$ is given by $x_n(\tilde{\omega})$, $n \in (-\infty, \infty)$
- The ensemble is built up by considering all possible values $\tilde{\omega}$ and their corresponding waveforms $x_n(\tilde{\omega})$. The rel. freq. of each waveform $x_n(\omega)$ is determined by $f_\omega(\omega)$.
- Usually, for simplicity, we can consider the random process as an infinite collection of RVs, e.g.,

$$\{X_n\}_{n=-\infty}^{\infty} = \{\dots, X_{-1}, X_0, X_1, \dots\} \quad \text{or} \quad \{X_n\}_{n=0}^{\infty} = \{X_0, X_1, \dots\}$$

- To completely specify the random process, we must specify
  - For continuous RVs $X_0, X_1, \dots$, its nth order joint pdf

$$f_{X_0 X_1 \dots X_n}(x_0, x_1, \dots x_n) \quad \forall n \in \mathbb{Z}^+.$$

  - For discrete RVs $X_0, X_1, \dots$, its nth order joint pmf.

$$p_{X_0 X_1 \dots X_n}(x_0, x_1, \dots x_n) \quad \forall n \in \mathbb{Z}^+$$

correlation and covariance functions.

- The autocorrelation function of a random process $\{X_n\}$ is

$$r_{XX}[n, m] = E[X_n X_m]$$

The autocovariance function of a random process $\{X_n\}$ is

$$c_{XX}[n, m] = E[(X_n - M_n)(X_m - M_m)]$$

- The crosscorrelation function btwn random processes $\{X_n\}$ and $\{Y_n\}$ is

$$r_{XY}[n, m] = E[X_n Y_m]$$

The crosscovariance function btwn random processes $\{X_n\}$ and $\{Y_n\}$ is

$$c_{XY}[n, m] = E[(X_n - M_{X,n})(Y_m - M_{Y,m})]$$

Strict-sense stationarity (SSS)

- A discrete-time random process $\{X_n\}$ is SSS if, for any finite k, m and $(\alpha_0, \alpha_1 \dots \alpha_k)$,

$$f_{X_0 X_1 \dots X_k}(\alpha_0, \alpha_1 \dots \alpha_k) = f_{X_m X_{m+1} \dots X_{m+k}}(\alpha_0, \alpha_1 \dots \alpha_k)$$

i.e. it has the same statistical characteristics irrespective of shifts along the time axis.

* If $X_0, X_1 \dots$ are discrete, we would use the joint pmf instead of the joint pdf to define SSS.

Wide sense stationarity (WSS)

- A discrete-time random process $\{X_n\}$ is WSS if all the following are satisfied

  (i) $E[X_n] = \mu \ \forall n$ (const. mean)

  (ii) $\text{Var}[X_n] < \infty$ / $E[X_n^2] < \infty \ \forall n$ (finite variance/power)

  (iii) $r_{XX}[n,m] \to r_{XX}[m-n]$ $\quad (r_{XX} = r_{XX}(m-n))$

- Two discrete-time random processes $\{X_n\}, \{Y_n\}$ are jointly WSS if the following are satisfied:

  (i) $E[X_n] = \mu_X$, $E[Y_n] = \mu_Y \ \forall n$ (const. mean)

  (ii) $\text{Var}[X_n] < \infty$ / $E[X_n^2] < \infty$, $\text{Var}[Y_n] < \infty$ / $E[Y_n^2] < \infty \ \forall n$ (finite variance/power)

  (iii) $r_{XX}[n,m] \to r_{XX}[m-n]$, $r_{YY}[n,m] \to r_{YY}[m-n]$, $r_{XY}[n,m] \to r_{XY}[m-n]$ $\quad (r = r(m-n))$

- For WSS $\{X_n\}$, $\boxed{r_{XX}[k] = r_{XX}[-k]}$, i.e. $r_{XX}[k]$ is even

$$r_{XX}[k] = E[X_n X_{n+k}] = E[X_{n+k} X_n] = r_{XX}[-k].$$

and $\boxed{\max |r_{XX}[k]| = r_{XX}[0]}$.

$$0 \le E[(X_{n+k} - aX_n)^2] = E[X_{n+k}^2] - 2a E[X_{n+k} X_n] + a^2 E[X_n^2] = r_{XX}[0](1+a^2) - r_{XX}[k](2a)$$

$$\therefore r_{XX}[0](1+a^2) \ge 2a\, r_{XX}[k]$$

setting $a=1 \to r_{XX}[0] \ge r_{XX}[k]$; setting $a=-1 \to r_{XX}[0] \ge -r_{XX}[k] \Rightarrow r_{XX}[0] \ge |r_{XX}[k]|$

- For jointly WSS $\{X_n\}, \{Y_n\}$, $\boxed{r_{XY}[k] = r_{YX}[-k]}$

$$r_{XY}[k] = E[X_n Y_{n+k}] = E[Y_{n+k} X_n] = r_{YX}[-k].$$

Power spectral density.

- For WSS $\{X_n\}$, the PSD function is the DTFT of the autocorrelation function $r_{XX}[k]$.

$$\boxed{S_X(f) = \sum_{k=-\infty}^{\infty} r_{XX}[k] e^{-j2\pi fk}} \qquad \boxed{r_{XX}[k] = \int_{-1/2}^{1/2} S_X(f) e^{j2\pi fk}}$$

- For jointly WSS $\{X_n\}, \{Y_n\}$, the cross PSD function is the DTFT of the cross-correlation function $r_{XY}[k]$

$$\boxed{S_{XY}(f) = \sum_{k=-\infty}^{\infty} r_{XY}[k] e^{-j2\pi fk}} \qquad \boxed{r_{XY}[k] = \int_{-1/2}^{1/2} S_{XY}(f) e^{j2\pi fk}}$$

- The PSD $S_X(f)$ is periodic w/ period 1

$$S_X(f+1) = \sum_{k=-\infty}^{\infty} r_{XX}[k] e^{-j2\pi(f+1)k} = \sum_{k=-\infty}^{\infty} r_{XX}[k] e^{-j2\pi fk} \underbrace{e^{-j2\pi k}}_{=1} = S_X(f)$$

- The PSD $S_X(f)$ is real-valued and even. It is also continuous if $\sum_{k=-\infty}^{\infty} |r_{XX}[k]| < \infty$.

$$S_X(f) = \sum_{k=-\infty}^{\infty} r_{XX}[k] e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} r_{XX}[k] \cos(2\pi fk) - j \underbrace{\sum_{k=-\infty}^{\infty} r_{XX}[k] \sin(2\pi fk)}_{\to 0} = \sum_{k=-\infty}^{\infty} \underbrace{r_{XX}[k]}_{\text{even}} \underbrace{\cos(2\pi fk)}_{\text{even}}$$

$\therefore S_X(f)$ is the sum of real-valued, even and continuous functions $\to S_X(f)$ is real, even and continuous.

- The PSD $S_X(f)$ can be regarded as the density of power — integrating gives power, so it must be non-negative.

- The cross PSD $S_{XY}(f)$ has the symmetry $\boxed{S_{XY}(f) = S_{YX}^*(f)}$

$$S_{XY}(f) = \sum_{k=-\infty}^{\infty} r_{XY}[k] e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} r_{YX}[-k] e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} r_{YX}[k] e^{+j2\pi fk} = \left[\sum_{k=-\infty}^{\infty} r_{YX}[k] e^{-j2\pi fk}\right]^* = S_{YX}^*(f)$$

- The cross PSD $S_{XY}(f)$ can be regarded as a measure of the coherence btwn the two processes for a certain frequency.

Linear systems and random processes.

- If the i/p $\{X_n\}$ of a LTI system w/ impulse resp $\{h_n\}$ is WSS, and $\sum_{n=-\infty}^{\infty}|h_n|<\infty$, then the o/p $\{Y_n\}$ is also WSS.

(i) Consider the mean $E[Y_n]$.

$$E[Y_n] = E\left[\sum_{k=-\infty}^{\infty} h_k X_{n-k}\right] = \sum_{k=-\infty}^{\infty} h_k E[X_{n-k}] = E[X_0]\sum_{k=-\infty}^{\infty} h_k \rightarrow \text{const. over all } n.$$

(ii) Consider the autocorrelation $r_{YY}[k]$.

$$r_{YY}[k] = E\left[\left(\sum_{i=-\infty}^{\infty} h_i X_{n-i}\right)\left(\sum_{j=-\infty}^{\infty} h_j X_{n+k-j}\right)\right] = E\left[\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j X_{n-i}X_{n+k-j}\right]$$

$$= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j E[X_{n-i}X_{n+k-j}] = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j r_{XX}[k+i-j] = \sum_{i=-\infty}^{\infty} h_i\left(\sum_{j=-\infty}^{\infty} h_j r_{XX}[(k+i)-j]\right)$$

$$= \sum_{i=-\infty}^{\infty} h_i\left(h_{k+i} * r_{XX}[k+i]\right) = \tilde{h}_k * h_k * r_{XX}[k] \rightarrow \text{only a function of } k.$$

where $\tilde{h}_k$ is the time-reversed sequence of h_k, i.e. $\tilde{h}_k = h_{-k}$.

(Note $a_n * b_n = \sum_{k=-\infty}^{\infty} a_k b_{n-k}$, $\tilde{a}_n * b_n = \sum_{k=-\infty}^{\infty} \tilde{a}_k b_{n-k} = \sum_{k=-\infty}^{\infty} a_{-k} b_{n-k} = \sum_{k=-\infty}^{\infty} a_k b_{n+k}$).

(iii) Consider the power $E[Y_n^2] = r_{YY}[0]$.

$$r_{YY}[0] = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j r_{XX}[i-j] \le \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j r_{XX}[0] = r_{XX}[0]\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j \le r_{XX}[0]\sum_{i=-\infty}^{\infty}|h_i|\sum_{j=-\infty}^{\infty}|h_j|$$

This is finite if $\sum_{i=-\infty}^{\infty}|h_i| < \infty$.

- The cross correlation $r_{XY}[k]$ and o/p autocorrelation $r_{YY}[k]$ can be expressed as convolutions.

$$\boxed{r_{XY}[k] = h_k * r_{XX}[k]} \qquad \boxed{r_{YY}[k] = \tilde{h}_k * h_k * r_{XX}[k]}$$

↳ $r_{XY}[k] = E[X_n Y_{n+k}] = E\left[X_n \sum_{i=-\infty}^{\infty} h_i X_{n+k-i}\right] = \sum_{i=-\infty}^{\infty} h_i E[X_n X_{n+k-i}] = \sum_{i=-\infty}^{\infty} h_i r_{XX}[k-i] = h_k * r_{XX}[k]$.

↳ Derivation for $r_{YY}[k] = \tilde{h}_k * h_k * r_{XX}[k]$ found above.

$r_{XX}[k] \rightarrow \boxed{h_k} \rightarrow r_{XY}[k] \rightarrow \boxed{h_{-k}} \rightarrow r_{YY}[k]$

- Taking the DTFT of the convolution form of the o/p autocorrelation $r_{YY}[k]$,

$$\text{DTFT}\left[r_{YY}[k] = \tilde{h}_k * h_k * r_{XX}[k]\right] \rightarrow \boxed{S_Y(f) = H^*(f)H(f)S_X(f) = |H(f)|^2 S_X(f)}$$

where $\text{DTFT}[\tilde{h}_k] = \sum_{k=-\infty}^{\infty}\tilde{h}_k e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} h_{-k} e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} h_k e^{j2\pi fk} = \left(\sum_{k=-\infty}^{\infty} h_k e^{-j2\pi fk}\right)^* = H^*(f)$.

Physical interpretation of PSD

- For a deterministic signal x_n, the instantaneous power is x_n^2 and the average power is $\lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N} x_n^2$.

- For a random process $\{X_n\}$, the expected instantaneous power is $E[X_n^2]$ and the expected average power is $E\left[\lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N} X_n^2\right] = \lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N} E[X_n^2]$.

- If $\{X_n\}$ is WSS, then $E[X_n^2] = r_{XX}[0]$ for all n, so the expected average power is $r_{XX}[0]$.

$$r_{XX}[0] = \int_{-1/2}^{1/2} S_X(f) e^{j2\pi f(0)}\,df = \int_{-1/2}^{1/2} S_X(f)\,df$$

i.e the area under $S_X(f)$ btwn $[-1/2, 1/2]$ gives the total power of $\{X_n\}$.

- To find the power of a WSS process $\{X_n\}$ in a particular freq. band, say $[f_1, f_2]$ s.t. $0 < f_1 < f_2 < 1/2$, consider passing $\{X_n\}$ through an ideal bandpass filter w/ TF $H(f) = \begin{cases} 1 & \text{if } f_1 \le |f| \le f_2 \\ 0 & \text{o/w} \end{cases}$, then find the o/p power $r_{YY}[0]$.

$$r_{YY}[0] = \int_{-1/2}^{1/2} S_Y(f)\,df = \int_{-1/2}^{1/2} |H(f)|^2 S_X(f)\,df = \int_{-f_2}^{-f_1} S_X(f)\,df + \int_{f_1}^{f_2} S_X(f)\,df = 2\int_{f_1}^{f_2} S_X(f)\,df.$$

→ The power of $\{X_n\}$ in a particular freq. band $f_1 \le f \le f_2$ is $\boxed{2\int_{f_1}^{f_2} S_X(f)\,df}$

Ergodic random process.

- For an ergodic random process, we can estimate expectations by performing time-averaging on a single sample function (rather than averaging over the whole ensemble).

(i) mean ergodic $\mu = E[X_n] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} X_n \approx \frac{1}{N} \sum_{n=0}^{N-1} X_n$ ← time average

(ii) correlation ergodic $R_{XX}[k] = E[X_n X_{n+k}] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} X_n X_{n+k} \approx \frac{1}{N} \sum_{n=0}^{N-1} X_n X_{n+k}$

ensemble average

- A necessary and sufficient condition for mean ergodicity is given by

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

A simpler sufficient condition for mean ergodicity is that

$C_{XX}[0] < \infty$ and $\lim_{N\to\infty} C_{XX}[N] = 0$

(N+S condition true → statement true, N+S condition false → statement false

S condition true → statement true, S condition false → statement indeterminate)

* Unless otherwise stated, we can always assume that all signals encountered are ergodic.

