Stochastic Processes

Topic 5

Fundamentals of Estimation

`nctuee09f`

Summary

In this topic, I will discuss:

- Fundamental Concept of Estimation
- Estimator Performance
- Sample Mean, Sample Variance and Gaussian Sample
- Interval Estimator
- T Random Variable
- Maximum Likelihood (ML) Estimation
- Least Squares Estimation
- Least Squares Using SVD
- Minimum Mean Squared Error (MMSE) Estimation
- Linear MMSE

Notation

We will use the following notation rules, unless otherwise noted, to represent symbols during this course.

- Boldface upper case letter to represent **MATRIX**
- Boldface lower case letter to represent **vector**
- Superscript $(\cdot)^T$ and $(\cdot)^H$ to denote transpose and hermitian (conjugate transpose), respectively
- Upper case italic letter to represent $RANDOM\ VARIABLE$

1 Estimation

Why Estimation?

(1) The parameter itself is of interest, such as the distance of an aircraft from the base of a radar system

(2) For the purpose of decision making
Knowledge of the parameter describes the statistical property, i.e. pdf, of observed (or measured) data $\mathbf{y}$, *e.g.*

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n},$$

where knowledge of $\boldsymbol{\theta}$ is essential to find the pdf of $\mathbf{y}$.

What is an Estimator?

An estimator $\hat{\boldsymbol{\theta}}$ is a ***function*** $g(\mathbf{y})$ of the observation vector $\mathbf{y}$ that estimates $\boldsymbol{\theta}$.

Example:
Let $Y_1, \cdots, Y_n$ be n observations with

$$y_i = \theta + \epsilon_i,$$

where θ is the unknown parameter we want to estimate, and ϵ_i's are measurement noises. A reasonable estimator for θ would be the sample mean

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Mathematic Model

(1) Model Formulation

In determining good estimators, the first step would be to mathematically and properly ***model*** the whole system, explicitly establishing the ***mathematical relation*** between the desired ***unknown quantities*** and the ***measured data***.

Example:
In the previous example, we have a model

$$Y_i = \theta + \epsilon_i,$$

where θ is the unknown parameter we want to estimate, Y_i is the ith measured data and ϵ_i's are measurement noises.
If the noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, the pdf of Y_i is given by

$$f_{Y_i}(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right).$$

(2) Generally, the measured data $\mathbf{y}$ can be a vector. In many applications, the measured data $\mathbf{y}$ is modeled to be ***linear*** with respect to the unknown parameter (denoted by $\boldsymbol{\theta}$), and can be expressed by

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n},$$

where $\mathbf{H}$ is commonly referred to as the *observation matrix* or *system matrix* and $\mathbf{n}$ is the measurement noise.

2 Estimator Performance

Questions asked to evaluate an estimator:

1. How close will $\hat{\theta}$ be to the real θ?
2. Are there any better estimators?

Typical Performance Measures:

(1) **Unbiased**
An estimator $\hat{\theta}$ for the parameter θ is said to be ***unbiased*** if $E[\hat{\theta}] = \theta$.

(2) **Consistent**
Let $\hat{\theta}_n$ be an estimator computed from n samples. Then, $\hat{\theta}_n$ is said to be ***consistent*** if

$$\lim_{n\to\infty} P[|\hat{\theta}_n - \theta| > \varepsilon] = 0 \quad \text{for every} \quad \varepsilon > 0. \tag{1}$$

(3) **Minimum mean squared error**
An estimator $\hat{\theta}$ is called a minimum mean square error (MMSE) estimator if

$$E[(\hat{\theta} - \theta)^2] \leq E[(\hat{\theta}' - \theta)^2]$$

for any other estimator $\hat{\theta}'$.

Remarks:

(a) The condition in (1) is also known as ***convergence in probability.***
In other words, $\hat{\theta}_n$ is consistent if it converges to θ in probability.

(b) How to check consistency of an unbiased estimator?

Chebyshev inequality states that for any arbitrary random variable X having mean $E[X]$ and finite variance $\text{Var}(X)$, we have

$$P\left[\,\left|X - E[X]\right| > k\right] \leq \frac{\text{Var}(X)}{k^2}, \quad \text{for any} \quad k > 0.$$

See page 205 in textbook for a proof.

3 Sample Mean and Sample Variance

Let $X_1, \cdots, X_n$ be i.i.d. random variables with mean $E[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. The sample mean

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an unbiased and consistent estimator for the mean μ. And, the sample variance

$$S_n^2 \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an unbiased and consistent estimator for the variance σ^2.

(1) Unbiasedness of sample mean

It is clear to see

$$E[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \mu.$$

(2) Consistency of sample mean

$\Rightarrow$ Use Chebyshev inequality

(3) Unbiasedness of sample variance

(4) Consistency of sample variance

This can be verified by examining whether $\lim_{n\to\infty} P[|S_n^2 - \sigma^2| > \varepsilon] = 0$. For that, we need to know the variance of the sample variance, which can be shown to be

$$\mathrm{Var}(S_n^2) = \frac{1}{n}\left[m_4 - \frac{n-3}{n-1}\sigma^2\right],$$

where $m_4 = E[(X_i - \mu)^4]$. It follows that, by inserting this result into the Chebyshev inequality,

$$\lim_{n\to\infty} P[|S_n^2 - \sigma^2| > \varepsilon] \le \lim_{n\to\infty} \frac{1}{n\varepsilon^2}\left[m_4 - \frac{n-3}{n-1}\sigma^2\right] = 0.$$

Remarks:

(1) The sample mean $\bar{X}_n$ is uncorrelated with the sequence of deviation $X_i - \bar{X}_n$ for $i = 1 \cdots n$.

(2) When $X_1 \cdots X_n$ are i.i.d Gaussian sample, $\bar{X}_n$ is "independent" with the sequence of deviation $X_i - \bar{X}_n$ for $i = 1 \cdots n$, due to

(a) $\text{Cov}(\bar{X}_n, X_i - \bar{X}_n) = 0$, and

(b) $\bar{X}_n$ and $X_i - \bar{X}_n$ are jointly Gaussian.

4 Gaussian Sample

(1) Let $X_1, \cdots, X_n$ be i.i.d. Gaussian random variables. We have shown that the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sequence of deviations $X_i - \bar{X}_n$, for $i = 1 \cdots n$ are independent.

We can deduce that, from the following theorem, $\bar{X}_n$ and the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ are independent.

