ELE8088: Control & Estimation Theory **QUB, 2021**

Handout 12: Introduction to Estimation Theory

Lecturer: Pantelis Sopasakis *Date:* ______

Topics: Bayes Theorem ∘ Minimum Variance Estimation.

12.1 Memory refresher

Let X be a real-valued random variable. The *cumulative distribution function* (cdf) of X is the function

$$F_X(x) = \underline{\hspace{5cm}}.$$

A function $p_X : \mathbb{R} \to \mathbb{R}$ is called the *probability density function* (pdf) of X if

$$\mathsf{P}[a \le X \le b] = \underline{\hspace{6cm}}.$$

Suppose X has a pdf p_X; then, by taking $a = b$ we conclude that

$$\mathsf{P}[X = a] = \underline{\hspace{2.5cm}}. \tag{12.1}$$

If X has a pdf, it is called a *continuous* random variable. Given the pdf of a continuous random variable, its cdf is

$$F_X(x) = \underline{\hspace{6cm}}.$$

We say that a real-value random variable follows the normal distribution, $\mathcal{N}(\mu, \sigma^2)$, if its pdf is

$$p_X(x) = \underline{\hspace{6cm}}. \tag{12.2}$$

The expectation of a real-valued random variable X with pdf p_X is

$$\mathbb{E}[X] = \underline{\hspace{5cm}},$$

provided that $\underline{\hspace{16cm}}$.

The *variance* of a real-valued (discrete or continuous) random variable is defined as

$$\mathrm{Var}[X] = \underline{\hspace{7cm}}. \tag{12.3}$$

If X is a continuous random variable with pdf p_X, the variance can be determined by

$$\mathrm{Var}[X] = \underline{\hspace{5cm}}. \tag{12.4}$$

If X is a discrete random variable on a space $\Omega = \{1, \ldots, n\}$, with expectation $\mu = \mathbb{E}[X]$, then

$$\mathrm{Var}[X] = \underline{\hspace{4cm}}.$$

Let $X = (X_1, \ldots, X_n)$ be a multivariate random variable. The *expectation* of X is

$$\mathbb{E}[X] = \underline{\hspace{5cm}}. \tag{12.5}$$

The *variance-covariance* matrix of X is defined as

$$\mathrm{Var}[X] = \underline{\hspace{5cm}}.$$

Given two n-dimensional random variables X, Y, their *cross-covariance matrix* is

$$\mathrm{Cov}[X, Y] = \underline{\hspace{5cm}}.$$

Let X, Y be two real-valued continuous random variables and $Z = (X, Y)$. Let $A = \left\{\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathbb{R}^2 : a \le x \le b, c \le y \le d\right\}$. Then

$$\mathsf{P}[Z \in A] = \underline{\hspace{6.5cm}}.$$

If $Z = (X, Y)$ is a continuous multivariate random variable with pdf p_Z, the *marginal pdf* of X is given by

$$p_X(x) = \underline{\hspace{5cm}}, \tag{12.6}$$

and the marginal pdf of Y is

$$p_Y(y) = \underline{\hspace{5cm}}. \tag{12.7}$$

12.2 Bayes Theorem

We have defined the conditional probability as

$$\mathsf{P}[A \mid B] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]}, \tag{12.8}$$

provided $\mathsf{P}[B] > 0$. Similarly,

$$\mathsf{P}[B \mid A] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[A]}, \tag{12.9}$$

provided $\mathsf{P}[A] > 0$. Combining the two equations

$$\boxed{\mathsf{P}[A \mid B] = \frac{\mathsf{P}[B \mid A]\mathsf{P}[A]}{\mathsf{P}[B]}} \tag{Bayes}$$

with $\mathsf{P}[B] > 0$ (it holds for $\mathsf{P}[A] = 0$). This is known as *Bayes's formula.*

We have shown (Exercise in previous section) that for $A \in \mathcal{F}$ and a sequence of events $(B_n)_n$ that *partitions* Ω, it is

$$\mathsf{P}[A] = \sum_n \mathsf{P}[A \cap B_n].$$

Wlog we assume $\mathsf{P}[B_n] > 0$ for all n. Then $\mathsf{P}[A \cap B_n] = \mathsf{P}[A \mid B_n]\mathsf{P}[B_n]$, so

$$\boxed{\mathsf{P}[A] = \sum_n \mathsf{P}[A \mid B_n]\mathsf{P}[B_n].}$$

This is known as the law of total probability (LTP).