Markov chains (MC)

Markov chains. (MC)

limited memory → Markov property.

- A random process $\{X_n\}$ is a MC if for all times $n \geq 0$,

$$P_{X_n|X_0 \ldots X_{n-1}}(i_n|i_0 \ldots i_{n-1}) = P_{X_n|X_{n-1}}(i_n|i_{n-1})$$

where the RVs $X_0, X_1, \ldots$ can take values in the state space $S = \{1, \ldots, L\}$

- The MC is completely defined by the transitional probability matrix $\underline{\underline{Q}}$ and the initial distribution of the chain $\underline{\lambda}$.

↳ The transitional probability matrix $\underline{\underline{Q}}$ has non-negative entries and each row sums to 1.

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

where Q_{i_{n-1}, i_n} denotes the conditional pmf of X_n given $(X_0, \ldots, X_{n-1}) = (i_0, \ldots, i_{n-1})$, i.e.

$$Q_{i_{n-1}, i_n} = P_{X_n|X_0 \ldots X_{n-1}}(i_n|i_0 \ldots i_{n-1})$$

↳ The initial distribution $\underline{\lambda}$ has non-negative entries and they sum to 1.

$$\lambda_i = P_{X_0}(i) \qquad i = 1, \ldots, L$$

* For MC, we make the notation compact by dropping the subscripts, e.g. $P_{X_0 \ldots X_n}(i_0 \ldots i_n) \to P(i_0 \ldots i_n)$

Marginals of a Markov chain.

- The marginal of a Markov chain can be expressed as

$$\boxed{P_{X_n}(i_n) = (\underline{\lambda}\underline{\underline{Q}}^n)_{i_n}}$$

i.e. $P_{X_n}(i_n) = \underset{1\times L}{[\ ]}\underset{L\times L}{[\ ]}^n = \underset{1\times L}{[\ ]}$ ← pick i_nth entry

- Starting from the joint pmf $p(i_0,\dots i_n)$.

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

$$\boxed{p(i_0\dots i_n) = p(i_n|i_{n-1})\,p(i_{n-1}|i_{n-2})\cdots p(i_1|i_0)\,p(i_0) = Q_{i_{n-1},i_n}Q_{i_{n-2},i_{n-1}}\cdots Q_{i_0,i_1}\lambda_{i_0}}$$

Marginalising by summing over $i_0\dots i_{n-1}$, (dot product b/w $\underline{\lambda}$ and ith col. of $\underline{\underline{Q}}$)

$$p(i_n) = \sum_{i_0\dots i_{n-1}\in S} p(i_0\dots i_n) = \sum_{i_1\dots i_{n-1}\in S} Q_{i_{n-1},i_n}\cdots Q_{i_1,i_2}\sum_{i_0\in S} Q_{i_0,i_1}\lambda_{i_0} = \sum_{i_1\dots i_{n-1}\in S} Q_{i_{n-1},i_n}\cdots Q_{i_1,i_2}(\underline{\lambda}\underline{\underline{Q}})_{i_1}$$

$$= \sum_{i_2\dots i_{n-1}\in S} Q_{i_{n-1},i_n}\cdots Q_{i_2,i_3}\sum_{i_1\in S} Q_{i_1,i_2}(\underline{\lambda}\underline{\underline{Q}})_{i_1} = \sum_{i_2\dots i_{n-1}\in S} Q_{i_{n-1},i_n}\cdots Q_{i_2,i_3}(\underline{\lambda}\underline{\underline{Q}}^2)_{i_2} = (\underline{\lambda}\underline{\underline{Q}}^n)_{i_n}.$$

Invariant distribution of a Markov chain.

- consider a MC w/ transition probability matrix $\underline{\underline{Q}}$ w/ state space S. The pmf $\underline{\pi} = (\pi_i : i\in S)$ is invariant for $\underline{\underline{Q}}$ if $\forall j\in S$.

$$\pi_j = \sum_{i\in S}\pi_i Q_{i,j} = (\underline{\pi}\underline{\underline{Q}})_j \qquad \text{or} \qquad \boxed{\underline{\pi} = \underline{\pi}\underline{\underline{Q}}}.$$

- the MC $(\underline{\pi},\underline{\underline{Q}})$ is SSS.

↳ $P(i_n) = (\pi\underline{\underline{Q}}^n)_{i_n} = (\underline{\pi}\underline{\underline{Q}}\underline{\underline{Q}}^{n-1})_{i_n} = (\pi\underline{\underline{Q}}^{n-1})_{i_n} = \dots = \pi_{i_n}$

$p(i_m,\dots,i_{m+k}) = Q_{i_{m+k-1},i_{m+k}}\cdots Q_{i_m,i_{m+1}}\pi_{i_m}$ → we have same joint pdf for any m → SSS.

- the probability distribution $\underline{\pi}$ is known as a stationary distribution.

Ergodic theorem for Markov chain.

- An irreducible MC refers to a chain where all state values in S communicate w/ each other.
- This means for any pair of states (i,j), the MC starting at i will eventually visit j and vice versa, in a finite no. of steps.
- We can identify communicating states by drawing a graph from the transition probability matrix $\underline{\underline{Q}}$.
- When the MC $(\underline{\lambda},\underline{\underline{Q}})$ w/ stationary distribution $\underline{\pi}$ is irreducible, then for any initial distribution $\underline{\lambda}$, the sample average converges to $\sum_{i\in S}\pi_i r(i)$, i.e.

$$\boxed{\frac{1}{n+1}\sum_{t=0}^{n} r(X_t) \to \sum_{i\in S}\pi_i r(i)}$$

Time series analysis

Time series analysis.

- A time series is a set of observations $y_n, n=0,1,\ldots$ arranged in increasing time.
- Typically observations are recorded at regular intervals. For start time t_0 and time interval Δ, the nth observation is recorded at $t_n = t_0 + n\Delta$.
- When given a novel time-series data $y_0, y_1, \ldots$, the workflow for time series analysis is
  1) Select a probability model to represent the data
  2) Estimate model parameters.
  3) Deploy — simulate or forecast.
- We will consider the probability model $Y_n = M_n + S_n + X_n$,
  where M_n is the trend — evolution of mean over time
  S_n is the seasonal component — sinusoids w/ certain period.
  X_n is the residual — zero-mean RV (not necessarily independent over time)

Fitting the trend M_n and seasonal component S_n

- Assuming the trend to be a kth order polynomial,

$$M_n = \alpha_0 + \alpha_1 n + \ldots \alpha_k n^k.$$

we use the data $y_0 \ldots y_N$ to estimate $(\alpha_0, \alpha_1 \ldots \alpha_k)$ via least squares. → so we get $\hat{m}_n$

$$\left(\frac{\partial}{\partial \alpha_0}, \frac{\partial}{\partial \alpha_1} \ldots \frac{\partial}{\partial \alpha_k}\right) \sum_{n=0}^{N} \left(y_n - \alpha_0 - \alpha_1 n - \alpha_k n^k\right)^2 = (0, 0 \ldots 0)$$

- Assuming the seasonal component is in the form

$$s_n = A\cos(2\pi f n + \phi) = \alpha_1 \cos(2\pi f n) + \alpha_2 \sin(2\pi f n)$$

we use the data $y_0 \ldots y_N$ to estimate (α_1, α_2) via least squares → so we get $\hat{s}_n$.

$$\left(\frac{\partial}{\partial \alpha_1}, \frac{\partial}{\partial \alpha_2}\right) \sum_{n=0}^{N} \left(\hat{y}_n - \alpha_1 \cos(2\pi f n) - \alpha_2 \sin(2\pi f n)\right)^2 = (0, 0)$$

where $\hat{y}_n$ is the detrended data, $\hat{y}_n = y_n - m_n$.

- Guessing the trend model will result in a biased estimate $\hat{m}_n$ s.t. $E[\hat{m}_n] \neq 0$.
- Instead of using least squares, the residual is modelled using AR / MA / ARMA models.

Time series models.

Autoregressive (AR) process.

- Let $\{w_n\}_{n=-\infty}^{\infty}$ be a sequence of RVs s.t. $E[w_n]=0\ \forall n$, $E[w_i w_j]=\begin{cases}\sigma^2 & \text{for } i=j\\ 0 & \text{for } i\neq j\end{cases}$.

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

$$\boxed{x_n=\left(\sum_{i=1}^{p} a_i x_{n-i}\right)+w_n}$$

where $a_1,\dots,a_p$ are const. and p is the order of the process.

- A real time-series x_n is modelled as a function of its previous values and random part.

$$x_n=\underbrace{a_1x_{n-1}+\dots+a_px_{n-p}}_{\text{predictable part}}+\underbrace{w_n}_{\text{random part}}$$

- Consider the AR(1) process, $x_n=ax_{n-1}+w_n$.

(i) The AR(1) process can be expanded in terms of w_k, $k\le n$.

$$x_n=ax_{n-1}+w_n=a(ax_{n-2}+w_{n-1})+w_n=\sum_{k=0}^{\infty}w_{n-k}\cdot a^k$$

$$\therefore\ \boxed{x_n=\sum_{k=0}^{\infty}w_{n-k}h_k=w_n*h_n,\ \text{where } h_k=a^k}$$

i.e. AR(1) is causal w/ impulse response $\{h_k\}_{k\ge0}$ [IIR filter]

(ii) The mean $E[x_n]$ is zero, $\boxed{E[x_n]=0}$.

$$E[x_n]=E\left[\sum_{k=0}^{\infty}w_{n-k}a^k\right]=\sum_{k=0}^{\infty}a^k\underbrace{E[w_{n-k}]}_{=0}=0$$

alternatively,
$E[x_n]=E[ax_{n-1}+w_n]$
$E[x_n]=aE[x_{n-1}]+E[w_n]$
$E[x_n](1-a)=0\ \rightarrow\ E[x_n]=0$

(iii) The variance $\mathrm{Var}[x_n]$ is given by $\boxed{\mathrm{Var}[x_n]=\frac{\sigma^2}{1-a^2}}$

$$\mathrm{Var}[x_n]=E[x_n^2]-\underbrace{E[x_n]^2}_{=0}=E\left[\left(\sum_{i=0}^{\infty}w_{n-i}a^i\right)\left(\sum_{j=0}^{\infty}w_{n-j}a^j\right)\right]=E\left[\left(\sum_{i=0}^{\infty}w_{n-i}^2a^{2i}\right)+\text{cross terms}\right]$$

$$=\sum_{i=0}^{\infty}a^{2i}E[w_{n-i}^2]=\sum_{i=0}^{\infty}a^{2i}\sigma^2=\frac{\sigma^2}{1-a^2},\ \text{provided } |a|<1$$

Note $E[w_iw_j]=0$ for $i\neq j$ → cross-terms are neglected.

alternatively
$\mathrm{Var}[x_n]=\mathrm{Var}[ax_{n-1}+w_n]$
$\mathrm{Var}[x_n]=a^2\mathrm{Var}[x_{n-1}]+\mathrm{Var}[w_n]$
$\mathrm{Var}[x_n](1-a^2)=\sigma^2\ \rightarrow\ \mathrm{Var}[x_n]=\frac{\sigma^2}{1-a^2}$

(iv) The autocorrelation $r_{xx}[k]$ is given by $\boxed{r_{xx}[k]=a^k\frac{\sigma^2}{1-a^2}}$

$x_{n-1}=ax_{n-2}+w_{n-1}$
→ x_{n-1}, w_n are independent.

$$r_{xx}[1]=E[x_{n-1}x_n]=E[x_{n-1}(ax_{n-1}+w_n)]=aE[x_{n-1}^2]+\underbrace{E[x_{n-1}w_n]}_{=0}=a\,\mathrm{Var}[x_n]=a\frac{\sigma^2}{1-a^2}$$

$$r_{xx}[2]=E[x_{n-2}x_n]=E[x_{n-2}(ax_{n-1}+w_n)]=aE[x_{n-2}x_{n-1}]+\underbrace{E[x_{n-2}w_n]}_{=0}=ar_{xx}[1]=a^2\frac{\sigma^2}{1-a^2}.$$

By inspection, we see $r_{xx}[k+1]=ar_{xx}[k]$, so $r_{xx}[k]=a^k\frac{\sigma^2}{1-a^2}$ for $k\ge0$

(v) The PSD $S_x(f)$ is given by $\boxed{S_x(f)=\frac{\sigma^2}{1+a^2-2a\cos(2\pi f)}}$

$$S_x(f)=\sum_{k=-\infty}^{\infty}a^{|k|}\frac{\sigma^2}{1-a^2}e^{-j2\pi fk}=\frac{\sigma^2}{1-a^2}\sum_{k=-\infty}^{\infty}a^{|k|}e^{-j2\pi fk}=\frac{\sigma^2}{1-a^2}\left[\sum_{k=0}^{\infty}a^ke^{-j2\pi fk}+\sum_{k=0}^{\infty}a^ke^{j2\pi fk}-1\right]$$

$$=\frac{\sigma^2}{1-a^2}\left[\frac{1}{1-ae^{-j2\pi f}}+\frac{1}{1-ae^{j2\pi f}}-1\right]=\frac{\sigma^2}{1-a^2}\left[\frac{1-ae^{j2\pi f}+1-ae^{-j2\pi f}-1+a^2+ae^{-j2\pi f}+ae^{j2\pi f}}{(1-ae^{-j2\pi f})(1-ae^{j2\pi f})}\right]$$

$$=\frac{\sigma^2}{(1-ae^{-j2\pi f})(1-ae^{j2\pi f})}=\frac{\sigma^2}{1+a^2-ae^{-j2\pi f}-ae^{j2\pi f}}=\frac{\sigma^2}{1+a^2-2a\cos(2\pi f)}$$

Alternatively, the PSD $S_x(f)$ can be found using $S_x(f)=|H(f)|^2S_w(f)$, where $H(f)=\mathrm{DTFT}[h_k]=\mathrm{DTFT}[a^k]$

$$H(f)=\sum_{k=-\infty}^{\infty}h_ke^{-j2\pi fk}=\sum_{k=0}^{\infty}a^ke^{-j2\pi fk}=\frac{1}{1-ae^{-j2\pi f}},\quad S_w(f)=\sum_{k=-\infty}^{\infty}r_{ww}[k]e^{-j2\pi fk}=\sigma^2\ \left(r_{ww}[k]=\begin{cases}\sigma^2 & \text{if } k=0\\ 0 & \text{o/w}\end{cases}\right)$$

$$\therefore\ S_x(f)=H(f)H^*(f)S_w(f)=\frac{\sigma^2}{(1-ae^{-j2\pi f})(1-ae^{j2\pi f})}$$

* (ii), (iii), (iv) → AR(1) process is WSS.