(2) **An important theorem**
Let $\mathbf{y}_1, \cdots, \mathbf{y}_n$ be independent random vectors. And, let $g_i(\mathbf{y}_i)$ be a function only of $\mathbf{y}_i$, $i = 1 \cdots n$. Then, the random variables $U_i \triangleq g_i(\mathbf{y}_i)$, $i = 1 \cdots n$, are mutually independent.

(a) Let's see how to apply this theorem to the above.

(b) Now we give a proof for a simple case that $n = 2$, and $\mathbf{y}_1 \triangleq Y_1$ and $\mathbf{y}_2 \triangleq Y_2$ are both scalar random variables.

Define
$$U_1 \triangleq g_1(Y_1) \quad \text{and} \quad U_2 \triangleq g_2(Y_2).$$
We can find the joint probability distribution of U_1 and U_2 given by

$$\begin{aligned} F_{U_1,U_2}(u_1, u_2) &= P[U_1 \leq u_1, U_2 \leq u_2] \\ &= P[g_1(Y_1) \leq u_1, g_2(Y_2) \leq u_2] \\ &= P[Y_1 \in \mathsf{A}, Y_2 \in \mathsf{B}] \\ &= P[Y_1 \in \mathsf{A}] \cdot P[Y_2 \in \mathsf{B}], \end{aligned}$$

where the last equality stands from the assumption of independence between Y_1 and Y_2, and A and B are two sets satisfying $\mathsf{A} = \{y_1 : g_1(y_1) \leq u_1\}$ and $\mathsf{B} = \{y_2 : g_2(y_2) \leq u_2\}$, respectively. It follows the joint pdf

$$\begin{aligned} f_{U_1,U_2}(u_1, u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} F_{U_1,U_2}(u_1, u_2) \\ &= \left(\frac{\partial}{\partial u_1} P[Y_1 \in \mathsf{A}] \right) \cdot \left(\frac{\partial}{\partial u_2} P[Y_2 \in \mathsf{B}] \right) \\ &= f_{U_1}(u_1) f_{U_2}(u_2). \end{aligned}$$

■

(3) The independence property between $\bar{X}_n$ and S_n^2 when $X_1 \cdots X_n$ are i.i.d. Gaussian with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$ allows us to

(a) verify that $\frac{(n-1)S_n^2}{\sigma^2}$ is ***chi-squared distributed*** with $n-1$ degrees of freedom and,

(b) Find the pdf of

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

$\longrightarrow$ Student's T random variable

$\longrightarrow$ Commonly used to specify ***confidence interval*** of the estimator of μ

5 Confidence Interval

(1) For an interval estimator $[L(\mathbf{x}), U(\mathbf{x})]$ of a parameter θ based on the observation $\mathbf{x}$, we say that the confidence coefficient of this interval is $1-\alpha$ if

$$P\Big[\theta \in [L(\mathbf{x}), U(\mathbf{x})]\Big] \geq 1-\alpha,$$

or we say $[L(\mathbf{x}), U(\mathbf{x})]$ is a $(1-\alpha)\times 100\%$ confidence interval if

$$P\Big[L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})\Big] = 1-\alpha.$$

Note: The random quantity here is the interval (based on the observation $\mathbf{x}$), not the parameter θ. That is, the probability statements $P\big[L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})\big]$ refers to $\mathbf{x}$, not θ. Specifically, to find the probability, we actually need to find

$$P\Big[L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})\Big] = P\Big[\mathbf{x} : L(\mathbf{x}) \leq \theta \text{ and } \theta \leq U(\mathbf{x})\Big].$$

(2) Confidence interval of the mean μ for two cases:

(a) Unknown mean, known variance
Let $X_1, \cdots, X_n$ be i.i.d. Gaussian variables with unknown mean μ and known variance σ^2. The sample mean is a Gaussian random variable with $\bar{X}_n \sim N(\mu, \sigma^2/n)$ and

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

We can specify an interval $[-z, z]$ within which the normalized sample mean has a probability

$$P\left[-z \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right] = Q(-z) - Q(z) = 1 - 2Q(z),$$

where $Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ is the standard Q-function. With simple algebraic efforts, the above can be rewritten as

$$P\left[\bar{X}_n - \frac{\sigma z}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{\sigma z}{\sqrt{n}}\right] = 1 - 2Q(z). \qquad (2)$$

This means the interval

$$[\bar{X}_n - \frac{\sigma z}{\sqrt{n}}, \bar{X}_n + \frac{\sigma z}{\sqrt{n}}]$$

contains μ with probability $1-2Q(z)$. By letting $\alpha = 2Q(z)$, we can find a corresponding $z_{\alpha/2}$ such that this interval is a $(1-\alpha)\times 100\%$ confidence interval for μ.

(b) Unknown mean and unknown variance
Let $X_1, \cdots, X_n$ be i.i.d. Gaussian variables with unknown mean μ and unknown variance σ^2. The confidence interval now becomes

$$[\bar{X}_n - \frac{S_n z}{\sqrt{n}}, \bar{X}_n + \frac{S_n z}{\sqrt{n}}],$$

where the variance σ^2 is replaced by the sample variance S_n^2. So, the probability of μ containing in this interval is

$$P\left[\bar{X}_n - \frac{S_n z}{\sqrt{n}} \le \mu \le \bar{X}_n + \frac{S_n z}{\sqrt{n}}\right] = P\left[-z \le \underbrace{\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}}_{\triangleq T} \le z\right].$$

The random variable involved in figuring out the above probability measure is

$$T \triangleq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

We need to find the pdf of T in order to specify the interval. The random variable T is called Student's T random variable. With some rearrangement, we see

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}},$$

where the numerator $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is a standard normal random variable independent with $\frac{(n-1)S_n^2}{\sigma^2}$, which is a chi-squared random variable with $n-1$ degree of freedom, in the denominator.

Next, we will see the following 3 things:

(1) What is chi-squared distribution?

(2) How to justify $\frac{(n-1)S_n^2}{\sigma^2}$ is chi-squared distributed?

(3) How to find the pdf of T?

