Module C: Estimation

Outline:

- Basic Estimation Theory: ML, MAP
- Conditional Expectation, and Mean Square Estimation
- Orthogonality Principle and LMMSE Estimator

Estimation Theory

- Main Question: Given an observation Y of a random variable X, how to estimate X?
- In other words, what is the best function g such that $\hat{X} = g(Y)$ is the best estimator? How to quantify "best"?
- More generally: given a sequence of observation of $\widehat{y}_1, \ldots, \widehat{y}_k$, how to estimate X?
- Example: Radar detection: Suppose that X is the radial distance of an aircraft from a radar station and $Y = X + Z$ is the radar's observed location where Z is independent of X and $Z \sim \mathcal{N}(0, \sigma^2)$. What is the best estimator $\widehat{X} = g(Y)$ of the location X?

Motivating Example

- Let X be a random variable which is uniformly distributed over $[0, \theta]$.
- We observe m samples of X denoted $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_m$.
- Problem: estimate θ given our observations.
- Let the samples be $\{1, 2, 1.5, 1.75, 2, 1.3, 0.8, 0.3, 1\}$.
- What is a good estimate of θ?
- Can we find a function $g(\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_m)$ which will map any set of m samples into an estimate of θ? Such a function is termed "estimator."
- We often treat the observations as random variables that depend on the quantities that we are trying to estimate.
- Case 1: The unknown quantity θ is assumed to be an unknown parameter/-constant with observation $\underline{X \sim \text{distribution}(\theta)}$
- Case 2: The unknown quantity θ is assumed to be a random variable.

Maximum Likelihood Estimation (θ is a parameter)

- We observe X which is assumed to be a random variable whose distribution depends on an unknown parameter θ.
- When X is continuous, its density $f_X(x;\theta)$.
- When X is discrete, its pmf $p_X(x;\theta)$.
- When the observation is $\widehat{x}$, we define Likelihood function as

$$\mathcal{L}(\theta|X=\widehat{x}) = \begin{cases} f_X(\widehat{x};\theta) & \text{when} \quad X \quad \text{is continuous,} \\ p_X(\widehat{x};\theta) & \text{when} \quad X \quad \text{is discrete.} \end{cases}$$

- The maximum likelihood estimate of θ when $X=\widehat{x}$ is

$$\hat{\theta}_{ML}(\widehat{x}) := \text{argmax}_\theta \quad \mathcal{L}(\theta|X=\widehat{x}).$$

- Thus, maximum likelihood estimate is the value of θ which maximizes the likelihood of observing $\widehat{x}$.

Log Likelihood Estimation

- We rarely estimate a quantity based on a single observation.
- Suppose we have N i.i.d observations, $\{\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_N\}$ each drawn from the same distribution.
- Likelihood function is then computed as

$$
\begin{aligned}
&\mathcal{L}(\theta | X_1 = \widehat{x}_1, X_2 = \widehat{x}_1, \ldots, X_N = \widehat{x}_n) = f_{X_1, X_2, \ldots, X_N}(\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_N; \theta) \\
&\quad = f_{X_1}(\widehat{x}_1; \theta) \times f_{X_2}(\widehat{x}_2; \theta) \ldots \times f_{X_N}(\widehat{x}_N; \theta) \quad \text{(due to independence of observations)} \\
&\quad = f_X(\widehat{x}_1; \theta) \times f_X(\widehat{x}_2; \theta) \ldots \times f_X(\widehat{x}_N; \theta) \quad \text{(each } X_i \text{ has identical distribution)} \\
&\quad = \prod_{i=1}^{N} f_X(\widehat{x}_i; \theta) = \prod_{i=1}^{N} \mathcal{L}(\theta | X_i = \widehat{x}_i).
\end{aligned}
$$

- Product term is difficult to maximize. However, we can compute the log-likelihood as

$$\log(\mathcal{L}(\theta | X_1 = \widehat{x}_1, X_2 = \widehat{x}_1, \ldots, X_N = \widehat{x}_n)) = \sum_{i=1}^{N} \log(f_X(\widehat{x}_i; \theta))$$

which is often easier to maximize with respect to θ.

