2102-502
Random Signals and Systems

Text "*Random Signals and Systems*", Richard E. Mortensen, Wiley. Chapter 1 – 6, 8 and 11.

Chapter 1
Discussion of Probability and Stochastic Processes

Probability is a quantity of the uncertainty that a random event will occur. Other similar quantities such as *fuzzy*, *plausibility*, *believe*, etc. will do the same job but with different properties.

Event is a set of some ***elementary events***. This random event occurred if one (*and only one*) of the elementary events in the set occurred. Two events can occur at the same time if they share the same elementary event that occurred.

Sample Space is a universal set that includes all possible elementary events. Any outcome must be member of this set but not all elementary events in this set may occur (some zero probability event).

Questions

- Is the probability really exist in real physical world in the same way as Newton's law of motion?
- What is the mechanism that tries to balance the number of Head and Tail in flipping a balanced coin for 10^6 times? God?
- Can someone prove or find the correct value of the probability for any event?

Answers

- Probability is just a mathematics concept or model that try to fit, explain and predict the complex, uncertain real world.
- We can use other models to fit and explain the real world as well, *e.g.*, fuzzy, if appropriated. There is no the genuine or best model.
- We can confirm the correctness of probability value up to some accuracy by comparing the *predictions* with the *experiments*. But we still do not know the true probabilities any way.
- The probability value will be *assigned* to all events and if this gives good enough predictions compared to the experiments or observations, then we can accept that assignment.

- The ***frequency of occurrence*** for large number of experiments can be thought of as the probability estimation as well as the physical interpretation of the probability itself. But in many cases, we can not perform many experiments and this type of interpretation may not be useful in real life. *For example,* to estimate the probability of one particular student getting A from this course, he must take this course for 100 times and if he gets A for 20 times then his probability is 0.2 !?!?!?!

Examples

- Flip 10 coins: {(TTT…T), (HTT…T), (THT…T), (HHT…T), …, (HHH…H)} with 1,024 events
- No. of Heads in flipping 10 coins: {0, 1, …, 10}
- Sum of 2 dices: {2, 3, …, 12}
- Times of coin flipping until 1st Head: {1, 2, …, ∞}
- Fractional part of body weight in kg: [0,1)
- Noise voltage: (-∞,∞), the whole real line

***Finite Case*:** We can assign arbitrary probabilities to all elementary events such that $0 \le \text{prob} \le 1$ and total sum of all probabilities is 1. Then we can find the probabilities of any set as the sum of elementary probabilities without any conflict.

***Infinite Case*:** The probability of each elementary event is $\to 0$ and we can only assign probabilities to some subset of elementary events. But we will not be able to find the probabilities of all subsets. Only some subsets, called *admissible subset*, can have the probabilities.

***Probability Trio* (Ω, $\mathcal{A}$, P)** is the defined entities to support the use of probability space.

Ω: sample space with *finite*, *countably infinite* or *uncountably infinite* number of elementary events.

$\mathcal{A}$: a family of *admissible subsets* (events) of Ω to which the probabilities can be assigned, also called ***Borel field*** or ***σ-field***.

- *Finite case*: all $2^{|\Omega|}$ subsets of Ω (power set)
- *Infinite case*: only some subsets of Ω

P: the probability measures (value) assigned to all admissible subsets satisfying the following *Axiom*

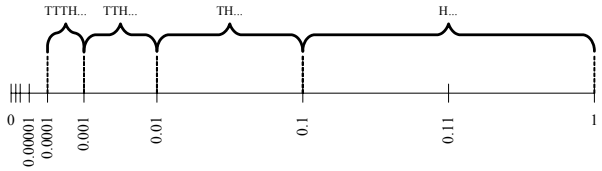
$$\begin{aligned} &P(\phi) = 0 \\ &P(\Omega) = 1 \\ &P(A_k) \ge 0,\ A_k \in \mathcal{A} \\ &P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k),\ \textit{for}\ \ A_i \cap A_j = \phi \end{aligned}$$

Examples

Flipping a coin infinite times will have a sample space of countably infinite number of outcomes, and each outcome can be represented by ∞ bits sequence. The probability of each sequence will be zero and we can only assign P to some subsets of sequences. One possible family of subsets is the power set of

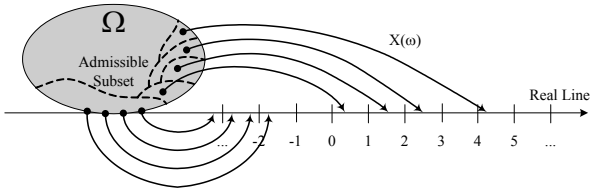
B = {H…, TH…, TTH…, TTTH…, TTTTH…, …}

which is also countably infinite set of subsets of Ω. The probabilities assigned to members (subsets) of B are 1/2 , 1/4 , 1/8 , 1/16 , 1/32 , …respectively if the coin is balanced. Note that the subsets are disjointed and the probability of any combinations (union) of these subsets will be the sum of their probabilities. But, still, there exists infinitely many subsets of Ω that we can not find the probabilities from this given probability trio. For example, subsets of all outcomes which is irrational number in binary representation, *e.g.*, HTHH… represented as 0.1011… .



Random Variables are the real numbers X assigned (or mapped) to any elementary events ω of Ω, or as a real function X(ω). A set of elementary events then can be represented by a set of real numbers with the same probability. For *infinite case*, we cannot assign random variables X(ω) arbitrarily. But the assignment for the ***admissible random variables*** must agree with the *admissible subset* $\mathcal{A}$ as following. The assignment must allow us to find the probabilities of all intervals, I = (-∞, a] for all a, on the real line.

If $X^{-1}(I) = \{\omega \in \Omega \mid X(\omega) \in I\}$ is a set of random events represented by real interval I. Then X is an *admissible random variable* if $X^{-1}(I) \in \mathcal{A}$ for all a. This will guarantee the existence of the probabilities ***cdf (Cumulative Distribution Function)*** defined as $F_X(a) = P\{-\infty < X \le a\} = P(I) = P\{X^{-1}(I)\}$.



If Ω is uncountably infinite, F_X will be differentiable and define ***pdf (Probability Density Function)*** as

$$f_X(x) = \frac{d}{dx} F_X(x)$$

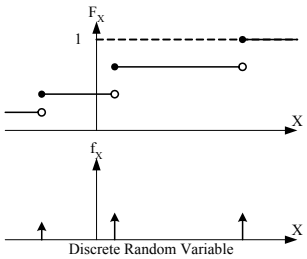
$$P\{a < X \le b\} = \int_a^b f_X(x)dx$$

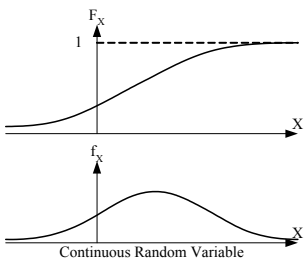
Properties of cdf

1. Non-decreasing;
   $a < b \Rightarrow F_X(a) \le F_X(b)$
2. $\lim_{x \to +\infty} F_X(x) = 1$
   $\lim_{x \to -\infty} F_X(x) = 0$
3. Continuous from the right, any discontinuity has upper value

Properties of pdf

1. Non-negative but can be greater than 1
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. Discontinuity in cdf will give delta function pdf





Expected Value or Mean *of a random variable is the best guess that gives minimum average square error.*

Discrete:

$$\mu = \sum_{\omega \in \Omega} X(\omega) P\{\omega \in \Omega \mid X(\omega) = a_k\} = E[X]$$

$$= \sum_{k=1}^{n} a_k P_k = \textit{Lebesque Integral}$$

Continuous:

$$\mu = \int_{-\infty}^{\infty} x f_X(x)dx = E[X]$$

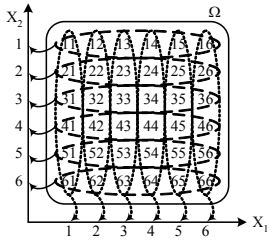
If X is a random variable then Y = g(X) is also a random variable. The ***expected value of function g(X)*** is defined as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx =$$

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y)dy = \mu_Y$$

It is possible to have many random variables defined on the same *probability trio* (Ω, $\mathcal{A}$, P). Define the ***Joint Cumulative Distribution Function*** as

$$F_{X_1,X_2,\ldots,X_n}(a_1,a_2,\ldots,a_n) = P\{-\infty < x_1 \le a_1, -\infty < x_2 \le a_2, \ldots, -\infty < x_n \le a_n\}$$



And if jointly differentiable with respect to all arguments then, define the ***Joint Density Function*** as

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$$

We can use ***vector-valued random variable*** notation for jointly distributed random variables as $\underline{X} = [x_1\ x_2\ \ldots\ x_n]^T$ with $f_{\underline{X}}(\underline{x})$ as joint density function.

Independence and Conditional Probability

Given *probability trio* (Ω, $\mathcal{A}$, P), let A, B $\in \mathcal{A}$ and if

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

we call A and B ***independent***. In general, if P(B) ≠ 0, define ***conditional probability*** of A given B as

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

If A and B are *independent* then P(A|B) = P(A) which means that event B has no influence on event A.

Bivariate: Given X,Y defined on the same (Ω, $\mathcal{A}$, P) with *joint density* $f_{XY}(x,y)$. Define ***marginal density*** as

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy : \textit{to eliminate } Y$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx : \textit{to eliminate } X$$

If X and Y are *independent* then $f_{XY}(x,y) = f_X(x)f_Y(y)$. Define ***conditional density*** of X given Y=y as

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

and if X and Y are *independent* then $f_{X|Y}(x|y) = f_X(x)$.

Multivariate: $X_1, X_2, \ldots, X_n$ defined on the same (Ω, $\mathcal{A}$, P) with *joint density* $f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$.
Let 1 < m < n, and define the ***conditional density*** of $X_{m+1}, X_{m+2}, \ldots, X_n$ given $X_1, X_2, \ldots, X_m$ as

$$f_{X_{m+1},\ldots,X_n|X_1,\ldots,X_m}(x_{m+1},x_{m+2},\ldots,x_n \mid x_1,x_2,\ldots,x_m) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)}{f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m)}$$

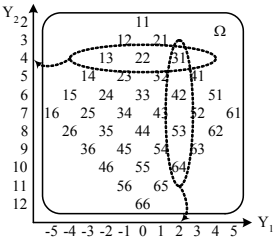
where the *marginal density*

$$f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_{m+1}dx_{m+2}\cdots dx_n$$

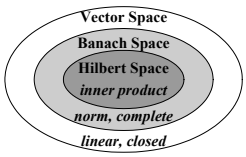
If $X_1, \ldots, X_m$ are *independent* from $X_{m+1}, \ldots, X_n$ then

$$f_{X_{m+1},X_{m+2},\ldots,X_n|X_1,X_2,\ldots,X_m} = f_{X_{m+1},X_{m+2},\ldots,X_n}$$

$P(Y_2=4|Y_1=2)$
$= P(Y_2=4,Y_1=2)/P(Y_1=2)$
$= (1/36)/(4/36)$
$= 1/4$
$\neq P(Y_2=4) = 3/36$



Hilbert Space of Second-Order Random Variables



Linear Vector Space:
closed, associative, commutative, distributive, identity, additive inverse

Banach Space: Complete ≡ *every Cauchy sequence converges in the normed vector space*. Norm of vector is a measure of distance or length of vector.

$$\|X\| \ge 0;\ \|X\| = 0 \ \textit{iff}\ X = 0;\ \|\alpha X\| = |\alpha| \cdot \|X\|;$$
$$\|X + Y\| \le \|X\| + \|Y\|$$

Hilbert Space: Inner product of two vectors measures the orientation between them. Normed vector space induced by inner product is complete.

$$\langle X, Y\rangle^* = \langle Y, X\rangle;\ \langle X, X\rangle > 0 \Rightarrow X \neq 0;$$
$$\langle \alpha X, Y\rangle = \alpha\langle X, Y\rangle;\ \langle X+Y, Z\rangle = \langle X, Z\rangle + \langle Y, Z\rangle;$$
$$\textit{Induced}\quad \textit{Norm}: \|X\| = \langle X, X\rangle^{1/2}$$

Second-Order Random Variable: random variables with finite *second moment* (*variance*, *power*, *size*) as

$$\mu_2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx < \infty$$

The vector space of *second-order random variables* with the inner product below is ***Hilbert Space***.

$$\begin{aligned} &Inner\ \ product: \langle X, Y \rangle = E[XY] \\ &Induced\ \ Norm: \|X\| = \langle X, X \rangle^{1/2} = \sqrt{E[X^2]} < \infty \\ &Schwarz\ \ inequality: |\langle X, Y \rangle| \le \|X\| \cdot \|Y\| \end{aligned}$$

Linearly Independent: if and only if

$$c_1X_1 + c_2X_2 + \ldots + c_nX_n = 0 \text{ implies } c_1 = c_2 = \ldots = c_n = 0$$

which means all X_k can not be expressed as a linear summation of other X's.

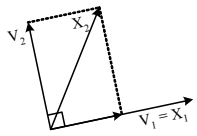
Statistical Independent: if and only if

$$F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} F_{X_k}(x_k)$$

If all X_k have $\mu = 0$ and $\mu_2 < \infty$ then statistical independence implies linearly independence too, but not conversely. For example, X and X^2 are linearly independent but not statistical independent.

Gram-Schmidt Orthogonalization

Given $X_1, X_2, \ldots, X_n$, find a set of orthogonal vectors (inner product = 0)



$$\begin{aligned} V_1 &= X_1 \\ V_2 &= X_2 - \frac{\langle X_2, V_1 \rangle}{\langle V_1, V_1 \rangle} \cdot V_1 \\ &\vdots \\ V_k &= X_k - \sum_{j=1}^{k-1} \frac{\langle X_k, V_j \rangle}{\langle V_j, V_j \rangle} \cdot V_j \end{aligned}$$

If all $X_1, X_2, \ldots, X_n$ are linearly independent, we will get $V_1, V_2, \ldots, V_n$ as an orthogonal set. Otherwise, V_k may equal 0 for some k then skip that X_k .

Random Processes

Definition: A ***Random Process*** or ***Stochastic Process*** is a family of random variables $\{X_t \mid t \in T\}$, all based on the same $(\Omega, \mathcal{A}, P)$. The set T is the parameter set of the random process.

We can interpret set T as time and a random process is a system that outputs X_t at each time $t \in T$. X_t can be statistical dependent or independent on the others.

X_t is random variable while x_t is the value of X_t (just one case of X_t that happened). When we said all X_t are identical, it means that all the random variables have the same *cdf* or *pdf* but not the same value x_t.

Examples

- Throwing 2 dices at a time, let X_t for t = 1,2,3, … is the sum of both dices thrown at each t. In this case t is discrete and all X_t are identical and statistical independent.
- ***Random walk***: let X_t for t = 1,2,3, … is the distance you have moved from the original position. At each time t, we toss a coin then move forward one step if getting H and backward one step for T. X_t will now depend on (and only on) X_{t-1}. Because X_t must be either $X_{t-1}+1$ or $X_{t-1}-1$.
- ***Wiener process***: is a random walk process with the lim $\Delta t \to 0$, $\Delta t = t_n - t_{n-1}$. Step size must be reduced proportionally and the process will be continuous in time. The *Wiener process* in 3-dimensional space is called ***Brownian process***.

discrete parameter (or ***time***)	if t is discrete
continuous parameter (or ***time***)	if t is continuous
discrete state (or ***value***)	if X is discrete
continuous state (or ***value***)	if X is continuous

Markov Processes are the random processes that the next $X_{t_{m+1}}$ and all futures will be statistically dependent on (and only on) the present X_{t_m} (not $X_{t_{m-1}}$ or earlier).

$$f_{X_{t_{m+1}}, X_{t_{m+2}}, \ldots, X_{t_n} | X_{t_1}, X_{t_2}, \ldots, X_{t_m}} = f_{X_{t_{m+1}}, X_{t_{m+2}}, \ldots, X_{t_n} | X_{t_m}}$$

Gaussian Processes are the processes that their joint *pdf* of all X_{t_n} are Gaussian distributions. Given any set $\{t_1, t_2, \ldots, t_n\}$ in ascending order from T for any integer n, there exists an nxn ***Autocovariance Matrix*** **C** and an n-dimensional vector $\underline{\boldsymbol{\mu}}$ such that

$$f_{\underline{X}}(\underline{X}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\tfrac{1}{2}(\underline{X} - \underline{\mu})^T \mathbf{C}^{-1} (\underline{X} - \underline{\mu})\right\}$$

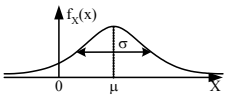
$$\underline{X} = \begin{bmatrix} X_{t_1} & X_{t_2} & \cdots & X_{t_n} \end{bmatrix}^T$$

$$\mathbf{C} = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T]$$

Chapter 2
1-D and 2-D Gaussian Distribution

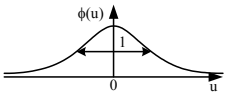
1-D Gaussian Distribution has the *pdf* as

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ \int_{-\infty}^{\infty}f(x)dx=1$$



Normalize x by using $u=\frac{(x-\mu)}{\sigma}$ and the normalized *pdf* of u will become

$$\phi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}};\ \int_{-\infty}^{\infty}\phi(u)du=1$$



Note that any *pdf* must have total area under curve =1. And because of symmetry, half of the area is

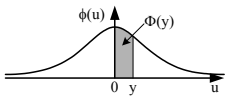
$$\int_{\mu}^{\infty}f(x)dx=\int_{0}^{\infty}\phi(u)du=\frac{1}{2}$$

To calculate probability of x in the interval [a,b)

$$P\{a\le X<b\}=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{a}^{b}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=\frac{1}{\sqrt{2\pi}}\int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma}e^{-\frac{u^2}{2}}du$$

We can create a table of integration

$$\Phi(y)=\int_{0}^{y}\phi(u)du$$



and calculate the probability of interval [a,b] as

$$P\{a\le X<b\}=\pm\Phi(\tfrac{|b-\mu|}{\sigma})\pm\Phi(\tfrac{|a-\mu|}{\sigma})$$ depended on a,b.

2-D Gaussian Distribution for two random variables with joint Gaussian *pdf* as

$$f_{X_1,X_2}(x_1,x_2)=\frac{|\mathbf{C}|^{-1/2}}{2\pi}\exp\left\{-\tfrac{1}{2}\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\mathbf{C}^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\right\}$$

$$\mathbf{C}=E\left[\begin{bmatrix}(x_1-\mu_1)\\ (x_2-\mu_2)\end{bmatrix}\begin{bmatrix}(x_1-\mu_1) & (x_2-\mu_2)\end{bmatrix}\right]$$

$$=\begin{bmatrix}E\left[(x_1-\mu_1)^2\right] & E\left[(x_1-\mu_1)(x_2-\mu_2)\right]\\ E\left[(x_2-\mu_2)(x_1-\mu_1)\right] & E\left[(x_2-\mu_2)^2\right]\end{bmatrix}$$

$$=\begin{bmatrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{bmatrix}=\textit{covariance matrix}$$

Covariance matrix **C** is symmetric ($c_{12}=c_{21}$) and also its inverse $\mathbf{C}^{-1}$ and $|\mathbf{C}|$ is its determinant.