6 T Distribution

(1) **Review of chi-squared distribution**
If $Z_1, \cdots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables, then

$$Y \triangleq \sum_{i=1}^{n} Z_i^2 \tag{3}$$

has the ***chi-squared distribution*** with n degrees of freedom, denoted by $Y \sim \chi_n^2$.
When $n = 1$, we have $Y = Z_1^2$ and the pdf is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{\frac{1}{2} e^{\frac{-y}{2}} (\frac{1}{2} y)^{\frac{1}{2}-1}}{\sqrt{\pi}} \sim \Gamma(\frac{1}{2}, \frac{1}{2}), \end{aligned}$$

which is exactly the ***Gamma*** pdf with parameter $(\frac{1}{2}, \frac{1}{2})$. We can recall that the pdf of a Gamma random variable X with $X \sim \Gamma(n, \lambda)$ is

$$f_X(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{n-1}}{\Gamma(n)} \quad \text{x} > 0,$$

where $\Gamma(n) = \int_0^\infty e^{-u} u^{n-1} du$ with $\Gamma(n) = (n-1)!$, $\Gamma(\frac{n}{2}) = (\frac{n}{2} - 1)!$, and $\Gamma(1/2) = \sqrt{\pi}$.

— The chi-squared random variable in (3) is a summation of n independent Gamma random variables each with parameter $(\frac{1}{2}, \frac{1}{2})$.

— Use the fact that if $X_1 \sim \Gamma(n_1, \lambda)$ is independent with $X_2 \sim \Gamma(n_2, \lambda)$, then $X_1 + X_2 \sim \Gamma(n_1 + n_2, \lambda)$.

Thus,

$$\begin{aligned} Y &\sim \Gamma\left(\underbrace{\frac{1}{2} + \cdots + \frac{1}{2}}_{=n/2}, \frac{1}{2} \right) \\ f_Y(y) &= \frac{\frac{1}{2} e^{\frac{-y}{2}} (\frac{1}{2} y)^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})} \quad y > 0. \end{aligned}$$

(2) The MGF for $Y \sim \chi_n^2$ is $M_Y(t) = (1 - 2t)^{\frac{-n}{2}}$. This can be shown by first finding the MGF of Z_i^2 in (3). And,

$$M_Y(t) = \left(M_{Z_i^2}(t) \right)^n .$$

7 Justifying $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

(1) To find a confidence interval for the mean of i.i.d. Gaussian sample $X_1, \cdots, X_n$ with unknown variance, we need to know the distribution of $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$, where $\bar{X}_n$ is the sample mean and S_n is the sample variance. The random variable

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}} \overset{d}{=} \frac{U}{\sqrt{V/(n-1)}}$$

is called the Student's T random variable with $n-1$ degrees of freedom, where $U \sim \mathcal{N}(0,1)$ is independent with $V \sim \chi_{n-1}^2$.

(2) Now, we want to justify that $\frac{(n-1)S_n^2}{\sigma^2}$ is indeed chi-squared distributed with $n-1$ degrees of freedom. With some algebraic efforts, we have

$$\begin{aligned}
\frac{(n-1)S_n^2}{\sigma^2} &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 = \sum_{i=1}^n \left(\underbrace{\frac{X_i - \mu}{\sigma}}_{\triangleq Z_i} - \underbrace{\frac{(\bar{X}_n - \mu)}{\sigma}}_{\triangleq \bar{Z}_n}\right)^2 \\
&= \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \\
&= \left(\sum_{i=1}^n Z_i^2\right) - \left(\sqrt{n}\bar{Z}_n\right)^2,
\end{aligned}$$

where $Z_i \triangleq \frac{X_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$ and $\sqrt{n}\bar{Z}_n \triangleq \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ is also a standard Gaussian random variable. Rearranging the above yields

$$\frac{(n-1)S_n^2}{\sigma^2} + \underbrace{\left(\sqrt{n}\bar{Z}_n\right)^2}_{\sim\chi_1^2} = \underbrace{\left(\sum_{i=1}^n Z_i^2\right)}_{\sim\chi_n^2},$$

where the right hand side is by definition a chi-squared random variable with n degrees of freedom and has MGF equal to $(1-2t)^{\frac{-n}{2}}$. Also, we know that $(\sqrt{n}\bar{Z}_n)^2$ is a chi-squared random variable with 1 degree of freedom. With the fact that $\bar{X}_n$ and S_n are statistically independent in Gaussian sample, we can conclude that the MGF of $V \triangleq \frac{(n-1)S_n^2}{\sigma^2}$ is

$$M_V(t) = (1-2t)^{-\frac{(n-1)}{2}},$$

suggesting that V is a chi-squared random variable with $n-1$ degrees of freedom.

8 T Distribution

The pdf of a Student's T random variable T_n with n degrees of freedom is given by (see also p. 231 in textbook)

$$f_{T_n}(t) = K_{st} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}, \tag{4}$$

where $K_{st} = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}}$.
(**Derivation:**)
T_n by definition can be expressed by

$$T_n = \frac{U}{\sqrt{V/n}}, \quad \text{where} \quad U \sim N(0,1), \quad V \sim \chi_n^2,$$

and U is independent with V. We can first write down the joint pdf for U and V as

$$\begin{aligned} f_{UV}(u,v) &= f_U(u) f_V(v) \qquad (5) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{1}{2} e^{-\frac{1}{2}v}\right) \left(\frac{1}{2}v\right)^{\frac{n}{2}-1}, \quad -\infty < u < \infty \quad 0 < v < \infty. \end{aligned}$$

The idea to find the pdf of T_n is through the concept of linear transformation and through (5). Now, by introducing an auxiliary function $S = V$, we have

$$\begin{cases} T_n = \frac{U}{\sqrt{V/n}} \\ S = V \end{cases}$$

and its joint pdf can be found by means of

$$f_{T_nS}(t,s) = \frac{1}{|\mathrm{J}|} f_{UV}(u,v)\Big|_{v=s,u=\sqrt{\frac{s}{n}}t}, \tag{6}$$

where

$$|\mathrm{J}| = \begin{vmatrix} \frac{\partial T_n}{\partial U} & \frac{\partial T_n}{\partial V} \\ \frac{\partial S}{\partial U} & \frac{\partial S}{\partial V} \end{vmatrix} = \begin{vmatrix} \sqrt{\frac{n}{V}} & \Delta \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{\mathrm{n}}{\mathrm{v}}},$$

where Δ is something we don't care. And, our final goal can be achieved by evaluating

$$f_{T_n}(t) = \int_{-\infty}^{\infty} f_{T_nS}(t,s)ds. \tag{7}$$

To be more specific, the result of carrying out (6) is

$$\begin{aligned} f_{T_nS}(t,s) &= \sqrt{\frac{v}{n}} f_{UV}(u,v)\Big|_{v=s,u=\sqrt{\frac{s}{n}}t} \\ &= \frac{1}{\sqrt{2\pi}\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} e^{-\left(\frac{1}{2}+\frac{t^2}{2n}\right)s} s^{\frac{n+1}{2}-1}. \qquad (8) \end{aligned}$$

It follows, by observing that (8) takes the form of Gamma distribution and change of variables, the result of (7) is (4).