Example

- Consider a random variable X defined as

$$X = \begin{cases} 1 & \text{with probability} \quad \theta \\ 0 & \text{with probability} \quad 1-\theta \end{cases} \quad , \qquad \theta \in [0,1].$$

- We observe $\{\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_N\}$ with each $\widehat{x}_i \in \{0,1\}$.
- Problem: find $\hat{\theta}_{ML}(\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_N)$
- The likelihood function $\mathcal{L}(\theta | X_1 = x_1, X_2 = x_2 .... X_N = x_n) = \quad .$
- The log-likelihood function $\log(\mathcal{L}(\theta | X_1 = x_1, X_2 = x_2 .... X_N = x_n)) = \quad .$
- Optimizing log-likelihood function with respect to θ yields
- ML Estimator $\hat{\theta}_{ML}(X_1, X_2, ...X_N)$ is a r.v that is function of $X_1, ...X_N$ given by

$$\hat{\theta}_{ML}(X_1, X_2, ...X_N) = \qquad .$$

- When X is a discrete random variable with p.m.f. $[\theta_1 \ \theta_2 \ ....\theta_N] = \theta$ with

$$\mathbb{P}(X = 1) = \theta_1, \qquad \mathbb{P}(X = 2) = \theta_2 \qquad ... \qquad \text{and so on.}$$

Then, the likelihood function $\mathcal{L}(\theta | X = i) = \theta_i$. What is the likelihood function after N observations?

Conditional distribution

- Recall that conditional probability of two events A and B is defined as

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

- Example: let X_1 : outcome of one coin toss with

$$X_1 = \begin{cases} 1, & \text{with probability} \quad p \\ 0, & \text{with probability} \quad 1-p. \end{cases}$$

- Let X_2 : be outcome of second coin toss, and X_2 has same distribution as X_1.

- Joint pmf: $p_{X_1X_2}(x_1, x_2) = \begin{cases} p^2 \quad \text{when} \quad (x_1, x_2) = (1,1) \\ p(1-p) \quad \text{when} \quad (x_1, x_2) = (1,0) \\ p(1-p) \quad \text{when} \quad (x_1, x_2) = (0,1) \\ (1-p)^2 \quad \text{when} \quad (x_1, x_2) = (0,0) \end{cases}$

- Conditional pmf of X_1 conditioned on X_2:

$$p_{X_1|X_2}(x_1|X_2 = x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)}.$$

- Conditional pmf of X_1 given $X_2 = 0$ is given by:

$$p_{X_1|X_2}(0|X_2 = 0) = \mathbb{P}(X_1 = 0|X_2 = 0) =$$
$$p_{X_1|X_2}(1|X_2 = 0) = \mathbb{P}(X_1 = 1|X_2 = 0) =$$

Conditional Distributions

- Consider two discrete random variables X and Y. Let X takes values from the set $\{x_1, \ldots, x_n\}$ and let Y takes values from the set $\{y_1, \ldots, y_m\}$.

- Conditional pmf of X given $Y = y_j$ is given by:

$$p_{X|Y}(x_i|Y = y_j) = \mathbb{P}(X = x_i|Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} \quad \forall i \in \{1, 2, \ldots, n\}.$$

- The numerator is obtained from the joint distribution of X and Y. The denominator is obtained from the marginal distribution of Y.

- For two continuous random variables X and Y conditional CDF is given by

$$F_{X|Y}(x|y) = \mathbb{P}(X \le x|Y \le y) = \frac{F_{X,Y}(x, y)}{F_Y(y)}.$$

- In this case, the conditional density is given by

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

Example

Consider two continuous random variables X and Y with joint density

$$f_{XY}(x,y) = \begin{cases} x+y & \text{if } 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine $\mathbb{P}(X < \frac{1}{4} | Y = \frac{1}{3})$ by deriving and using the conditional density of X given Y.

Example

Consider a random variable X whose density is given by

$$f_X(x) = \begin{cases} 1 & , \ 0 \leq x \leq 1 \\ 0 & , \ \text{otherwise} \end{cases}$$

The conditional density of Y given $X = x$ is given by

$$f_{Y|X=x}(y) = \begin{cases} \frac{1}{1-x}, & x \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Determine the marginal density of Y.

Maximum A-Posteriori (MAP) Estimation

- ML estimators assume θ to be an unknown parameter. If instead θ is a r.v with some distribution that is known, we use a Bayesian approach to estimate θ.