- Higher order AR models can capture finer details and short-term fluctuations.

Moving average process (MA)

- Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of RVs s.t. $E[W_n]=0$ $\forall n$, $E[W_iW_j]=\begin{cases}\sigma^2 & \text{for } i=j\\ 0 & \text{for } i\neq j\end{cases}$.

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

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

where $b_1 \ldots b_q$ are const. and q is the order of the process

- A real time series X_n is modelled as a function of its previous values and a random part.

predictable part · random part

$$X_n = b_1 W_{n-1} + \ldots + b_q W_{n-q} + W_n$$

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

(i) The MA(q) process can be written in terms of W_k, $k \le n$.

$$X_n = \sum_{k=0}^{\infty} W_{n-k} h_k = W_n * h_n \text{, where } h_k = \begin{cases}1 & \text{if } k=0\\ b_k & \text{if } k\in[1,q]\\ 0 & \text{o/w}\end{cases}$$

i.e. MA(q) is causal w/ impulse response $\{h_k\}_{k=0}^{q}$ [FIR filter]

(ii) The mean $E[X_n]$ is zero, $E[X_n]=0$.

$$E[X_n] = E\left[\sum_{k=0}^{q} W_{n-k} h_k\right] = \sum_{k=0}^{q} h_k \cancel{E[W_{n-k}]}^{0} = 0.$$

(iii) The variance $\mathrm{Var}[X_n]$ is given by $\mathrm{Var}[X_n] = \sigma^2(1+b_1^2+\ldots+b_q^2)$

$$\mathrm{Var}[X_n] = E[X_n^2] - \cancel{E[X_n]^2}^{0} = E\left[\left(\sum_{i=1}^{q} b_i W_{n-i} + W_n\right)\left(\sum_{j=1}^{q} b_j W_{n-j} + W_n\right)\right] = E\left[\sum_{i=1}^{q} b_i^2 W_{n-i}^2\right] + W_n^2 + \text{cross terms}$$

$$= \left(\sum_{i=1}^{q} b_i^2 E[W_{n-i}^2]\right) + E[W_n^2] = \left(\sum_{i=1}^{q} b_i^2 \sigma^2\right) + \sigma^2 = \sigma^2(1+b_1^2+\ldots+b_q^2)$$

(iv) The autocorrelation $r_{XX}[k]$ is given by $r_{XX}[k] = \sigma^2 \sum_{i=-\infty}^{\infty} h_i h_{i+k}$

$$r_{XX}[k] = E[X_n X_{n+k}] = E\left[\left(\sum_{i=-\infty}^{\infty} W_{n-i} h_i\right)\left(\sum_{j=-\infty}^{\infty} W_{n+k-j} h_j\right)\right] = E\left[\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j W_{n-i} W_{n+k-j}\right] = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h_i h_j E[W_{n-i} W_{n+k-j}]$$

$E[W_{n-i}W_{n+k-j}] = \begin{cases}\sigma^2 & \text{if } j=i+k\\ 0 & \text{o/w}\end{cases}$ so $r_{XX}[k] = \sigma^2 \sum_{i=-\infty}^{\infty} h_i h_{i+k}$

Note $r_{XX}[k]=0$ for $k>q$ (since either h_i or h_{i+k} is zero).

(v) The PSD $S_X(f)$ is given by $S_X(f) = \left(1+\sum_{k=1}^{q} b_k e^{j2\pi fk}\right)\left(1+\sum_{k=1}^{q} b_k e^{-j2\pi fk}\right)\sigma^2$

$S_X(f) = |H(f)|^2 S_W(f)$, where $H(f) = \mathrm{DTFT}[\{h_k\}] = \mathrm{DTFT}[\{1, b_1 \ldots b_q\}]$

$$H(f) = \sum_{k=0}^{q} h_k e^{-j2\pi fk} = \sum_{k=0}^{q} h_k e^{-j2\pi fk} = 1 + \sum_{k=1}^{q} b_k e^{-j2\pi fk}, \quad S_W(f) = \sum_{k=-\infty}^{\infty} r_W[k] e^{-j2\pi fk} = \sigma^2 \left(r_W[k] = \begin{cases}\sigma^2 & \text{if } k=0\\ 0 & \text{o/w}\end{cases}\right)$$

$$S_X(f) = H(f)H^*(f)S_W(f) = \left(1+\sum_{k=1}^{q} b_k e^{-j2\pi fk}\right)\left(1+\sum_{k=1}^{q} b_k e^{j2\pi fk}\right)\sigma^2$$

* (ii), (iii), (iv) → MA(q) process is WSS

Autoregressive-moving average process (ARMA).

- Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of RVs s.t. $E[W_n]=0$ $\forall n$, $E[W_iW_j]=\begin{cases}\sigma^2 & \text{for } i=j\\ 0 & \text{for } i\neq j\end{cases}$.

The ARMA(p,q) process $\{X_n\}_{n=-\infty}^{\infty}$ is defined as.

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

where $a_1 \ldots a_p$, $b_1 \ldots b_q$ are const. and p/q are the order of the IIR/FIR respectively.

- The ARMA(p,q) process can be written in terms of W_k, $k \le n$.

$$X_n = \sum_{k=0}^{\infty} W_{n-k} h_k = W_n * h_n \text{. where } h_k = \begin{cases}1 & \text{if } k=0\\ a_k+b_k & \text{if } k\in[1,r]\\ a_k & \text{o/w}\end{cases}.$$

i.e. ARMA(p,q) is causal w impulse response $\{h_k\}_{k=0}^{\infty}$ [IIR filter]

- The ARMA(p,r) process is WSS.

AR, MA, ARMA processes as causal filters.

- AR, MA, ARMA process can act as causal filters w/ impulse response $\{h_k\}$, subject to the i/p $\{W_n\}_{n=-\infty}^{\infty}$, where $E[W_n]=0$ and $E[W_iW_j]=\begin{cases}\sigma^2 & \text{for } i=j\\ 0 & \text{for } i\neq j\end{cases}$.
- Consider the random process $\{W_n\}_{n=-\infty}^{\infty}$.
  - (i) Mean $E[W_n]$: $E[W_n]=0$, i.e. const. $\forall n$.
  - (ii) Power $E[W_n^2]$: $E[W_n^2]=\sigma^2<\infty$ i.e. finite power
  - (iii) Autocorrelation $r_{WW}[k]$: $r_{WW}[k]=E[W_nW_{n+k}]=0=f(k)$, i.e. only depends on k
  - → $\{W_n\}_{n=-\infty}^{\infty}$ is WSS.
- Provided for the causal filter, its impulse response $\{h_n\}$ satisfies $\sum_{n=-\infty}^{\infty}|h_i|<\infty$, then given a WSS i/p $\{X_n\}$, its o/p $\{Y_n\}$ is also WSS → AR, MA, ARMA processes must be WSS.

Random phase cosine (harmonic process)

- For the random phase cosine, the amplitude a and frequency f_0 are known, but the phase Φ is unknown and random (could correspond to an unknown delay). It has the form

$$X_n=a\cos(2\pi f_0 n+\Phi)$$

where a, f_0 are const. and $\Phi\sim\text{Unif}[0,2\pi)$, i.e. its pdf is $p_\Phi(\phi)=\begin{cases}1/2\pi & \text{if } \phi\in[0,2\pi)\\ 0 & \text{o/w}\end{cases}$

* To generate $\{X_n\}$, we need to first sample Φ to get $\Phi=\hat{\Phi}$, then set $X_n=\cos(2\pi f_0 n+\hat{\Phi})$ for $n=0,1,\dots$

- Consider the harmonic process, $X_n=a\cos(2\pi f_0 n+\Phi)$
  - (i) The mean $E[X_n]$ is zero, $\boxed{E[X_n]=0}$

    $E[X_n]=E[a\cos(2\pi f_0 n+\Phi)]=E[a\cos(2\pi f_0 n)\cos(\Phi)-a\sin(2\pi f_0 n)\sin(\Phi)]=a\cos(2\pi f_0 n)E[\cos(\Phi)]-a\sin(2\pi f_0 n)E[\sin(\Phi)]=0$

    since $E[\cos(\Phi)]=E[\sin(\Phi)]=0$ for uniform distribution Φ.
  - (ii) The variance $\text{Var}[X_n]$ is given by $\boxed{\text{Var}[X_n]=\frac{1}{2}a^2}$.

    $\text{Var}[X_n]=E[X_n^2]-\cancel{E[X_n]^2}=E[a^2\cos^2(2\pi f_0 n+\Phi)]=E[\frac{a^2}{2}(1+\cos(4\pi f_0 n+2\Phi))]=\frac{a^2}{2}+\frac{a^2}{2}\cancel{E[\cos(4\pi f_0 n+2\Phi)]}^{\to 0}=\frac{a^2}{2}$
  - (iii) The autocorrelation $r_{XX}[k]$ is given by $\boxed{r_{XX}[k]=\frac{1}{2}a^2\cos(2\pi f_0 k)}$

    $r_{XX}[k]=E[X_nX_{n+k}]=E[(a\cos(2\pi f_0 n+\Phi)a\cos(2\pi f_0(n+k)+\Phi)]=E[\frac{a^2}{2}(\cos(2\pi f_0(2n+k)+2\Phi)+\cos(2\pi f_0 k)]$

    $=\frac{a^2}{2}\cancel{E[\cos(2\pi f_0(2n+k)+2\Phi)]}^{\to 0}+\frac{a^2}{2}\cos(2\pi f_0 k)=\frac{a^2}{2}\cos(2\pi f_0 k)$
  - (iv) The PSD $S_X(f)$ is given by $\boxed{S_X(f)=\frac{a^2\pi}{2}(\delta(2\pi(f+f_0))+\delta(2\pi(f+f_0)))}$

    $S_X(f)=\sum_{k=-\infty}^{\infty}r_{XX}[k]e^{-j2\pi fk}=\sum_{k=-\infty}^{\infty}\frac{a^2}{2}\cos(2\pi f_0 k)e^{-j2\pi fk}=\frac{a^2}{4}\sum_{k=-\infty}^{\infty}(e^{-j2\pi(f-f_0)k}+e^{-j2\pi(f+f_0)k})$

    $=\frac{a^2}{4}[2\pi\delta(2\pi(f-f_0))+2\pi\delta(2\pi(f+f_0))]=\frac{a^2\pi}{2}(\delta(2\pi(f-f_0))+\delta(2\pi(f+f_0)))$ ← over $f\in[-1/2,1/2]$, and periodic w/ period of 1.

* (i), (ii), (iii) → harmonic process is WSS.

- Consider the inversion formula for PSD, $r_{XX}[k]=\int_{-1/2}^{1/2}S_X(f)e^{j2\pi fk}df=\int_{-1/2}^{1/2}\underset{\text{even}}{S_X(f)}\underset{\text{even}}{\cos(2\pi fk)}df+\int_{-1/2}^{1/2}\underset{\text{odd}}{S_X(f)}\underset{\text{odd}}{\sin(2\pi fk)}df$

  $\therefore r_{XX}[k]=2\int_0^{1/2}S_X(f)\cos(2\pi fk)df\xrightarrow{\text{divide into } K \text{ intervals, width } 1/2K}2\sum_{i=1}^{K}S_X(f_i)\cos(2\pi f_i k)\frac{1}{2K}=\frac{1}{K}\sum_{i=1}^{K}S_X(f_i)\cos(2\pi f_i k)$

i.e. we have a sum of random phase cosines w/ freq. f_i and amplitude a s.t. $\frac{a^2}{2}=\frac{S_X(f_i)}{K}\to a=\sqrt{\frac{2S_X(f_i)}{K}}$

$\therefore\ \boxed{X_n=\sum_{i=1}^{K}\sqrt{\frac{2S_X(f_i)}{K}}\cos(2\pi f_i k+\Phi_i)}$ where Φ_i iid $\sim\text{Unif}[0,2\pi)$

White noise process.

- A WSS process $\{X_n\}$ is termed white noise if

$$C_{XX}[k] = \sigma^2 \delta[k]$$

where $\delta[k]$ is the discrete impulse function, $\delta[k] = \begin{cases} 1 & \text{if } k=0 \\ 0 & \text{o/w.} \end{cases}$

- The PSD $S_X(f)$ for white noise is given by $S_X(f) = \sigma^2 + \mu^2 2\pi\delta(f)$

$$S_X(f) = \sum_{k=-\infty}^{\infty} r_{XX}[k] e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} (C_{XX}[k] + \mu^2) e^{-j2\pi fk} = \sum_{k=-\infty}^{\infty} \sigma^2 \delta[k] e^{-j2\pi fk} + \sum_{k=-\infty}^{\infty} \mu^2 e^{-j2\pi fk} = \sigma^2 + \mu^2 2\pi\delta(f).$$

for zero mean white noise, $\mu = 0$, so $S_X(f) = \sigma^2$, i.e. flat across all freq. ← $\{w_n\}_{n=-\infty}^{\infty}$ for ARMA is zero-mean white noise

- If the values x_n are drawn independently from a Gaussian distribution $N(0, \sigma^2)$, we have white Gaussian noise (WGN), the nth order pdf for WGN process is

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} N(x_i | 0, \sigma^2)$$ ← simple product of x_i are independent

where $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ is the univariate normal pdf.

- The WGN process is SSS, since

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} N(x_i | 0, \sigma^2) = f_{X_{1+k} X_{2+k} \ldots X_{n+k}}(x_1, x_2 \ldots x_n)$$

Optimal filtering and signal detection

Optimal filtering

PSD and transfer function in terms of Ω.