Remarks:

(1) Let's go back to our initial intention to find a confidence interval of μ with unknown variance. The probability of μ containing in the interval $[\bar{X}_n - \frac{S_n z}{\sqrt{n}}, \bar{X}_n + \frac{S_n z}{\sqrt{n}}]$ is

$$\begin{aligned} P\left[\bar{X}_n - \frac{S_n z}{\sqrt{n}} \le \mu \le \bar{X}_n + \frac{S_n z}{\sqrt{n}}\right] &= P\left[-z \le T_{n-1} \le z\right] \\ &= F_{T_{n-1}}(z) - F_{T_{n-1}}(-z) \\ &= 1 - 2F_{T_{n-1}}(-z), \end{aligned}$$

where the last equality comes from the fact that T distribution is symmetric (c.f.(4)). When $\alpha = 2F_{T_{n-1}}(-z)$ is specified, we can find a corresponding $z_{\alpha/2}$ such that the interval

$$\left[\bar{X}_n - \frac{S_n z}{\sqrt{n}}, \bar{X}_n + \frac{S_n z}{\sqrt{n}}\right]$$

is a $(1 - \alpha) \times 100\%$ confidence interval for μ.

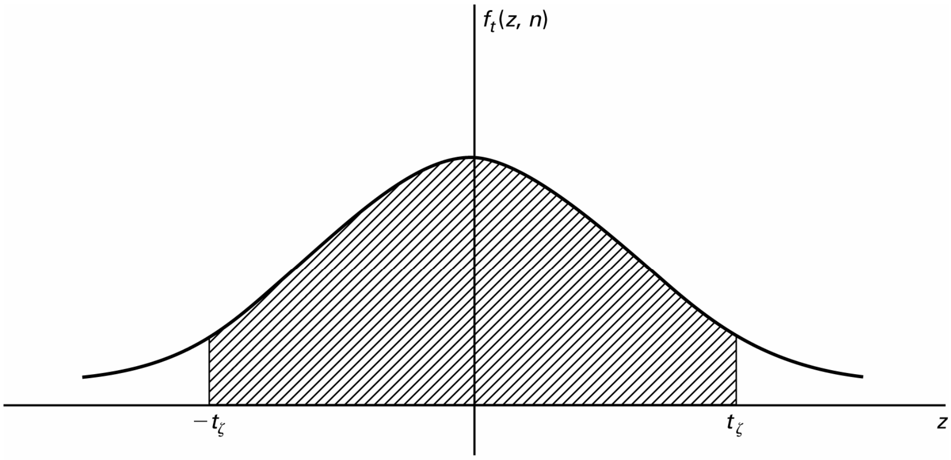

Figure 4.8-1

The numbers t_ζ. The area between $-t_\zeta$ and t_ζ is $1-2\zeta$.

Figure 1: The pdf of T random variable.

(2) T random variable also has a ***bell shape*** pdf symmetric with respect to the origin, but its bell is wider and shorter than standard normal. This implies, for a fixed confidence level $1-\alpha$, it is expected to have a wider (i.e. less precise) interval for μ when the variance is not known as compared to the case of known variance. This follows the intuition.

(3) As the number of observations n increases, the sample variance gets closer to the true variance in the sense that sample variance is a consistent estimator. As a result, the interval estimator will become narrower with increasing n. The interval estimator, i.e. $[\bar{X}_n - \frac{S_n z}{\sqrt{n}}, \bar{X}_n + \frac{S_n z}{\sqrt{n}}]$, with unknown variance will approach that with known variance, i.e. $[\bar{X}_n - \frac{\sigma z}{\sqrt{n}}, \bar{X}_n + \frac{\sigma z}{\sqrt{n}}]$,. In fact, T_n converges to standard normal in distribution when $n \to \infty$.

9 Maximum Likelihood Estimation

(1) **(Likelihood Function)**

Let $f_{\mathbf{x}}(\mathsf{x};\theta)$ be the joint pdf or pmf of the sample $\mathbf{x} = [X_1, X_2, \cdots, X_n]^T$. Then, given that $\mathbf{x} = \mathsf{x}^*$ is observed, the function of the unknown and ***deterministic*** parameter θ defined by

$$\mathrm{L}(\theta|\mathsf{x}^*) \triangleq f_{\mathbf{x}}(\mathsf{x}^*;\theta)$$

is called the ***likelihood function*** of θ given $\mathbf{x} = \mathsf{x}^*$.

(2) The ***maximum likelihood estimate*** (MLE) of θ by observing a sample $\mathbf{x} = [X_1, X_2, \cdots, X_n]^T$ is determined through

$$\begin{aligned} \hat{\theta}_{ML}(\mathbf{x}) &= \arg\max_{\theta} \mathrm{L}(\theta|\mathbf{x}) \\ &= \arg\max_{\theta} f_{\mathbf{x}}(\mathbf{x};\theta) \end{aligned}$$

Remarks:

- It should be noted that the parameter to be estimated in MMSE is modeled as random, while here the parameter to be estimated in MLE is non-random (deterministic).
- Obtaining an MLE involves (i) specifying the likelihood function, and (ii) finding the parameter value that maximizes the function.
- If the likelihood function is differentiable, possible candidates for the MLE are the values of $\theta_1, \cdots, \theta_k$ for a certain k that solves

  $$\frac{\partial}{\partial \theta_i} f_{\mathbf{x}}(\mathbf{x};\theta) = 0, \quad i = 1 \cdots k.$$

  Besides, we need to check the boundaries of the domain of θ as well.
- Points at which the first derivatives are 0 may be local or global ***minima***, local or global ***maxima***, or ***inflection points***. Our job in obtaining MLE is to find a ***global maximum***.
- In many cases, it is easier to work with the differentiation of the natural logarithm of $\mathrm{L}(\theta|\mathbf{x})$, $\log \mathrm{L}(\theta|\mathbf{x})$, known as the ***log likelihood***. Finding a θ that maximizes the likelihood function is the same thing as finding a θ that maximizes the log likelihood, since the log function is strictly increasing in $(0, \infty)$.