- We assume prior distribution: $f_\theta(\theta)/p_\theta(\theta)$ of θ that is known to us beforehand.

- Conditional distribution: $f_{X|\theta}(x|\theta)$ is also as some to be known. The distribution of the observed quantity is known if the unknown parameter is exactly known.

- Once we observe $X = \widehat{x}$, we find posterior distribution using Baye's law as:

$$\begin{aligned} f_{\theta|X}(\theta|X = \widehat{x}) &= \frac{f_{\theta,X}(\theta, \widehat{x})}{f_X(\widehat{x})} \\ &= \frac{f_{X|\theta}(\widehat{x}|\theta) f_\theta(\theta)}{f_X(\widehat{x})} \\ &= \frac{f_{X|\theta}(\widehat{x}|\theta) f_\theta(\theta)}{\int_\theta f_{X|\theta}(\widehat{x}|\theta) f_\theta(\theta) d\theta}. \end{aligned}$$

- The MAP estimate is defined as:

$$\hat{\theta}_{\mathsf{MAP}}(\widehat{x}) = \mathsf{argmax}_\theta \quad f_{\theta|X}(\theta|X = \widehat{x}) = \mathsf{argmax}_\theta \quad f_{X|\theta}(\widehat{x}|\theta) f_\theta(\theta),$$

  which is the mode of the posterior distribution.

Example (Previous year End Semester Question)

Suppose Θ is a random parameter, and given $\Theta = \theta$, the observed quantity Y has conditional density

$$f_{Y|\Theta}(y|\theta) = \frac{\theta}{2} e^{-\theta|y|}, y \in \mathbb{R}.$$

1. Find the Maximum Likelihood (ML) estimate of Θ based on the observation $Y = -0.5$.

   Suppose further that Θ has prior density given by $f_\Theta(\theta) = \frac{1}{\theta}, 1 \leq \theta \leq e$ (and $f_\Theta(\theta) = 0$ for $\theta < 1$ and $\theta > e$.). Then,

2. find the Maximum A-Posteriori (MAP) estimate of Θ based on the observation $Y = -0.5$.

Answer: $\widehat{\Theta}_{ML}(Y = -0.5) = 2, \widehat{\Theta}_{MAP}(Y = -0.5) = 1$.

Mean Square Estimation Theory

- The best is subjective and need to set a criteria. One popular criteria is *Mean Square Error (MSE)*.

- For measurements $X_1, \ldots, X_k$ of a random variable X, we define the MSE of (a measurable) an estimator (function) $g : \mathbb{R}^k \to \mathbb{R}$ to be

$$\mathbb{E}[|g(X_1, \ldots, X_k) - X|^2].$$

- In this setting, we view $\mathbb{E}[|U - X|^2]$ as the squared *distance* of random variables U and X.

- Once we fix the MSE criteria for the best estimator, then the problem of finding the best MSE estimator for X based on the measurements $X_1, \ldots, X_k$ can be formulated as:

$$\arg \min_{g:\mathbb{R}^k \to \mathbb{R}} \mathbb{E}[|g(X_1, \ldots, X_k) - X|^2].$$

- Any g that minimizes the above criteria is called a Minimum Mean Square Error (MMSE) estimator.

- When solving for MMSE, we always assume that all the random variables involved have finite mean and variance.

MMSE

- In practice: finding the MMSE *might be* hard.
- We can restrict our attention to special classes of functions g.
- Let $k = 0$, and suppose that we want to find the best *constant* c that estimates X. Note that in this case, we view c as a constant random variable.

$$\text{objective: finding } c \in \operatorname{argmin}_c \mathbb{E}[|X - c|^2]. \tag{1}$$

- Let $\bar{X} = \mathbb{E}[X]$. Then,

$$\begin{aligned}\mathbb{E}[|X - c|^2] &= \mathbb{E}[|X - \bar{X} + \bar{X} - c|^2] \\ &= \mathbb{E}[|X - \bar{X}|^2 + 2(\bar{X} - c)\mathbb{E}[(X - \bar{X})] + (\bar{X} - c)^2 \\ &= \mathbb{E}[(X - \bar{X})^2] + \mathbb{E}[(\bar{X} - c)^2].\end{aligned}$$

- Therefore, (1) is minimized when $c = \bar{X}$ and MMSE value is going to be Var(X).
- **Estimation theory interpretation of mean and variance**: The best constant MMSE estimator of X is $\mathbb{E}[X]$ and the corresponding MMSE value is Var(X).