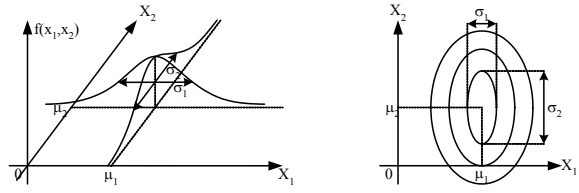
$$\mathbf{C}^{-1}=\frac{1}{|\mathbf{C}|}\begin{bmatrix}c_{22} & -c_{12}\\ -c_{12} & c_{11}\end{bmatrix};\ |\mathbf{C}|=c_{11}c_{22}-c_{12}^2$$

X_1 and X_2 are ***uncorrelated*** (or ***orthogonal***) if $c_{12}=E\{(x_1-\mu_1)(x_2-\mu_2)\}=0$. In case of Gaussian, *uncorrelated* implies *statistical independent*.

$$f_{X_1,X_2}(x_1,x_2)=\frac{1}{2\pi\sqrt{c_{11}c_{22}}}\exp\left\{-\tfrac{1}{2}\left(\frac{(x_1-\mu_1)^2}{c_{11}}+\frac{(x_2-\mu_2)^2}{c_{22}}\right)\right\}$$

$$=\frac{1}{\sqrt{2\pi c_{11}}}\exp\left\{-\frac{(x_1-\mu_1)^2}{2c_{11}}\right\}\frac{1}{\sqrt{2\pi c_{22}}}\exp\left\{-\frac{(x_2-\mu_2)^2}{2c_{22}}\right\}$$

$$=f_{X_1}(x_1)f_{X_2}(x_2)$$



2-D Gaussian with $c_{12}=c_{21}=0$, $c_{11}=\sigma_1^2$, $c_{22}=\sigma_2^2$

To calculate the probability of a rectangular area, let

$$\underline{\mathbf{X}}=\begin{bmatrix}x_1\\ x_2\end{bmatrix},\ \underline{\boldsymbol{\mu}}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$$

$$P\{a_1\le x_1<b_1,a_2\le x_2<b_2\}=\int_{a_1}^{b_1}\int_{a_2}^{b_2}f(x_1,x_2)dx_1dx_2$$

$$=\frac{|\mathbf{C}|^{-1/2}}{2\pi}\int_{a_1}^{b_1}\int_{a_2}^{b_2}\exp\left\{-\tfrac{1}{2}\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)^{\mathrm{T}}\mathbf{C}^{-1}\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)\right\}dx_1dx_2;$$

And again, the total volume $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x_1,x_2)dx_1dx_2=1$

In 2-D, the symmetry may not be straight forward. If $\mu_1=\mu_2=0$ and X_1, X_2 are *uncorrelated* ($c_{12}=0$), the volume under *pdf* surface over one quadrant will be

$$\int_{0}^{\infty}\int_{0}^{\infty}f(x_1,x_2)dx_1dx_2=\int_{0}^{\infty}f(x_1)dx_1\int_{0}^{\infty}f(x_2)dx_2=\frac{1}{4}$$

This is not true if X_1 and X_2 are *correlated* ($c_{12}\neq 0$) as in the following example.

Integrating over a quadrant **, let $\mu_1 = \mu_2 = 0$ and** $\mathbf{C} = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}$ ***then*** $|\mathbf{C}| = 1$; $\mathbf{C}^{-1} = \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}$

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left\{ -\tfrac{1}{2} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\}$$

$$= \frac{1}{2\pi} \exp\left\{ -\tfrac{1}{2} \left(2x_1^{\,2} - 6x_1x_2 + 5x_2^{\,2} \right) \right\}$$

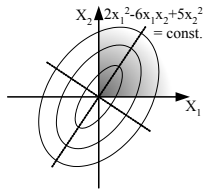
$$= \frac{1}{2\pi} \exp\left\{ -\left(x_1 - \tfrac{3}{2}x_2 \right)^2 - \tfrac{1}{4}x_2^{\,2} \right\}$$

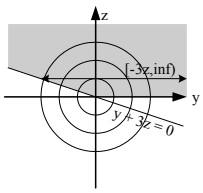
Let $y = x_1 - \frac{3}{2}x_2$; $z = \frac{1}{2}x_2$ then the Jacobian equals

$$J = \begin{vmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} \\ \frac{\partial z}{\partial x_1} & \frac{\partial z}{\partial x_2} \end{vmatrix} = \begin{vmatrix} 1 & -\frac{3}{2} \\ 0 & \frac{1}{2} \end{vmatrix} = \tfrac{1}{2}$$ and the integration will be

$$I = \int_0^\infty \int_0^\infty f(x_1, x_2) dx_1 dx_2 = \int_0^\infty \int_{-3z}^\infty f(y, z) 2 dy dz$$

$$= \frac{1}{\pi} \int_0^\infty \int_{-3z}^\infty \exp\left(-y^2 - z^2 \right) \, dy dz = \tan^{-1}\left(-\tfrac{1}{3} \right) = 0.4488$$





Chapter 3
Multi-dimensional Gaussian Distribution

Let $\underline{\mathbf{X}} = [X_1\ X_2\ \dots\ X_n]^T$ be an n-dim vector of random variables, $\underline{\boldsymbol{\mu}} = [\mu_1\ \mu_2\ \dots\ \mu_n]^T$ a vector of constants and $\mathbf{C} = [c_{ij}]$, i, j = 1,2, … , n a ***positive definite symmetric*** matrix ($\underline{\mathbf{x}}^T\mathbf{C}\underline{\mathbf{x}} \geq 0$ for all $\underline{\mathbf{x}}$). The n-dim Gaussian *pdf* is

$$f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\tfrac{1}{2} \overbrace{(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})^T \mathbf{C}^{-1} (\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})}^{\text{always} \geq 0, \text{ if } \mathbf{C} \text{ positive definite}} \right\}$$

$|\mathbf{C}|$ = determinant of $\mathbf{C}$. Let $|d\underline{\mathbf{x}}| = dx_1 dx_2 \dots dx_n$ then

$$P\left\{ a_1 \leq x_1 < b_1, \dots, a_n \leq x_n < b_n \right\} = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(\underline{\mathbf{x}}) |d\underline{\mathbf{x}}|$$

Theorem: If X is an n-dim random vector with n-dim joint Gaussian distribution described above, then its *mean* $E\{\underline{\mathbf{X}}\} = \underline{\boldsymbol{\mu}}$ and
covariance matrix $E\{(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})^T\} = \mathbf{C}$

$\mathbf{C}$ must always be *positive definite matrix* because it is square of real number and *symmetric matrix* because the multiplication is commutative.

All positive definite symmetric matrix can be factored as $\mathbf{C} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ where

$$\mathbf{D} = \begin{bmatrix} d_1 & & & \mathbf{0} \\ & d_2 & & \\ & & \ddots & \\ \mathbf{0} & & & d_n \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & & & & \mathbf{0} \\ l_{23} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}$$

is a diagonal and lower triangular matrix respectively. $\mathbf{L}$ has all its diagonal elements equal 1 so that $|\mathbf{L}| = 1$, and $|\mathbf{C}| = |\mathbf{D}|$. Then $\mathbf{C}^{-1} = (\mathbf{L}^T)^{-1}\mathbf{D}^{-1}\mathbf{L}^{-1}$.
If we substitute the random vector $\underline{\mathbf{X}}$ by $\underline{\mathbf{Y}}$ with $\underline{\mathbf{y}} = \mathbf{L}^{-1}(\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}})$ which has unit Jacobian $\Rightarrow |d\underline{\mathbf{x}}| = |d\underline{\mathbf{y}}|$ then

$$P\left\{ \underline{\mathbf{a}} \leq \underline{\mathbf{x}} < \underline{\mathbf{b}} \right\} = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n}$$

$$\exp\left\{ -\tfrac{1}{2} \overbrace{(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})^T (\mathbf{L}^T)^{-1}}^{\underline{\mathbf{y}}^T} \mathbf{D}^{-1} \overbrace{\mathbf{L}^{-1}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})}^{\underline{\mathbf{y}}} \right\} |d\underline{\mathbf{x}}|$$

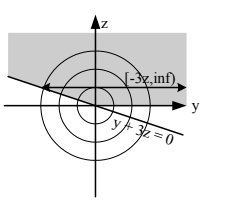
$$= \frac{|\mathbf{D}|^{-1/2}}{(2\pi)^{n/2}} \int_{\alpha_1}^{\beta_1} \int_{\alpha_2}^{\beta_2} \cdots \int_{\alpha_n}^{\beta_n} \exp\left\{ -\tfrac{1}{2} \underline{\mathbf{y}}^T \mathbf{D}^{-1} \underline{\mathbf{y}} \right\} \ |d\underline{\mathbf{y}}|$$

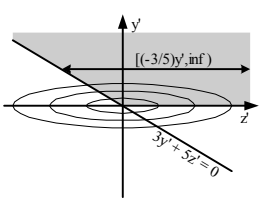
α_i and β_i are not constant but functions of y_k, k<i.

From the same example from chapter 2, we get $\mathbf{C} = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ with $\mathbf{L} = \begin{bmatrix} 1 & 0 \\ \frac{3}{5} & 1 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} 5 & 0 \\ 0 & \frac{1}{5} \end{bmatrix}$ and $\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{3}{5} & 1 \end{bmatrix}$, $\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & 5 \end{bmatrix}$. Instead of using $\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} 1 & -\frac{3}{2} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ to *decorrelate* y and z, we use $\begin{bmatrix} y' \\ z' \end{bmatrix} = L^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{3}{5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Rightarrow \begin{cases} y' = x_1 \\ z' = -\frac{3}{5}x_1 + x_2 \end{cases}$ and get the new *uncorrelated* joint Gaussian *pdf* as $f(y', z') = \frac{1}{2\pi} \exp\left\{ -\frac{1}{10}y'^2 - \frac{5}{2}z'^2 \right\}$. Then

$$I = \int_0^\infty \int_0^\infty f(x_1, x_2) dx_1 dx_2 = \int_0^\infty \int_{-\frac{3}{5}y}^\infty f(y', z') \ dz'dy'$$

$$= \frac{1}{2\pi} \int_0^\infty \int_{-\frac{3}{5}y}^\infty \exp\left(-\tfrac{1}{10}y'^2 - \tfrac{5}{2}z'^2 \right) \ dz'dy' = 0.4488$$





Conditional Density Function

Let n+m random variables have joint Gaussian *pdf* with $\underline{\mathbf{x}} = [x_1\ x_2\ \dots\ x_n]^T$ and $\underline{\mathbf{y}} = [y_1\ y_2\ \dots\ y_m]^T$ as the as the first n and last m random variables respectively. The *covariance matrix* $\mathbf{C}$ can be partitioned as $\mathbf{C} = \begin{bmatrix} \mathbf{C_{XX}} & \mathbf{C_{XY}} \\ \mathbf{C_{YX}} & \mathbf{C_{YY}} \end{bmatrix}$ where $\mathbf{C}$ is (n+m)×(n+m), $\mathbf{C_{XX}}$ is n×n, $\mathbf{C_{XY}} = \mathbf{C_{YX}}^T$ is n×m and $\mathbf{C_{YY}}$ is m×m matrix. The (n+m)-dim *mean* vector is $\begin{bmatrix} \underline{\boldsymbol{\mu}}_X \\ \underline{\boldsymbol{\mu}}_Y \end{bmatrix}$ where $\underline{\boldsymbol{\mu}}_X$ and $\underline{\boldsymbol{\mu}}_Y$ are n and m-dim *mean* vectors for $\underline{\mathbf{x}}$, $\underline{\mathbf{y}}$ respectively.

$$f(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{(n+m)/2}} \cdot \exp\left\{ -\tfrac{1}{2} \begin{bmatrix} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T & (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \end{bmatrix} \mathbf{C}^{-1} \begin{bmatrix} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X) \\ (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \end{bmatrix} \right\}$$

Marginal integration:

$$f_1(\underline{\mathbf{x}}) = \int_{-\infty}^{\infty} \overbrace{\cdots}^{m} \int_{-\infty}^{\infty} f(\underline{\mathbf{x}},\underline{\mathbf{y}}) |d\underline{\mathbf{y}}|$$

$$f_2(\underline{\mathbf{y}}) = \int_{-\infty}^{\infty} \overbrace{\cdots}^{n} \int_{-\infty}^{\infty} f(\underline{\mathbf{x}},\underline{\mathbf{y}}) |d\underline{\mathbf{x}}|$$

Conditional Density Function of $\underline{\mathbf{x}}$ Given $\underline{\mathbf{y}}$

$$f_C(\underline{\mathbf{x}} \mid \underline{\mathbf{y}}) = \frac{f(\underline{\mathbf{x}},\underline{\mathbf{y}})}{f_2(\underline{\mathbf{y}})}$$

Both $f(\underline{\mathbf{x}},\underline{\mathbf{y}})$ and $f_2(\underline{\mathbf{y}})$ are Gaussian so $f_C(\underline{\mathbf{x}}|\underline{\mathbf{y}})$ will also be Gaussian. $f_C(\underline{\mathbf{x}}|\underline{\mathbf{y}})$ is the n-dim *pdf* and must have

$$\int_{-\infty}^{\infty} \overbrace{\cdots}^{n} \int_{-\infty}^{\infty} f_C(\underline{\mathbf{x}} \mid \underline{\mathbf{y}}) |d\underline{\mathbf{x}}| = 1$$

$f_C(\underline{\mathbf{x}}|\underline{\mathbf{y}})$ must be of the form

$$f_C(\underline{\mathbf{x}} \mid \underline{\mathbf{y}}) = \frac{|\mathbf{P}|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\tfrac{1}{2} (\underline{\mathbf{x}}-\underline{\mathbf{m}})^T \mathbf{P}^{-1} (\underline{\mathbf{x}}-\underline{\mathbf{m}}) \right\}$$

Both n×n *covariance matrix* $\mathbf{P}$ and n-dim *mean* vector $\underline{\mathbf{m}}$ can be functions of the given $\underline{\mathbf{y}}$ (in fact only $\underline{\mathbf{m}}$ is).

How to get $\mathbf{P}$ and $\underline{\mathbf{m}}$ from the known parameters $\mathbf{C_{XX}}$, $\mathbf{C_{XY}}$, $\mathbf{C_{YY}}$, $\underline{\boldsymbol{\mu}}_X$, $\underline{\boldsymbol{\mu}}_Y$ and given value $\underline{\mathbf{y}}$.

Matrix Inversion Lemma (proof in text pp. 35-38)

From $\mathbf{C} = \begin{bmatrix} \mathbf{C_{XX}} & \mathbf{C_{XY}} \\ \mathbf{C_{YX}} & \mathbf{C_{YY}} \end{bmatrix}$ we have $\mathbf{C}^{-1} = \begin{bmatrix} \mathbf{A_{XX}} & \mathbf{A_{XY}} \\ \mathbf{A_{YX}} & \mathbf{A_{YY}} \end{bmatrix}$

Where

$$\begin{aligned} \mathbf{A_{XX}} &= \left(\mathbf{C_{XX}} - \mathbf{C_{XY}}\mathbf{C_{YY}}^{-1}\mathbf{C_{YX}}\right)^{-1} \\ &= \mathbf{C_{XX}}^{-1} + \mathbf{C_{XX}}^{-1}\mathbf{C_{XY}}\mathbf{A_{YY}}\mathbf{C_{YX}}\mathbf{C_{XX}}^{-1} \\ \mathbf{A_{XY}} &= -\mathbf{A_{XX}}\mathbf{C_{XY}}\mathbf{C_{YY}}^{-1} = -\mathbf{C_{XX}}^{-1}\mathbf{C_{XY}}\mathbf{A_{YY}} \\ \mathbf{A_{YX}} &= \mathbf{A_{XY}}^T \\ \mathbf{A_{YY}} &= \left(\mathbf{C_{YY}} - \mathbf{C_{YX}}\mathbf{C_{XX}}^{-1}\mathbf{C_{XY}}\right)^{-1} \\ &= \mathbf{C_{YY}}^{-1} + \mathbf{C_{YY}}^{-1}\mathbf{C_{YX}}\mathbf{A_{XX}}\mathbf{C_{XY}}\mathbf{C_{YY}}^{-1} \end{aligned}$$

$$\textit{mean}: \mathrm{E}\left\{\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{Y}} \end{bmatrix}\right\} = \int_{-\infty}^{\infty} \overbrace{\cdots}^{n+m} \int_{-\infty}^{\infty} \begin{bmatrix} \underline{\mathbf{x}} \\ \underline{\mathbf{y}} \end{bmatrix} f(\underline{\mathbf{x}},\underline{\mathbf{y}}) |d\underline{\mathbf{x}}| |d\underline{\mathbf{y}}| = \begin{bmatrix} \underline{\boldsymbol{\mu}}_X \\ \underline{\boldsymbol{\mu}}_Y \end{bmatrix}$$

$$\textit{covariance}: \mathrm{E}\left\{ \begin{bmatrix} (\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}_X) \\ (\underline{\mathbf{Y}}-\underline{\boldsymbol{\mu}}_Y) \end{bmatrix} \begin{bmatrix} (\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}_X)^T & (\underline{\mathbf{Y}}-\underline{\boldsymbol{\mu}}_Y)^T \end{bmatrix} \right\}$$

$$= \int_{-\infty}^{\infty} \overbrace{\cdots}^{n+m} \int_{-\infty}^{\infty} \begin{bmatrix} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)(\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T & (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)(\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \\ (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)(\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T & (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)(\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \end{bmatrix} \cdot f(\underline{\mathbf{x}},\underline{\mathbf{y}}) |d\underline{\mathbf{x}}| |d\underline{\mathbf{y}}|$$

$$= \begin{bmatrix} \mathbf{C_{XX}} & \mathbf{C_{XY}} \\ \mathbf{C_{YX}} & \mathbf{C_{YY}} \end{bmatrix} = \mathbf{C}$$

Because $f_2(\underline{\mathbf{y}})$ is Gaussian, its *pdf* will be

$$f_2(\underline{\mathbf{y}}) = \frac{|\mathbf{C_{YY}}|^{-1/2}}{(2\pi)^{m/2}} \exp\left\{ -\tfrac{1}{2} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \mathbf{C_{YY}}^{-1} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \right\}$$

Then the conditional *pdf* of $\underline{\mathbf{x}}$ given $\underline{\mathbf{y}}$ is

$$f_C(\underline{\mathbf{x}} \mid \underline{\mathbf{y}}) = \frac{f(\underline{\mathbf{x}},\underline{\mathbf{y}})}{f_2(\underline{\mathbf{y}})} = \frac{|\mathbf{C}|^{-1/2}}{|\mathbf{C_{YY}}|^{-1/2}} \cdot \frac{(2\pi)^{m/2}}{(2\pi)^{(n+m)/2}} \cdot \frac{\exp\left\{ -\tfrac{1}{2} \begin{bmatrix} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T & (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \end{bmatrix} \mathbf{C}^{-1} \begin{bmatrix} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X) \\ (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \end{bmatrix} \right\}}{\exp\left\{ -\tfrac{1}{2} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \mathbf{C_{YY}}^{-1} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \right\}}$$

$$= \frac{\left(|\mathbf{C}|/|\mathbf{C_{YY}}|\right)^{-1/2}}{(2\pi)^{n/2}} \cdot \exp\{ -\tfrac{1}{2} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T \mathbf{A_{XX}} (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X) - (\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}_X)^T \mathbf{A_{XY}} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) - \tfrac{1}{2} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \mathbf{A_{YY}} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) + \tfrac{1}{2} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y)^T \mathbf{C_{YY}}^{-1} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \}$$

$$= k \cdot \exp\{ \overbrace{-\tfrac{1}{2} \underline{\mathbf{x}}^T \mathbf{A_{XX}} \underline{\mathbf{x}}}^{\textit{Quadratic in } \underline{\mathbf{x}}} + \overbrace{\underline{\mathbf{x}}^T \left(\mathbf{A_{XX}}\underline{\boldsymbol{\mu}}_X - \mathbf{A_{XY}} (\underline{\mathbf{y}}-\underline{\boldsymbol{\mu}}_Y) \right)}^{\textit{Linear in } \underline{\mathbf{x}}} + \dots \textit{other terms without } \underline{\mathbf{x}} \}$$

Conditional Mean and Covariance

By matching the coefficients of quadratic and linear terms of $\underline{\mathbf{x}}$ to find $\mathbf{P}$ and $\underline{\mathbf{m}}$, we get

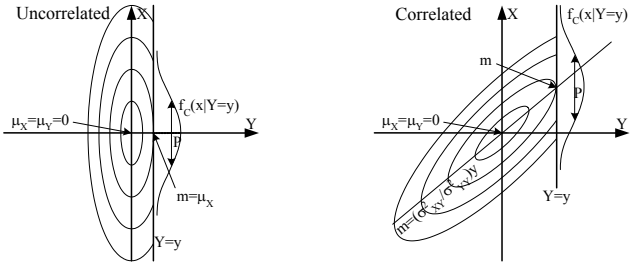
conditional covariance matrix:

$$\mathbf{P} = \mathbf{A}_{XX}^{-1} = \mathbf{C}_{XX} - \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}$$

conditional mean:

$$\underline{\mathbf{m}} = \underline{\boldsymbol{\mu}}_X + \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}(\underline{\mathbf{y}} - \underline{\boldsymbol{\mu}}_Y)$$

The conditional mean $\underline{\mathbf{m}}$ is linear function of $\underline{\mathbf{y}}$ while the conditional covariance matrix is constant.