Example:
Let $X_1 \cdots X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$ with σ^2 known. The likelihood function of θ given $\mathsf{x} = [X_1 = \mathsf{x}_1, X_2 = \mathsf{x}_2, \cdots, X_n = \mathsf{x}_n]$ is

$$\begin{aligned} \mathrm{L}(\theta|\mathsf{x}) &= f_{\mathsf{x}}(\mathsf{x};\theta) = \prod_{i=1}^{n} f_{X_i}(\mathsf{x}_i;\theta) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathsf{x}_i - \theta)^2\right). \end{aligned}$$

And, the log likelihood function is

$$\log \mathrm{L}(\theta|\mathsf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathsf{x}_i - \theta)^2.$$

After taking the derivative, we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log \mathrm{L}(\theta|\mathsf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathsf{x}_i - \theta).$$

So, one possible candidate of MLE is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We still need to check
(i) whether or not θ is a maximum, and
(ii) boundaries of θ.

$\Rightarrow$
(i) The second derivative $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log \mathrm{L}(\theta|\mathsf{x}) = -\frac{n}{\sigma^2} < 0$. So, $\hat{\theta}$ is indeed a maximum.
(ii) Check boundaries $\theta \to \infty$ and $\theta \to -\infty$. It is straightforward to examine

$$\lim_{\theta\to\infty} \mathrm{L}(\theta|\mathsf{x}) = \lim_{\theta\to-\infty} \mathrm{L}(\theta|\mathsf{x}) = 0.$$

From (i) and (ii), we can conclude

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which is the sample mean of $X_1 \cdots X_n$. ■

10 Properties of MLE

Maximum likelihood estimation is perhaps the most widely used technique to find an estimate of unknown deterministic parameters due to the following nice properties.

(1) MLE is ***consistent***

$$\lim_{n\to\infty} P\big[|\hat{\theta}_{ML}(n) - \theta| > \varepsilon\big] = 0 \quad \forall\, \varepsilon > 0.$$

(2) MLE is ***asymptotically Gaussian***

$$\hat{\theta}_{ML}(n) \sim \text{Gaussian} \quad \text{as } n \to \infty.$$

(3) MLE is ***asymptotically efficient***
The asymptotic efficiency says that as $n \to \infty$

$$E[|\hat{\theta}_{ML}(n) - \theta|^2] \leq E[|\hat{\theta} - \theta|^2]$$

for any other estimators $\hat{\theta}$ of θ.

(4) MLE is ***invariant***
Suppose we know $\hat{\theta}_{ML}$ and would like to find the MLE of $\tau = g(\theta)$ for ***any*** functions $g(\cdot)$. The invariant property says that

$$\hat{\tau}_{ML} = g(\hat{\theta}_{ML}).$$

11 MLE for Gaussian Linear Model

Consider the linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w},$$

where $\mathbf{H}$ is a "known" $n \times p$ observation matrix and $\mathbf{w}$ is a noise vector of dimension $n \times 1$ with joint pdf $\mathcal{N}(0, \mathbf{K})$. Then, the maximum likelihood estimator for θ is given by

$$\hat{\boldsymbol{\theta}}_{ML} = \left(\mathbf{H}^T\mathbf{K}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{K}^{-1}\mathbf{y}. \tag{9}$$

Remarks:

(1) Use the following facts to justify (9).

- The derivative of the quadratic form $q(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ with respect to $\mathbf{x}$ is
$$\frac{dq(\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$
- Let $\mathbf{a}$ and $\mathbf{x}$ be two n-vectors. With $y = \mathbf{a}^T\mathbf{x}$, we have
$$\frac{dy}{d\mathbf{x}} = \mathbf{a}.$$
- Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{A}$ be two n-vectors and an $n \times n$ matrix, respectively. With $q = \mathbf{y}^T\mathbf{A}\mathbf{x}$, we have
$$\frac{dq}{d\mathbf{x}} = \mathbf{A}^T\mathbf{y}.$$

(2) When the noise vector $\mathbf{w}$ has uncorrelated entries, the MLE becomes

$$\hat{\boldsymbol{\theta}}_{ML} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y},$$

which is the ***least-squares*** estimator of $\boldsymbol{\theta}$.

(3) The MLE in (9) is a Gaussian random vector. Furthermore, it is an unbiased as well as the most ***efficient*** estimator ***within the class of linear estimators***.

— An unbiased estimator $\hat{\theta}$ of a scalar deterministic parameter θ is said to be more *efficient* than any other unbiased estimator $\hat{\theta}'$ if

$$\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\hat{\theta}').$$

— An unbiased estimator $\hat{\boldsymbol{\theta}}$ of a vector deterministic parameter $\boldsymbol{\theta}$ is said to be more ***efficient*** than any other vector unbiased estimator $\hat{\boldsymbol{\theta}}'$ if

$$\mathbf{K}_{\hat{\theta}} \leq \mathbf{K}_{\hat{\theta}'}$$

where the inequality for the matrix means $\mathbf{K}_{\hat{\theta}} - \mathbf{K}_{\hat{\theta}'}$ is a ***negative semi-definite*** matrix (or, $\mathbf{K}_{\hat{\theta}'} - \mathbf{K}_{\hat{\theta}}$ is a ***positive semi-definite*** matrix), and $\mathbf{K}_{\hat{\theta}}$ and $\mathbf{K}_{\hat{\theta}'}$ are the covariance matrix of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}'$, respectively.

12 Difference between MLE and MLD

The difference between maximum likelihood estimation (MLE) and maximum likelihood detection (MLD) can be explained by the fundamental differences between estimation and detection.

Detection

→ Decide among a finite set of alternatives whether a phenomenon is present or not.

→ Example
The receiver's task in a binary communication link is to decide whether the transmitter sends a 0 or a 1, which is a typical detection problem.

Estimation

→ Similarity to detection
Find out an unknown parameter based on the observations.

→ Difference
In estimation, the unknown parameters (may or may not be random) take value in a continuum of alternatives.

→ Example
The receiver needs to estimate possible unknown phase ranging from $[-\pi, \pi]$ in order to do a better job in detection. We need to find out a value of the unknown phase in the continuous domain $[-\pi, \pi]$.