Conditional Expectation

Example: Let X, Y be discrete r.v with $(X, Y \in \{1, 2\})$ and joint pmf:

$$\mathbb{P}[X=1, Y=1] = \frac{1}{2}, \quad \mathbb{P}[X=1, Y=2] = \frac{1}{10}$$
$$\mathbb{P}[X=2, Y=1] = \frac{1}{10}, \quad \mathbb{P}[X=2, Y=2] = \frac{3}{10}$$

- Determine the marginal pmf of X and Y.
- Show that the conditional pmf of X given $Y = 1$ is

$$\mathbb{P}[X|Y=1] = \begin{cases} \frac{5}{6} & \text{if} \quad X = 1 \\ \frac{1}{6} & \text{if} \quad X = 2. \end{cases}$$

- We can then compute

$$\mathbb{E}[X|Y=1] = \sum_{x \in X} x \mathbb{P}[X=x|Y=1] = \quad .$$

- Similarly, show that the conditional pmf of X given $Y = 2$ is

$$\mathbb{P}[X|Y=2] = \begin{cases} \frac{1}{4} & \text{if} \quad X = 1 \\ \frac{3}{4} & \text{if} \quad X = 2. \end{cases}$$

- Then, $\mathbb{E}[X|Y=2] = \quad$.
- We can view $\mathbb{E}[X|Y]$ as a function of Y as

$$g(Y) = \mathbb{E}[X|Y] = \begin{cases} \mathbb{E}[X|Y=1] & \text{with probability} \quad \mathbb{P}[Y=1] \\ \mathbb{E}[X|Y=2] & \text{with probability} \quad \mathbb{P}[Y=2] \end{cases}$$

- Now, determine $\mathbb{E}[g(Y)]$.
- Determine $\mathbb{E}[X]$. What do you notice?

Conditional Expectation

- If the value of Y is specified, then $\mathbb{E}[X|Y=y]$ is a scalar.
- Otherwise, $\mathbb{E}[X|Y]$ **is a random variable which is a function of** Y.

  if for $\omega_1 \neq \omega_2, Y(\omega_1) = Y(\omega_2) \Rightarrow \mathbb{E}[X|Y = Y(\omega_1)] = \mathbb{E}[X|Y = Y(\omega_2)]$.

- For two continuous random variables X, Y,

$$\mathbb{E}[X|Y=y] = \int_x x f_{X|Y}(x \mid Y=y)dx = \int_x x \frac{f_{X,Y}(x,y)}{f_Y(y)} dx.$$

- Similarly,

$$\mathbb{E}[h(X)|Y=y] = \int_x h(x) f_{X|Y}(x, Y=y)dx$$
$$\mathbb{E}[l(X,Y)|Y=y] = \int_x l(x,y) f_{X|Y}(x, Y=y)dx$$

- If the value of Y is not specified, $\mathbb{E}[l(X,Y)|Y]$ is a random variable.

Example

Let X and Y be two random variables and independent with

$$X = \begin{cases} 1 & \text{with probability} \quad \frac{1}{2}, \\ 0 & \text{with probability} \quad \frac{1}{2}. \end{cases}$$

Let Y have the same distribution as X. Let $Z = X + Y$.

- Determine the pmf of Z.
- Find conditional distribution and expectation of X when $z = 1$ and $z = 2$.
- Find conditional distribution and expectation of z when $X = 1$.

Properties of Conditional Expectation

- Linearity: $\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]$ a.e.
- Monotonicty: $X \leq Y \Rightarrow \mathbb{E}[X|Z] \leq \mathbb{E}[Y|Z]$ a.e.
- Identity: $\mathbb{E}[Y|Y = y] = y$. What is the conditional distribution of Y when its value is specified? Determine $\mathbb{E}[Y|Y]$ and $\mathbb{E}[g(Y)]$.
- Independence: Suppose X and Y are indepdent. Then,

  $$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \int_x x f_{x|Y=y}(x \mid Y = y)dx \\ &= \int_x x \frac{f_{xy}(x,y)}{f_Y(y)}dx = \int_x x f_X(x)dx = \mathbb{E}[X]\end{aligned}$$

  independent of the value of $Y = y$.