- The PSD and TF can be expressed in terms of the normalised freq. Ω. (in radians per sample).

  where $\Omega = \omega T$, and T is the sampling interval of the discrete-time process.

$$S_x(e^{j\Omega}) = \sum_{k=-\infty}^{\infty} r_{xx}[k] e^{-jk\Omega} \qquad r_{xx}[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(e^{j\Omega}) e^{jk\Omega}$$

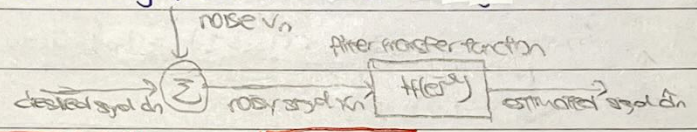
$$S_{xy}(e^{j\Omega}) = \sum_{k=-\infty}^{\infty} r_{xy}[k] e^{-jk\Omega} \qquad r_{xy}[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xy}(e^{j\Omega}) e^{jk\Omega}$$

$$P_{[-\pi,\pi]} = r_{xx}[0] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(e^{j\Omega}) d\Omega \qquad P_{(\omega_1,\omega_2)} = \frac{1}{\pi}\int_{\omega_1 T}^{\omega_2 T} S_x(e^{j\Omega}) d\Omega$$

+ In this section, we explicitly write the PSD/TF as a function of $e^{j\Omega}$ to stress that it is always periodic w/ a period of 2π. → looks like a Z-transform, w/ $z = e^{j\Omega}$

Optimal filtering.

- Optimal filtering is where we design filters that are optimally adapted to the statistical characteristics of a random process.
- The Wiener filter can optimally estimate a desired signal d_n given the noisy observations x_n, w some assumptions about the statistics of the random signal and noise processes.
- The general Wiener filtering problem can be expressed as a block diagram.



  i.e. we have $x_n = d_n + v_n$ and $\hat{d}_n = x_n * h_n$.

- We define the error signal ε_n as

$$\varepsilon_n = d_n - \hat{d}_n = d_n - x_n * h_n$$

  and the mean-squared error (MSE) J as

$$J = E[\varepsilon_n^2]$$

  where the expectation is wrt the random signal d_n and random noise v_n.

- The Wiener filter minimises the MSE J wrt the filter coefficients $\{h_n\}$.
- The Wiener filter assume the following.

  (i) $\{x_n\}$ and $\{d_n\}$ are jointly WSS, i.e. $r_{xx}[n,m] \to r_{xx}[n-m]$; $r_{dd}[n,m] \to r_{dd}[n-m]$, $r_{xd}[n,m] \to r_{xd}[n-m]$

  (ii) $\{d_n\}$ and $\{v_n\}$ have zero mean, i.e. $E[d_n] = 0$, $E[v_n] = 0$

+ The Wiener filter is only the optimal linear estimator for stationary signals

The general Wiener filter.

- In the general case, we can filter the observed signal x_n w/ an infinite-dimensional filter, having a non-causal impulse response $\{h_p\}_{-\infty}^{\infty}$. our estimate of the desired signal $\hat{d}_n$ is therefore

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

- We want to minimise the MSE J wrt the impulse response values h_q.

A sufficient condition for a min. is that simultaneously $\forall q \in (-\infty, \infty)$,

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

Noting that $\frac{\partial \varepsilon_n}{\partial h_q} = \frac{\partial}{\partial h_q}\left[d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right] = -x_{n-q}$, we get the condition (orthogonality principle)

$$E\left[2\varepsilon_n \frac{\partial \varepsilon_n}{\partial h_q}\right] = E[2\varepsilon_n x_{n-q}] = 0 \quad \rightarrow \quad \boxed{E[\varepsilon_n x_{n-q}] = 0 \;\; \forall q \in (-\infty, \infty)}$$

Substituting ε_n in $E[\varepsilon_n x_{n-q}]$ gives

$$E[\varepsilon_n x_{n-q}] = E\left[\left(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right) x_{n-q}\right] = E[d_n x_{n-q}] - \sum_{p=-\infty}^{\infty} h_p E[x_{n-p} x_{n-q}] = r_{xd}[q] - \sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p]$$

$$r_{xd}[q] - \sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = 0 \quad \rightarrow \quad \boxed{\sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = r_{xd}[q], \;\; \forall q \in (-\infty, \infty)}$$

which are collectively known as the Wiener–Hopf eqns.

- The simplest way to solve for the infinite no. of unknowns h_p in the Wiener–Hopf eqns is to transform everything into the frequency domain.

Rewriting the Wiener–Hopf eqns as a convolution,

$$h_q * r_{xx}[q] = r_{xd}[q] \quad \forall q \in (-\infty, \infty).$$

Taking the DTFT on both sides,

$$H(e^{j\Omega}) S_x(e^{j\Omega}) = S_{xd}(e^{j\Omega}) \quad \rightarrow \quad \boxed{H(e^{j\Omega}) = \frac{S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})}}$$

- The MSE of the optimal filter, J_{min}, can be found using (i) time or (ii) frequency domain

(i). $J = E[\varepsilon_n^2] = E\left[\varepsilon_n\left(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right)\right] = E[\varepsilon_n d_n] - \sum_{p=-\infty}^{\infty} h_p E[\varepsilon_n x_{n-p}]$.

For the optimal filter, the orthogonality principle states $E[\varepsilon_n x_{n-p}] = 0 \; \forall p \in (-\infty, \infty)$, so

$$\boxed{J_{min} = E[\varepsilon_n d_n] = E\left[\left(d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}\right) d_n\right] = E[d_n^2] - \sum_{p=-\infty}^{\infty} h_p E[d_n x_{n-p}] = r_{dd}[0] - \sum_{p=-\infty}^{\infty} h_p r_{xd}[p]}$$

(ii) $\varepsilon_n \varepsilon_{n+k} = \left(d_n - \sum_{i=-\infty}^{\infty} h_i x_{n-i}\right)\left(d_{n+k} - \sum_{j=-\infty}^{\infty} h_j x_{n+k-j}\right) = d_n d_{n+k} - d_n \sum_{j=-\infty}^{\infty} h_j x_{n+k-j} - d_{n+k} \sum_{i=-\infty}^{\infty} h_i x_{n-i} + \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h_i h_j x_{n-i} x_{n+k-j}$

$$r_{\varepsilon}[k] = E[\varepsilon_n \varepsilon_{n+k}] = r_{dd}[k] - \sum_{j=-\infty}^{\infty} h_j r_{dx}[k-j] - \sum_{i=-\infty}^{\infty} h_i r_{dx}[-i-k] + \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h_i h_j r_{xx}[k+i-j]$$

$$= r_{dd}[k] - h_k * r_{dx}[k] - \tilde{h}_k * r_{xd}[k] + h_k * \tilde{h}_k * r_{xx}[k]$$

$$S_{\varepsilon}(e^{j\Omega}) = S_d(e^{j\Omega}) - H(e^{j\Omega}) S_{dx}(e^{j\Omega}) - H^*(e^{j\Omega}) S_{xd}(e^{j\Omega}) + H(e^{j\Omega}) H^*(e^{j\Omega}) S_x(e^{j\Omega})$$

$$J = E[\varepsilon_n^2] = r_{\varepsilon}[0] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\varepsilon}(e^{j\Omega}) d\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[S_d(e^{j\Omega}) - H(e^{j\Omega}) S_{dx}(e^{j\Omega}) - H^*(e^{j\Omega}) S_{xd}(e^{j\Omega}) + |H(e^{j\Omega})|^2 S_x(e^{j\Omega})\right] d\Omega$$

For the optimal filter, solving the Wiener–Hopf eqns give $H(e^{j\Omega}) = \frac{S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})}$.

$$H^*(e^{j\Omega}) S_{xd}(e^{j\Omega}) = \frac{S_{xd}^*(e^{j\Omega}) S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})} = \frac{|S_{xd}(e^{j\Omega})|^2}{S_x(e^{j\Omega})}; \quad |H(e^{j\Omega})|^2 S_x(e^{j\Omega}) = \left|\frac{S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})}\right|^2 = \frac{|S_{xd}(e^{j\Omega})|^2}{S_x(e^{j\Omega})}.$$

$$\rightarrow \boxed{J_{min} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[S_d(e^{j\Omega}) - H(e^{j\Omega}) S_{dx}(e^{j\Omega})\right] d\Omega}$$

+ If the random process is ergodic, we estimate $r_{xx}[k]$, $r_{xd}[k]$ from time averages

∝ In general, the optimal filter yielded is non-causal → not implementable in practice

The FIR Wiener filter

- For the FIR Wiener filter, it has a causal p-th order finite impulse response $\{h_p\}_{p=0}^{P}$. Our estimate of the desired signal, $\hat{d}_n$, is therefore

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

- As before, we want to minimise the MSE J wrt the impulse response values h_q.

Orthogonality principle: $E[\varepsilon_n x_{n-q}] = 0 \quad \forall q \in [0,P]$

Wiener-Hopf eqns: $\sum_{p=0}^{P} h_p r_{xx}[q-p] = r_{xd}[q], \quad \forall q \in [0,P]$

- The Wiener-Hopf eqns are now a finite set of simultaneous eqns that can be solved in the time domain. The eqns may be written in matrix form as

$$\underline{\underline{R}}_x \underline{h} = \underline{r}_{xd}$$

where $\underline{h} = \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_P \end{bmatrix}$, $\underline{r}_{xd} = \begin{bmatrix} r_{xd}[0] \\ r_{xd}[1] \\ \vdots \\ r_{xd}[P] \end{bmatrix}$, $\underline{\underline{R}}_x = \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[P] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[P-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[P] & r_{xx}[P-1] & \cdots & r_{xx}[0] \end{bmatrix}$

(Note the correlation matrix $\underline{\underline{R}}_x$ is symmetric and has constant diagonals).

The coefficient vector can be found by matrix inversion,

$$\underline{h} = \underline{\underline{R}}_x^{-1} \underline{r}_{xd}$$

- As before, the MSE of the optimal filter J_{min}, is given by

$$J_{min} = r_{dd}[0] - \sum_{p=0}^{P} h_p r_{xd}[p] = r_{dd}[0] - \underline{r}_{xd}^T \underline{h} = r_{dd}[0] - \underline{r}_{xd}^T \underline{\underline{R}}_x^{-1} \underline{r}_{xd}$$

Uncorrelated signal and noise processes,

- Typically, the environmental noise v_n is independent of the desired signal $d_n \rightarrow$ uncorrelated, i.e.

$$r_{dv}[k] = E[d_n v_{n+k}] \qquad k \in (-\infty, \infty)$$

- In this situation, the correlation functions $r_{xd}[k]$ and $r_{xx}[k]$ can be simplified.

(i) $r_{xd}[k] = E[x_n d_{n+k}] = E[(d_n + v_n) d_{n+k}] = E[d_n d_{n+k}] + \cancel{E[v_n d_{n+k}]}^{\to 0} = r_{dd}[k]$.

$$r_{xd}[k] = r_{dd}[k] \quad \hookrightarrow \quad S_{xd}(e^{j\Omega}) = S_d(e^{j\Omega})$$

(ii) $r_{xx}[k] = E[x_n x_{n+k}] = E[(d_n + v_n)(d_{n+k} + v_{n+k})] = E[d_n d_{n+k}] + \cancel{E[d_n v_{n+k}]}^{\to 0} + \cancel{E[v_n d_{n+k}]}^{\to 0} + E[v_n v_{n+k}] = r_{dd}[k] + r_{vv}[k]$

$$r_{xx}[k] = r_{dd}[k] + r_{vv}[k] \quad \hookrightarrow \quad S_x(e^{j\Omega}) = S_d(e^{j\Omega}) + S_v(e^{j\Omega})$$

- For the general Wiener filter, the optimal filter TF $H(e^{j\Omega})$ becomes.

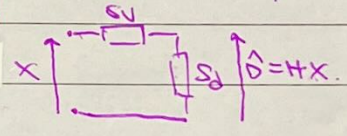
$$H(e^{j\Omega}) = \frac{S_{xd}(e^{j\Omega})}{S_x(e^{j\Omega})} = \frac{S_d(e^{j\Omega})}{S_d(e^{j\Omega}) + S_v(e^{j\Omega})} = \frac{1}{1 + 1/\rho(\Omega)}$$

$|H(e^{j\Omega})| \in [0,1]$

$\rho(\Omega) \gg 1 \rightarrow |H(e^{j\Omega})| \rightarrow 1$

$\rho(\Omega) \approx 0 \rightarrow |H(e^{j\Omega})| \rightarrow 0$.

where $\rho(\Omega) = \frac{S_d(e^{j\Omega})}{S_v(e^{j\Omega})}$ is the (freq.-dependent) SNR power ratio.



- The MSE of the optimal filter, J_{min} is now reduced to.

$$J_{min} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(S_d(e^{j\Omega}) - H(e^{j\Omega}) S_d(e^{j\Omega})\right) d\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_d(e^{j\Omega}) \left(1 - \frac{1}{1+1/\rho(\Omega)}\right) d\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_d(e^{j\Omega}) \frac{1}{1+\rho(\Omega)} d\Omega$$

- In practice, the correlation functions are estimated as (assuming signal is ergodic).

$$r_{xx}[k] \approx \frac{1}{N}\sum_{n=0}^{N-1} x_n x_{n+k}, \qquad r_{dd}[k] = r_{xx}[k] - r_{vv}[k] = \begin{cases} r_{xx}[k] - \sigma_v^2 & \text{if } k = 0 \\ r_{xx}[k] & \text{o/w} \end{cases}$$

* We should make sure $r_{xx}[k]$ and $r_{dd}[k]$ are valid autocorrelation sequences, since they are only empirical estimates.

Validity of autocorrelation sequences

- For the autocorrelation $r_{xx}[k]$ to form a valid sequence for constructing the correlation matrix $\underline{\underline{R}}_x$, the necessary condition is the resulting $\underline{\underline{R}}_x$ is positive-semidefinite, ie.

$$\underline{a}^T \underline{\underline{R}}_x \underline{a} \geq 0$$

for any length $p+1$ vector $\underline{a}$.