Conditional Mean ≡ Bayesian Estimation

For any joint *pdf* (not necessary Gaussian) $f(\underline{\mathbf{x}},\underline{\mathbf{y}})$, if $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ are correlated, knowing $\underline{\mathbf{Y}}=\underline{\mathbf{y}}$ will give some informations on $\underline{\mathbf{X}}$. We then can estimate $\underline{\mathbf{X}}$ from $\underline{\mathbf{Y}}$.

Estimation of $\underline{\mathbf{X}}$: $\quad \underline{\hat{\mathbf{X}}} = \underline{\mathbf{g}}(\underline{\mathbf{Y}})$

Estimation error: $\quad \underline{\mathbf{e}} = \underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}}$

Loss function: $\quad L(\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}})$ is a real, non-negative, convex function (*i.e.*, has minimum). We want to find the optimal function $\underline{\mathbf{g}}(\cdot)$ that minimize $L(\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}})$. If $L(\underline{\mathbf{e}})$ is $\|\underline{\mathbf{e}}\|^2$, square of Euclidean norm, this is called ***Minimum Mean Square Error (MMSE)*** criterion.

minimize: $\quad E\left\{L(\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}})\right\} = E\left\{\left\|\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}}\right\|^2\right\} = E\left\{\left\|\underline{\mathbf{X}} - \underline{\mathbf{g}}(\underline{\mathbf{Y}})\right\|^2\right\}$

$$= \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} \left\|\underline{\mathbf{x}} - \underline{\mathbf{g}}(\underline{\mathbf{y}})\right\|^2 f(\underline{\mathbf{x}},\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|\left|d\underline{\mathbf{y}}\right|$$

$$= \int_{\mathbb{R}^m}\left\{\int_{\mathbb{R}^n}\left[\underbrace{\underline{\mathbf{x}}^T\underline{\mathbf{x}}}_{\boxed{A}} - \underbrace{2\underline{\mathbf{g}}^T(\underline{\mathbf{y}})\underline{\mathbf{x}}}_{\boxed{B}} + \underbrace{\underline{\mathbf{g}}^T(\underline{\mathbf{y}})\underline{\mathbf{g}}(\underline{\mathbf{y}})}_{\boxed{C}}\right] f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|\right\} f_2(\underline{\mathbf{y}})\left|d\underline{\mathbf{y}}\right|$$

Define ***conditional mean*** **:** $\underline{\mathbf{m}}(\underline{\mathbf{y}}) = \int_{\mathbb{R}^n} \underline{\mathbf{x}}\, f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|$

and

Define ***total variance*** **:** $V(\underline{\mathbf{y}}) = \int_{\mathbb{R}^n} \left\|\underline{\mathbf{x}} - \underline{\mathbf{m}}(\underline{\mathbf{y}})\right\|^2 f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|$

Find $\int \boxed{A} f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|$, from the identity,

$$\underline{\mathbf{x}}^T\underline{\mathbf{x}} = \underbrace{\left[\underline{\mathbf{x}} - \underline{\mathbf{m}}(\underline{\mathbf{y}})\right]^T\left[\underline{\mathbf{x}} - \underline{\mathbf{m}}(\underline{\mathbf{y}})\right]}_{\searrow} + \underbrace{2\underline{\mathbf{m}}^T(\underline{\mathbf{y}})\left[\underline{\mathbf{x}} - \underline{\mathbf{m}}(\underline{\mathbf{y}})\right]}_{\downarrow} + \underbrace{\underline{\mathbf{m}}^T(\underline{\mathbf{y}})\underline{\mathbf{m}}(\underline{\mathbf{y}})}_{\downarrow}$$

$$\int_{\mathbb{R}^n} \underline{\mathbf{x}}^T\underline{\mathbf{x}}\, f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right| = V(\underline{\mathbf{y}}) + 2\underline{\mathbf{m}}^T(\underline{\mathbf{y}})\cancel{\left[\underline{\mathbf{m}}(\underline{\mathbf{y}}) - \underline{\mathbf{m}}(\underline{\mathbf{y}})\right]} + \left\|\underline{\mathbf{m}}(\underline{\mathbf{y}})\right\|^2$$

Find $\int \boxed{B} f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|$

$$\int_{\mathbb{R}^n} -2\underline{\mathbf{g}}^T(\underline{\mathbf{y}})\underline{\mathbf{x}}\, f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right| = -2\underline{\mathbf{g}}^T(\underline{\mathbf{y}})\underline{\mathbf{m}}(\underline{\mathbf{y}})$$

Find $\int \boxed{C} f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right|$

$$\int_{\mathbb{R}^n} \left\|\underline{\mathbf{g}}(\underline{\mathbf{y}})\right\|^2 f_C(\underline{\mathbf{x}}\,|\,\underline{\mathbf{y}})\left|d\underline{\mathbf{x}}\right| = \left\|\underline{\mathbf{g}}(\underline{\mathbf{y}})\right\|^2$$

Substitute part $\boxed{A}$, $\boxed{B}$ and $\boxed{C}$ and get $E\left\{L(\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}})\right\}$

$$= \int_{\mathbb{R}^m}\left[V(\underline{\mathbf{y}}) + \underbrace{\left\|\underline{\mathbf{m}}(\underline{\mathbf{y}})\right\|^2 - 2\underline{\mathbf{g}}^T(\underline{\mathbf{y}})\underline{\mathbf{m}}(\underline{\mathbf{y}}) + \left\|\underline{\mathbf{g}}(\underline{\mathbf{y}})\right\|^2}_{\downarrow}\right] f_2(\underline{\mathbf{y}})\left|d\underline{\mathbf{y}}\right|$$

$$= \underbrace{\int_{\mathbb{R}^m} V(\underline{\mathbf{y}}) f_2(\underline{\mathbf{y}})\left|d\underline{\mathbf{y}}\right|}_{\text{not depend on } \underline{\mathbf{g}}(\underline{\mathbf{y}})} + \underbrace{\int_{\mathbb{R}^m} \left\|\underline{\mathbf{m}}(\underline{\mathbf{y}}) - \underline{\mathbf{g}}(\underline{\mathbf{y}})\right\|^2 f_2(\underline{\mathbf{y}})\left|d\underline{\mathbf{y}}\right|}_{\text{non-negative, min if } \underline{\mathbf{g}}(\underline{\mathbf{y}})=\underline{\mathbf{m}}(\underline{\mathbf{y}})}$$

We can see that the optimum Bayesian MMSE estimator of $\underline{\mathbf{X}}$ from given $\underline{\mathbf{Y}}$ is the *conditional mean*,

$$\underline{\hat{\mathbf{X}}}_{opt} = \underline{\mathbf{m}}(\underline{\mathbf{Y}})$$

Gaussian case: $\underline{\mathbf{m}}(\underline{\mathbf{y}}) = \underline{\boldsymbol{\mu}}_X + \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}(\underline{\mathbf{y}} - \underline{\boldsymbol{\mu}}_Y)$

and $V(\underline{\mathbf{y}})$ is constant (not depend on $\underline{\mathbf{y}}$).

$V(\underline{\mathbf{y}})$ = trace of $\mathbf{P}$ = tr $\mathbf{P}$ = $p_{11} + p_{22} + \ldots + P_{nn}$
= sum of eigen value of $\mathbf{P}$
= total variance (scalar) of $\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}}$ because P is also the covariance matrix of $\underline{\mathbf{X}} - \underline{\hat{\mathbf{X}}}$

Notations

$E\{\underline{\mathbf{X}}|\underline{\mathbf{Y}}=\underline{\mathbf{y}}\} = \underline{\mathbf{m}}(\underline{\mathbf{y}})$: *not random variables*

but

$E\{\underline{\mathbf{X}}|\underline{\mathbf{Y}}\} = \underline{\mathbf{m}}(\underline{\mathbf{Y}})$: *are random variables*

In case of Gaussian,

$\underline{\mathbf{m}} = \underline{\boldsymbol{\mu}}_X + \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}(\underline{\mathbf{y}} - \underline{\boldsymbol{\mu}}_Y)$: *expected value of* $\underline{\mathbf{X}}$

but

$\underline{\hat{\mathbf{X}}} = \underline{\boldsymbol{\mu}}_X + \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}(\underline{\mathbf{Y}} - \underline{\boldsymbol{\mu}}_Y)$: *new random variables*

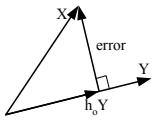
Linear MMSE Estimator

Consider simplest case, let X and Y are two correlated random variables with arbitrary joint *pdf*. If we only want the estimator $\hat{X}_{linear}$ from given Y=y to be linear function of y, we call the optimum estimator $\hat{X}_{linear,opt}$ as Linear MMSE Estimator. The solution will be in linear form, $\hat{X}_{linear,opt} = h_o Y$, assuming all zero means. To solve the optimum h_o, we can use the Hilbert space of second order random variables to find the solution. Because $\hat{X}_{linear}$ is linear functions of Y, it must be in the subspace spanned by Y. Then the optimum vector in this subspace that gives the shortest error vector will be the orthogonal projection of vector X on this subspace. Then the error vector $X - \hat{X}_{linear}$ must be orthogonal to Y,

$$0 = \left\langle \left(X - \hat{X}_{linear,opt} \right), Y \right\rangle = \left\langle (X - h_o Y), Y \right\rangle$$

$$0 = \langle X, Y \rangle - \langle h_o Y, Y \rangle = \langle X, Y \rangle - h_o \langle Y, Y \rangle$$

$$h_o = \frac{\langle X, Y \rangle}{\langle Y, Y \rangle} = \frac{C_{XY}}{C_{YY}} \Rightarrow \hat{X}_{linear,opt} = \frac{C_{XY}}{C_{YY}} \cdot Y$$



For Gaussian, Linear MMSE is also Bayesian MMSE estimator (*conditional mean*, the best of all)

Chapter 4
Finite Random Sequences

Random Process: is a family $\{X_t | t \in T\}$ of random variables, all defined on the same (Ω, $\mathcal{A}$, **P**). T is parameter set of random process. For EE's applications, we can treat set T as a set of time.

Random sequence ≡ ***Discrete time process***

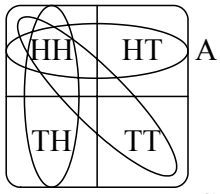
Successive Viewpoint

Let $X_1, X_2, \ldots, X_n$ be a sequence of n Gaussian random variables. For each X_k , we must have *pdf* as

$$f(x_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

μ_k = mean and σ_k^2 = variance of X_k. If all X_k's are ***mutually independent*** random variables, *i.e.*, all pairs, triples, quadruples, … of X_k's are independent, then only μ_k and σ_k^2, k=1,2,…,n are enough to define this random sequence (or discrete time random process). Mutual independence means that we can not know anything about X_k from knowing all other X's. For Gaussian distribution, *pairwise* independence implies *mutual* independence. This is not true for other cases.

P(A) = P(B) = P(C) = 1/2
P(A∩B) = P(A)P(B) = 1/4
P(B∩C) = P(B)P(C) = 1/4
P(C∩A) = P(C)P(A) = 1/4
P(A∩B∩C) = 1/4
≠ P(A)P(B)P(C) = 1/8



A, B, C are pairwise independent but not mutually independent. Knowing that one of them has happened will not tell anything about two others. But knowing that two of them has happened will fix the last one.

That is, for Gaussian: $E\{(X_i-\mu_i)(X_j-\mu_j)\} = 0$ for $i \neq j$, being pairwise uncorrelated ⇒ pairwise independent ⇒mutually independent. This type of processes are simple but not useful, because we cannot make any statistical inference. The normalized sequence of uncorrelated Gaussian random variables $W_1, W_2, \ldots, W_n$ with all $\mu_k = 0$ and $\sigma_k^2 = 1$ is called ***Unit White Gaussian Noise*** (***u.w.g.n.***). All W_k's are ***independent identically distributed*** (***i.i.d.***) random variables.

In general, the sequence of X_k's are not independent and we need more parameters to describe the process. The *covariances* $c_{ij} = E\{(X_i-\mu_i)(X_j-\mu_j)\}$, with $c_{ij} = c_{ji}$, will give the correlations of each pairs of X's and c_{ii} is the variances of X_i. The *covariance matrix* $\mathbf{C} = [c_{ij}]$.

We can create a new random sequence by using linear combinations of random variables from other random sequence and get different characteristics (*mean*, *covariance*). The simplest case is by using *u.w.g.n.* to create a sequence with given μ_i and c_{ij}. That is given sequence $\{W_i\}$ with

$$E\{W_i\} = 0,\ E\{W_i^2\} = 1 \text{ and } E\{W_i W_j\} = 0 \text{ for } i \neq j$$

we want to create a new sequence $\{X_i\}$ with

$$E\{X_i\} = \mu_i \text{ and } E\{(X_i-\mu_i)(X_j-\mu_j)\} = c_{ij}$$

By factoring **C** into

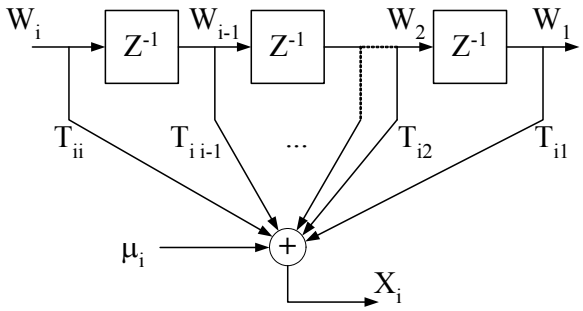
$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n1} & \cdots & c_{nn} \end{bmatrix} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{L}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{L}^T$$

$$= \left(\mathbf{L}\mathbf{D}^{1/2}\right)\left(\mathbf{D}^{1/2}\mathbf{L}\right)^T = \mathbf{T}\mathbf{T}^T, \text{ where}$$

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{d_{11}} & & & \mathbf{0} \\ & \sqrt{d_{22}} & & \\ & & \ddots & \\ \mathbf{0} & & & \sqrt{d_{nn}} \end{bmatrix}, \mathbf{T} = \begin{bmatrix} T_{11} & & & \mathbf{0} \\ T_{21} & T_{22} & & \\ \vdots & \vdots & \ddots & \\ T_{n1} & T_{n1} & \cdots & T_{nn} \end{bmatrix}$$

we can create the sequence $\{X_i\}$ as

$$\begin{aligned} X_1 &= T_{11}W_1 + \mu_1 \\ X_2 &= T_{21}W_1 + T_{22}W_2 + \mu_2 \\ \vdots &= \vdots \\ X_k &= \sum_{j=1}^{k} T_{kj}W_j + \mu_k \end{aligned}$$



Causal Linear Time-varying System (Filter)
with random signal input at time t = i

Each time we feed the system with random sequence $\{W_i\}$, we will get random sequence $\{X_i\}$. Each output sequence is called a ***realization*** of $\{X_i\}$ from sample space of all possible sequence ***called ensemble of realizations***.

For each sequence $\{X_i\}$, we can not find any statistics out of it. We can find the time average on a sequence as $\frac{1}{n}\sum X_i$ but this is not the expectation $E\{X_i\}$. To find the expectation (or statistical average) we must average over the *ensemble*, not the time. To get the mean of X_1, we must get many many realizations $\{X_i\}$ and take the average only the first value of each.

Examples If all $\mu_k=0$ for k =1,2,…,n but all X_k's are not independent (*e.g.*, all X_k equals kX_1). Then the average of X_k over one sequence will not be 0. But the average of X_k over the *ensemble* (many sequences) will give all $\mu_k=0$.

The expectation operator $E\{\cdot\}$ uses only the *ensemble* average, not time average.

Simultaneous Viewpoint

We treat a whole random sequence as a n-dimensional random vector instead of n random variables.
For u.w.g.n vector $\underline{\mathbf{W}} = [W_1\ W_2\ \dots\ W_n]^T$, we have

$$f_{\underline{W}}(\underline{\mathbf{w}}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\left\|\underline{\mathbf{w}}\right\|^2\right)$$

$$\boldsymbol{\mu}_{\underline{W}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T = \underline{\mathbf{0}} \quad (\textit{zero vector})$$

$$\mathbf{C}_{\underline{W}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I} \quad (\textit{identity matrix})$$

To generate $\underline{\mathbf{X}} = [X_1\ X_2\ \dots\ X_n]^T$ from $\underline{\mathbf{W}}$ with

$$E\{\underline{\mathbf{X}}\} = \underline{\boldsymbol{\mu}} = [\mu_1\ \mu_2\ \dots\ \mu_n]^T,\ E\{(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})^T\} = \mathbf{C} = \mathbf{T}\mathbf{T}^T,$$

we get $\underline{\mathbf{X}} = \mathbf{T}\underline{\mathbf{W}} + \underline{\boldsymbol{\mu}}$.

Proof:

$E\{\underline{\mathbf{X}}\} = \mathbf{T}E\{\underline{\mathbf{W}}\} + \underline{\boldsymbol{\mu}} = \underline{\boldsymbol{\mu}}$ and
$E\{(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}})^T\} = E\{(\mathbf{T}\underline{\mathbf{W}})(\mathbf{T}\underline{\mathbf{W}})^T\} = \mathbf{T}E\{\underline{\mathbf{W}}\,\underline{\mathbf{W}}^T\}\mathbf{T}^T = \mathbf{T}\mathbf{T}^T = \mathbf{C}$

Simultaneous viewpoint has more powerful tools for analysis. But if $n\to\infty$ then the successive view point may be more practical.

Chapter 5
Stationary Random Sequences

We will extend random sequence to infinity,
$-\infty\ \dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots\ \infty \Rightarrow X_k$ for $-\infty < k < \infty$.
$\mu_k = E[X_k]$, $c_{ij} = E[(X_i-\mu_i)(X_j-\mu_j)]$
From simultaneous viewpoint, $\{X_k\}$ is just a vector in infinite-dimensional space and C is an operator in that space. An important class of doubly infinite sequence is ***stationary random sequence***:

All statistical parameters are time invariant (do not change under time translation)

For Gaussian sequence, only 2 statistical parameters needed to be defined, *mean* and *covariance*. If both are time invariant, the sequence is stationary.
mean stationary: if $\mu = E[X_i] = E[X_{i+k}]$ = constant,
covariance stationary: if $c_{ij} = c_{i+k,j+k}$, for all i, j, k.
That is $E[(X_i-\mu_i)(X_j-\mu_j)] = E[(X_{i+k}-\mu_{i+k})(X_{j+k}-\mu_{j+k})]$ and if mean is also stationary, $E[X_iX_j] = E[X_{i+k}X_{j+k}]$.

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mu \\ \mu \\ \mu \\ \vdots \end{bmatrix}, \mathbf{c} = \begin{bmatrix} c_{00} & c_{01} & c_{02} & c_{03} \\ c_{01} & c_{00} & c_{01} & \searrow \\ c_{02} & c_{01} & c_{00} & \searrow \\ c_{03} & \searrow & \searrow & \searrow \end{bmatrix}$$ *equal diagonal values*

For EE's applications, we use t as time instead of i, j and use the notation as {X(t)} for a discrete-time random process. For stationary random process {X(t)}

define: ***mean*** $\mu_X = E[X(t)]$ and
autocovariance $c_{XX}(\tau) = E\{[X(t+\tau) - \mu_X][X(t) - \mu_X]\}$

Note that both are not functions of t (absolute time) so that the values will not change with time t. τ is just the time interval parameter.