  In other words,

  $$\mathbb{E}[X \mid Y] = \int_y \mathbb{E}[X \mid Y = y] f_Y(y)dy = \mathbb{E}[X] \int_y f_Y(y)dy = \mathbb{E}[X].$$

  Similarly, $\mathbb{E}[g(X) \mid Y] = \mathbb{E}[g(X)]$.
- $\mathbb{E}[Xg(Y)|Y] = g(Y)\mathbb{E}[X|Y]$.

Tower Property and Orthogonality

Tower Property:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Proof:

$$\begin{aligned}
\mathbb{E}_Y[\mathbb{E}[X|Y]] &= \int_y \mathbb{E}[X|Y=y] f_Y(y) dy \\
&= \int_y \Big(\int_x x f_{X|Y}(x \mid Y=y) dx \Big) f_Y(y) dy \\
&= \int_y \int_x x \underbrace{f_{X|Y}(x \mid Y=y) f_Y(y)}_{f_{xy}(x,y)} dy dx \\
&= \int_x x \underbrace{\left(\int_y f_{XY}(x,y) dy \right)}_{=: f_X(x)} dx \\
&= \int_x x f_X(x) dx = \mathbb{E}[X]
\end{aligned}$$

Orthogonality: for any measurable function g,

$$\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = 0.$$

That is, $(X - \mathbb{E}[X|Y])$ is orthogonal to any function $g(Y)$ of Y.
Proof:

Minimum Mean Square Estimator (MMSE)

Proposition: Let $g(Y)$ be an estimator of X, and the mean square estimation error be defined as $\mathbb{E}[(X - g(Y))^2]$. Then,

$$\mathbb{E}[(X - \mathbb{E}[X|Y])^2] \leq \mathbb{E}[(X - g(Y))^2], \quad \text{for all measurable} \quad g.$$

Proof:

$$\mathbb{E}\left[(X - g(Y))^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y] - g(Y))^2\right]$$

$$=$$

$L_2(\Omega, \mathcal{F}, \mathbb{P})$ **Space of Random Variables**

- We define $L_2(\Omega, \mathcal{F}, \mathbb{P})$ (or simply L_2) to be the set of random variables with finite second moment, i.e., $L_2 = \{X \mid \mathbb{E}[X^2] < \infty\}$.
- Properties of L_2:
  - L_2 is a linear subspace of random variables:
    (i) $aX \in L_2$ for all $X \in L_2$ and $a \in \mathbb{R}$ as $\mathbb{E}[(aX)^2] = a^2\mathbb{E}[X^2] < \infty$, and
    (ii) $X + Y \in L_2$ for all $X, Y \in L_2$
  - **The most important property**: L_2 is an inner-product space. For any two random variables $X, Y \in L_2$, let us define their inner product

$$X \cdot Y := \mathbb{E}[XY].$$

  - Then this operation satisfies the axioms of an inner product:
    (i) $X \cdot X = \mathbb{E}[X^2] \geq 0$.
    (ii) $X \cdot X = 0$ iff $X = 0$ almost surely.
    (iii) *linearity*: $(\alpha X + Y) \cdot Z = X \cdot Z + \alpha Y \cdot Z$.
- Therefore, L_2 is a normed vector space, with the norm $\|\cdot\|$ defined by

$$\|X\| := \sqrt{X \cdot X} = \sqrt{\mathbb{E}[X^2]}.$$

- Similarly, we have $\|X - Y\|^2 := (X - Y) \cdot (X - Y) = E[(X - Y)^2]$.

L_2-norm and L_2 convergence

- Since L_2 is a normed space, we can define a new limit of random variables:

  Definition 1. *We say that a sequence $\{X_k\}$ converges in L_2 (or in MSE sense) to X if* $\lim_{k\to\infty} \|X - X_k\| = 0$.

- Note that $\lim_{k\to\infty} \|X - X_k\| = 0$ iff $\lim_{k\to\infty} \mathbb{E}[|X - X_k|^2] = 0$.

- **Definition:** We say that $H \subseteq L_2$ is a linear subspace if

  (i) for any $X, Y \in H$, we have $X + Y \in H$, and

  (ii) for any $X \in H$ and $a \in \mathbb{R}$, $aX \in H$.