13 Least Squares

Consider the linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w},$$

where $\mathbf{H}$ is a "known" $m \times n$ observation matrix, $\boldsymbol{\theta}$ is an $n \times 1$ unknown parameter which may or may not be random, and $\mathbf{w}$ is a noise vector.
Then, the least-squares estimator for θ that minimizes the 2-norm

$$||\mathbf{y} - \mathbf{H}\boldsymbol{\theta}||^2 = (\mathbf{y} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{H}\boldsymbol{\theta})$$

is given by

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}. \tag{10}$$

Remarks:

(1) Note that when $\mathbf{H}$ is square and non-singular, the least-squares estimator is reduced to

$$\hat{\boldsymbol{\theta}}_{LS} = \mathbf{H}^{-1}\mathbf{y}.$$

(2) The matrix $\mathbf{H}^\dagger \triangleq \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ is called the pseudo-inverse of $\mathbf{H}$.

(3) The matrix $\mathbf{H}^T\mathbf{H}$ must be non-singular for (10) to hold true, which requires $\mathbf{H}$ being of full rank. In practice, we solve the least-squares problems using the following system of normal equations:

$$\left(\mathbf{H}^T\mathbf{H}\right)\hat{\boldsymbol{\theta}}_{LS} = \mathbf{H}^T\mathbf{y}.$$

(4) Let $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\theta}}_{LS}$. From the normal equations we will find

$$\mathbf{H}^T\tilde{\mathbf{y}} = \mathbf{0}.$$

This is known as the ***orthogonality condition***.

(5) The minimum least-squares is found as

$$\begin{aligned} J_{min} &= ||\mathbf{y} - \mathbf{H}\boldsymbol{\theta}_{LS}||^2 \\ &= \mathbf{y}^T\left(\mathbf{I} - \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\right)\mathbf{y} \end{aligned}$$

14 Geometric Interpretations

The least-squares problem for the linear model

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

can be interpreted geometrically, from the concept of distance by matrix 2-norm.

(1) The received signal $\mathbf{y} \in \mathbb{R}^m$.
If the matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$ for $m \geq n$ is full-rank, then the range space of $\mathbf{H}$ is $\mathbb{R}^n$, which is a subspace of $\mathbb{R}^m$.

(2) The LS estimate $\boldsymbol{\theta}_{LS}$ is the vector that renders $\hat{\mathbf{s}} = \mathbf{H}\boldsymbol{\theta}_{LS}$ the ***orthogonal projection*** of the vector $\mathbf{y}$ onto the subspace spanned by the column vectors of $\mathbf{H}$, i.e. the range of $\mathbf{H}$. The orthogonal projection is given by

$$\hat{\mathbf{s}} = \underbrace{\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T}_{\triangleq \mathbf{P}} \cdot \mathbf{y},$$

where $\mathbf{P} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ is the projection matrix of any vector in $\mathbb{R}^m$, such as $\mathbf{y}$, onto the range of $\mathbf{H}$.

a) Idempotent $\mathbf{P} = \mathbf{P}^2$

b) Symmetric $\mathbf{P} = \mathbf{P}^T$

c) $\mathbf{P}^{\perp} \triangleq \mathbf{I} - \mathbf{P}$ is also a projection matrix. We have

$$J_{min} = ||\mathbf{P}^{\perp}\mathbf{y}||^2.$$

15 Least Squares Using SVD

The LS estimate can be computed in terms of the SVD of the matrix $\mathbf{H}$. More specifically, the SVD for $\mathbf{H}$ is

$$\mathbf{H} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}^H,$$

where $\mathbf{U}$ and $\mathbf{V}$ are $m \times m$ and $n \times n$ unitary matrices, respectively, and

$$\mathbf{D} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right],$$

with $\text{rank}(\mathbf{H}) = r$. Then, we have the least-square estimate given by

$$\hat{\boldsymbol{\theta}}_{LS} = \mathbf{V} \left[\begin{array}{c|c} \boldsymbol{\Sigma}_r^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \mathbf{U}^H \cdot \mathbf{y}.$$

16 Minimum Mean-Squared Error (MMSE) Estimation

(1) **Orthogonality Principle**
For random vectors $\mathbf{x}$ and $\mathbf{y}$ with ***arbitrary*** distributions, the orthogonality principle states that $\mathbf{x} - E[\mathbf{x}|\mathbf{y}]$ is orthogonal to $k(\mathbf{y})$ for any function $k(\cdot)$.

Recall that orthogonality between random vectors $\mathbf{x} - E[\mathbf{x}|\mathbf{y}]$ and $k(\mathbf{y})$ means

$$E\left[\Big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\Big) \cdot k^T(\mathbf{y})\right] = \mathbf{0}$$

with all the vectors, including the zero vector, having proper dimensions. We can see this by carrying out

$$\begin{aligned} E\left[\Big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\Big) \cdot k^T(\mathbf{y})\right] &= E\left[\mathbf{x}k^T(\mathbf{y})\right] - E\Big[E\left[\mathbf{x}|\mathbf{y}\right]k^T(\mathbf{y})\Big] \\ &= E\left[\mathbf{x}k^T(\mathbf{y})\right] - E\Big[E\left[\mathbf{x}k^T(\mathbf{y})|\mathbf{y}\right]\Big] \\ &= E\left[\mathbf{x}k^T(\mathbf{y})\right] - E\left[\mathbf{x}k^T(\mathbf{y})\right] \\ &= \mathbf{0}. \end{aligned}$$

■

We can consider $E[\mathbf{x}|\mathbf{y}]$ as the orthogonal projection of $\mathbf{x}$ onto the space spanned by all the functions of $\mathbf{y}$.

(2) **Fundamental Theorem**

Suppose we want to estimate an unknown random vector $\mathbf{x}$ based on the observation vector $\mathbf{y}$ through a rule $g(\mathbf{y})$. The estimator that minimizes $E\left[||\mathbf{x} - g(\mathbf{y})||^2\right]$ is called the minimum mean squared error (MMSE) estimator, and is given by

$$\boxed{g_{mmse}(\mathbf{y}) = \arg\min_{g(\mathbf{y})} E\left[||\mathbf{x} - g(\mathbf{y})||^2\right] = E[\mathbf{x}|\mathbf{y}]} \tag{11}$$

Proof:

We will show the fundamental theorem by means of 2 different approaches, one with the orthogonality principle and the other with direct manipulations of the cost function $E\left[||\mathbf{x} - g(\mathbf{y})||^2\right]$.