White Noise Input to Discrete-Time System

Let {V(t)} be a u.w.g.n discrete-time random process,

$$\mu_V = 0 \quad , \quad c_{VV}(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \neq 0 \end{cases}$$

We can see that u.w.g.n. is a stationary process. We can generate new random process {X(t)} from {V(t)} by using ***discrete-time convolution***. This is equivalent to filtering the u.w.g.n. V(t) with a filter. Let the filter has ***impulse response*** h(t) which must be *bounded* and *square summable*, $\sum_{t=0}^{\infty} |h(t)|^2 < \infty$. For causal system in real life, h(t) is zero for all negative time and has to be defined for non-negative time only.

The filter's output is

$$\textit{convolution}: \quad X(t) = \sum_{n=0}^{\infty} h(n)V(t-n)$$

$$\textit{mean}: \quad \mu = E[X(t)] = \sum_{n=0}^{\infty} h(n)\,\cancel{E[V(t-n)]} = 0$$

$$\textit{autocovariance function}: \; c_{XX}(\tau) = E[X(t+\tau)X(t)]$$

$$= E\left[\left(\sum_{m=0}^{\infty} h(m)V(t+\tau-m)\right)\left(\sum_{n=0}^{\infty} h(n)V(t-n)\right)\right]$$

$$= E\left[\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h(m)h(n)V(t+\tau-m)V(t-n)\right]$$

$$= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h(m)h(n)E[V(\underbrace{t+\tau-m}_{\searrow})V(\underbrace{t-n}_{\swarrow})]$$

$$= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h(m)h(n)c_{VV}(\overbrace{n+\tau-m}^{\textit{difference}})$$

$c_{VV} = 1$ only if its argument = 0. Take the summation over m first and c_{VV} will be all zero except only when $m = n + \tau$. Finally, we get

$$c_{XX}(\tau) = \sum_{n=0}^{\infty} h(n+\tau)h(n)$$

Because h(t) is causal impulse response, *i.e.*, h(n) = 0 for n < 0, we can show that $c_{XX}(-\tau)=c_{XX}(\tau)$. For τ = -5,

$$c_{XX}(-5) = \sum_{n=0}^{\infty} h(n-5)h(n)$$

$$= \sum_{n=5}^{\infty} h(n-5)h(n) \qquad : \textit{h(negative) = 0}$$

$$= \sum_{m=0}^{\infty} h(m)h(m+5) \qquad : \textit{let m = n - 5}$$

$$= c_{XX}(5)$$

Cross-covariance Function

Let {X(t)} and {Y(t)} are two random processes with $\mu_X(t)$ and $\mu_Y(t)$ respectively. Define ***cross-covariance***

$$c_{XY}(t_1,t_2) = E\{[X(t_1) - \mu_X(t_1)][Y(t_2) - \mu_Y(t_2)]\}$$

Even if both X and Y are stationary, they may not be jointly stationary. For Gaussian process to become stationary, c_{XY} must depend only on $\tau = t_2 - t_1$,

$$c_{XY}(\tau) = E\{[X(t+\tau) - \mu_X(t+\tau)][Y(t) - \mu_Y(t)]\}.$$

From previous example of generating sequence {X(t)} from sequence {V(t)} we have,

$$c_{XV}(\tau) = E[X(t+\tau)V(t)]$$

$$= E\left[\sum_{n=0}^{\infty} h(n)V(t+\tau-n)V(t)\right]$$

$$= \sum_{n=0}^{\infty} h(n)E[V(t+\tau-n)V(t)]$$

$$c_{XV}(\tau) = \sum_{n=0}^{\infty} h(n)c_{VV}(\tau-n)$$

From this, we can interpret $c_{XV}(t)$ as the output signal of the filter, with impulse response h(t), when we put $c_{VV}(t)$ as input. For u.w.g.n. {V(t)}, its $c_{VV}(t)$ is delta function, $c_{VV}(\tau-n) = 1$ only if $\tau-n = 0 \Rightarrow$ $c_{XV}(\tau) = h(\tau)$.

$$c_{XX}(\tau) = E\left[\sum_{n=0}^{\infty} h(n)V(t+\tau-n)X(t)\right]$$

$$= \sum_{n=0}^{\infty} h(n)E[V(t+\tau-n)X(t)]$$

$$c_{XX}(\tau) = \sum_{n=0}^{\infty} h(n)c_{VX}(\tau-n)$$

If *stationary*, $c_{VX}(\tau) = E[V(t+\tau)X(t)]$
$= E[X(t)V(t+\tau)]$
$= E[X(t-\tau)V(t)]$
$= c_{XV}(-\tau)$

So that for stationary {X(t)} and {V(t)},

$$c_{VX}(\tau\text{-}n) = c_{XV}(n\text{-}\tau)$$

Then

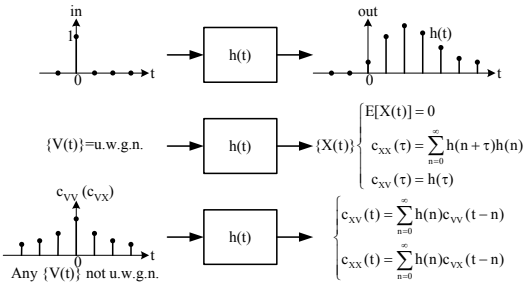
$$c_{XX}(\tau) = \sum_{n=0}^{\infty} h(n)c_{XV}(n-\tau)$$

$$= \sum_{n=0}^{\infty} h(n)\sum_{\lambda=0}^{\infty} h(\lambda)c_{VV}(n-\tau-\lambda)$$

$$\boxed{c_{XX}(\tau) = \sum_{n=0}^{\infty}\sum_{\lambda=0}^{\infty} h(\lambda)h(n)c_{VV}(n-\tau-\lambda)}$$

Again, for u.w.g.n. {V(t)}, $c_{VV}(\tau\text{-}n) = 1$ only if $\tau\text{-}n = 0$ and the summation over n will be all zero except only when $n = \tau + \lambda$. In this case, we get the same as before

$$\boxed{c_{XX}(\tau) = \sum_{\lambda=0}^{\infty} h(\lambda)h(\tau+\lambda)}$$



Power Spectral Density

Let X(t) be a discrete time waveform (deterministic or random), if we pass X(t) through a narrow band filter, center at frequency ω, the power output per Δω of the filter is called ***power spectral density*** (***psd***) of X(t). In the case of deterministic waveform this is $|X(\omega)|^2$, *the square of discrete Fourier transform* of X(t). But for random waveform {X(t)}, we can not use $|X(\omega)|^2$ of single realization waveform because each different realization will give different result. The power spectral density of random signal must be defined on the statistics of random signal, not the signal itself.

Definition: The *power spectral density* of stationary discrete time random process is the discrete Fourier transform of its *autocorrelation* function

$$\textbf{\textit{psd}} = \mathcal{F}[r_{XX}(\tau)] = \sum_{\tau=-\infty}^{\infty} r_{XX}(\tau)e^{-j\omega\tau}$$

where ***autocorrelation*** $r_{XX}(\tau) = E[X(t+\tau)X(t)]$

The *autocorrelation* $r_{XX}(\tau)$ and *autocovariance* $c_{XX}(\tau)$ will be the same if $\mu_X = 0$. If $\mu_X \neq 0$, X(t) will have D.C. component equals to μ_X and

$$r_{XX}(\tau) = E\{([X(t+\tau)-\mu_X]+\mu_X)([X(t)-\mu_X]+\mu_X)\}$$
$$= c_{XX}(\tau) + \mu_X^2$$

Let $\boxed{\phi_{XX}(\omega) = \mathcal{F}[c_{XX}(\tau)] = \sum_{\tau=-\infty}^{\infty} c_{XX}(\tau)e^{-j\omega\tau}}$ be the A.C. component of *power spectral density* and the D.C. component is $\mathcal{F}[\mu_X^2] = \sum_{\tau=-\infty}^{\infty} \mu_X^2 e^{-j\omega\tau} = 2\pi\mu_X^2\delta(\omega)$. Then

$$\boxed{\begin{aligned} \textit{power spectral density} &= \mathcal{F}[r_{XX}(\tau)] = \mathcal{F}[c_{XX}(\tau)+\mu_X^2] \\ &= \phi_{XX}(\omega) + \mathcal{F}[\mu_X^2] \\ &= \underbrace{\phi_{XX}(\omega)}_{\textit{A.C. power}} + \underbrace{2\pi\mu_X^2\delta(\omega)}_{\textit{D.C. power}} \end{aligned}}$$

Note that the D.C. *power spectral density* is a delta function $\delta(\omega)$ which is an infinite impulse at $\omega = 0$ with area under the curve equals 1. This means that the signal power is concentrated at just one frequency. If X(t) is periodic (deterministic) signal, its *psd* will also be a delta function $\delta(\omega\text{-}\omega_o)$ but at frequency ω_o.

We can see that $\phi_{XX}(\omega)$ must have non-negative value, $\phi_{XX}(\omega) \geq 0$ for all ω, because the power density at any frequency must be positive. This is also the result of **Bochner's theorem** that $\phi_{XX}(\omega) \geq 0$ is necessary and sufficient conditions for matrix C of X(t) to become non-negative definite.

$$X(t) \underset{\frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{j\omega\tau}d\omega}{\overset{\sum_{t=-\infty}^{\infty} X(t)e^{-j\omega t}}{\rightleftarrows}} X(\omega)$$

$$E[X(t)X(t+\tau)] \downarrow \qquad \downarrow |\cdot|^2$$

$$c_{XX}(\tau) \underset{\frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_{XX}(\omega)e^{j\omega\tau}d\omega}{\overset{\sum_{\tau=-\infty}^{\infty} c_{XX}(\tau)e^{-j\omega\tau}}{\rightleftarrows}} \phi_{XX}(\omega)$$

Review of Constant Parameter Discrete-Time Deterministic Linear System Theory

Let y(t) be a deterministic discrete-time sequence, for t = 0, 1, 2, …, ∞. Define its Z-transform as

$$Y(z) = \sum_{t=0}^{\infty} y(t)z^{-t}$$

compare to DFT : $Y(\omega) = \sum_{t=-\infty}^{\infty} y(t)e^{-j\omega t}$ which gives the signal analysis of Y(ω) as linear combinations of all frequencies $e^{-j\omega t}$ at amplitude y(t).

Z-transform is the tools for describing the system operations on inputs to get outputs. In Z-transform, each z^{-t} has the physical meaning of delaying signal by t time steps.

Consider a sequence y(t+1), for t = 0, 1, 2, …, ∞. This is a left shifted version of y(t) and delete out y(0). The Z-transform of this sequence becomes

$$\sum_{t=0}^{\infty} y(t+1)z^{-t} = \sum_{s=1}^{\infty} y(s)z^{-(s-1)} = z\sum_{s=1}^{\infty} y(s)z^{-s}$$
$$= z[Y(z) - y(0)]$$

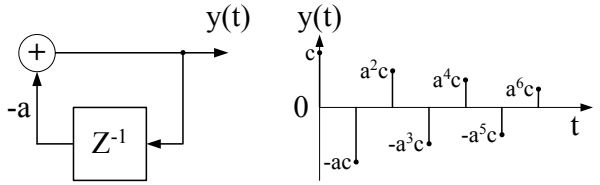
The operation of z in front of the bracket means left shifted, opposite to z^{-1} which is right shifted or delay.

Example

Linear Difference Equation: $y(t+1) + a\,y(t) = 0$
with *initial condition*: $y(0) = c$
Z-transform: $z[Y(z) - y(0)] + a\,Y(z) = 0$

$$Y(z) = \frac{c\,z}{z+a} = \frac{c}{1+\frac{a}{z}} = c\left[1 - \frac{a}{z} + \frac{a^2}{z^2} - \frac{a^3}{z^3} + \ldots\right]$$

$$Y(z) = c\sum_{t=0}^{\infty}(-a)^t z^{-t} \Rightarrow y(t) = c(-a)^t$$



Let signal u(t) pass a discrete-time linear system (filter) with impulse response h(t), the output y(t) will be the convolution:

$$y(t) = \sum_{n=0}^{\infty} h(n)u(t-n)$$

$$Y(z) = \sum_{t=0}^{\infty} z^{-t}\left(\sum_{n=0}^{\infty} h(n)u(t-n)\right) \quad : \textit{Take Z-transform}$$

$$= \sum_{n=0}^{\infty} h(n)\sum_{t=0}^{\infty} u(t-n)z^{-t}$$

$$= \sum_{n=0}^{\infty} h(n)\sum_{r=-n}^{\infty} u(r)z^{-(r+n)} \quad : \textit{Let } r = t - n$$

$$= \sum_{n=0}^{\infty} h(n)\sum_{r=0}^{\infty} u(r)z^{-(r+n)} \quad : \textit{u(t) = 0 for t < 0}$$

$$= \left(\sum_{n=0}^{\infty} h(n)z^{-n}\right)\left(\sum_{r=0}^{\infty} u(r)z^{-r}\right)$$

$$= H(z)U(z)$$

$$\textit{Discrete Time Fourier Transform}: F(\omega) = \sum_{t=-\infty}^{\infty} f(t)e^{-j\omega t}$$

$$\textit{Z-Transform}: F(z) = \sum_{t=0}^{\infty} f(t)z^{-t}$$

Input-Output Relations for Spectral Densities

Analogy to $\phi_{XX}(\omega)$ for $c_{XX}(\tau)$, define the ***cross-spectral density*** for cross-covariance $c_{XY}(\tau)$ as

$$\phi_{XY}(\omega) = \mathcal{F}[c_{XY}(\tau)] = \sum_{\tau=-\infty}^{\infty} c_{XY}(\tau)e^{-j\omega\tau}$$

. Then,

$$\phi_{XX}(\omega) = \sum_{\tau=-\infty}^{\infty} c_{XX}(\tau)e^{-j\omega\tau}$$
$$= \sum_{\tau=-\infty}^{\infty}\left(\sum_{n=0}^{\infty} h(n)c_{VX}(\tau-n)\right)e^{-j\omega\tau}$$
$$= \sum_{n=0}^{\infty} h(n)\sum_{\tau=-\infty}^{\infty} c_{VX}(\tau-n)e^{-j\omega\tau}$$
$$= \sum_{n=0}^{\infty} h(n)\sum_{\lambda=-\infty}^{\infty} c_{VX}(\lambda)e^{-j\omega(\lambda+n)} \qquad \textit{let } \tau - n = \lambda$$
$$= \left(\sum_{n=0}^{\infty} h(n)e^{-j\omega n}\right)\left(\sum_{\lambda=-\infty}^{\infty} c_{VX}(\lambda)e^{-j\omega\lambda}\right)$$
$$= H(e^{j\omega})\phi_{VX}(\omega)$$

Similarly, $\phi_{XV}(\omega) = H(e^{j\omega})\phi_{VV}(\omega)$

And because $c_{VX}(\tau) = c_{XV}(-\tau) \Rightarrow$ $\phi_{VX}(\omega) = \phi_{XV}(-\omega)$

Then $\phi_{XX}(\omega) = H(e^{j\omega})H(e^{-j\omega})\phi_{VV}(\omega) = |H(e^{j\omega})|^2\phi_{VV}(\omega)$

For the u.w.g.n. input v(t), $c_{VV}(\tau)$ is delta function and $\phi_{VV}(\omega) = 1$ (constant for all frequencies). Then

$$\phi_{XX}(\omega) = H(e^{j\omega})H(e^{-j\omega}) = |H(e^{j\omega})|^2$$

Note that we use $H(e^{j\omega})$ instead of $H(\omega)$ to emphasize that $H(e^{j\omega})$ is in fact the H(z) with $z = e^{j\omega}$ (evaluate z on the unit circle). H(z) is called ***sample-data transfer function***.

And $\phi_{XX}(\omega)$ must be non-negative real value for all ω and symmetry with ω. But $\phi_{XV}(\omega)$ can be complex value and non-symmetry.

Given required $c_{XX}(\tau)$ or $\phi_{XX}(\omega)$, how to find the filter that outputs X(t) from u.w.g.n. input, is conceptually by taking the square root of given $c_{XX}(\tau)$ or $\phi_{XX}(\omega)$ to get h(t) or H(z). But there are many solutions, most of them are not *realizable* and more than one *realizable* solutions. The *realizable* filter is the causal filter so that the impulse response must be zero for all $t < 0$. For discrete-time linear filter, this also means that all poles must be inside the unit circle of z-plane. The ***spectral factorization*** is to find the causal filter with *transfer function* H(z) that has $|H(e^{j\omega})|^2 = \phi_{XX}(\omega)$.

Factorization of Rational Spectral Densities

Definition: $\phi_{XX}(\omega)$ is called ***rational*** if and only if it can be written as a ratio of two polynomials in $e^{j\omega}$.

Theorem: Given *rational* $\phi_{XX}(\omega)$, there exists rational function H(z) with all poles and zeros inside the unit circle in z-plane such that $\phi_{XX}(\omega) = |H(e^{j\omega})|^2$.

This is equivalent to factorize the covariance matrix **C** into $\mathbf{TT}^T$or factorize the autocovariance $c_{XX}(\tau)$ into $\sum_{n=0}^{\infty} h(n)h(\tau+n)$.

Example

If we start from simple FIR filter (all zeros, no pole) with impulse response: h(0)=1, h(1)=-0.60, h(2)=0.25 and h(t) = 0 for all t > 2 or t < 0. Then

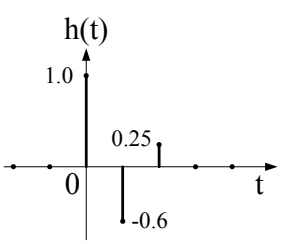
$$H(z) = 1 - 0.60z^{-1} + 0.25z^{-2}$$
$$H(e^{j\omega}) = 1 - 0.60e^{-j\omega} + 0.25e^{-2j\omega}$$
$$H^*(e^{j\omega}) = H(e^{-j\omega}) = 1 - 0.60e^{j\omega} + 0.25e^{2j\omega}$$

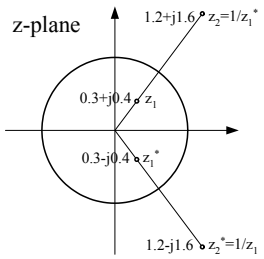
So, $\phi_{XX}(\omega) = |H(e^{j\omega})|^2 = H(e^{j\omega})H(e^{-j\omega})$

$= 1.4225 - 0.75(e^{j\omega} + e^{-j\omega}) + 0.25(e^{j2\omega} + e^{-j2\omega})$

$= 1.4225 - 1.5\cos(\omega) + 0.5\cos(2\omega)$, for $-\pi \le \omega \le \pi$

But if $\phi_{XX}(\omega)$ is given and we want to reverse the calculations to find h(t), we must factor $\phi_{XX}(\omega)$ as





$\phi_{XX}(\omega) = e^{-j2\omega}[.25e^{j4\omega} - .75e^{j3\omega} + 1.4225e^{j2\omega} - .75e^{j\omega} + .25]$

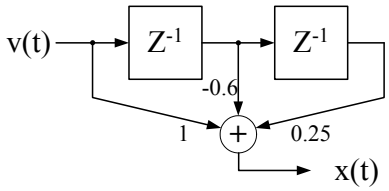
$= 0.25e^{-j2\omega}(e^{j2\omega} - 0.6e^{j\omega} + 0.25)(e^{j2\omega} - 2.4e^{j\omega} + 4)$

$= 0.25e^{-j2\omega}(e^{j\omega} - z_1)(e^{j\omega} - z_1^*)(e^{j\omega} - z_2)(e^{j\omega} - z_2^*)$

where $z_1 = 0.3 + j0.4$ and $z_2 = 1.2 + j1.6 = 1/z_1^*$

We see that, apart from conjugate pairs, the zeros will also be in reciprocal pairs. If one zero is inside the unit circle, its reciprocal will be outside the unit circle but at the conjugate angle. Because when we take the conjugate of |H(z)| with $z = e^{j\omega}$, z^* will be $e^{-j\omega} = 1/e^{j\omega}$. $\phi_{XX}(\omega) = |H(e^{j\omega})|^2 = H(e^{j\omega})H(e^{-j\omega})$ will always have extra reciprocal zeros than the original zeros of H(z).