(since $\underline{\underline{R}}_x$ is symmetric, this means its eigenvalues λ_i must satisfy $\lambda_i \geq 0 \ \forall i$)

- consider the vector $\underline{x}_n = [x_n \ x_{n-1} \ \dots \ x_{n-p}]^T$ for real-valued $x_n, \underline{a}$, $\underline{a}^T \underline{x}_n = \underline{x}_n^T \underline{a}$.

$$0 \leq (\underline{a}^T \underline{x}_n)^2 = (\underline{a}^T \underline{x}_n)(\underline{x}_n^T \underline{a}) = \underline{a}^T (\underline{x}_n \underline{x}_n^T) \underline{a}$$

The (i,j)th element of $\underline{x}_n \underline{x}_n^T$ is $x_{n-i+1} x_{n-j+1}$ and $E[x_{n-i+1} x_{n-j+1}] = r_{xx}[i-j] \rightarrow E[\underline{x}_n \underline{x}_n^T] = \underline{\underline{R}}_x$.

$\therefore 0 \leq E[\underline{a}^T (\underline{x}_n \underline{x}_n^T) \underline{a}] = \underline{a}^T E[(\underline{x}_n \underline{x}_n^T)] \underline{a} = \underline{a}^T \underline{\underline{R}}_x \underline{a} \rightarrow \underline{\underline{R}}_x$ must be positive-semidefinite.

Extending the Wiener filter

- To extend the Wiener filter beyond the regular noise reduction case, we simply replace the desired signal d w/ what we want to predict/estimate, then redefine the filter.
- In general, the FIR formula

$$\underline{h} = \underline{\underline{R}}_x^{-1} \underline{r}_{xd}$$

applies for any form of the (unobserved) desired signal d and the (observed) "noisy" signal x, provided we can calculate the necessary correlation functions.

- Some standard examples include:
  - ↳ Prediction of a signal, $x_n = u_n + v_n$, $d_n = u_{n+p}$
  - ↳ Smoothing of a signal, $x_n = u_n + v_n$, $d_n = u_{n-p}$
  - ↳ Deconvolution. $x_n = h_{an} * u_n + v_n = \sum_{q=0}^{M} h_q u_{n-q} + v_n$, $d_n = u_n$.

* In all cases, the Wiener filter estimate takes the same form as before

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

and the error is defined as before

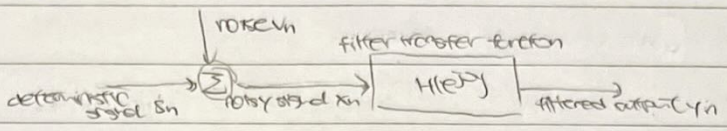
$$\epsilon_n = d_n - \hat{d}_n .$$

Note that only d is different, and we need to redefine an expression for $r_{xd}[k] / S_{xd}(e^{j\omega})$.

Signal detection

Signal detection

- The matched filter is an optimal FIR filter that can detect a deterministic signal s_n, $n=0,1,\dots N-1$, from the noisy observations x_n.
- The general signal detection problem can be expressed as a block diagram



i.e. we have $x_n = s_n + v_n$ and $y_n = x_n * h_n$

- In vector form, we have

$$\boxed{\underline{x} = \underline{s} + \underline{v}}$$

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

- Using a $(N-1)$th order FIR filter, the o/p of the filter at time $N-1$ is given by

$$\boxed{y_{N-1} = \sum_{m=0}^{N-1} h_m x_{N-1-m} = \underline{h}^T \underline{\tilde{x}} = \underline{h}^T(\underline{\tilde{s}} + \underline{\tilde{v}}) = \underline{h}^T\underline{\tilde{s}} + \underline{h}^T\underline{\tilde{v}} = \sum_{m=0}^{N-1} h_m s_{N-1-m} + \sum_{m=0}^{N-1} h_m v_{N-1-m} = y_{N-1}^s + y_{N-1}^n}$$

where y_{N-1}^s is the o/p from the signal only part; y_{N-1}^n is the o/p from the noise only part.

- We define the o/p SNR as

$$\boxed{SNR = \frac{E[|y_{N-1}^s|^2]}{E[|y_{N-1}^n|^2]} = \frac{E[|\underline{h}^T\underline{\tilde{s}}|^2]}{E[|\underline{h}^T\underline{\tilde{v}}|^2]} = \frac{|\underline{h}^T\underline{\tilde{s}}|^2}{E[|\underline{h}^T\underline{\tilde{v}}|^2]}}$$

- The matched filter maximises the o/p SNR, wrt the filter coefficients $\{h_n\}$ to give the best possible chance of detecting the signal s_n.

Signal output energy.

- The signal o/p energy is given by $y_{N-1}^{s\,2} = |\underline{h}^T\underline{\tilde{s}}|^2 = (\underline{h}^T\underline{\tilde{s}})(\underline{\tilde{s}}^T\underline{h}) = \underline{h}^T(\underline{\tilde{s}}\underline{\tilde{s}}^T)\underline{h}$.
- Consider the matrix $\underline{\underline{M}} = \underline{\tilde{s}}\underline{\tilde{s}}^T$. We want to find its eigenvectors $\underline{v}$ and eigenvalues λ.

$\underline{v} = \underline{s}$: $\quad \underline{\underline{M}}\,\underline{\tilde{s}} = (\underline{\tilde{s}}\underline{\tilde{s}}^T)\underline{\tilde{s}} = \underline{\tilde{s}}(\underline{s}^T\underline{\tilde{s}}) = (\underline{\tilde{s}}^T\underline{\tilde{s}})\underline{\tilde{s}}$

i.e. $\underline{v}_0 = \frac{\underline{\tilde{s}}}{|\underline{\tilde{s}}|}$ is an eigenvector and $\lambda_0 = \underline{\tilde{s}}^T\underline{\tilde{s}}$ is the corresponding eigenvalue.

$\underline{v} = \{\underline{v}' : \underline{\tilde{s}}^T\underline{v}' = 0\}$: $\quad \underline{\underline{M}}\,\underline{v}' = (\underline{\tilde{s}}\underline{\tilde{s}}^T)\underline{v}' = \underline{\tilde{s}}(\underline{\tilde{s}}^T\underline{v}') = 0$

i.e. $v = v'$ is an eigenvector w/ $\lambda = 0$. (There are $N-1$ orthonormal vectors orthogonal to $\underline{\tilde{s}}$: $\underline{v}_1, \underline{v}_2 \dots \underline{v}_{N-1}$)

- We can represent any filter coefficient vector $\underline{h}$ as a linear combination of these eigenvectors.

$$\underline{h} = \alpha\underline{v}_0 + \beta\underline{v}_1 + \gamma\underline{v}_2 + \dots + \dots \underline{v}_{N-1}$$

$$\rightarrow \underline{\underline{M}}\,\underline{h} = \underline{\underline{M}}(\alpha\underline{v}_0 + \beta\underline{v}_1 + \gamma\underline{v}_2 + \dots + \dots \underline{v}_{N-1}) = \alpha\underline{\underline{M}}\,\underline{v}_0 + \beta\underline{\underline{M}}\,\underline{v}_1 + \gamma\underline{\underline{M}}\,\underline{v}_2 + \dots + \dots \underline{\underline{M}}\,\underline{v}_{N-1} = \alpha(\underline{\tilde{s}}\underline{\tilde{s}}^T)\underline{v}_0$$

- The signal o/p energy is therefore.

$$\underline{h}^T\underline{\tilde{s}}\underline{\tilde{s}}^T\underline{h} = \underline{h}^T\underline{\underline{M}}\,\underline{h} = \underline{h}^T(\alpha\underline{\tilde{s}}\underline{\tilde{s}}^T\underline{v}_0) = (\alpha\underline{\tilde{s}}^T\underline{\tilde{s}})\,\underline{h}^T\underline{v}_0 = (\alpha\underline{\tilde{s}}^T\underline{\tilde{s}})(\alpha\underline{v}_0 + \beta\underline{v}_1 + \gamma\underline{v}_2 + \dots + \dots \underline{v}_{N-1})^T\underline{v}_0$$

$$\therefore \boxed{y_{N-1}^{s\,2} = |\underline{h}^T\underline{\tilde{s}}|^2 = \alpha^2\underline{\tilde{s}}^T\underline{\tilde{s}}}$$

since $\underline{v}_0^T\underline{v}_0 = 1$ and $\underline{v}_0^T\underline{v}' = 0$ (orthogonality).

Noise output energy

- The expected noise o/p energy is given by $E[y_{N-1}^2] = E[\underline{h}^T \underline{V}^2] = E[\underline{h}^T \underline{V} (\underline{V}^T \underline{h})] = \underline{h}^T E[\underline{V}\underline{V}^T] \underline{h}$
- considering the case where the noise is white, for time indices $i = 0, 1 \ldots N-1$ and $j = 0, 1 \ldots N-1$,

$$E[v_i v_j] = \begin{cases} \sigma_v^2 & \text{if } i = j \\ 0 & \text{o/w} \end{cases}$$

$\therefore E[\underline{V}\underline{V}^T] = \sigma_v^2 \underline{\underline{I}}$, where $\underline{\underline{I}}$ is the $N \times N$ identity matrix.

- The expected noise o/p energy is therefore

$$E[y_{N-1}^2] = \underline{h}^T E[\underline{V}\underline{V}^T] \underline{h} = \sigma_v^2 \underline{h}^T \underline{h}$$

Expanding $\underline{h}$ in terms of the eigenvectors of $\underline{\underline{M}} = \underline{\underline{S}}\underline{\underline{S}}^T$, $(\underline{h} = \alpha \underline{v}_0 + \beta \underline{v}_1 + \gamma \underline{v}_2 + \ldots + \ldots \underline{v}_{N-1})$, we have

$$E[y_{N-1}^2] = \sigma_v^2 \underline{h}^T \underline{h} = \sigma_v^2 (\alpha^2 + \beta^2 + \gamma^2 + \ldots)$$

SNR maximisation.

- The SNR may now be expressed as

$$SNR = \frac{|\underline{h}^T \underline{S}|^2}{E[|\underline{h}^T \underline{V}|^2]} = \frac{\alpha^2 \underline{S}^T \underline{S}}{\sigma_v^2 (\alpha^2 + \beta^2 + \gamma^2 + \ldots)}$$

- By inspection, scaling $\underline{h}$ by some factor ρ will not change the SNR, since both the numerator and numerator will both scale equally by ρ^2. → we can fix $|\underline{h}| = |\alpha \underline{v}_0 + \beta \underline{v}_1 + \gamma \underline{v}_2 + \ldots + \ldots \underline{v}_{N-1}| = 1$, then max.
- Setting $|\underline{h}| = 1$, we have $(\alpha^2 + \beta^2 + \gamma^2 + \ldots) = 1$, so the SNR becomes

$$SNR = \frac{\alpha^2 \underline{S}^T \underline{S}}{\sigma_v^2}$$

- The largest possible value of α given that $|\underline{h}| = 1$ corresponds to $\alpha = 1$ (i.e. $\beta = \gamma = \ldots = 0$).

This means the optimal filter coefficient vector $\underline{h}$ is

$$\underline{h} = 1 \, \underline{v}_0 = \frac{\underline{S}}{|\underline{S}|^2}$$

time-reversed signal, normalised.

and the corresponding SNR is

$$SNR = \frac{|\underline{S}|^2}{\sigma_v^2}$$

↑ depends on energy of signal s and noise v, as expected.

Practical implementation of the matched filter.

- For a signal $\underline{s}$ of length N, to optimise a filter $\underline{h}$, we considered a batch of data $\underline{x}$ of length N.
- In practice, we would run this optimised filter over a much longer length of data $\{x_n\}$, which contains $\underline{s}$ at some unknown position, then find the time at which max. energy occurs.
- The time at which max. energy occurs is the pt at which $\underline{s}$ can be detected, based on some threshold.
- Matched filtering maximises the best SNR obtainable significantly. Before filtering,

$$SNR_{max} = \frac{\text{max signal value}^2}{\text{average noise energy}} = \frac{\max s_n}{\sigma_v^2}$$

Estimation theory and inference

Estimation theory and inference methods.

Estimation and inference

- In estimation theory, we start w/ a vector of signal measurements $\underline{x} = [x_0\ x_1\ \dots\ x_{N-1}]^T$ and some unknown quantities (parameters) that we wish to infer, $\underline{\theta} = [\theta_0\ \theta_1 \dots \theta_{p-1}]^T$ (usually $p \ll N$)
- We typically suppose the probability distribution of the data $\underline{x}$ can be expressed as a likelihood function $p(\underline{x}|\underline{\theta})$, which represents how "likely" diff. realisations of the observed data would be if we knew $\underline{\theta}$.
- In many problems, we may treat $\underline{\theta}$ as a random vector, and a prior pdf can be formulated for $\underline{\theta}$, $p(\underline{\theta})$. The prior represents prior belief about likely parameter configurations prior to any data observations.
- The prior distribution can be used to regularise the inference problem by constraining the parameter search to responsible parts of the domain of $\underline{\theta}$ — Bayesian inference.
- In some cases, we may not wish to consider $\underline{\theta}$ as a random quantity. → we can only rely on the likelihood function — classical inference.
- Estimation is the task of forming an estimate $\hat{\theta}$ which is in some sense close to the true $\underline{\theta}$. Inference has estimation as a special case, and studies the whole prob. distribution of the unknown.

Simple estimators and their properties

- It is useful to consider some fundamental properties of estimators — their mean and variance.
- An estimator $\hat{\theta}$ is said to be unbiased if its expectation satisfies

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

- An estimator $\hat{\theta}$ is said to be consistent if it is unbiased and its variance satisfies

$$\lim_{N\to\infty} \mathrm{Var}[\hat{\theta}] = 0$$

- The defn of unbiased estimators and their variance can be used to measure the performance of any proposed estimation scheme.
- Estimators can be designed for a given problem specifically to lead to no bias and min. variance. (min. variance unbiased (MVU) estimators)
- Consider a set of N independent samples of a RV X which has mean μ and standard deviation σ

$$\{x_1, x_2, x_3 \dots x_N\}.$$

An estimator for the mean $\hat{\mu}$, is obtained by taking the arithmetic average of the samples.