I. (From orthogonality principle)

$$\begin{aligned}
E\left[||\mathbf{x} - g(\mathbf{y})||^2\right] &= E\left[||\mathbf{x} - E[\mathbf{x}|\mathbf{y}] + E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})||^2\right] \\
&= E\left[\left|\left|\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\right|\right|^2\right] + E\left[\left|\left|E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\right|\right|^2\right] \\
&+ \underbrace{E\left[\left(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\right)\left(E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\right)^T\right]}_{(A)} \\
&+ \underbrace{E\left[\left(E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\right)\left(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\right)^T\right]}_{(B)}.
\end{aligned}$$

Since $E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})$ is a function only of the vector $\mathbf{y}$, we know that according to the orthogonality principle, (A) and (B) in the above are zero vectors. Therefore, we have the mean squared error (MSE)

$$E\left[||\mathbf{x} - g(\mathbf{y})||^2\right] = E\left[\left|\left|\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\right|\right|^2\right] + E\left[\left|\left|E[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\right|\right|^2\right].$$

Our goal is to find a rule $g(\mathbf{y})$ that minimizes the above mean squared error. It is evident that

$$g_{mmse}(\mathbf{y}) = E[\mathbf{x}|\mathbf{y}]$$

satisfies the minimum MSE criterion, and the resulting MSE is

$$\begin{aligned}
\text{MSE} &= E\big[||\mathbf{x} - g_{mmse}(\mathbf{y})||^2\big] \\
&= E\left[\big|\big|\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big|\big|^2\right] \\
&= E\left[\operatorname{tr}\Big(\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)^T\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)\Big)\right] \\
&= E\left[\operatorname{tr}\Big(\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)^T\Big)\right] \\
&= \operatorname{tr}\Big(E\left[\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)^T\right]\Big) \\
&= \operatorname{tr}\Big(E\Big[E\left[\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)\big(\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big)^T\Big|\mathbf{y}\right]\Big]\Big) \\
&= \operatorname{tr}\Big(E\left[\mathbf{K}_{\mathbf{x}|\mathbf{y}}\right]\Big).
\end{aligned}$$

II. Another way to show the fundamental theorem of estimation theory is by direct manipulations of the MSE as follows:

$$
\begin{aligned}
E\left[||\mathbf{x}-g(\mathbf{y})||^2\right] &= \int\int ||\mathbf{x}-g(\mathbf{y})||^2 f_{\mathbf{xy}}(\mathsf{x},\mathsf{y})d\mathsf{x}d\mathsf{y} \\
&= \int\int ||\mathbf{x}-g(\mathbf{y})||^2 f_{\mathbf{x}|\mathbf{y}}(\mathsf{x}|\mathsf{y})f_{\mathbf{y}}(\mathsf{y})d\mathsf{x}d\mathsf{y} \\
&= \int \underbrace{\left(\int ||\mathbf{x}-g(\mathbf{y})||^2 f_{\mathbf{x}|\mathbf{y}}(\mathsf{x}|\mathsf{y})d\mathsf{x}\right)}_{=E\left[||\mathbf{x}-g(\mathbf{y})||^2\middle|\mathbf{y}\right]} f_{\mathbf{y}}(\mathsf{y})d\mathsf{y} \\
&= \int E\left[||\mathbf{x}-g(\mathbf{y})||^2\middle|\mathbf{y}\right] f_{\mathbf{y}}(\mathsf{y})d\mathsf{y}.
\end{aligned}
$$

Since the joint pdf $f_{\mathbf{y}}(\mathsf{y})$ is everywhere non-negative, minimizing the MSE $E\left[||\mathbf{x}-g(\mathbf{y})||^2\right]$ by choosing a proper $g(\mathbf{y})$ is equivalent to minimizing the conditional MSE $E\left[||\mathbf{x}-g(\mathbf{y})||^2\middle|\mathbf{y}\right]$ with the same $g(\mathbf{y})$, i.e.,

$$
\arg\min_{g(\mathbf{y})} E\left[||\mathbf{x}-g(\mathbf{y})||^2\right] = \arg\min_{g(\mathbf{y})} E\left[||\mathbf{x}-g(\mathbf{y})||^2|\mathbf{y}\right].
$$

So, we can turn our focus to the conditional MSE. Carrying out the conditional MSE yields

$$
\begin{aligned}
E\left[||\mathbf{x}\right. &\left.-g(\mathbf{y})||^2|\mathbf{y}\right] \\
&= E\left[\left(\mathbf{x}-g(\mathbf{y})\right)^T\left(\mathbf{x}-g(\mathbf{y})\right)|\mathbf{y}\right] \\
&= E\left[\mathbf{x}^T\mathbf{x}|\mathbf{y}\right] - E\left[g(\mathbf{y})^T\mathbf{x}|\mathbf{y}\right] - E\left[\mathbf{x}^T g(\mathbf{y})|\mathbf{y}\right] + E\left[g(\mathbf{y})^T g(\mathbf{y})|\mathbf{y}\right] \\
&= E\left[\mathbf{x}^T\mathbf{x}|\mathbf{y}\right] - g(\mathbf{y})^T E\left[\mathbf{x}|\mathbf{y}\right] - E\left[\mathbf{x}^T|\mathbf{y}\right]g(\mathbf{y}) + g(\mathbf{y})^T g(\mathbf{y}).
\end{aligned}
$$

With further inspection, we find that the above result is in a quadratic form with respect to $g(\mathbf{y})$. It follows that

$$
\begin{aligned}
E\left[||\mathbf{x}\right. &- \left.g(\mathbf{y})||^2|\mathbf{y}\right] \\
&= \left(g(\mathbf{y})-E\left[\mathbf{x}|\mathbf{y}\right]\right)^T\left(g(\mathbf{y})-E\left[\mathbf{x}|\mathbf{y}\right]\right) + E\left[||\mathbf{x}||^2|\mathbf{y}\right] - \left|\left|E\left[\mathbf{x}|\mathbf{y}\right]\right|\right|^2.
\end{aligned}
$$

The conditional MSE, and therefore the objective MSE, is minimized when

$$
g_{mmse}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right].
$$

■

Remarks:

(1) Although the MMSE estimator has a simple form $E[\mathbf{x}|\mathbf{y}]$, finding it requires the knowledge of conditional pdf $f_{\mathbf{x}|\mathbf{y}}(\mathsf{x}|\mathsf{y})$, which is often difficult to obtain.