How do we know which one is the original one? We don't. Both can be the original zeros of H(z). So both $H(z) = z^{-2}(z^2 - 0.6z + 0.25) = 1 - 0.6z^{-1} + 0.25z^{-2}$ or $H(z) = 0.25z^{-2}(z^2 - 2.4z + 4) = 0.25 - 0.6z^{-1} + z^{-2}$ will have the same $\phi_{XX}(\omega)$. But we choose the one with zeros inside the unit circle because it can be inverted.



Now, for the filter with both poles and zeros, the factorization will be the same for both nominator and denominators of H(z). But now we must choose the poles inside the unit circle only. The filter with poles outside the unit circle will be unstable or non-causal and can not be realizable.

In conclusion, we must select the inside half of the reciprocal pairs for both poles and zeros for the H(z) and the other half belongs to $H^*(z)$.

For filters with poles, the impulse response h(t) will be infinitely long, so that the filter is called ***infinite impulse response*** (**IIR**) filter. For example, for |a| < 1

$$\frac{1}{\left(1+az^{-1}\right)} = 1 - az^{-1} + a^2z^{-2} - a^3z^{-3} + \ldots = \sum_{t=0}^{\infty}(-a)^t z^{-t}$$

The polynomial degree 1 of pole is equivalent to polynomial degree ∞ of zeros. In implementations, the zeros are from delayed input sum but the poles are from delayed feedback output sum.

Chapter 6
Continuous Time Stationary Gaussian Processes

Notations

X(t) one realization (waveform, function) of stochastic (random) process

{X(·)} ensemble of waveforms of random process

{X(t)} ensemble of random variables evaluated at time t (scalar random variable)

stochastic process: ensemble of all possible random waveforms with probability distribution for them

Define $\mu(t) = E[X(t)]$

$c(t,s) = E\{[X(t) - \mu(t)][X(s) - \mu(s)]\}$

Now t is real variable. Let $t_1, t_2, \ldots, t_n$ be a set of times. The value of random process {X(·)} at these times are n-dimensional vector $[X(t_1)\ X(t_2)\ \ldots\ X(t_n)]^T$. If this random vector has n-dimensional Gaussian *pdf*

$$f_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\tfrac{1}{2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})^T \mathbf{C}^{-1}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})\right\}$$

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mu(t_1) \\ \mu(t_2) \\ \vdots \\ \mu(t_n) \end{bmatrix}, \mathbf{C} = \begin{bmatrix} c(t_1,t_1) & c(t_1,t_2) & \cdots & c(t_1,t_n) \\ c(t_2,t_1) & c(t_2,t_2) & \cdots & c(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ c(t_n,t_1) & c(t_n,t_2) & \cdots & c(t_n,t_n) \end{bmatrix}$$

C must be symmetric, non-negative definite matrix. If it is true for any set of times, this is ***Gaussian process***. In general, the techniques , *e.g.*, spectral factorization, in chapter 5 and 6 can be used for any second-order processes. Because it uses only second order statistics in the processing (mean and covariance). Gaussian is just the simplest case of all second order process with the same second-order parameters.

Covariance and Spectral Density Functions

- ***strict sense stationary***: all moments of random process are time invariant (stationary)
- ***wide sense stationary***: only the first two moments are required to be stationary, also called ***second-order stationary***

Definition A second-order random process is called stationary if and only if

$$\mu(t) = \mu$$

and $$c(t,s) = c(t-s) = c(\tau)$$

Let $\{X(\cdot)\}$ and $\{Y(\cdot)\}$ be a zero-mean second-order stationary process.

- ***autocovariance***: $c_{XX}(\tau) = E[X(t+\tau)X(t)]$
- ***crosscovariance***: $c_{XY}(\tau) = E[X(t+\tau)Y(t)]$
- ***power spectral density***: $\phi_{XX}(\omega) = \int_{-\infty}^{\infty} c_{XX}(\tau)e^{-j\omega\tau}d\tau$
- ***cross spectral density***: $\phi_{XY}(\omega) = \int_{-\infty}^{\infty} c_{XY}(\tau)e^{-j\omega\tau}d\tau$
- ***inverse Fourier transform***:

$$c_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_{XX}(\omega)e^{j\omega\tau}d\omega$$

$$c_{XY}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_{XY}(\omega)e^{j\omega\tau}d\omega$$

- ***symmetry***: $c_{XX}(-\tau) = c_{XX}(\tau)$ and $\phi_{XX}(-\omega) = \phi_{XX}(\omega)$
  $c_{XY}(-\tau) = c_{YX}(\tau)$ and $\phi_{XY}(-\omega) = \phi_{YX}(\omega)$
- ***non-negativity***: $\phi_{XX}(\omega) \geq 0$, $-\infty < \omega < \infty$
- ***finite power***:

$$c_{XX}(0) = E[X^2(t)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_{XX}(\omega)d\omega < \infty$$

The non-negativity comes from Bochner theorem that this is the necessary and sufficient condition for non-negative covariance matrix **C**.

Note that symmetric and non-negative definite **C** does not imply the Gaussian distribution, any second-order distribution must have these same properties.

Laplace Transform and Linear System Theory

Let h(t) be the impulse response of a continuous-time time-invariant linear dynamic system (filter).

Laplace transform: $$H(s) = \int_0^{\infty} h(t)e^{-st}dt$$

H(s) is called ***transfer function*** of the system (filter). The relationship between Laplace and Z transform with Fourier transform is that the first two are for system description (or processing) but the last one is for signal analysis. For the causal system in real life, it can not take the future input value for computing the output. That's why the integration or summation for the system (Laplace or Z transform) will be only from 0 to ∞. But for the signal analysis (or decomposition), the Fourier integral can cover the whole -∞ to ∞. Apart from this, all these transforms are related by

$$z = e^{j\omega} = e^{s}$$

Because of the nonlinear relationship between s and z, we can not convert polynomial of one variable into polynomial of other variable, only approximation.

Let u(t) and y(t) be two deterministic signals defined for $t \geq 0$. Their Laplace transforms are

$$U(s) = \int_0^{\infty} u(t)e^{-st}dt \quad \textit{and} \quad Y(s) = \int_0^{\infty} y(t)e^{-st}dt$$

If u(t) is input of filter with impulse response h(t). And y(t) is output of the filter. Let u(t) = 0 for t < 0.

$$y(t) = \int_0^{\infty} h(r)u(t-r)dr \quad \textit{and} \quad Y(s) = H(s)U(s)$$

Input-Output Relations for Stochastic Processes

Let $\{X(\cdot)\}$ be the input and $\{Y(\cdot)\}$ be the output of filter with impulse response h(t). Then for each realization of X(t) we have $$Y(t) = \int_0^{\infty} h(r)X(t-r)dr$$

To find the statistics of output related to input, we first find *cross-covariance* and *autocovariance* functions. Then we can find the *power spectral density* and *cross-spectral density* from their Fourier transforms.

$$c_{XY}(\tau) = E[X(t+\tau)Y(t)] = E[X(t+\tau)\int_0^\infty h(r)X(t-r)dr]$$

$$= \int_0^\infty h(r)E[X(t+\tau)X(t-r)]dr = \int_0^\infty h(r)c_{XX}(\tau+r)dr$$

$$c_{YX}(\tau) = c_{XY}(-\tau) = \int_0^\infty h(r)c_{XX}(r-\tau)dr$$

$$= \int_0^\infty h(r)c_{XX}(\tau-r)dr = h(\tau) * c_{XX}(\tau) \quad \boxed{c_{XX} \text{ symmetry}}$$

$$\phi_{YX}(\omega) = H(j\omega)\phi_{XX}(\omega)$$

$$\phi_{XY}(\omega) = \phi_{YX}(-\omega) = H(-j\omega)\phi_{XX}(-\omega) = H^*(j\omega)\phi_{XX}(\omega)$$

$$c_{YY}(\tau) = E[Y(t+\tau)Y(t)] = E[Y(t+\tau)\int_0^\infty h(r)X(t-r)dr]$$

$$= \int_0^\infty h(r)E[Y(t+\tau)X(t-r)]dr = \int_0^\infty h(r)c_{YX}(r+\tau)dr$$

$$= \int_0^\infty h(r)c_{YX}(r-\tau)dr \quad \boxed{c_{YY} \text{ symmetry}}$$

$$= \int_0^\infty h(r)c_{XY}(\tau-r)dr \quad \boxed{c_{XY}(\tau) = c_{YX}(-\tau)}$$

$$= h(\tau) * c_{XY}(\tau)$$

$$\phi_{YY}(\omega) = H(j\omega)\phi_{XY}(\omega) = H(j\omega)H^*(j\omega)\phi_{XX}(\omega)$$

$$= |H(j\omega)|^2 \phi_{XX}(\omega)$$

Spectral Factorization and Paley-Wiener Criterion

If we want to create the stochastic process $\{Y(\cdot)\}$ with *power spectral density* $\phi_{YY}(\omega)$, how to find the transfer function $H(j\omega)$ of the filter that output $\{Y(\cdot)\}$ from known input process $\{X(\cdot)\}$. The simplest $\{X(\cdot)\}$ is white noise $\{V(\cdot)\}$ with $\phi_{VV}(\omega) = 1$, $-\infty < \omega < \infty$. The inverse Fourier transform of 1 is ***Dirac delta function***

$$\boxed{\delta(t) = \begin{cases} \infty & , t = 0 \\ 0 & , elsewhere \end{cases} \quad and \quad \int_{-\infty}^{\infty} \delta(t)dt = 1}$$

So, $c_{VV}(\tau) = \delta(\tau)$ will have infinite power (variance).

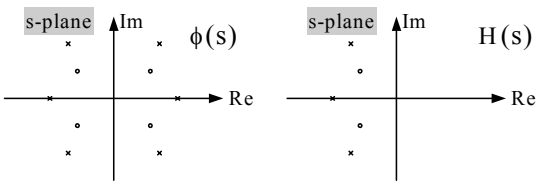
$$c_{VV}(0) = E[V^2(t)] = \int_{-\infty}^{\infty} \phi_{VV}(\omega)d\omega = \delta(0) = \infty$$

This means that there is no true white noise in real life even though we always use it in theoretical analysis. In practice, the signal always have limited power and the white noise we used is actually band-limited, *i.e.*, $\phi_{VV}(\omega) = 1$ up to some frequency ω_{max} which is much higher than operating frequency. Beyond that $\phi_{VV}(\omega)$ will go down to zero. If we apply V(t) as input , then the output power spectral density will be

$$\boxed{\phi_{YY}(\omega) = |H(j\omega)|^2}$$

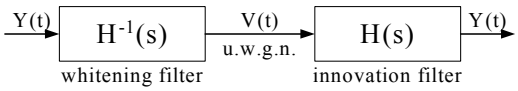
Again, obviously $\phi_{YY}(\omega) \geq 0$ and $\phi_{YY}(-\omega) = \phi_{YY}(\omega)$.

But not all non-negative symmetric $\phi_{YY}(\omega)$ can be factored into $|H(j\omega)|^2$. If it can be factored, Y(t) is ***realizable*** and H(s) must be the transfer function of a (continuous-time) ***stable causal linear system***. That is H(s) must have no pole on the right half of s-plane. But all non-negative symmetric ***rational*** $\phi_{YY}(\omega)$ can always be factored. All the poles and zeros of $\phi_{YY}(\omega)$ must have opposite pairs, *e.g.*, a+jb opposite to -a-jb. Because when we take the conjugate of H(s) and evaluate with $s = j\omega$, s^* will become $-j\omega = -s$ and will create opposite poles and zeros to the original ones.



Then, H(s) will be the rational function with all left half plane poles and zeros of $\phi_{YY}(\omega)$, and the all the right half belongs to $H^*(s)$. In fact, we can also choose some or all right half plane zeros instead and still get the stable causal filter with the same *power spectral density*. But it will not be invertible. The system H(s) with all left half plane zeros (and poles) is called ***minimum phase system***.

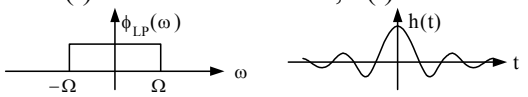
The inverse system $H^{-1}(s)$ will accept input $\{Y(\cdot)\}$ and output u.w.g.n so this system is called ***whitening filter*** while H(s) is called ***innovation filter***.



For non-rational $\phi_{YY}(\omega)$, H(s) will not be rational and the system (filter) will be infinite order. The necessary and sufficient condition for $\phi_{YY}(\omega)$ to be factored as $|H(j\omega)|^2$ and H(s) is stable causal system is the

Paley-Wiener criterion: $\boxed{\int_{-\infty}^{\infty} \frac{|\log \phi(\omega)|}{1+\omega^2} d\omega < \infty}$

This is more specific condition in frequency domain than just the non-negativity at all frequencies for the causal system. Obviously, $\phi(\omega)$ must be non-negative. But more than that, it can not be 0 or ∞ continuously in some intervals of frequency. So the ideal low pass, band pass or high pass signals, which have the spectral density in the stop band equals to zero, will not satisfy Paley-Wiener criterion and can not be realizable. These signal will have non-causal impulse response h(t) in time domain too, $h(t) \neq 0$ for $t < 0$.



Ergodic Process

Stochastic process is a set of random variables X_t or X(t) where t is a parameter (discrete or continuous, finite or infinite set). If we imagine that a random process is a set of dices, each dice outcome is X(t). The whole vector $\underline{\mathbf{X}}_t$ of outcome of all dices is one random object (realization, sequence, waveform). To find statistical parameters (mean, variance) of each X_t (each dice), we need many many trials (realizations) and find the statistics over these *ensemble average*.

If the process is stationary, these parameters will not change with the dices we are observing (*translation invariant*). But we still need many many trials for finding the statistics. We cannot tell the mean of each dice from just one outcome, even all of them has them same mean. The statistical average over a realization, called *time average*, may not give the correct answer. For example, if all the dices are locked to turn up the same face as each other, then the time average will be the same as throwing a dice just once to find its mean.

But if time average statistics are the same as ensemble average statistics, then we call that random process is ***Ergodic***. Obviously, all *ergodic process* must also be *stationary* but not inversely.

Define $$\overline{X}(T_1,T_2)=\frac{1}{T_2-T_1}\int_{T_1}^{T_2}X(t)dt$$

as time average of stationary process $\{X(\cdot)\}$. $\overline{X}(T_1,T_2)$ is also a random variable because it is linear function of random variables $\{X(t)\}$. For Gaussian process, $\overline{X}(T_1,T_2)$ will also be Gaussian with mean

$$\begin{aligned}E\left\{\overline{X}(T_1,T_2)\right\}&=\frac{1}{T_2-T_1}\int_{T_1}^{T_2}E\{X(t)\}\,dt\\&=\frac{1}{T_2-T_1}\int_{T_1}^{T_2}\mu_X dt=\mu_X\end{aligned}$$

We can see that $\overline{X}(T_1,T_2)$ is a good estimator for μ_X because its mean equals μ_X. This property is called ***unbias*** estimator or statistics. But $\overline{X}(T_1,T_2)$ is random and its variance will determine how close to μ_X in average.

$$\begin{aligned}\mathrm{Var}\left(\overline{X}(T_1,T_2)\right)&=E\left\{\left[\overline{X}(T_1,T_2)-\mu_X\right]^2\right\}\\&=E\left\{\overline{X}^2(T_1,T_2)\right\}-2E\left\{\overline{X}(T_1,T_2)\right\}\mu_X+\mu_X^2\\&=E\left\{\overline{X}^2(T_1,T_2)\right\}-\mu_X^2\end{aligned}$$

where, $$E\left\{\overline{X}^2(T_1,T_2)\right\}=\frac{1}{(T_2-T_1)^2}E\left\{\left[\int_{T_1}^{T_2}X(t)dt\right]^2\right\}$$

$$=\frac{1}{(T_2-T_1)^2}\int_{T_1}^{T_2}\int_{T_1}^{T_2}E\{X(t)X(\lambda)\}\,dtd\lambda$$

$$=\frac{1}{(T_2-T_1)^2}\int_{T_1}^{T_2}\int_{T_1}^{T_2}r_{XX}(t-\lambda)dtd\lambda$$

$$=\frac{1}{(T_2-T_1)^2}\int_{T_1}^{T_2}\int_{T_1}^{T_2}c_{XX}(t-\lambda)dtd\lambda+\mu_X^2$$

then, $$\mathrm{Var}\left(\overline{X}(T_1,T_2)\right)=\frac{1}{(T_2-T_1)^2}\int_{T_1}^{T_2}\int_{T_1}^{T_2}c_{XX}(t-\lambda)dtd\lambda$$

If we know $c_{XX}(\tau)$, we can choose T_1, T_2 to get the variance as small as we need and $\overline{X}(T_1,T_2)$ can be almost the same as μ_X . But in practice we do not know $c_{XX}(\tau)$ and all other ensemble average statistics, that is why we try to use time average. But if we know that the integral of $c_{XX}(\tau)$ grows slower than $(T_2\text{-}T_1)^2$ as $T_1\to\text{-}\infty$ and $T_2\to\infty$, then the $\mathrm{Var}(\overline{X})$ will be 0 as $T\to\infty$ and

$$\overline{X}=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)dt\to\mu_X$$

We can see that $c_{XX}(\tau)$ must decrease fast enough in the ergodic process which means that X(t) and X(t+τ) must be almost independent of each other if τ is large enough. Then we can split one realization into several independent short segments as if they were different realizations. Then the ensemble averages between them are, in fact, time average. Similarly, we can define the time-average estimator for c_{XX} or r_{XX} as

$$R_{XX}(\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t+\tau)X(t)dt$$

In summary, the process is ergodic if and only if $\overline{X}=\mu_X$ and $R_{XX}(\tau)=r_{XX}(\tau)$

Power Spectra for Deterministic Signals

In many applications, the signals are not truly random but we may need the signal spectrum from observed unknown signals. We can also use the same technique as for random signal. That is, finding the time-average autocorrelation function first and calculates its Fourier transform to get the spectrum even if the signals are not random at all. Deterministic signal can be finite *energy* or finite *power* signal. The random signals and periodic signals must always be *power* signals. The finite duration signals must always be *energy* signals.

To deal with (deterministic) energy signals, we must re-define all the power terms into energy by removing the time averaging as

$$\mathcal{P} = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x^2(t)dt \quad ;\textit{signal power}$$

$$= \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} x^2(t)dt \quad ;\textit{periodic, } \tau = \textit{period}$$

$$R_{xx}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt \quad ;\textit{autocorrelation}$$

$$S_{xx}(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau)e^{-j\omega\tau}d\tau \quad ;\textit{power spectrum}$$

$$\mathcal{E} = \int_{-\infty}^{\infty} x^2(t)dt \quad ;\textit{signal energy}$$

$$\rho_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt \quad ;\textit{energy autocorrelation}$$

$$\phi_x(\omega) = \int_{-\infty}^{\infty} \rho_x(\tau)e^{-j\omega\tau}d\tau \quad ;\textit{energy spectrum}$$

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \quad ;\textit{signal Fourier transform}$$

The relationship between $X(j\omega)$ and $\phi_x(\omega)$ is

$$\begin{aligned}\phi_x(\omega) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t)x(t+\tau)e^{-j\omega\tau}dtd\tau \\ &= \int_{-\infty}^{\infty} x(t)\int_{-\infty}^{\infty} x(\sigma)e^{-j\omega(\sigma-t)}d\sigma dt \quad ;\textit{let } \sigma = t+\tau \\ &= \int_{-\infty}^{\infty} x(t)e^{j\omega t}dt\int_{-\infty}^{\infty} x(\sigma)e^{-j\omega\sigma}d\sigma \\ &= X(-j\omega)X(j\omega) = |X(j\omega)|^2\end{aligned}$$

The *energy spectrum* is simply the magnitude squared of the Fourier transform of the signal itself.