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The expectation of the estimator $\hat{\mu}$, $E[\hat{\mu}]$ is given by

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu \rightarrow \text{unbiased}.$$

The variance of the estimator $\hat{\mu}$, $\mathrm{Var}[\hat{\mu}]$ is given by

$$\mathrm{Var}[\hat{\mu}] = E[(\hat{\mu} - E[\hat{\mu}])^2] = E[\hat{\mu}^2] - E[\hat{\mu}]^2 = \left(\mu^2 + \frac{\sigma^2}{N}\right) - \mu^2 = \frac{\sigma^2}{N} \rightarrow \text{tends to } 0 \text{ as } N \to \infty.$$

where $E[\hat{\mu}^2] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right)\right] = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} E[x_i x_j]$, and $E[x_i x_j] = \begin{cases} E[x_i^2] = \mu^2 + \sigma^2 & i = j \\ E[x_i]E[x_j] = \mu^2 & i \neq j \end{cases}$.

$$\therefore E[\hat{\mu}^2] = \frac{1}{N^2}\left[(N^2 - N)\mu^2 + N(\mu^2 + \sigma^2)\right] = \frac{1}{N^2}(N^2\mu^2 + N\sigma^2) = \mu^2 + \frac{\sigma^2}{N}.$$

The general linear model / linear regression model.

- In the linear model, it is assumed that the data x are generated as a linear function of the parameters θ w/ 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.

- This can be written for the whole vector $\underline{x}$ as

$$\underline{x} = G\underline{\theta} + \underline{e}, \quad \text{where} \quad G = \begin{bmatrix} \leftarrow g_0^T \rightarrow \\ \leftarrow g_1^T \rightarrow \\ \vdots \\ \leftarrow g_{N-1}^T \rightarrow \end{bmatrix}$$

- The choice of the design matrix G will lead to a wide range of possible models, for example,

① constant level in noise

$$x_n = \theta + e_n \quad \leftrightarrow \quad \underline{x} = G\underline{\theta} + \underline{e}, \text{ where } G = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

② sinusoidal model (1 frequency)

$$x_n = a\cos(n\omega) + b\sin(n\omega) + e_n \quad \leftrightarrow \quad \underline{x} = G\underline{\theta} + \underline{e}, \text{ where } G = [c(\omega)\ s(\omega)],\ \underline{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$$

(note $c(\omega) = [\cos(0)\ \cos(\omega) \dots \cos((N-1)\omega)]^T$, $s(\omega) = [\sin(0)\ \sin(\omega) \dots \sin((N-1)\omega)]^T$)

③ sinusoidal model (J frequencies)

$$x_n = \sum_{j=1}^{J} a_j \cos(n\omega_j) + b_j \sin(n\omega_j) + e_n \quad \leftrightarrow \quad \underline{x} = G\underline{\theta} + \underline{e}, \text{ where } G = [c(\omega_1)\ s(\omega_1) \dots c(\omega_J)\ s(\omega_J)],\ \underline{\theta} = \begin{bmatrix} a_1 \\ b_1 \\ \vdots \\ a_J \\ b_J \end{bmatrix}$$

(note $c(\omega_j) = [\cos(0)\ \cos(\omega_j) \dots \cos((N-1)\omega_j)]^T$, $s(\omega_j) = [\sin(0)\ \sin(\omega_j) \dots \sin((N-1)\omega_j)]^T$)

④ Autoregressive model.

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n \quad \leftrightarrow \quad \underline{x} = G\underline{a} + \underline{e}, \text{ where } G = \begin{bmatrix} x_{-1} & x_{-2} & \dots & x_{-(p)} & x_0 \\ x_0 & x_{-1} & \dots & x_{-(p-2)} & x_{-(p)} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{N-2} & x_{N-3} & \dots & x_{N-p} & x_{N-1} \end{bmatrix}$$

(Note G is not fixed before making the measurements, but is made up of observed data pts.

→ cannot directly generate a vector of data using the formula $\underline{x} = G\underline{a} + \underline{e}$ → generated sequentially)

Einstein-Wiener-Khinchin theorem.

- The Einstein-Wiener-Khinchin thm. states that the power spectrum is equal the expected value of the time-normalised DTFT-squared of the signal values, i.e.

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

- consider a time-windowed version of a signal x_n, w/ duration of $2N+1$ samples, denoted by x_n^N,

$$x_n^N = w_n^N x_n, \quad \text{where } w_n^N = \begin{cases} 1 & \text{if } -N \le n \le N \\ 0 & \text{o/w} \end{cases}$$

Its DTFT $X^N(e^{j\omega})$ is just the DFT of a frame of $2N+1$ data pts + linear phase shift

$$\text{DTFT}[x_n^N] = X^N(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w_n^N x_n e^{j\omega n}.$$

$$\therefore |X^N(e^{j\omega})|^2 = X^N(e^{j\omega}) X^N(e^{j\omega})^* = \text{DTFT}[x_n^N * \tilde{x}_n^N] = \text{DTFT}\left[\sum_{m=-\infty}^{\infty} x_m w_m^N x_{m-n} w_{m-n}^N\right]$$

(ensures we have finite energy)

Dividing both sides by the window duration $2N+1$ and taking expectations,

$$E\left[\frac{1}{2N+1} |X^N(e^{j\omega})|^2\right] = E\left[\frac{1}{2N+1} \text{DTFT}\left[\sum_{m=-\infty}^{\infty} x_m w_m^N x_{m-n} w_{m-n}^N\right]\right] = \frac{1}{2N+1} \text{DTFT}\left[\sum_{m=-\infty}^{\infty} r_{xx}[n] w_m^N w_{m-n}^N\right] = \text{DTFT}[r_{xx}[n] * c[n]]$$

where $c[n] = \frac{1}{2N+1} \sum_{m=-\infty}^{\infty} w_m^N w_{m-n}^N$ is the deterministic autocorrelation function of the window function w_n.

Noting that $\text{DTFT}[r_{xx}[n]] = S_x(e^{j\Omega})$ and defining $\text{DTFT}[\epsilon[n]] = T(e^{j\Omega})$, we have

$$E\left[\frac{1}{2N+1}|X(e^{j\Omega})|^2\right] = \text{DTFT}[r_{xx}[n]\,\epsilon[n]] = S_x(e^{j\Omega}) * T(e^{j\Omega})$$

As N increases, $\epsilon[n]$ become flatter and $T(e^{j\Omega})$ tends to a delta function, so the limit as $N\to\infty$ is

$$\lim_{N\to\infty} E\left[\frac{1}{2N+1}|X(e^{j\Omega})|^2\right] = \lim_{N\to\infty} \text{DTFT}[r_{xx}[n]\,\epsilon[n]] = \text{DTFT}[r_{xx}[n]] = S_x(e^{j\Omega})$$

i.e. power spectrum is prop. to expected value of the DTFT squared of the data.

Useful linear algebra/vector calculus properties

- Denote $\nabla J = \frac{\partial J}{\partial \underline{x}}$ as the vector gradient for a column vector $\underline{x}$, we have.

① $J = \underline{b}^T\underline{x} \rightarrow \boxed{\nabla J = \underline{b}}$ for a column vector $\underline{b}$.

$$J = \underline{b}^T\underline{x} = \sum_i b_i x_i \rightarrow \frac{\partial J}{\partial x_i} = b_i \rightarrow \frac{\partial J}{\partial \underline{x}} = \underline{b}.$$

② $J = \underline{x}^T\underline{\underline{B}}\,\underline{x} \rightarrow \boxed{\nabla J = (\underline{\underline{B}} + \underline{\underline{B}}^T)\,\underline{x}}$ for a matrix $\underline{\underline{B}}$.

$$J = \underline{x}^T\underline{\underline{B}}\,\underline{x} = \sum_i x_i \sum_j B_{ij} x_j \rightarrow \frac{\partial J}{\partial x_i} = 2B_{ii}x_i + \sum_{j\neq i}(x_j B_{ji} + B_{ij}x_j) = \sum_j (B_{ij} + B_{ji})x_j$$

$$\therefore \frac{\partial J}{\partial \underline{x}} = (\underline{\underline{B}} + \underline{\underline{B}}^T)\,\underline{x}$$

* for symmetric $\underline{\underline{B}}$, $\underline{\underline{B}}^T = \underline{\underline{B}}$, so $\boxed{\frac{\partial J}{\partial \underline{x}} = (\underline{\underline{B}} + \underline{\underline{B}})\,\underline{x} = 2\underline{\underline{B}}\,\underline{x}}$

- For a matrix in the form $\underline{\underline{B}} = \underline{\underline{G}}^T\underline{\underline{G}}$, it is positive semi-definite, i.e. $\underline{x}^T\underline{\underline{B}}\,\underline{x} \geq 0$ for any $\underline{x}$.

$$0 \leq (\underline{\underline{G}}\underline{x})^T(\underline{\underline{G}}\underline{x}) = \underline{x}^T\underline{\underline{G}}^T\underline{\underline{G}}\,\underline{x} = \underline{x}^T\underline{\underline{B}}\,\underline{x}$$

If $\underline{\underline{G}}$ is full rank, then $\underline{\underline{G}}\underline{x} \neq 0$ for any non-zero $\underline{x}$, so $\underline{\underline{B}}$ is positive definite, i.e. $\underline{x}^T\underline{\underline{B}}\,\underline{x} > 0$ for any $\underline{x} \neq 0$

- The global min. of $J = \underline{x}^T\underline{\underline{B}}\,\underline{x} - 2\underline{b}^T\underline{x}$ wrt. $\underline{x}$ for a symmetric positive definite $\underline{\underline{B}}$ is $\boxed{\underline{x} = \underline{\underline{B}}^{-1}\underline{b}}$

(i) Differentiating J using vector calculus.

$$\frac{\partial J}{\partial \underline{x}} = 2\underline{\underline{B}}\,\underline{x} - 2\underline{b} = 0 \rightarrow \underline{x} = \underline{\underline{B}}^{-1}\underline{b}$$

(ii) Completing the square

$$J = \underline{x}^T\underline{\underline{B}}\,\underline{x} + 2\underline{b}^T\underline{x} = (\underline{x} - \underline{\underline{B}}^{-1}\underline{a})^T\underline{\underline{B}}(\underline{x} - \underline{\underline{B}}^{-1}\underline{a}) - \underline{a}^T\underline{\underline{B}}^{-1}\underline{a} \rightarrow \underline{x} = \underline{\underline{B}}^{-1}\underline{a}.$$

Linear estimation of the general linear model

Least squares (LS) estimator $\underline{\hat{\theta}}^{LS}$

- For the general linear model $\underline{x} = \underline{\underline{H}}\underline{\theta} + \underline{e}$, linear estimation finds the best fit by min. the error J,

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

Expanding using $\underline{e} = \underline{x} - \underline{\underline{H}}\underline{\theta}$,

$$\underline{x}^T\underline{\underline{H}}\underline{\theta} - \underline{\theta}^T\underline{\underline{H}}^T\underline{\underline{H}}\underline{\theta} + \underline{x}^T\underline{x} = (\underline{\underline{H}}\underline{\theta} - \underline{x})^T(\underline{\underline{H}}\underline{\theta} - \underline{x}) = \underline{e}^T\underline{e} = J$$

The global min. of J wrt $\underline{\theta}$ can be found to be $\boxed{\underline{x}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1} = \underline{\hat{\theta}}^{LS}}$

(i) $\frac{\partial J}{\partial \underline{\theta}} = -2\underline{x}^T\underline{\underline{H}} + 2\underline{\theta}^T\underline{\underline{H}}^T\underline{\underline{H}} + 0 = 0 \rightarrow \underline{x}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1} = \underline{\hat{\theta}}^{LS}$

(ii) $\underline{x}^T\underline{x} + \underline{x}^T\underline{\underline{H}}\left(\underline{x}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1}\right) - \left(\underline{x}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1}\right)\underline{\underline{H}}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1}\underline{\underline{H}}^T\underline{x} - \underline{\theta}^T\underline{\underline{H}}^T\underline{\underline{H}}\underline{\theta} + \underline{x}^T\underline{x} = J = \underline{x}^T\underline{\underline{H}}\underline{\theta} - \underline{\theta}^T\underline{\underline{H}}^T\underline{\underline{H}}\underline{\theta} + \underline{x}^T\underline{x}$

$$\therefore \underline{x}^T\underline{\underline{H}}(\underline{\underline{H}}^T\underline{\underline{H}})^{-1} = \underline{\hat{\theta}}^{LS}$$

Properties of the linear estimator

- The OLS estimator of the general linear model is a linear estimator since it is in the form $\underline{\hat{\theta}}^{OLS} = \underline{C}\underline{x}$.
- To investigate its properties, we consider its mean and variance:

① Mean, $E[\underline{\theta}^{OLS}]$

$$E[\underline{x}] = E[\underline{G}\underline{\theta} + \underline{\varepsilon}] = E[\underline{G}\underline{\theta}] + \cancel{E[\underline{\varepsilon}]}^{0} = \underline{G}\underline{\theta}$$

$$\therefore E[\underline{\theta}^{OLS}] = E[(\underline{G}^T\underline{G})^{-1}\underline{G}^T\underline{x}] = (\underline{G}^T\underline{G})^{-1}\underline{G}^T E[\underline{x}] = (\underline{G}^T\underline{G})^{-1}\underline{G}^T\underline{G}\underline{\theta} = \underline{\theta} \rightarrow \text{unbiased}.$$

② covariance, $cov[\underline{\theta}^{OLS}] = E[(\underline{\theta}^{OLS} - E[\underline{\theta}^{OLS}])(\underline{\theta}^{OLS} - E[\underline{\theta}^{OLS}])^T]$

Define the OLS matrix term as $\underline{C} = (\underline{G}^T\underline{G})^{-1}\underline{G}^T$, and consider a general unbiased linear estimator

$$\underline{\hat{\theta}} = \underline{D}\underline{x} \qquad \text{where } \underline{D} = \underline{C} + \underline{\Delta}.$$

For $\underline{\hat{\theta}} = \underline{D}\underline{x}$ to be unbiased, we req.

$$E[\underline{\hat{\theta}}] = E[\underline{D}\underline{x}] = \underline{\theta} \rightarrow E[\underline{D}\underline{x}] = E[(\underline{C}+\underline{\Delta})\underline{x}] = (\underline{C}+\underline{\Delta})E[\underline{x}] = (\underline{C}+\underline{\Delta})\underline{G}\underline{\theta} = \underline{\theta} + \underline{\Delta}\underline{G}\underline{\theta} = \underline{\theta}$$

$\therefore \underline{\Delta}\underline{G} = 0$, since it must work for any $\underline{\theta}$.