(2) When $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, the estimator that minimizes the MSE is

$$E[\mathbf{x}|\mathbf{y}] = \mathbf{m_x} + \mathbf{K_{xy}}\mathbf{K_y}^{-1}\big(\mathbf{y} - \mathbf{m_y}\big),$$

where $\mathbf{m_x} = E[\mathbf{x}]$, $\mathbf{m_y} = E[\mathbf{y}]$, $\mathbf{K_{xy}} = E[(\mathbf{x} - \mathbf{m_x})(\mathbf{y} - \mathbf{m_y})^{\mathbf{T}}]$, and $\mathbf{K_y} = E[(\mathbf{y} - \mathbf{m_y})(\mathbf{y} - \mathbf{m_y})^{\mathbf{T}}]$. And, the MSE is given by

$$\mathrm{MSE} = \mathrm{tr}\left(\mathbf{K_x} - \mathbf{K_{xy}}\mathbf{K_y}^{-1}\mathbf{K_{yx}}\right).$$

17 Linear MMSE

(1) **Why linear MMSE?**
It is often desirable to find an MMSE estimator constrained to be a linear function of the observations, due to reasons such as easier implementations of linear systems and, as mentioned, difficulties in finding $E[\mathbf{x}|\mathbf{y}]$.

(2) **Problem Formulation**
Suppose now $\mathbf{x}$ and $\mathbf{y}$ are not necessarily jointly Gaussian random vectors, and we know $\mathbf{m_x}$, $\mathbf{m_y}$, $\mathbf{K_{xy}}$, and $\mathbf{K_y}$. In this case, the estimator that takes the form

$$g(\mathbf{y}) = \mathbf{A} \cdot \mathbf{y} + \mathbf{b}$$

and minimizes the MSE at the same time is given by

$$g_{lmmse}(\mathbf{y}) = \mathbf{m_x} + \mathbf{K_{xy}}\mathbf{K_y}^{-1}\big(\mathbf{y} - \mathbf{m_y}\big) \triangleq L[\mathbf{x}|\mathbf{y}]$$

Proof:
We start with proving

$$E\left[\Big(\mathbf{x} - L[\mathbf{x}|\mathbf{y}]\Big) \cdot \big(\mathbf{A}\mathbf{y} + \mathbf{b}\big)^T\right] = \mathbf{0}, \tag{12}$$

for all matrices $\mathbf{A}$ and vectors $\mathbf{b}$, which is an extension of the orthogonality principle to the case of LMMSE. This can be easily shown by

$$\begin{aligned} E\Big[\Big(\mathbf{x} \;-\; & L[\mathbf{x}|\mathbf{y}]\Big) \cdot \big(\mathbf{A}\mathbf{x} + \mathbf{b}\big)^T\Big] \\ &= E\left[\Big(\mathbf{x} - \mathbf{m_x} - \mathbf{K_{xy}}\mathbf{K_y}^{-1}(\mathbf{y} - \mathbf{m_y})]\Big) \cdot \Big(\mathbf{A}(\mathbf{y} - \mathbf{m_y}) + \mathbf{b}'\Big)^T\right] \\ &= \mathbf{K_{xy}}\mathbf{A}^T - \mathbf{K_{xy}}\mathbf{K_y}^{-1}\mathbf{K_y}\mathbf{A}^T \\ &= \mathbf{0}, \end{aligned}$$

where $\mathbf{b}' = \mathbf{A}\mathbf{m_y} + \mathbf{b}$. The above extended orthogonality principle says that $L[\mathbf{x}|\mathbf{y}]$ is the orthogonal projection of $\mathbf{x}$ onto the space spanned by any *linear* functions of $\mathbf{y}$.

Next, with a similar procedure to what we've done in proving the general MMSE, we have

$$\begin{aligned} E\big[||\mathbf{x} - g(\mathbf{y})||^2\big] &= E\big[||\mathbf{x} - L[\mathbf{x}|\mathbf{y}] + L[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})||^2\big] \\ &= E\left[\big|\big|\mathbf{x} - L[\mathbf{x}|\mathbf{y}]\big|\big|^2\right] + E\left[\big|\big|L[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\big|\big|^2\right] \\ &+ \underbrace{E\left[\big(\mathbf{x} - L[\mathbf{x}|\mathbf{y}]\big)\big(L[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\big)^T\right]}_{(A)} \\ &+ \underbrace{E\left[\big(L[\mathbf{x}|\mathbf{y}] - g(\mathbf{y})\big)\big(\mathbf{x} - L[\mathbf{x}|\mathbf{y}]\big)^T\right]}_{(B)}, \end{aligned}$$

where (A) and (B) are zero vectors according to (12). We then can assure that

$$g_{lmmse}(\mathbf{y}) = L[\mathbf{x}|\mathbf{y}] = \mathbf{m_x} + \mathbf{K_{xy}}\mathbf{K_y}^{-1}\big(\mathbf{y} - \mathbf{m_y}\big).$$

■

Remark:
The MSE of LMMSE is generally larger than that of MMSE, since

$$\begin{aligned} E\big[||\mathbf{x} - L[\mathbf{x}|\mathbf{y}]||^2\big] &= E\big[||\mathbf{x} - E[\mathbf{x}|\mathbf{y}] + E[\mathbf{x}|\mathbf{y}] - L[\mathbf{x}|\mathbf{y}]||^2\big] \\ &= E\left[\big|\big|\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big|\big|^2\right] + E\left[\big|\big|L[\mathbf{x}|\mathbf{y}] - E[\mathbf{x}|\mathbf{y}]\big|\big|^2\right] \\ &\geq E\left[\big|\big|\mathbf{x} - E[\mathbf{x}|\mathbf{y}]\big|\big|^2\right], \end{aligned}$$

with the equality holds when $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian random vectors.

Example:
Suppose we want to estimate X from the observation of

$$Y = X + Z,$$

where $X \sim \mathcal{N}(0, \sigma_X^2)$ is independent $Z \sim \mathcal{N}(0, \sigma_Z^2)$. We know the MMSE estimate of X is

$$\hat{X}_{mmse} = E[X|Y].$$

Since X and Y are jointly Gaussian (by showing $aX + bY$ is a Gaussian random variable for any a and b), we have

$$\begin{aligned} \hat{X}_{mmse} &= E[X|Y] = m_X + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}(Y - m_Y) \\ &= \mathbf{K}_{XY}\mathbf{K}_Y^{-1}Y = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}Y. \end{aligned}$$

Also, by symmetry, we can obtain $\hat{Z}_{mmse} = \frac{\sigma_Z^2}{\sigma_X^2+\sigma_Z^2}Y$, giving

$$\hat{X}_{mmse} + \hat{Z}_{mmse} = Y.$$

This indicates that the estimation splits the observation between signal and noise according to their variances (i.e, average power or energy). Intuitively, when $E[X^2] > E[Z^2]$, we want to attribute the major part of Y to X, and the math tells us it is so indeed.