- **Definition:** We say that $H \subseteq L_2$ is closed if for any sequence $\{X_k\}$ with

$$\lim_{m,n\to\infty} \|X_m - X_n\|^2 = \lim_{m,n\to\infty} \mathbb{E}[|X_m - X_n|^2] = 0,$$

  we have $\lim_{k\to\infty} X_k \overset{L_2}{\to} X$ for some random variable $X \in L_2$.

- Showing linear subspace is easy, but closedness might be hard.

- Important Cases:

  1. For random variables $X_1, \ldots, X_k \in L_2$, the set $H = \{\alpha_1 X_1 + \ldots + \alpha_k X_k \mid \alpha_i \in \mathbb{R}\}$ is a closed linear subspace.
  2. For any random variables $X_1, \ldots, X_k \in L_2$, the set $H = \{\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \mid \alpha_i \in \mathbb{R}\}$ is a closed linear subspace.

Orthogonality Principle

Theorem 1. *Let H be a closed linear subspace of L_2 and let $X \in L_2$. Then,*

a. *There exists a unique (up to almost sure equivalence) random variable $Y^\star \in H$ such that*

$$\|Y^\star - X\|^2 \leq \|Z - X\|^2, \quad \text{for all } Z \in H.$$

b. *Let W be a random variable. $W = Y^\star$ a.e. if and only if $W \in H$ and*

$$\mathbb{E}[(X - W)Z] = 0, \quad \text{for all } Z \in H.$$

Note:

- $Y^\star$ is called the projection of X on the subspace H and is denoted by $\Pi_H(X)$.
- Two random variables X, Y are orthogonal, $X \perp Y$, if $\mathbb{E}[XY] = 0$.
- Relate MSE estimator with the above theorem.

Linear Minimum Mean Square Error (LMMSE) Estimation

- Let Y be a measurement of X and we want to find an estimate of X which is a linear function of Y minimizing the mean square error. The estimator is of the form: $\widehat{X}_{\texttt{LMSE}}(Y) = aY + b$. The goal is to find coefficients $a^*, b^* \in \mathbb{R}$ such that

$$\|X - (a^*Y + b^*)\| \leq \|X - (aY + b)\|, \quad \text{for any } a, b \in \mathbb{R}.$$

- Let $\mathcal{L}(Y) := \{Z \mid Z = aY + b, \quad a, b \in \mathbb{R}\}$ be the set of random variables that are linear functions of Y. One can show that $\mathcal{L}(Y)$ is a closed linear subspace.
- Then, $\widehat{X}_{\texttt{LMSE}}(Y) = \Pi_{\mathcal{L}(Y)}(Y)$.
- From orthogonality property, we know that $\mathbb{E}[(X - \widehat{X}_{\texttt{LMSE}}(Y))Z] = 0$ for all $Z \in \mathcal{L}(Y)$.
- Show that the coefficients a^*, b^* satisfy

$$a^* = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(Y)}, \quad b^* = \mathbb{E}[X] - a^*\mathbb{E}[Y].$$

- Thus, the LMMSE estimate

$$\hat{X}(Y) := a^*Y + b^* = a^*(Y - \mathbb{E}[Y]) + \mathbb{E}[X] = \mathbb{E}[X] + \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(Y)}(Y - \mathbb{E}[Y]).$$

- We can verify that $(X - \hat{X}) \perp (\alpha Y + \beta)$ for all $\alpha, \beta \in \mathbb{R}$.
- What is the mean square estimation error?

Derivation of LMMSE Coefficients

LMMSE Coefficients for Multiple Observations

- Let $Y = [Y_1, \quad \ldots, \quad Y_k]^\top$ be measurements available to us.
- We wish to determine $\widehat{X}_{\texttt{LMSE}}(Y) = a_0 + \sum_{i=1}^k a_i Y_i = \Pi_{\mathcal{L}(Y)}$.
- The goal is to find coefficients that minimize the mean square error

$$\min_{a_0, a_1, \ldots, a_k} \mathbb{E}[(X - (a_0 + \sum_{i=1}^k a_i Y_i))^2].$$

- Due to the orthogonality property, the LMMSE estimator satisfies

$$\mathbb{E}[(X - (a_0^* + \sum_{i=1}^k a_i^* Y_i))Z] = 0 \qquad \forall Z \in \mathcal{L}(Y).$$

- We need to cleverly choose $k+1$ elements from $\mathcal{L}(Y)$ to set up a system of $k+1$ linear equations and solve for the coefficients.