Consider the outer product $\underline{x}^T\underline{x}$, and take its expectation,

$\underline{\varepsilon}$ is zero mean white noise w/ variance σ_e^2.

$$\underline{x}\underline{x}^T = (\underline{G}\underline{\theta}+\underline{\varepsilon})(\underline{G}\underline{\theta}+\underline{\varepsilon})^T = \underline{G}\underline{\theta}\underline{\theta}^T\underline{G}^T + \underline{\varepsilon}\underline{\varepsilon}^T + \underline{\varepsilon}\underline{\theta}^T\underline{G}^T + \underline{G}\underline{\theta}\underline{\varepsilon}^T \rightarrow E[\underline{x}\underline{x}^T] = \underline{G}\underline{\theta}\underline{\theta}^T\underline{G}^T + \sigma_e^2\underline{I}$$

(zero mean)

For the linear estimator $\underline{\hat{\theta}} = \underline{D}\underline{x}$, the expectation of the outer product $\underline{\hat{\theta}}\underline{\hat{\theta}}^T$, $E[\underline{\hat{\theta}}\underline{\hat{\theta}}^T]$ is

$$E[\underline{\hat{\theta}}\underline{\hat{\theta}}^T] = E[\underline{D}\underline{x}\underline{x}^T\underline{D}^T] = \underline{D}E[\underline{x}\underline{x}^T]\underline{D}^T = \underline{D}(\underline{G}\underline{\theta}\underline{\theta}^T\underline{G}^T + \sigma_e^2\underline{I})\underline{D}^T = \underline{\theta}\underline{\theta}^T + \sigma_e^2\underline{D}\underline{D}^T.$$

since $\underline{D}\underline{G}\underline{\theta} = (\underline{C}+\underline{\Delta})\underline{G}\underline{\theta} = \underline{\theta} + \cancel{\underline{\Delta}\underline{G}}^{0}\underline{\theta} = \underline{\theta} \rightarrow \underline{D}\underline{G}\underline{\theta}\underline{\theta}^T\underline{G}^T\underline{D}^T = \underline{\theta}\underline{\theta}^T$.

$\underline{\Delta}\underline{G} = 0 \Rightarrow \underline{\Delta}^T = 0$, $-(\underline{G}^T\underline{G})^{-1}\cancel{\underline{G}^T\underline{\Delta}^T}^{0} = 0$

$$\therefore cov[\underline{\hat{\theta}}] = E[\underline{\hat{\theta}}\underline{\hat{\theta}}^T] - \underline{\theta}\underline{\theta}^T = \underline{\theta}\underline{\theta}^T + \sigma_e^2\underline{D}\underline{D}^T - \underline{\theta}\underline{\theta}^T = \sigma_e^2(\underline{C}+\underline{\Delta})(\underline{C}+\underline{\Delta})^T = \sigma_e^2(\underline{C}\underline{C}^T + \underline{C}\underline{\Delta}^T + \underline{\Delta}\underline{C}^T + \underline{\Delta}\underline{\Delta}^T)$$

$$= \sigma_e^2(\underline{C}\underline{C}^T + \underline{\Delta}\underline{\Delta}^T) = \sigma_e^2((\underline{G}^T\underline{G})^{-1} + \underline{\Delta}\underline{\Delta}^T) = cov[\underline{\theta}^{OLS}] + \sigma_e^2\underline{\Delta}\underline{\Delta}^T$$

since w/ $cov[\underline{\theta}^{OLS}] = cov[\underline{\hat{\theta}}]|_{\underline{\Delta}=0} = \sigma_e^2(\underline{G}^T\underline{G})^{-1}$

→ Denoting $\theta_{i,i}$ as the ith diagonal element of $cov[\underline{\hat{\theta}}]$, and by construction, the diagonal elements of $\underline{\Delta}\underline{\Delta}^T$ are also ≥ 0 → $var[\hat{\theta}_i] \geq var[\theta_i^{OLS}]$ for each $i = 0, \ldots, p-1$, w/ equality when $\underline{\Delta} = 0$.

- This means the OLS estimator $\underline{\theta}^{OLS}$ is the MVU estimator of $\underline{\theta}$. Such an estimator is a Best Linear Unbiased Estimator (BLUE).
- By considering the expression for $cov[\underline{\hat{\theta}}]$, we can see that $\underline{\theta}^{OLS}$ is the unique BLUE for the general linear model.
- The only assumption used is that the noise process $\{e_n\}$ is zero mean and white (we did not assume any probability distribution for $\{e_n\}$).
- However, if we do know the distribution of $\{e_n\}$, it is possible that the linear estimator can be beaten by a nonlinear estimation method.
- If in addition, we have prior probability information about $\underline{\theta}$, $p(\underline{\theta})$, then a Bayesian estimator can given even better mean squared performance at the cost of some bias in the estimate.

Likelihood-based inference

Error sequence $\underline{e}$ and the linear Gaussian model

- We assume the error sequence $\underline{e}$ to be drawn from an iid noise distribution w/ pdf $p_e(e)$, so

$$p_{\underline{e}}(\underline{e}) = p_e(e_0)\,p_e(e_1)\dots p_e(e_{N-1})$$

and we also assume the noise process is white (though not necessarily zero mean).

- For the linear Gaussian model (Gauss-Markov model), p_e is the zero-mean Gaussian distribution.

$$p_e(x) = N(x \mid M=0, \sigma^2=\sigma_e^2) = \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{x^2}{2\sigma_e^2}}$$

The error sequence $\underline{e}$ thus has a joint pdf $p(\underline{e})$, given by

$$p(\underline{e}) = \prod_{n=0}^{N-1} N(e_n \mid M=0, \sigma^2=\sigma_e^2) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{e_n^2}{2\sigma_e^2}} = \frac{1}{(2\pi\sigma_e^2)^{N/2}} e^{-\frac{1}{2\sigma_e^2}\sum_{n=0}^{N-1} e_n^2} = \frac{1}{(2\pi\sigma_e^2)^{N/2}} e^{-\frac{\underline{e}^T\underline{e}}{2\sigma_e^2}} = N(\underline{e} \mid \underline{M}=\underline{0}, \underline{C}_{\underline{e}}=\sigma_e^2\underline{I})$$

- Note the multivariate Gaussian density function for a length N random column vector $\underline{x}$, $f_{\underline{x}}(\underline{x})$ is

$$f_{\underline{x}}(\underline{x}) = \frac{1}{(2\pi)^{N/2}|\underline{C}_{\underline{x}}|^{1/2}} \exp\left(-\tfrac{1}{2}(\underline{x}-\underline{M})^T \underline{C}_{\underline{x}}^{-1}(\underline{x}-\underline{M})\right)$$

where $\underline{M} = E[\underline{x}]$ is the mean vector and $\underline{C}_{\underline{x}} = E[(\underline{x}-\underline{M})(\underline{x}-\underline{M})^T]$ is the covariance matrix.

(Elementwise, we have $[\underline{C}_{\underline{x}}]_{ij} = C_{x_i x_j}$, the covariance btwn elements x_i and x_j.)

Maximum likelihood (ML) estimator $\underline{\theta}^{ML}$

- The ML estimator treats the parameters $\underline{\theta}$ as unknown constant which we incorporate no prior information. The observed data $\underline{x}$ is considered random and we can often obtain the pdf for $\underline{x}$ when $\underline{\theta}$ is known.

- This pdf is termed the likelihood $L(\underline{x};\underline{\theta})$, defined as:

$$L(\underline{x};\underline{\theta}) = p(\underline{x}\,|\,\underline{\theta}).$$

- The ML estimate for $\underline{\theta}$ is then the value of $\underline{\theta}$ which max the likelihood for given observation $\underline{x}$.

$$\boxed{\underline{\theta}^{ML} = \arg\max_{\underline{\theta}} p(\underline{x}\,|\,\underline{\theta})}.$$

i.e. the ML soln corresponds to the parameter vector which would have generated $\underline{x}$ w/ highest probability.

- We typically max the log-likelihood function $l(\underline{x};\underline{\theta}) = \log(L(\underline{x};\underline{\theta}))$ rather than $L(\underline{x};\underline{\theta})$ itself for convenience. Since $\log(\cdot)$ is a monotonically increasing function, the two solns are identical.

- For the linear Gaussian model, $\underline{x} = \underline{\underline{G}}\underline{\theta} + \underline{e}$, we do a vector change of variable $\underline{e} \to \underline{x}$ to get the likelihood $p(\underline{x}\,|\,\underline{\theta})$. (We are conditioning on $\underline{\theta}$ → treat $\underline{\underline{G}}\underline{\theta}$ as a const. term for change of variable)

$$\frac{\partial x_i}{\partial e_j} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases} \quad \to \quad \underline{\underline{J}}^T = \underline{\underline{I}} \quad \to \quad p(\underline{x}\,|\,\underline{\theta}) = p_e(\underline{x}-\underline{\underline{G}}\underline{\theta})\left|\underline{\underline{I}}\right| = p_e(\underline{x}-\underline{\underline{G}}\underline{\theta})$$

Expanding this out, we get

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

$$\therefore l(\underline{x};\underline{\theta}) = \log L(\underline{x};\underline{\theta}) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{1}{2\sigma_e^2}(\underline{x}-\underline{\underline{G}}\underline{\theta})^T(\underline{x}-\underline{\underline{G}}\underline{\theta}) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{1}{2\sigma_e^2}\underline{e}^T\underline{e} = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{J}{2\sigma_e^2}$$

→ Maximising $l(\underline{x};\underline{\theta})$ wrt $\underline{\theta}$ is equivalent to minimising $J = \underline{e}^T\underline{e}$ wrt $\underline{\theta}$ → same as OLS estimation, so

$$\boxed{\underline{\theta}^{ML} = \underline{\theta}^{OLS} = (\underline{\underline{G}}^T\underline{\underline{G}})^{-1}\underline{\underline{G}}^T\underline{x}}$$

- The OLS soln is a special case of the ML soln where the error process $\{e_n\}$ is zero-mean, independent and Gaussian w/ fixed variance.

Estimating the noise variance

- The noise variance σ_e^2 of the linear Gaussian model can be estimated by ML.
- consider the log-likelihood function at $\underline{\theta} = \underline{\theta}^{ML}$, now also considered as a function of σ_e^2.

$$l(\underline{x}; \underline{\theta} = \underline{\theta}^{ML}, \sigma_e^2) = \log L(\underline{x}; \underline{\theta} = \underline{\theta}^{ML}, \sigma_e^2) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{1}{2\sigma_e^2}(\underline{x} - \underline{\underline{\theta}}\underline{\theta}^{ML})^T(\underline{x} - \underline{\underline{\theta}}\underline{\theta}^{ML}) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{J^{ML}}{2\sigma_e^2}$$

where $J^{ML} = (\underline{x} - \underline{\underline{\theta}}\underline{\theta}^{ML})^T(\underline{x} - \underline{\underline{\theta}}\underline{\theta}^{ML})$ is the min. squared error term corresponding to ML optimisation.

Differentiating wrt σ_e^2 and setting to zero,

$$\frac{\partial I(\underline{x}; \underline{\theta} = \underline{\theta}^{ML}, \sigma_e^2)}{\partial \sigma_e^2} = -\frac{N/2}{\sigma_e^2} + \frac{J^{ML}}{2(\sigma_e^2)^2} = 0 \quad \rightarrow \quad \boxed{\sigma_e^{2ML} = \frac{J^{ML}}{N}}$$

i.e. the noise variance σ_e^{2ML} is estimated as the mean squared error at the ML parameter soln.

Bayesian inference

Bayesian methods

- The ML method treat parameters $\underline{\theta}$ as unknown const. If we instead treat the parameters $\underline{\theta}$ as RVs, we can assign a prior pdf $p(\underline{\theta})$. (key diff btwn likelihood-based / Bayesian inference)
- The prior pdf $p(\underline{\theta})$ should broadly express some prior knowledge abt. the rel. probability of diff. parameter values $\underline{\theta}$ before the data $\underline{x}$ are observed. (If nothing abt the parameters $\underline{\theta}$ is known a priori, $p(\underline{\theta})$ should not express initial preference for a set of parameters over others).
- the precise form assigned to $p(\underline{\theta})$ req. careful consideration since misleading results can be obtained from erroneous priors.
- We can use Bayes' thm. to find the posterior probability $p(\underline{\theta}|\underline{x})$, which can be used to estimate random parameters $\underline{\theta}$ from a random vector $\underline{x}$ of observations.

$$p(\underline{\theta}|\underline{x}) = \frac{p(\underline{x}|\underline{\theta})\,p(\underline{\theta})}{p(\underline{x})}$$

(Note $p(\underline{x}|\underline{\theta})$ is the likelihood; $p(\underline{\theta})$ is the prior; $p(\underline{x})$ is the marginal likelihood)

- The generation of the posterior from the prior when data $\underline{x}$ is observed can be thought of as a refinement to any previous knowledge abt. the parameters $\underline{\theta}$. ($p(\underline{\theta}) \xrightarrow{\text{data } \underline{x}} p(\underline{\theta}|\underline{x})$).
- The marginal likelihood $p(\underline{x})$ is useful in model selection probs., but is const. for any given observation $\underline{x}$ → ignored if we only want to optimise the posterior probabilities wrt $\underline{\theta}$.

$$\boxed{p(\underline{\theta}|\underline{x}) \propto p(\underline{x}|\underline{\theta})\,p(\underline{\theta})}$$

- $p(\underline{x})$ could be calculated using $p(\underline{x}) = \int p(\underline{x}, \underline{\theta})\,d\underline{\theta} = \int p(\underline{x}|\underline{\theta})\,p(\underline{\theta})\,d\underline{\theta}$, and effectively serves as the normalising const. for the posterior density $p(\underline{\theta}|\underline{x})$.
- All the distributions above are implicitly conditioned upon all prior modelling assumptions, so some texts adopt the notation $p(\underline{\theta}|\underline{x}, M)$, where M denotes all of the modelling and distributional assumptions that are being made (not just the assumed prior $p(\underline{\theta})$).
- LS / ML methods only provide a pt-estimate of $\underline{\theta}$, whereas Bayesian methods produce a posterior pdf w values defined for all $\underline{\theta}$ → fully interpretable probability distribution

Maximum a posteriori (MAP) estimator $\underline{\theta}^{MAP}$

- The MAP estimate for $\underline{\theta}$ is the value of $\underline{\theta}$ which max. the posterior for given observations $\underline{x}$ and prior $p(\underline{\theta})$

$$\underline{\theta}^{MAP} = \arg\max_{\underline{\theta}} p(\underline{\theta}|\underline{x})$$

i.e. the MAP estn corresponds to the parameter vector $\underline{\theta}$ inferred from $\underline{x}$ w/ highest probability

- For the linear Gaussian model, suppose the prior $p(\underline{\theta})$ is the multivariate Gaussian, i.e.

$$p(\underline{\theta}) = N(\underline{\theta}|\underline{m}_0, \underline{C}_0) = \frac{1}{(2\pi)^{p/2}|\underline{C}_0|^{1/2}} \exp\left(-\frac{1}{2}(\underline{\theta}-\underline{m}_0)^T \underline{C}_0^{-1}(\underline{\theta}-\underline{m}_0)\right)$$

where p is the no. of parameters in $\underline{\theta}$.