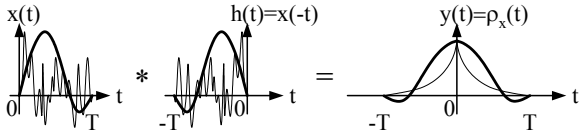
For (deterministic) periodic signals, which are power signals, we must use the time-average functions in the same way as for random signals. But because of the periodicity, the autocorrelation will be periodic and both the Fourier transform and power spectral density will be delta function at harmonic frequencies.

Matched Filter is the filter with transfer function as the conjugate Fourier transform of the signal to be matched. If the signal is x(t) and its Fourier transform is $X(j\omega)$, then the transfer function will be $X(-j\omega)$.

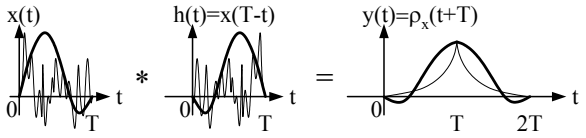
Then the filter impulse response will be the mirror signal x(-t). The output spectrum $Y(j\omega)$ of the filter with matched input x(t) will be

$$X(j\omega)X(-j\omega) = |X(j\omega)|^2 = \phi_x(\omega)$$

which is the signal energy spectrum. Therefore y(t) will be the signal autocorrelation function $\rho_x(t)$.



This filter will give the highest peak output for the matched input x(t), all other signals with the same energy will give lower output than this. But this filter is not causal and the filter must be modified. If x(t) has finite duration T, then we can shift h(t) to the right by T to make the filter causal, h(t) = x(T-t). And the transfer function will be $X(-j\omega)e^{-j\omega T}$, or just add linear phase shift. In this case, the output $y(t) = \rho_x(t+T)$ will give the peak output at time T instead.



Chapter 8
Karhunen-Loeve and Fourier Series Expansion of Continuous-Time Processes

We have already known how to expand deterministic signals into infinite series, *e.g.*, Fourier series, Taylor series. Now we want to expand random signals into infinite series, but in this case, the coefficients will be random variables, *e.g.*, different realizations will give different sets of coefficients. Then we want to know the statistics of these random coefficients. If we deal only with zero mean, second-order random variables, we can use the ***Hilbert space of second-order random variables*** for our analysis. Orthogonal in this space is statistical uncorrelated and also statistical independent for Gaussian case. We will also use the ***Hilbert space of Time Functions on an Interval*** for deterministic part of the signals analysis too. The basis functions for the expansions will be orthogonal in this space.

$L_2[0,T]$: ***The Hilbert space of all square integrable functions*** for the time interval $0 \le t \le T$ is the space of all waveform f(t) with

$$\int_0^T f^2(t)dt < \infty$$

This space has infinite dimensions and any waveform f(t) can be thought of as a vector. We can choose the basis waveforms so that it has simpler representations. For example, f(t) = sin(t) for $0 \le t \le 2\pi$ will be simple form if we use the basis of Fourier series expansion which are sin(nT) for n = 0, 1, …, ∞. Then f(t) will align on one axis with only one coefficient. But if we use the basis of Taylor series, t^n for n = 0, 1, …, ∞ , then f(t) still span on all infinite dimensional space. The inner product and induced norm of $L_2[0,T]$ is

$$\textit{inner product}, \quad \langle f,g\rangle = \int_0^T f(t)g(t)dt$$

$$\textit{induced norm}, \quad \|f\| = \langle f,f\rangle^{1/2} = \left[\int_0^T f^2(t)dt\right]^{1/2}$$

- In n-dim space, if we multiply n-dim vector $\underline{\mathbf{x}}$ by an n×n matrix A, we will get another n-dim vector $\underline{\mathbf{y}}$ in the same space. Matrix A is called an ***operator*** (vector function of vector that maps a vector to another vector in the same space), $\underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}}$.
- In ∞-dim space, **A** will be ∞×∞ matrix and the matrix multiply by a vector gives a new ∞-dim vector with each element as the inner products of each row of the matrix with the vector.

$$\textit{integral operator}, \quad g(s) = \int_0^T K(s,t)f(t)dt$$

$$\textit{abstract form}, \quad g = Lf$$

K(s,t) is called ***kernel function*** of ***integral operator***. K(s,t) is a real function for $0 \le s \le T$, $0 \le t \le T$, in analogy to a square matrix of ∞×∞ dimensions. Where s is row index and t column index of K(s,t).

- In n-dim space, we have ***eigenvalue***/***eigenvector*** of *operator* (or matrix) **A** such that $\mathbf{A}\underline{\mathbf{x}} = \lambda\underline{\mathbf{x}}$
- In ∞-dim space, we have ***spectrum*** of any linear *operator* equivalent to a set of *eigenvalue*, which can be continuous or discrete set. The *eigenvector* is the waveform or signal that if input to the *operator* will result in the same waveform with different size. The amplitude gain is, in fact, the eigenvalue for that eigenvector (or eigen-waveform).

In *unbounded linear system*, we can see that the only waveforms that input to a filter and get output as the same waveform are all the sinusoidal waveform at all frequencies. These are the eigenvectors of the system and the gain at all frequencies, which is its frequency response, are in fact the eigenvalues. This is why we call these *eigenvalues* as the *spectrum* of an *operator*.

The class of operator that we are interested in, is the class of ***compact, self adjoint operator*** which has the properties of its kernel equivalent to symmetric and positive definite function. So that, the auto covariance function c(t,s) can be used as a kernel function of a compact, self adjoint operator.

Definition A linear operator L in Hilbert space will be ***self adjoint operator*** if and only if, for all pairs of vectors f and g, we will get $\langle f, Lg\rangle = \langle Lf, g\rangle$. For our case of real function in $L_2[0,T]$, *self adjoint operator* means that the kernel is symmetric, K(t,s) = K(s,t).

Definition A linear operator L in Hilbert space will be ***compact operator*** if, for any orthonormal (O.N.) set of basis $\{e_n\}$ such that $\langle e_n, e_m\rangle = \begin{cases} 0 & , n \ne m \\ 1 & , n = m \end{cases}$, then

$$\|Le_n\| \to 0 \quad as \quad n \to \infty$$

For example, the kernel K(s,t) = δ(s-t), which gives the *identity operator* equivalent to identity matrix I_n in n-dim space, is not a *compact operator*. Because for all n, $Le_n = e_n$ will not converge to zero. While the kernel that has eigenvalues as $\lambda_n = \frac{1}{n}$, n = 1, 2, … ∞

is *compact operator*. Imagine the kernel as an ∞×∞ diagonal matrix with $\lambda_n = \frac{1}{n}$ on its diagonal. Then

$$Le_n = \frac{1}{n}e_n \to 0 \quad as \quad n \to \infty.$$

But compact operators can also be asymmetric (not self adjoint), for example, $Le_n = \frac{1}{n}e_{n+1}$. This operator has rank zero and no eigenvalue.

$$\begin{bmatrix} 1 & & & \mathbf{0} \\ & 1 & & \\ & & 1 & \\ & & & \ddots \\ \mathbf{0} & & & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & & & \mathbf{0} \\ & \tfrac{1}{2} & & \\ & & \tfrac{1}{3} & \\ & & & \ddots \\ \mathbf{0} & & & \tfrac{1}{n} \end{bmatrix} \quad \begin{bmatrix} 0 & & & \mathbf{0} \\ 1 & 0 & & \\ & \tfrac{1}{2} & 0 & \\ & & \tfrac{1}{3} & \ddots \\ \mathbf{0} & & \ddots & 0 \end{bmatrix}$$

$Le_n = e_n$ $\qquad$ $Le_n = \frac{1}{n}e_n$ $\qquad$ $Le_n = \frac{1}{n}e_{n+1}$

Theorem Every compact, self-adjoint operators which map Hilbert space into itself has at least one eigenvalue and at most *countably infinite* eigenvalues.

Theorem The eigenvectors of a compact, self-adjoint operator are ***mutually orthogonal***.

These orthogonal eigenvectors (waveforms) can be used as the othonormal basis functions for expansion.

The Karhunen-Loeve Expansion

Given a second-order process $\{X(\cdot)\}$ with auto-covariance function c(t,s), the symmetric positive definite matrix, if we use c(t,s) as a kernel of integral operator, the set of non-zero eigenvalues $\{\lambda_n\}$ will all be positive. $\{\phi_n\}$ is a set of normalized eigenfunctions (eigenvectors) for each λ_n that satisfy $L\phi_n = \lambda_n\phi_n$ and are mutually orthogonal because of the compact, self-adjoint operator.

$$\int_0^T c(s,t)\phi_n(t)dt = \lambda_n\phi_n(s) \qquad ,0 \le s,t \le T$$

$$and \quad \langle \phi_n, \phi_m \rangle = \int_0^T \phi_n(t)\phi_m(t)dt = \begin{cases} 1 & ,n = m \\ 0 & ,n \ne m \end{cases}$$

If we use $\{\phi_n\}$ as orthonormal set of basis functions (vectors) for expanding the sample function $X(\cdot)$ of the process $\{X(\cdot)\}$, we have each projections,

$$\alpha_n = \langle X, \phi_n \rangle = \int_0^T \phi_n(t)X(t)dt \quad ,n = 1,2,3,...$$

α_n are functions of random variables and also random variables. Combine all the projections back to get,

$$X(t) = \sum_{n=1}^{\infty} \alpha_n\phi_n(t)$$

In fact the above expansion is true for all orthonormal basis vectors $\{\phi_n\}$ such as Fourier series. The special property of using eigenfunctions of c(t,s) as the basis vectors is that it gives statistical independent (for zero mean Gaussian random variables) α_n as following.

$$\begin{aligned}
E\{\alpha_n\alpha_m\} &= E\left\{\left[\int_0^T \phi_n(t)X(t)dt\right]\left[\int_0^T \phi_m(s)X(s)ds\right]\right\} \\
&= E\left\{\int_0^T\int_0^T \phi_n(t)\phi_m(s)X(t)X(s)dtds\right\} \\
&= \int_0^T\int_0^T \phi_n(t)\phi_m(s)E\{X(t)X(s)\}dtds \\
&= \int_0^T\int_0^T \phi_n(t)\phi_m(s)c(t,s)dtds \\
&= \int_0^T\left[\int_0^T \phi_n(t)c(t,s)dt\right]\phi_m(s)ds \\
&= \int_0^T \lambda_n\phi_n(s)\phi_m(s)ds \\
&= \lambda_n\langle \phi_n, \phi_m \rangle = \begin{cases} \lambda_n & ,n = m \\ 0 & ,n \ne m \end{cases}
\end{aligned}$$

This expansion is called ***Karhunen-Loeve*** expansion. Its basis functions (vectors) are *orthonormal* in the Hilbert space of square integrable functions, $L_2[0,T]$, and its expansion coefficients are also *orthogonal* (statistical independent) in the Hilbert space of second-order random variables. All other orthogonal expansions will not have both properties together. The mean $E\{\alpha_n\} = 0$ and and its variance $E\{\alpha_n^2\} = \lambda_n$.

If c(t,s) is jointly continuous in both t and s. The expansion will converge ***statistically in the mean square sense, pointwise in time***. This means that the mean square error $\to 0$ as $n \to \infty$ for each t.

$$E\left\{\left[X(t) - \sum_{n=1}^{N} \alpha_n\phi_n(t)\right]^2\right\} \quad : \textit{MSE of the first N terms}$$

$$= E\left\{X^2(t) - 2\sum_{n=1}^{N} \alpha_n X(t)\phi_n(t) + \left[\sum_{n=1}^{N} \alpha_n\phi_n(t)\right]^2\right\}$$

$$= c(t,t) - 2\sum_{n=1}^{N} E\{\alpha_n X(t)\}\phi_n(t) + E\left\{\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m\phi_n(t)\phi_m(t)\right\}$$

$$= c(t,t) - 2\sum_{n=1}^{N} E\left\{\int_0^T \phi_n(s)X(s)ds X(t)\right\}\phi_n(t) + \sum_{n=1}^{N}\sum_{m=1}^{N} E\{\alpha_n\alpha_m\}\phi_n(t)\phi_m(t)$$

$$= c(t,t) - 2\sum_{n=1}^{N}\int_0^T \phi_n(s)E\{X(s)X(t)\}ds\phi_n(t) + \sum_{n=1}^{N} \lambda_n\phi_n^2(t)$$

$$= c(t,t) - 2\sum_{n=1}^{N}\int_0^T \phi_n(s)c(t,s)ds\phi_n(t) + \sum_{n=1}^{N} \lambda_n\phi_n^2(t)$$

$$= c(t,t) - 2\sum_{n=1}^{N} \lambda_n\phi_n^2(t) + \sum_{n=1}^{N} \lambda_n\phi_n^2(t)$$

$$= c(t,t) - \sum_{n=1}^{N} \lambda_n\phi_n^2(t) \quad \to \quad 0 \quad as \quad N \to \infty$$

This comes from ***Mercer's theorem*** which states that $c(t,s) = \sum_{n=1}^{\infty} \lambda_n\phi_n(t)\phi_n(s)$ and so, $c(t,t) = \sum_{n=1}^{\infty} \lambda_n\phi_n^2(t)$

This is equivalent to the ***eigenvalue decomposition*** of symmetric matrix into its own eigenvectors $\underline{\phi}_n$ and eigenvalues λ_n. The ***modal matrix*** $U = [\underline{\phi}_1 \ \underline{\phi}_2 \ \cdots \ \underline{\phi}_n]$ is an orthonormal matrix of normalized eigenvectors.

$$\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T$$

$$= \begin{bmatrix} \underline{\phi}_1 & \underline{\phi}_2 & \cdots & \underline{\phi}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_n \end{bmatrix} \begin{bmatrix} \underline{\phi}_1^T \\ \underline{\phi}_2^T \\ \vdots \\ \underline{\phi}_n^T \end{bmatrix}$$

Mercer's theorem is applicable onlty for jointly continuous autocovariance function c(t,s). If c(t,s) is not jointly continuous, the convergence of Karhunen-Loeve expansion will not be pointwise in t, but in the sense of $L_2[0,T]$. This means that at some points of time, it may not have zero error. But the integration of error square from 0 to T will converge to zero instead.

That is $\mathrm{E}\left\{\int_0^T\left[\mathrm{X}(t)-\sum_{n=1}^{N}\alpha_n\phi_n(t)\right]^2 dt\right\}\to 0$ as $N\to\infty$.

This requires the signal energy $\mathrm{E}\left\{\int_0^T \mathrm{X}^2(t)dt\right\}<\infty$

or equivalently, $\int_0^T \mathrm{E}\left\{\mathrm{X}^2(t)\right\}dt=\int_0^T c(t,t)dt<\infty$.

This is the trace of autocovariance function (sum of covariance matrix diagonal elements) which equals to sum of all its eigenvalues. Then,

$$\int_0^T \mathrm{E}\left\{\mathrm{X}^2(t)\right\}dt=\int_0^T c(t,t)dt=\sum_{n=1}^{\infty}\mathrm{E}\{\alpha_n^2\}=\sum_{n=1}^{\infty}\lambda_n<\infty$$

If we sum only up to α_N, the reconstructed waveform will have the total energy $\sum_{n=1}^{N}\lambda_n$ out of $\sum_{n=1}^{\infty}\lambda_n$ and the mean square error is just the energy we disregarded.

Example: ***Expansion of Brownian Motion***

Brownian motion is the limited case of random walk as $\Delta t\to 0$. In term of white noise process $\{V(\cdot)\}$, the Brownian process $\{B(\cdot)\}$ is simply the time integral of white noise,

$$B(t)=\int_0^t V(\tau)d\tau$$

This is the accumulation of random steps in random walk process. If $\{V(\cdot)\}$ has zero mean then $\{B(\cdot)\}$ too.

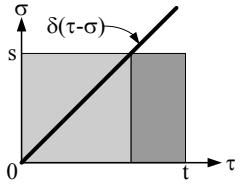
$$c_{BB}(t,s)=\mathrm{E}\{B(t)B(s)\}=\mathrm{E}\left\{\left[\int_0^t V(\tau)d\tau\right]\left[\int_0^s V(\sigma)d\sigma\right]\right\}$$

$$=\mathrm{E}\left\{\int_0^t\int_0^s V(\tau)V(\sigma)d\tau d\sigma\right\}$$

$$=\int_0^t\int_0^s \mathrm{E}\{V(\tau)V(\sigma)\}d\tau d\sigma$$

$$=\int_0^t\int_0^s \delta(\tau-\sigma)d\tau d\sigma$$

$$=\min(t,s)=\begin{cases}s & ,0\le s\le t\\ t & ,0\le t\le s\end{cases}$$



Consider the K-L expansion of B(t) for $0\le t\le T=1$. To find the eigenvalues of $c_{BB}(t,s)$ which must satisfy the ***eigenvalues equation*** $\int_0^1 c_{BB}(t,s)\phi_n(t)dt=\lambda_n\phi_n(s)$

$$\int_0^s t\phi_n(t)dt+s\int_s^1\phi_n(t)dt=\lambda_n\phi_n(s) \qquad (1)$$

$\frac{d}{ds}$ *both sides*; $\int_s^1\phi_n(t)dt=\lambda_n\frac{d}{ds}\phi_n(s)$ (2)

$\frac{d}{ds}$ *again*; $-\phi_n(s)=\lambda_n\frac{d^2}{ds^2}\phi_n(s)$

From (1) we get $\phi_n(0)=0$ and from (2) $\phi'_n(1)=0$. This is wave equation with boundary value problem,

$$\phi_n''(s)+\frac{1}{\lambda_n}\phi_n(s)=0 \quad with \quad \phi_n(0)=\phi_n'(1)=0$$

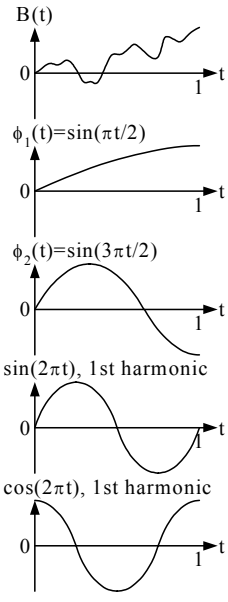
And the solutions for n=1, 2, …∞ are

$$\phi_n(s)=\sqrt{2}\sin\frac{(2n-1)\pi s}{2} \quad and \quad \lambda_n=\left(\frac{2}{(2n-1)\pi}\right)^2$$

Then the Brownian process has the K-L expansion as

$$B(t)=\sum_{n=1}^{\infty}\alpha_n\sqrt{2}\sin\frac{(2n-1)\pi t}{2}$$

and the mutually orthogonal random variables $\{\alpha_n\}$ will have $\mathrm{E}\{\alpha_n\}=0$ and $\mathrm{E}\{\alpha_n^2\}=\lambda_n=\left(\frac{2}{(2n-1)\pi}\right)^2$.



Note that the basis waveforms of K-L expansion are not the same as of Fourier series, which are all the harmonic frequencies. But the K-L basis waveforms are matched to the Brownian process better by choosing the odd sub-harmonic frequencies. They will all start from zero and end with peak amplitude. Fourier series will always gives the same amplitude at both the start and end points while B(t) always start from zero but very unlikely to end there. Fewer number of K-L coefficients can approximate B(t) better than all other expansions.

Equivalence in Finite Dimensional Space

All the expansions are just the change of basis vectors for representing the unknown vector. This can be thought of as a change of variable technique. We may try to use the new representation (coefficients) that are uncorrelated by using $\mathbf{C} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ technique. But the basis vectors in this case are not orthogonal, that is $\mathbf{L}^T\mathbf{L} \neq \mathbf{I}_n$. The K-L expansion uses the $\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T$ eigenvalue decomposition which has the orthogonal basis vectors, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$. But the matrix $\mathbf{U}$ is not lower triangular, so that it cannot be computed in successive viewpoint. Followings are the equivalent of K-L expansion using matrix for finite dimensional space.