Derivation of LMMSE Coefficients

- Hint: Choose 1 and $Y_i - \mathbb{E}[Y_i]$ for all $i \in \{1, 2, \ldots, k\}$.
- If $Z = 1$, then orthogonality yields

$$\mathbb{E}[(X - (a_0^* + \sum_{i=1}^{k} a_i^* Y_i))] = 0.$$

- If $Z = Y_j - \mathbb{E}[Y_j]$, then orthogonality yields

$$\mathbb{E}[(X - (a_0^* + \sum_{i=1}^{k} a_i^* Y_i))(Y_j - \mathbb{E}[Y_j])] = 0.$$

Derivation of LMMSE Coefficients

- Finally, from the above analysis, we obtain

$$\begin{bmatrix} a_1^* \\ a_2^* \\ \vdots \\ a_k^* \end{bmatrix} = [\mathtt{Cov}(Y)]^{-1}\mathtt{Cov}(X,Y).$$

- The LMMSE is given by

$$\begin{aligned}
\widehat{X}_{\mathtt{LMSE}}(Y) &= a_0^* + \sum_{i=1}^{k} a_i^* Y_i \\
&= \mathbb{E}[X] + \sum_{i=1}^{k} a_i^*(Y_i - \mathbb{E}[Y_i]) \\
&= \mathbb{E}[X] + (a^*)^\top [Y - \mathbb{E}[Y]] \\
&= \mathbb{E}[X] + \mathtt{Cov}(X,Y)^\top [\mathtt{Cov}(Y)]^{-1}[Y - \mathbb{E}[Y]].
\end{aligned}$$

- When X is also a random vector $\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$, the LMMSE is given by

$$\widehat{X}_{\mathtt{LMSE}}(Y) = \begin{bmatrix} \widehat{X}_{1,\mathtt{LMSE}}(Y) \\ \widehat{X}_{2,\mathtt{LMSE}}(Y) \\ \vdots \\ \widehat{X}_{n,\mathtt{LMSE}}(Y) \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1] + \mathtt{Cov}(X_1,Y)^\top [\mathtt{Cov}(Y)]^{-1}[Y - \mathbb{E}[Y]] \\ \mathbb{E}[X_2] + \mathtt{Cov}(X_2,Y)^\top [\mathtt{Cov}(Y)]^{-1}[Y - \mathbb{E}[Y]] \\ \vdots \\ \mathbb{E}[X_n] + \mathtt{Cov}(X_n,Y)^\top [\mathtt{Cov}(Y)]^{-1}[Y - \mathbb{E}[Y]] \end{bmatrix}.$$

Example (Previous year End-Sem Question)

X is a three-dimensional random vector with $E[X] = 0$ and autocorrelation matrix R_X with elements $r_{ij} = (-0.80)^{|i-j|}$. Use X_1 and X_2 to form a linear estimate of X_3 : $\hat{X}_3 = a_1 X_1 + a_2 X_2$, i.e., determine a_1 and a_2 that minimizes mean-square error.

MMSE and LMMSE Estimator Comparison

- An estimator $\widehat{X}(Y)$ is **unbiased** if $\mathbb{E}[\widehat{X}(Y)] = \mathbb{E}[X]$.
  - Is MMSE estimator unbiased?
  - Is LMMSE estimator unbiased?

- Among MMSE and LMMSE estimators, which one has smaller estimation error?

- If X and Y are uncorrelated, what does the LMMSE estimator give us? What about MMSE estimator?

- What do you need to know to determine MMSE and LMMSE estimators?

- What if $\texttt{Cov}(Y)$ is not invertible?

- When X and Y are jointly Gaussian,

$$\widehat{X}_{\texttt{LMMSE}}(Y) = \widehat{X}_{\texttt{MMSE}}(Y)$$
$$\iff \mathbb{E}[X|Y] = \mathbb{E}[X] + \texttt{Cov}(X,Y)^\top [\texttt{Cov}(Y)]^{-1}[Y - \mathbb{E}[Y]].$$

  Conditional expectation of X given Y is a linear function of Y.