(Note if we don't know much a priori, we usually select $\underline{m}_0 = 0$ and $\underline{C}_0 = \infty \underline{I}_p$)

The likelihood $p(\underline{x}|\underline{\theta})$ takes the same form as before for the ML estimator, so the posterior $p(\underline{\theta}|\underline{x})$ is

$$p(\underline{\theta}|\underline{x}) \propto p(\underline{\theta})\, p(\underline{x}|\underline{\theta}) = \frac{1}{(2\pi)^{p/2}|\underline{C}_0|^{1/2}} \exp\left(-\frac{1}{2}(\underline{\theta}-\underline{m}_0)^T \underline{C}_0^{-1}(\underline{\theta}-\underline{m}_0)\right) \cdot \left(\frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}}\right) \exp\left(-\frac{1}{2\sigma_\epsilon^2}(\underline{\theta}\underline{F}-\underline{x})^T(\underline{\theta}\underline{F}-\underline{x})\right)$$

Considering $-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x}))$, we have

$$-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x})) = (\underline{\theta}\underline{F}-\underline{x})^T(\underline{\theta}\underline{F}-\underline{x}) + (\underline{\theta}-\underline{m}_0)^T \underline{C}_0^{-1}(\underline{\theta}-\underline{m}_0) + \text{const.}$$

$$= \underline{\theta}\underline{F}^T\underline{F}\underline{\theta}^T - \underline{\theta}\underline{F}^T\underline{x} - \underline{x}^T\underline{F}\underline{\theta}^T + \underline{x}^T\underline{x} + \underline{\theta}\underline{C}_0^{-1}\underline{\theta}^T - \underline{\theta}\underline{C}_0^{-1}\underline{m}_0 - \underline{m}_0^T\underline{C}_0^{-1}\underline{\theta}^T + \underline{m}_0^T\underline{C}_0^{-1}\underline{m}_0 + \text{const.}$$

(i) Differentiating wrt $\underline{\theta}$ and setting to zero,

$$\frac{\partial(-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x})))}{\partial \underline{\theta}} = 2\underline{\theta}\underline{F}^T\underline{F} - 2\underline{x}^T\underline{F} + 0 + 2\underline{\theta}\underline{C}_0^{-1} - 2\underline{m}_0^T\underline{C}_0^{-1} + 0 = 0$$

$$\therefore (2\sigma_\epsilon^2) \quad \underline{x}^T\underline{F} + \underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 = \underline{\theta}(\underline{F}^T\underline{F} + \underline{C}_0^{-1}\sigma_\epsilon^2) \quad \rightarrow \quad \boxed{\underline{\theta}^{MAP} = (\underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{x}^T\underline{F})(\underline{F}^T\underline{F} + \underline{C}_0^{-1}\sigma_\epsilon^2)^{-1}}$$

(ii) Grouping terms in the form $\underline{\theta}\underline{A}\underline{\theta}^T - \underline{\theta}\underline{B}^T$ then completing the square

$$-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x})) = \underline{\theta}(\underline{F}^T\underline{F} + \underline{C}_0^{-1}\sigma_\epsilon^2)\underline{\theta}^T - \underline{\theta}(\underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{x}^T\underline{F})^T + \text{const.}$$

$$= \underline{\theta}(\underline{F}^T\underline{F} + \underline{C}_0^{-1}\sigma_\epsilon^2)\underline{\theta}^T - \underline{\theta}(\underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{x}^T\underline{F}) - (\underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{x}^T\underline{F})^T\underline{\theta}^T + \text{const.}$$

$$= (\underline{\theta}\underline{A}^T - \underline{q}^T\underline{A}^{-T})\underline{A}(\underline{\theta}^T - \underline{A}^{-1}\underline{q}) - \underline{q}^T\underline{A}^{-1}\underline{q} + \text{const.} = (\underline{\theta}\underline{A}^T - \underline{b}^T)\underline{A}(\underline{\theta}^T - \underline{A}^{-1}\underline{b}) + \text{const.}$$

$\therefore \underline{\theta}^{MAP} = \underline{A}^{-1}\underline{b}$, where $\underline{A} = (\underline{F}^T\underline{F} + \underline{C}_0^{-1}\sigma_\epsilon^2)$, $\underline{b} = (\underline{m}_0^T\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{x}^T\underline{F})$

(Note we assumed $\underline{F}^T\underline{F}$ is invertible, which is the case if $\underline{F}$ is full rank)

- By completing the square, for $-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x}))$ and considering the exponent of $p(\underline{\theta}|\underline{x})$,

$$-2\sigma_\epsilon^2 \log(p(\underline{\theta}|\underline{x})) = (\underline{\theta}-\underline{\theta}^{MAP})^T \underline{A}\,(\underline{\theta}-\underline{\theta}^{MAP}) + \text{const.}$$

$\therefore$ Exponent of $p(\underline{\theta}|\underline{x}) = -\frac{1}{2\sigma_\epsilon^2}(\underline{\theta}-\underline{\theta}^{MAP})^T \underline{A}\,(\underline{\theta}-\underline{\theta}^{MAP}) \rightarrow$ multivariate Gaussian $N(\underline{m}_\theta^{post}, \underline{C}_\theta^{post})$,

$$\underline{m}_\theta^{post} = \underline{\theta}^{MAP}, \quad \underline{C}_\theta^{post} = \left(\frac{\underline{A}}{2\sigma_\epsilon^2}\right)^{-1} = \sigma_\epsilon^2 \underline{A}^{-1}$$

since the remaining terms do not depend on $\underline{\theta}$, so the multivariate density function must be proportional

We can conclude the posterior $p(\underline{\theta}|\underline{x})$ is itself a multivariate Gaussian

$$\boxed{p(\underline{\theta}|\underline{x}) = N(\underline{\theta}|\underline{\theta}^{MAP}, \sigma_\epsilon^2 \underline{A}^{-1})}$$

- Comparing $\underline{\theta}^{MAP} = (\underline{C}_0^{-1}\sigma_\epsilon^2 + \underline{F}^T\underline{F})^{-1}(\underline{C}_0^{-1}\sigma_\epsilon^2\underline{m}_0 + \underline{F}^T\underline{x})$ w/ $\underline{\theta}^{ML} = (\underline{F}^T\underline{F})^{-1}\underline{F}^T\underline{x}$, we can see in the limit $\underline{C}_0^{-1} = 0$, $\underline{\theta}^{MAP} = \underline{\theta}^{ML}$

($\underline{C}_0^{-1} = 0 \rightarrow \sigma_{\theta_i}^2 \rightarrow \infty$, i.e. the prior tends to be flat/uniform (equiprobable $\underline{\theta}$) $\rightarrow$ impose no prior information)

- The ML estn is a special case of the MAP estn, where a uniform prior is assigned to $\underline{\theta}$.

- The MAP estimate also tends to the ML-estimate when the likelihood is strongly peaked around its max, compared w the prior (i.e. $\sigma_\epsilon^2 \ll \sigma_\theta^2$). This is the case for large N.

Minimum mean-squared error (MMSE) estimator. θ^{MMSE}

- consider a cost function $C(\hat{\underline{\theta}},\underline{\theta})$ which expresses the cost of estimating the parameter as $\hat{\underline{\theta}}$ when the true value $\underline{\theta}$. ($C(\hat{\underline{\theta}},\underline{\theta})$ should be non-negative and usually satisfies $C(\hat{\underline{\theta}},\underline{\theta})=0$).
- The form of the cost function $C(\hat{\underline{\theta}},\underline{\theta})$ depends on the req. of the problem. A cost of 0 indicates that the estimate is perfect for our req. (not necessarily $\hat{\underline{\theta}}=\underline{\theta}$), while the costs indicate poorer estimates.
- The expected cost over all of the unknown parameters, conditioned on the observed data $\underline{x}$ is

$$E[C(\hat{\underline{\theta}},\underline{\theta})|\underline{x}=\underline{x}] = \int_{\underline{\theta}} C(\hat{\underline{\theta}},\underline{\theta})\, p(\underline{\theta}|\underline{x})\, d\underline{\theta}$$

- The MMSE estimate for $\underline{\theta}$ is the value of $\hat{\underline{\theta}}$ which min. the average squared error for given data $\underline{x}$.

$$\theta^{MMSE} = \underset{\hat{\underline{\theta}}}{\operatorname{argmin}}\, E[(\hat{\underline{\theta}}-\underline{\theta})^2|\underline{x}=\underline{x}]$$

(Note the cost function $C(\hat{\underline{\theta}},\underline{\theta})$ is $C(\hat{\underline{\theta}},\underline{\theta})=(\hat{\underline{\theta}}-\underline{\theta})^2$)

Any small bias introduced can be tolerated provided the MSE is low.

- In general, the MSE can be expressed as.

$$J = E[(\hat{\underline{\theta}}-\underline{\theta})^2|\underline{x}=\underline{x}] = \int_{\underline{\theta}} (\hat{\underline{\theta}}-\underline{\theta})^2 p(\underline{\theta}|\underline{x})\, d\underline{\theta}$$

Differentiating wrt $\hat{\underline{\theta}}$ and setting to zero,

$$\frac{dJ}{d\hat{\underline{\theta}}} = \frac{d}{d\hat{\underline{\theta}}}\int_{\underline{\theta}} (\hat{\underline{\theta}}-\underline{\theta})^2 p(\underline{\theta}|\underline{x})\, d\underline{\theta} = \int_{\underline{\theta}} \frac{d}{d\hat{\underline{\theta}}}(\hat{\underline{\theta}}-\underline{\theta})^2 p(\underline{\theta}|\underline{x})\, d\underline{\theta} = \int_{\underline{\theta}} 2(\hat{\underline{\theta}}-\underline{\theta})\, p(\underline{\theta}|\underline{x})\, d\underline{\theta} = 0$$

$$\therefore \int_{\underline{\theta}} \hat{\underline{\theta}}\, p(\underline{\theta}|\underline{x})\, d\underline{\theta} = \int_{\underline{\theta}} \underline{\theta}\, p(\underline{\theta}|\underline{x})\, d\underline{\theta} \quad \rightarrow \quad \boxed{\hat{\underline{\theta}}^{MMSE} = E[\underline{\theta}|\underline{x}=\underline{x}] = \int_{\underline{\theta}} \underline{\theta}\, p(\underline{\theta}|\underline{x})\, d\underline{\theta}}$$

since $\int_{\underline{\theta}} \hat{\underline{\theta}}\, p(\underline{\theta}|\underline{x})\, d\underline{\theta} = \hat{\underline{\theta}} \int_{\underline{\theta}} p(\underline{\theta}|\underline{x})\, d\underline{\theta} = \hat{\underline{\theta}}$

- For the linear Gaussian model, the posterior distribution is $p(\underline{\theta}|\underline{x}) = N(\underline{\theta}|\underline{\theta}^{MAP}, \sigma_\varepsilon^2 \underline{\underline{\Theta}}^{-1})$, so

$$\hat{\underline{\theta}}^{MMSE} = E[\underline{\theta}|\underline{x}=\underline{x}] = \underline{\theta}^{MAP} = (\underline{\underline{G}}^T\underline{\underline{G}} + \sigma_\varepsilon^2\underline{\underline{\Theta}}^{-1})^{-1}(\underline{\underline{G}}^T\underline{x} + \sigma_\varepsilon^2\underline{\underline{\Theta}}^{-1}\underline{\mu}_\theta)$$

Summary of estimators

① Least squares (θ^{LS})

✓: Req. no knowledge of probability distributions
✓: Usually simplest to implement
✓: Guarantee of performance as BLUE estimator (not necessarily better than non-linear estimators)
X: cannot incorporate prior knowledge of $p(\underline{\theta})$

② Maximum likelihood (θ^{ML})

✓: Guarantee of performance when N is large
X: Req. knowledge of noise probability distribution
X: Cannot incorporate prior knowledge of $p(\underline{\theta})$.
X: Harder to implement than LS for non-Gaussian case.

③ Bayesian (θ^{MAP}, θ^{MMSE})

✓: Incorporates prior knowledge of $p(\underline{\theta})$.
✓: Guarantee of performance for any N (given correct $p(\underline{\theta})$)
X: Req. knowledge of noise probability distribution
X: Req. knowledge of prior distribution $p(\underline{\theta})$.
X: Can be harder to implement than LS/ML