Gaussian pdf: $$f(\underline{\mathbf{b}}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\tfrac{1}{2}\underline{\mathbf{b}}^T\mathbf{C}^{-1}\underline{\mathbf{b}}\right]$$

covariance matrix: $\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T$, $|\mathbf{U}| = 1$, $\Lambda = \text{diag}\,\lambda_i$

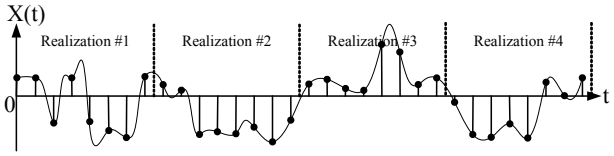
change of variables: $\underline{\boldsymbol{\alpha}} = \mathbf{U}^{-1}\underline{\mathbf{b}} \Rightarrow \underline{\mathbf{b}}^T\mathbf{C}^{-1}\underline{\mathbf{b}} = \underline{\boldsymbol{\alpha}}^T\Lambda^{-1}\underline{\boldsymbol{\alpha}}$

uncorrelated pdf: $$f(\underline{\boldsymbol{\alpha}}) = \frac{|\Lambda|^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\tfrac{1}{2}\underline{\boldsymbol{\alpha}}^T\Lambda^{-1}\underline{\boldsymbol{\alpha}}\right]$$

orthonormal basis: $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$

Application for 1-D signal compression

Karhunen-Loeve expansion also has other names such as, ***principal component transform***, ***Hotelling transform***, ***eigenvector transform***, ***Karhunen-Loeve transform*** (***KLT***). In DSP applications, we use finite dimensional process. Given 1-D signal, *e.g.*, speech waveform, we always assume stationarity and also ergodicity on it. Then we can break the waveform into several pieces with only n points in each waveform. We will treat each piece of n points waveform as one realization of n-dim random sequence. Then we will have several realizations and can find the ensemble average mean μ and covariance matrix $\mathbf{C}$ from them, but in fact all come from just one big realization (time average). If $\mathbf{C}$ can be decomposed into $\mathbf{U}\Lambda\mathbf{U}^T$, then the transform matrix is $\mathbf{U}^{-1}$. If $\underline{\mathbf{x}}$ is an unknown n-dim random vector (waveform), then the transform of $\underline{\mathbf{x}}$ is $\underline{\boldsymbol{\alpha}} = \mathbf{U}^{-1}\underline{\mathbf{x}}$ and the random variables α_n of vector $\underline{\boldsymbol{\alpha}}$ will be mutually independent with variances $E\{\alpha_n^2\} = \lambda_n$.



$$n = 8,\ \underline{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_8 \end{bmatrix},\ \mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T,\ \Lambda = \begin{bmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_8 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{18} \\ u_{21} & u_2 & \cdots & u_{28} \\ \vdots & \vdots & \ddots & \vdots \\ u_{81} & u_{82} & \cdots & u_{88} \end{bmatrix} = \begin{bmatrix} \underbrace{\underline{\mathbf{u}}_1 \quad \underline{\mathbf{u}}_2 \quad \cdots \quad \underline{\mathbf{u}}_8}_{8\ \text{eigenvectors}} \end{bmatrix},\ |\mathbf{U}| = 1$$

$$\underline{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_8 \end{bmatrix} = \mathbf{U}^{-1}\underline{\mathbf{x}},\ \textit{and}\ \underline{\mathbf{x}} = \mathbf{U}\underline{\boldsymbol{\alpha}} = \sum_{n=1}^{8} \alpha_n \underline{\mathbf{u}}_n$$

The variances σ^2 of x_i are stationary and all equal to the diagonal value of matrix $\mathbf{C}$ because all of x_i are identical. While the variance of α_i, the eigenvalues λ_i , are all different. But $\sum_{i=1}^{8}\lambda_i = \sum_{i=1}^{8}\sigma^2 = 8\sigma^2 =$ total signal energy. When some λ_i is bigger than σ^2, some other λ_i must be smaller and can be discard as zero. Then the representation of each segment of $\underline{\mathbf{x}}$ by $\underline{\boldsymbol{\alpha}}$ can reduce the data bit down. This property is ***called energy packing*** and KLT is the best energy packing of all.

Fourier Series Expansion of Random Signal

The Fourier series basis functions are $e^{j2\pi nt/T}/T$ for n^{th} harmonics with the fundamental frequency $f_0 = 1/T$. The basis functions are orthonormal in $L_2[0,T]$ but the Fourier series coefficients are not mutually orthogonal (uncorrelated). Let $\{X(\cdot)\}$ be zero mean, second-order process.

F.S. coefficients: $$\zeta_m = \int_0^T X(t)e^{-j2\pi mt/T}dt/T$$

F.S. expansion: $$X(t) = \sum_{m=-\infty}^{\infty} \zeta_m e^{j2\pi mt/T}/T$$

cross-covariance: $E\{\zeta_m\zeta_n^*\}$

$$= E\left\{\int_0^T\int_0^T X(t)X(s)e^{-j2\pi mt/T}e^{+j2\pi ns/T}dtds/T^2\right\}$$

$$= \int_0^T\int_0^T c_{XX}(t,s)e^{-j2\pi mt/T}e^{+j2\pi ns/T}dtds/T^2 = \gamma_{mn}$$

γ_{mn} is a ***2-D Fourier series coefficient*** of $c_{XX}(t,s)$. For arbitrary $c_{XX}(t,s)$, the cross-covariance γ_{mn} will not be delta function (ζ_m and ζ_n will be correlated). But if we force $\gamma_{mn} = \beta_m\delta_{mn}$ to find the properties of X(t) for this condition, then $c_{XX}(t,s)$ must be stationary. From the 2-D inverse Fourier series of γ_{mn} ,

$$c_{XX}(t,s) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \gamma_{mn} e^{+j2\pi mt/T} e^{-j2\pi ns/T}$$

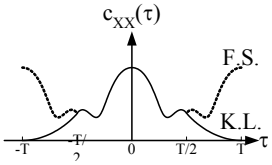
$$= \sum_{m=-\infty}^{\infty} \beta_m e^{+j2\pi m(t-s)/T}$$

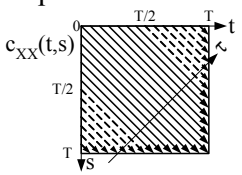
$$c_{XX}(\tau) = \sum_{m=-\infty}^{\infty} \beta_m e^{+j2\pi m\tau/T}$$

This is 1-D Fourier series expansion of $c_{XX}(t,s)$ but over the period of only T. And its F.S. coefficients are

$$\beta_m = \int_{-T/2}^{T/2} c_{XX}(\tau) e^{-j2\pi m\tau/T} d\tau / T$$

Note that we shift the integration interval to [-T/2,T/2] in stead of [0,T] to maintain the symmetry of $c_{XX}(\tau)$. This shows that to get uncorrelated F.S. expansion of X(t), *i.e.*, γ_{mn} as delta function, X(t) must have $c_{XX}(\tau)$ as a periodic function of period T, *i.e.*, $c_{XX}(\tau)$ defined only for $\tau \in$ [-T/2,T/2] then repeat itself. But for any signal X(t) of length T, its $c_{XX}(\tau)$ will have the length of 2T, for $\tau \in$ [-T,T]. So, $c_{XX}(\tau)$ for $|\tau| > T/2$ must be a replica of $|\tau| < T/2$. This is not required for K.L.





Chapter 11
Discrete-Time Kalman Filtering

For non-stationary process, the concept of Power Spectrum has no meaning and cannot be used any more. We must solve the problems in ***time domain***. The tools we used are ***state-space method***. This method describe the internal states of the system by differential or difference equation for continuous and discrete time respectively as,

$$\textit{continuous time}: \underset{n\times 1}{\dot{\underline{\mathbf{X}}}(t)} = \underset{n\times n}{\mathbf{A}(t)}\underset{n\times 1}{\underline{\mathbf{X}}(t)} + \underset{n\times r}{\mathbf{B}(t)}\underset{r\times 1}{\underline{\mathbf{V}}(t)}$$

$$\textit{discrete time}: \underline{\mathbf{X}}(t+1) = \mathbf{A}(t)\underline{\mathbf{X}}(t) + \mathbf{B}(t)\underline{\mathbf{V}}(t)$$

$\{\underline{\mathbf{V}}(t)\}$ is $r\times 1$ vector of zero mean white Gaussian noise with nonstationary $r\times r$ covariance matrix

$$E[\underline{\mathbf{V}}(t)\underline{\mathbf{V}}^T(t)] = \mathbf{Q}(t)$$

$\{\underline{\mathbf{X}}(t)\}$ is $n\times 1$ vector of internal states of the system at each time t. The initial state $\underline{\mathbf{X}}(0)$ is Gaussian with zero mean and $n\times n$ covariance matrix $\mathbf{P}_0$ which is independent of $\{\underline{\mathbf{V}}(t)\}$. After the initial random state, the state will evolve randomly governed by the state equations above. But we can not observe the state $\underline{\mathbf{X}}(t)$ directly. Instead, we can only measure the output

according to each state as $\underset{l\times 1}{Y(t)} = \underset{l\times n}{\underline{\mathbf{h}}^T(t)}\underset{n\times 1}{\underline{\mathbf{X}}(t)} + \underset{l\times 1}{N(t)}$.

Where N(t) is white Gaussian measurement noise, independent of $\underline{\mathbf{X}}(t)$ and $\underline{\mathbf{V}}(t)$, with zero mean and nonstationary variance R(t). In general, Y(t) can also be vector. But for simplicity, we will consider only the scalar output case above. Given the time-varying, deterministic parameters $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\underline{\mathbf{h}}(t)$, stochastic parameters $\mathbf{P}_0$, $\mathbf{Q}(t)$, $R(t)$ and observation sequence of output $Y(t)$: ***what is the MMSE estimation of the internal state*** $\hat{\underline{\mathbf{X}}}(t)$ based on all informations available until that time t. The solution is *Bayesian Estimation* which is the ***conditional mean***.

$$\hat{\underline{\mathbf{X}}}(t) = E\{\underline{\mathbf{X}}(t) \mid Y(t), Y(t-1), Y(t-2), \ldots, Y(1)\}$$

Let $\underline{\mathbf{Z}}(t) = \begin{bmatrix} Y(1) & Y(2) & \cdots & Y(t) \end{bmatrix}^T$ is a random vector with variable dimension of measured output Y until time t. $\underline{\mathbf{Z}}(t)$ will also be a Gaussian random vector with zero mean and $t\times t$ covariance matrix $\mathbf{C}_t = E\{\underline{\mathbf{Z}}(t)\underline{\mathbf{Z}}^T(t)\}$ (known by computations).

If we de-correlate using $\mathbf{C}_t = \mathbf{L}_t\mathbf{D}_t\mathbf{L}_t^T$ factorization and transform $\underline{\mathbf{Z}}(t)$ into a new uncorrelated random vector

$$\underline{\zeta}(t) = \begin{bmatrix} \nu(1) & \nu(2) & \cdots & \nu(t) \end{bmatrix}^T = \mathbf{L}_t^{-1}\underline{\mathbf{Z}}(t)$$

$$\text{with } E\left\{\underline{\zeta}(t)\underline{\zeta}^T(t)\right\} = \mathbf{D}_t : t\times t \text{ diagonal matrix}$$

Then $\{\nu(t)\}$ will be a sequence of uncorrelated random variables ***causally computable*** from $\{Y(t)\}$. This dues to the lower triangular $\mathbf{L}_t$ and the fact that $\underline{\mathbf{Z}}(t+1)$ add just one more element to $\underline{\mathbf{Z}}(t)$, then $\mathbf{C}_{t+1}$ and $\mathbf{L}_{t+1}$ will have one more row and column added to $\mathbf{C}_t$ and $\mathbf{L}_t$. And so $\underline{\zeta}(t+1)$ will only have one more element added to $\underline{\zeta}(t)$ without changing the rest too.

Let $\underset{n\times t}{\mathbf{C}_{X\underline{Z}}(t)} = E\left\{\underset{n\times 1}{\underline{\mathbf{X}}(t)}\underset{l\times t}{\underline{\mathbf{Z}}^T(t)}\right\}$, $\underset{n\times t}{\mathbf{C}_{X\underline{\zeta}}(t)} = E\left\{\underset{n\times 1}{\underline{\mathbf{X}}(t)}\underset{l\times t}{\underline{\zeta}^T(t)}\right\}$

The conditional mean for Gaussian process is

$$E\{\underline{\mathbf{X}} \mid \underline{\mathbf{Y}}\} = \underline{\mu}_X + \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}(\underline{\mathbf{Y}} - \underline{\mu}_Y)$$

Then $\underset{n\times 1}{\hat{\underline{\mathbf{X}}}(t)} = E\{\underline{\mathbf{X}}(t) \mid \underline{\mathbf{Z}}(t)\} = \underset{n\times t}{\mathbf{C}_{X\underline{Z}}}\underset{t\times t}{\mathbf{C}_t^{-1}}\underset{t\times 1}{\underline{\mathbf{Z}}(t)}$

and also $\underset{n\times 1}{\hat{\underline{\mathbf{X}}}(t)} = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t)\} = \underset{n\times t}{\mathbf{C}_{X\underline{\zeta}}}\underset{t\times t}{\mathbf{D}_t^{-1}}\underset{t\times 1}{\underline{\zeta}(t)}$ too.

Because $\underline{\mathbf{Z}}(t)$ and $\underline{\zeta}(t)$ gives the same information.

We will use $\underline{\zeta}(t)$ because of simpler diagonal $\mathbf{D}_t$. Let

$$\mathbf{C}_{X\zeta}(t) = \underset{n\times t}{\left[\underline{\alpha}_{t1} \quad \underline{\alpha}_{t2} \quad \cdots \quad \underline{\alpha}_{tt}\right]}, \quad \underline{\alpha}_{tk} = \underset{n\times 1}{E\left\{\underline{\mathbf{X}}(t)\nu(k)\right\}}$$

and then $$\hat{\underline{\mathbf{X}}}(t) = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t)\} = \sum_{k=1}^{t} \underline{\alpha}_{tk} \frac{\nu(k)}{d_k}$$

If we use conditional information only up to time (t-1) to compute conditional mean of X(t), then

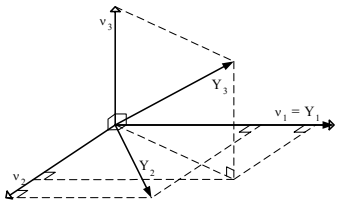
$$E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t-1)\} = \sum_{k=1}^{t-1} \underline{\alpha}_{tk} \frac{\nu(k)}{d_k} \text{ and}$$

Update Formula

$$\hat{\underline{\mathbf{X}}}(t) = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t)\} = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t-1)\} + \underline{\alpha}_{tt} \frac{\nu(t)}{d_t}$$

This means that if we already had $E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t-1)\}$ and receive one more data Y(t) which also gives new $\nu(t)$, then we can update to new $\hat{\underline{\mathbf{X}}}(t)$ by adding the latest term. We do not need to keep all the past value of Y(t) to be able to compute $\hat{\underline{\mathbf{X}}}(t)$. $\nu(t)$ is called ***innovation of Y(t)*** and $\{\nu(t)\}$ ***innovation sequence***.

We can view $\{\nu(t)\}$ as the output of Gram-Schmidt orthogonalization of $\{y(t)\}$ in Hilbert space of second order random variables. This processing is causal and $\nu(t)$ is the component of Y(t) which is orthogonal to the subspace spanned by all previous Y(1) to Y(t-1), or equivalently, $\nu(1)$ to $\nu(t-1)$.



Propagate Formula

If we already knew $E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t-1)\}$, called predictor, and then update with the new information from Y(t) or $\nu(t)$ to get $E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t)\}$ or $\hat{\underline{\mathbf{X}}}(t)$, called estimator, then how to get ***one-step prediction*** $E\{\underline{\mathbf{X}}(t+1) \mid \underline{\zeta}(t)\}$. This is where the state space model comes in. From state equation, $\underline{\mathbf{X}}(t+1) = \mathbf{A}(t)\underline{\mathbf{X}}(t) + \mathbf{B}(t)\underline{\mathbf{V}}(t)$, take the conditional mean using information up to time t,

$$E\left\{\underline{\mathbf{X}}(t+1)\middle|\underline{\zeta}(t)\right\} = E\left\{\mathbf{A}(t)\underline{\mathbf{X}}(t) + \mathbf{B}(t)\underline{\mathbf{V}}(t)\middle|\underline{\zeta}(t)\right\}$$

$$= \mathbf{A}(t)E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t)\right\} + \mathbf{B}(t)E\left\{\cancel{\underline{\mathbf{V}}(t)\middle|\underline{\zeta}(t)}\right\}$$

Because $\underline{\mathbf{V}}(t)$ is independent of $\underline{\mathbf{X}}(t)$ and N(t). But $\underline{\mathbf{Y}}(t)$ depends on $\underline{\mathbf{X}}(t)$ and N(t) so that $\underline{\mathbf{V}}(t)$ and $\underline{\mathbf{Y}}(t)$ are independent. Then $\underline{\mathbf{Z}}(t)$ and $\underline{\zeta}(t)$ will also be independent of $\underline{\mathbf{V}}(t)$ and $E\{\underline{\mathbf{V}}(t) \mid \underline{\zeta}(t)\} = E\{\underline{\mathbf{V}}(t)\} = 0$. Finally we get the ***propagate formula*** :

$$E\left\{\underline{\mathbf{X}}(t+1)\middle|\underline{\zeta}(t)\right\} = \mathbf{A}(t)E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t)\right\} = \mathbf{A}(t)\hat{\underline{\mathbf{X}}}(t)$$

or $$E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t-1)\right\} = \mathbf{A}(t-1)E\left\{\underline{\mathbf{X}}(t-1)\middle|\underline{\zeta}(t-1)\right\}$$

Kalman Filter Equations

Combine both update and propagate equation together

$$\hat{\underline{\mathbf{X}}}(t) = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t)\} = E\{\underline{\mathbf{X}}(t) \mid \underline{\zeta}(t-1)\} + \underline{\alpha}_{tt} \frac{\nu(t)}{d_t}$$

$$= \mathbf{A}(t-1)E\{\underline{\mathbf{X}}(t-1) \mid \underline{\zeta}(t-1)\} + \frac{1}{d_t}\underline{\alpha}_{tt}\nu(t)$$

$$\hat{\underline{\mathbf{X}}}(t) = \mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1) + \frac{1}{d_t}\underline{\alpha}_{tt}\nu(t)$$

This recursive formula will calculate new estimator $\hat{\underline{\mathbf{X}}}(t)$ from old estimator $\hat{\underline{\mathbf{X}}}(t-1)$ after receiving new information $\nu(t)$ or Y(t).

The vector $\frac{1}{d_t}\underline{\alpha}_{tt}$ is called ***Kalman gain*** which takes care of the random part in the state equation. It tells us how much to adjust $\hat{\underline{\mathbf{X}}}(t)$ for the randomness of $\underline{\mathbf{X}}(t)$ based on the available information from $\nu(t)$ or Y(t). These randomness will always give some error in the estimation of $\underline{\mathbf{X}}(t)$. We can calculate the *Kalman gain* and their mean square error of the prediction or estimation for each time t in advance from the system parameters $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\underline{\mathbf{h}}(t)$, $\mathbf{P}_0$, $\mathbf{Q}(t)$, R(t). Define ***prediction error***:

$$\underline{\Delta}(t) = \underline{\mathbf{X}}(t) - E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t-1)\right\} \text{ with}$$

$$\mathbf{M}(t) = E\left\{\underline{\Delta}(t)\underline{\Delta}^T(t)\right\}$$

estimation error:

$$\underline{\theta}(t) = \underline{\mathbf{X}}(t) - E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t)\right\} = \underline{\mathbf{X}}(t) - \hat{\underline{\mathbf{X}}}(t) \text{ with}$$

$$\mathbf{P}(t) = E\left\{\underline{\theta}(t)\underline{\theta}^T(t)\right\}$$

$\mathbf{M}(t)$ and $\mathbf{P}(t)$ are the error covariance matrices of predictor and estimator at time t, the smaller the better. How to compute them including the *Kalman gain* $= \frac{1}{d_t}\underline{\alpha}_{tt}$ in terms of $\mathbf{M}(t)$ and $\mathbf{P}(t)$ are following.

The key concept of innovation $\nu(t)$ is that it is the component of Y(t) which is new (unexpected, uncorrelated, orthogonal) to our knowledge from all the past measurement $\underline{\mathbf{Z}}(t-1)$ or $\underline{\zeta}(t-1)$. So that the prediction $\hat{Y}(t)$ based on past measurement is just the conditional mean of Y(t) given $\underline{\zeta}(t-1)$. This also has the meaning as Y(t) projected on (t-1)-dim subspace.

$$\hat{Y}(t) = E\left\{Y(t)\middle|\underline{\zeta}(t-1)\right\} = E\left\{\underline{\mathbf{h}}^T(t)\underline{\mathbf{X}}(t) + N(t)\middle|\underline{\zeta}(t-1)\right\}$$

$$= \underline{\mathbf{h}}^T(t)E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t-1)\right\} + E\left\{N(t)\middle|\underline{\zeta}(t-1)\right\} \qquad \boxed{1}$$

$$= \underline{\mathbf{h}}^T(t)\mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1)$$

Then, $\nu(t) = Y(t) - \hat{Y}(t) = Y(t) - \underline{\mathbf{h}}^T(t)\mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1)$, and substitute in *Kalman equation* to get,

$$\hat{\underline{\mathbf{X}}}(t) = E\{\underline{\mathbf{X}}(t)\,|\,\underline{\zeta}(t)\}$$

$$= \mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1) + \frac{1}{d_t}\underline{\alpha}_{tt}\left[Y(t) - \underline{\mathbf{h}}^T(t)\mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1)\right]$$

From $\boxed{1}$, the relationship between $\nu(t)$ and $\underline{\Delta}(t)$ is,

$$\nu(t) = Y(t) - \hat{Y}(t)$$

$$= \underline{\mathbf{h}}^T(t)\underline{\mathbf{X}}(t) + N(t) - \underline{\mathbf{h}}^T(t)E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t-1)\right\}$$

$$= \underline{\mathbf{h}}^T(t)\left[\underline{\mathbf{X}}(t) - E\left\{\underline{\mathbf{X}}(t)\middle|\underline{\zeta}(t-1)\right\}\right] + N(t)$$

and get $\boxed{\nu(t) = \underline{\mathbf{h}}^T(t)\underline{\Delta}(t) + N(t) = \underline{\Delta}^T(t)\underline{\mathbf{h}}(t) + N(t)}$

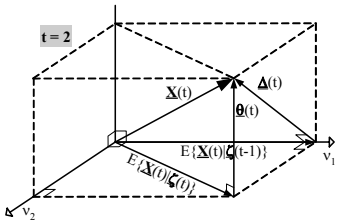
Then we can find $\underline{\alpha}_{tt}$ and d_t in terms of $\mathbf{M}(t)$ as

$$\underline{\alpha}_{tt} = E\left\{\underline{\mathbf{X}}(t)\nu(t)\right\} = E\left\{\underline{\mathbf{X}}(t)\underline{\Delta}^T(t)\right\}\underline{\mathbf{h}}(t) + E\left\{\underline{\mathbf{X}}(t)N(t)\right\}$$

$$\boxed{\underline{\alpha}_{tt} = E\left\{\underline{\Delta}(t)\underline{\Delta}^T(t)\right\}\underline{\mathbf{h}}(t) = \mathbf{M}(t)\underline{\mathbf{h}}(t)}$$

$$d_t = E\left\{\nu^2(t)\right\} = \underline{\mathbf{h}}^T(t)E\left\{\underline{\Delta}(t)\underline{\Delta}^T(t)\right\}\underline{\mathbf{h}}(t) + E\left\{N^2(t)\right\}$$

$$\boxed{d_t = \underline{\mathbf{h}}^T(t)\mathbf{M}(t)\underline{\mathbf{h}}(t) + R(t)}$$



To calculate $\mathbf{M}(t)$, we first find the relationship of $\underline{\Delta}(t)$ and $\underline{\theta}(t)$ as, $\underline{\Delta}(t) = \underline{\mathbf{X}}(t) - E\left\{\underline{\mathbf{X}}(t)\,|\,\underline{\zeta}(t-1)\right\}$

$$= \mathbf{A}(t-1)\underline{\mathbf{X}}(t-1) + \mathbf{B}(t-1)\underline{\mathbf{V}}(t-1) - \mathbf{A}(t-1)\hat{\underline{\mathbf{X}}}(t-1)$$

$$= \mathbf{A}(t-1)\underline{\theta}(t-1) + \mathbf{B}(t-1)\underline{\mathbf{V}}(t-1)$$

From $\mathbf{M}(t) = E\left\{\underline{\Delta}(t)\underline{\Delta}^T(t)\right\}$ and $\mathbf{P}(t) = E\left\{\underline{\theta}(t)\underline{\theta}^T(t)\right\}$, we can relate $\mathbf{M}(t)$ to $\mathbf{P}(t\text{-}1)$ as

$$\boxed{\mathbf{M}(t) = \mathbf{A}(t-1)\mathbf{P}(t-1)\mathbf{A}^T(t-1) + \mathbf{B}(t-1)\mathbf{Q}(t-1)\mathbf{B}^T(t-1)}$$

To complete the recursion loop, we must be able to calculate $\mathbf{P}(t)$ from $\mathbf{M}(t)$. From Kalman equation,

$$\theta(t) = \underline{\mathbf{X}}(t) - \hat{\underline{\mathbf{X}}}(t) = \underline{\mathbf{X}}(t) - E\{\underline{\mathbf{X}}(t)\,|\,\underline{\zeta}(t)\}$$

$$= \underline{\mathbf{X}}(t) - \overbrace{E\{\underline{\mathbf{X}}(t)\,|\,\underline{\zeta}(t-1)\}}^{\text{prediction}} - \overbrace{\frac{1}{d_t}\underline{\alpha}_{tt}\nu(t)}^{\text{update}}$$

$$= \Delta(t) - \frac{1}{d_t}\underline{\alpha}_{tt}\nu(t)$$

$$E\left\{\theta(t)\underline{\mathbf{X}}^T(t)\right\} = E\left\{\Delta(t)\underline{\mathbf{X}}^T(t)\right\} - \frac{1}{d_t}\underline{\alpha}_{tt}E\left\{\nu(t)\underline{\mathbf{X}}^T(t)\right\}$$

$$P(t) = M(t) - \overbrace{\frac{1}{d_t}\underline{\alpha}_{tt}\underline{\alpha}_{tt}^T}^{\text{always} \geq 0}$$

Substitute $\underline{\alpha}_{tt}$ and d_t in terms of $\mathbf{M}(t)$ and finally get

$$\boxed{\mathbf{P}(t) = \mathbf{M}(t) - \frac{\mathbf{M}(t)\underline{\mathbf{h}}(t)\underline{\mathbf{h}}^T(t)\mathbf{M}^T(t)}{\underline{\mathbf{h}}^T(t)\mathbf{M}(t)\underline{\mathbf{h}}(t) + R(t)}}$$

We can start from $\mathbf{P}(0) = \mathbf{P}_0$ to find $\mathbf{M}(1)$ and then find $\mathbf{P}(1)$, $\mathbf{M}(2)$, $\mathbf{P}(2)$, … . Then we can calculate all other parameters such as *Kalman gain* directly.

Example X(t+1) = 0.5X(t) + V(t)
Y(t) = X(t) + N(t)

Where X(0) = 0, V(t) and N(t) are independent zero mean u.w.g.n. This is a scalar and stationary case with A(t) = 0.5, B(t) = 1, Q(t) = 1, R(t) = 1, h(t) = 1, $P_0 = 0$ (because X(0) is deterministic). We will get

$$M(t) = 0.5P(t-1)0.5 + (1)(1)(1) = 0.25P(t-1) + 1$$

$$P(t) = M(t) - \frac{M^2(t)}{M(t)+1} = \frac{M(t)}{M(t)+1}$$

Substitute M(t): $P(t) = \dfrac{P(t-1)+4}{P(t-1)+8}$; $P(0) = P_0 = 0$

Kalman gain: $k(t) = \dfrac{M(t)}{M(t)+1} = P(t)$

$$\hat{X}(t) = 0.5\hat{X}(t-1) + P(t)[Y(t) - 0.5\hat{X}(t-1)]$$

$$= [1 - P(t)]0.5\hat{X}(t-1) + P(t)Y(t)$$

Because this is stationary and with $t \to \infty$, P(t) will converge to

$$P = \frac{P+4}{P+8} \quad \Rightarrow \quad P^2 + 7P - 4 = 0 \quad \Rightarrow \quad P = \frac{\sqrt{65}-7}{2} \approx 0.531$$

The Kalman filter will reduce to time-invariant filter,

$$\hat{X}(t) = 0.234\hat{X}(t-1) + 0.531Y(t)$$

Even with infinite observations of Y(t), the estimation $\hat{X}(t)$ still has error $\theta(t)$ with variance P = 0.531.

Solutions to selected problems

#9/p.79 $\phi(\omega) = \dfrac{1+\cos\omega}{2.125-\cos 2\omega} = \dfrac{1+\dfrac{e^{j\omega}+e^{-j\omega}}{2}}{\dfrac{17}{8}-\dfrac{e^{j2\omega}+e^{-j2\omega}}{2}}$

$$= \frac{4(e^{j\omega}+2+e^{-j\omega})}{(-4e^{j2\omega}+17-4e^{-j2\omega})}$$

To find filter transfer function,

$$\left|H^2(z)\right| = \frac{4(z+2+z^{-1})}{(-4z^2+17-4z^{-2})} = \frac{2(z+1)}{(4z^2-1)}\cdot\frac{2(z^{-1}+1)}{(4z^{-2}-1)}$$

$$H(z) = \frac{2(z+1)}{(4z^2-1)} = \frac{z^{-1}(1+z^{-1})}{2(1-\frac{1}{4}z^{-2})}$$

To find filter impulse response,

$$H(z) = \frac{(2z)^{-1}\left(1+z^{-1}\right)}{\left(1-(2z)^{-2}\right)}$$

$$= (2z)^{-1}[1+z^{-1}][1+(2z)^{-2}+(2z)^{-4}+(2z)^{-6}+\ldots]$$

$$= 2^{-1}z^{-1}+2^{-1}z^{-2}+2^{-3}z^{-3}+2^{-3}z^{-4}+2^{-5}z^{-5}+2^{-5}z^{-6}+\ldots$$

$$h(t) = 0, 2^{-1}, 2^{-1}, 2^{-3}, 2^{-3}, 2^{-5}, 2^{-5}, \ldots; \textit{ for } t = 0,1,2,\ldots$$

We can find autocovariance function from h(t) by,

$$c(\tau) = \sum_{\tau=0}^{\infty} h(\lambda)h(\lambda+\tau)$$

But this is suitable for finite h(t) or finding c(t) at some value of t only. We can also find autocovariance function from $\phi(\omega)$ by,

$$c(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\phi(\omega)e^{j\omega\tau}d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{4(e^{j\omega}+2+e^{-j\omega})}{(-4e^{j2\omega}+17-4e^{-j2\omega})}e^{j\omega\tau}d\omega$$

Let $z = e^{j\omega} \rightarrow dz = je^{j\omega}d\omega = jzd\omega$ and $e^{j\omega\tau} = z^{\tau}$ and the integral becomes contour integral on unit circle. The direction will be counter clockwise for $\tau \geq 0$.

$$c(\tau) = \frac{1}{2\pi}\oint_{|z|=1}\frac{4(z+2+z^{-1})}{(-4z^2+17-4z^{-2})}z^{\tau}\frac{dz}{jz}$$

$$= \frac{1}{2\pi j}\oint_{|z|=1}\frac{-(z+1)^2z^{\tau}}{(z-\frac{1}{2})(z+\frac{1}{2})(z-2)(z+2)}dz$$

$$= \frac{1}{2\pi j}\oint_{|z|=1}\left\{\frac{A}{(z-\frac{1}{2})}+\frac{B}{(z+\frac{1}{2})}+\frac{C}{(z-2)}+\frac{D}{(z+2)}\right\}dz$$

$$= A + B$$

A and B are the residues at poles $z_P = \pm 1/2$ which are enclosed by the contour integral.

To find the residues (or the coefficients of the partial fraction above), we multiply the integrand by $(z - z_P)$ and evaluate the integrand at $z = z_P$ as followings.

$$A = \left.\frac{-(z+1)^2z^{\tau}}{\cancel{(z-\frac{1}{2})}(z+\frac{1}{2})(z-2)(z+2)}\cdot\cancel{(z-\tfrac{1}{2})}\right|_{z=1/2}$$

$$= \frac{-(\frac{3}{2})^2(\frac{1}{2})^{\tau}}{(1)(\frac{-3}{2})(\frac{5}{2})} = (\tfrac{3}{5})(2)^{-\tau}$$

$$B = \left.\frac{-(z+1)^2z^{\tau}}{(z-\frac{1}{2})\cancel{(z+\frac{1}{2})}(z-2)(z+2)}\cdot\cancel{(z+\tfrac{1}{2})}\right|_{z=-1/2}$$

$$= \frac{-(\frac{1}{2})^2(\frac{-1}{2})^{\tau}}{(-1)(\frac{-5}{2})(\frac{3}{2})} = (\tfrac{-1}{15})(-2)^{-\tau}$$

For negative τ, the integral direction will be clockwise and enclose the pole outside unit circle instead. But because $c(\tau)$ is symmetry, so we can just use the $|\tau|$.

$$c(\tau) = (\tfrac{3}{5})(2)^{-|\tau|} + (\tfrac{-1}{15})(-2)^{-|\tau|}$$

$$= \left[\frac{9-(-1)^{|\tau|}}{15}\right](2)^{-|\tau|}$$

$c(0) = \frac{8}{15}$, $c(\pm2) = \frac{8}{15}\cdot\frac{1}{4}$, $c(\pm4) = \frac{8}{15}\cdot\frac{1}{16}$, $c(\pm6) = \frac{8}{15}\cdot\frac{1}{64}$

$c(\pm1) = \frac{2}{3}\cdot\frac{1}{2}$, $c(\pm3) = \frac{2}{3}\cdot\frac{1}{8}$, $c(\pm5) = \frac{2}{3}\cdot\frac{1}{32}$, $c(\pm7) = \frac{2}{3}\cdot\frac{1}{128}$

Kalman Filter Example A box of resistors has the mean and variance of the resistance as μ_X and σ_X^2 respectively. Pick up ***only one*** resistor and measure the resistance ***many times*** using an unbiased noisy ohm-meter with variance of measurement error σ_n^2. Assume independent measurement errors and Gaussian pdf for all. Let X be the true resistance and and Y(t) the measured value at time t = 1, 2, 3, … respectively. Formulate the Kalman filter to estimate the $\hat{X}(t)$ at each time t in terms of μ_X , σ_X^2, σ_n^2, y(t) and t. If σ_X^2 or σ_n^2 is very large then $\hat{X}(t)$ will reduce to simple form, explain the meaning and reason for each case.

Static system, scalar, no state noise case:

$$\boxed{\begin{aligned} X(t+1) &= X(t) \\ Y(t) &= X(t) + N(t) \end{aligned}}$$

With A(t) = 1 , B(t) = 0 , h(t) = 1 , $P(0) = \sigma_X^2$, Q(t) = N.A. , $R(t) = \sigma_n^2$

$$M(t) = A(t-1)P(t-1)A^T(t-1) + B(t-1)Q(t-1)B^T(t-1)$$
$$= P(t-1)$$

and

$$P(t) = M(t) - \frac{\alpha_{tt}\alpha_{tt}^T}{d_t} = M(t) - \frac{M^2(t)}{M(t)+\sigma_n^2} = \frac{M(t)\sigma_n^2}{M(t)+\sigma_n^2}$$

then

$$P(t)=\frac{P(t-1)\sigma_n^2}{P(t-1)+\sigma_n^2} \quad \Rightarrow \quad \frac{1}{P(t)}=\frac{1}{P(t-1)}+\frac{1}{\sigma_n^2}$$

Start from $P(0)=\sigma_X^2$,

$$\frac{1}{P(1)}=\frac{1}{\sigma_X^2}+\frac{1}{\sigma_n^2} \quad \Rightarrow \quad \frac{1}{P(2)}=\frac{1}{\sigma_X^2}+\frac{2}{\sigma_n^2} \quad \Rightarrow \cdots$$

$$\Rightarrow \quad \frac{1}{P(t)}=\frac{1}{\sigma_X^2}+\frac{t}{\sigma_n^2}$$

and get

$$M(t)=P(t-1)=\frac{1}{\frac{1}{\sigma_X^2}+\frac{t-1}{\sigma_n^2}}$$

Find Kalman gain,

$$\alpha_{tt}=M(t)h(t)=M(t)$$

$$d_t=h^T(t)M(t)h(t)+R(t)=M(t)+\sigma_n^2$$

$$\text{Kalman gain}=k(t)=\frac{\alpha_{tt}}{d_t}=\frac{M(t)}{M(t)+\sigma_n^2}$$

$$=\frac{1}{1+\frac{\sigma_n^2}{M(t)}}=\frac{1}{1+\sigma_n^2\left[\frac{1}{\sigma_X^2}+\frac{t-1}{\sigma_n^2}\right]}$$

$$=\frac{1}{\sigma_n^2/\sigma_X^2+t}=\frac{1}{S+t} \quad \text{for} \quad S=\sigma_n^2/\sigma_X^2$$

The MMSE estimator from Kalman equation,

$$\hat{X}(t)=(1-k(t))\hat{X}(t-1)+k(t)Y(t)$$

$$=\frac{S+t-1}{S+t}\hat{X}(t-1)+\frac{1}{S+t}Y(t)$$

$$(S+t)\hat{X}(t)=(S+t-1)\hat{X}(t-1)+Y(t)$$

From the initial estimator $\hat{X}(0)=E[X]=\mu_X$ (the unconditional mean),

$$(S+1)\hat{X}(1)=S\hat{X}(0)+Y(1)$$

$$(S+2)\hat{X}(2)=(S+1)\hat{X}(1)+Y(2)=S\hat{X}(0)+Y(1)+Y(2)$$

$$\vdots$$

$$(S+t)\hat{X}(t)=S\hat{X}(0)+\sum_{m=1}^{t}Y(m)=S\mu_X+\sum_{m=1}^{t}Y(m)$$

then $\hat{X}(t)=\frac{S}{S+t}\mu_X+\frac{1}{S+t}\sum_{m=1}^{t}Y(m)$

$$=\frac{\sigma_n^2}{\sigma_n^2+t\sigma_X^2}\mu_X+\frac{\sigma_X^2}{\sigma_n^2+t\sigma_X^2}\sum_{m=1}^{t}Y(m)$$

$$=\frac{\sigma_n^2/t}{\sigma_n^2/t+\sigma_X^2}\mu_X+\frac{\sigma_X^2}{\sigma_n^2/t+\sigma_X^2}\left[\frac{1}{t}\sum_{m=1}^{t}Y(m)\right]$$

$$=\frac{\sigma_n^2/t}{\sigma_n^2/t+\sigma_X^2}\mu_X+\frac{\sigma_X^2}{\sigma_n^2/t+\sigma_X^2}\overline{Y}$$

for $\overline{Y}=\frac{1}{t}\sum_{m=1}^{t}Y(m)$

Interpretation of the result:

- If σ_n^2 is very large then $\hat{X}(t)=\mu_X$. Because the measurements are too noisy to be useful and the filter will rely only on a priori knowledge of X.
- If σ_X^2 is very large then $\hat{X}(t)=\overline{Y}$. Because X can be very far from μ_X and the filter have to rely on the measurement value only which is just the averaging.
- For general cases, the filter will try to balance the two sources of information about X according to their variances so that $\hat{X}(t)$becomes the best estimator.