By LTP, for $A, B \in \mathcal{F}$,

$$\mathsf{P}[A] = \mathsf{P}[A \mid B] \cdot \mathsf{P}[B] + \mathsf{P}[A \mid B^c] \cdot \mathsf{P}[B^c].$$

12.2.1 Example: Do I have COVID?

Note: the values given below are arbitrary. A certain test for COVID-19 has *sensitivity* 99%, that is

$$\mathsf{P}[\text{Positive} \mid \text{COVID}] = 0.99. \tag{12.10}$$

The test has *specificity* 98%, that is

$$\mathsf{P}[\text{Negative} \mid \text{Healthy}] = 0.98. \tag{12.11}$$

The *prevalence* of COVID is 10%, i.e.,

$$\mathsf{P}[\text{COVID}] = 0.1. \tag{12.12}$$

What is the probability that a random person tests positive? By the law of total probability:

$$\begin{aligned}\mathsf{P}[\text{Positive}] &= \mathsf{P}[\text{Pos} \mid \text{COVID}] \cdot \mathsf{P}[\text{COVID}] + \mathsf{P}[\text{Pos} \mid \text{Healthy}] \cdot \mathsf{P}[\text{Healthy}] \\ &= \mathsf{P}[\text{Pos} \mid \text{COVID}] \cdot \mathsf{P}[\text{COVID}] \\ &\qquad + (1 - \mathsf{P}[\text{Neg} \mid \text{Healthy}]) \cdot (1 - \mathsf{P}[\text{COVID}]) \\ &= 0.99 \cdot 0.1 + (1 - 0.98) \cdot (1 - 0.1) = 0.117. \end{aligned} \tag{12.13}$$

Suppose you test positive. What is the probability you have COVID-19?

$$\mathsf{P}[\text{COVID} \mid \text{Positive}] = \frac{\mathsf{P}[\text{Positive} \mid \text{COVID}] \cdot \mathsf{P}[\text{COVID}]}{\mathsf{P}[\text{Positive}]} = \frac{0.99 \cdot 0.1}{0.117} = 84.6\%$$

In Bayesian statistics, probability expresses a degree of *belief*. How much should you believe that you have COVID-19? Answer:

$$\mathsf{P}[\text{COVID}] = 10\%.$$

Suppose you tested positive. How much should you believe that you have COVID-19? Answer:

$$\mathsf{P}[\text{COVID} \mid \text{Positive}] = 84.6\%.$$

Note that our belief changes as data become available.

$$\mathsf{P}[\text{COVID} \mid \text{Positive}] = \frac{\mathsf{P}[\text{Positive} \mid \text{COVID}] \cdot \mathsf{P}[\text{COVID}]}{\mathsf{P}[\text{Positive}]}$$

Some terminology:

- $\mathsf{P}[\text{COVID}]$ is called the *prior*
- $\mathsf{P}[\text{COVID} \mid \text{Positive}]$ is the *posterior*
- $\mathsf{P}[\text{Positive} \mid \text{COVID}]$ is the *likelihood*

12.2.2 Bayes theorem for continuous random variables

Let X and Y be two real-valued continuous random variables. Then,

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \tag{12.14}$$

for all y such that $p_Y(y) > 0$. Similarly,

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, \tag{12.15}$$

for all x such that $p_X(x) > 0$. Combining the two equations

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)p_X(x)}{p_Y(y)}, \tag{12.16}$$

provided $p_Y(y) > 0$. Recall that the marginal pdf p_Y is given by

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)\mathrm{d}x = \int_{-\infty}^{\infty} p_{Y|X}(y \mid x)p_X(x)\mathrm{d}x \tag{12.17}$$

Example:[1] Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N \sim \mathcal{N}(0, \sigma_N^2)$ be independent and

$$Y = X + N. \tag{12.18}$$

We measure $Y = y$. Find $p_{X|Y}(x \mid y)$. Therefore,

$$p_{X|Y}(x \mid y) = \frac{p_N(y-x)p_X(x)}{\int_{-\infty}^{\infty} p_N(y-x)p_X(x)\mathrm{d}x}$$

Reminder: $N \sim \mathcal{N}(0, \sigma_N^2)$ means that $p_N(n) = \frac{1}{\sigma_N\sqrt{2\pi}} e^{-\frac{n^2}{2\sigma_N^2}}$.

$$= \frac{\frac{1}{\sigma_N\sqrt{2\pi}} \exp\left[-\frac{(y-x)^2}{2\sigma_N^2}\right] \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma_X^2}\right]}{\int_{-\infty}^{\infty} \frac{1}{\sigma_N\sqrt{2\pi}} \exp\left[-\frac{(y-x)^2}{2\sigma_N^2}\right] \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma_X^2}\right] \mathrm{d}x}$$

a few cups of coffee later...

$$p_{X|Y}(x \mid y) = \frac{1}{\sqrt{2\pi \frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}}} \exp\left[-\frac{\left(x - \frac{y\sigma_X^2}{\sigma_X^2+\sigma_N^2}\right)^2}{2\frac{\sigma_X^2\sigma_N^2}{\sigma_X^2+\sigma_N^2}}\right]. \tag{12.19}$$

[1]This is Exercise 3.1 in Anderson and Moore.

The conditional random variable $X \mid Y = y$ is normally distributed with mean

$$\mathbb{E}[X \mid Y = y] = \frac{y\sigma_X^2}{\sigma_X^2 + \sigma_N^2}, \tag{12.20}$$

and variance

$$\mathrm{Var}[X \mid Y = y] = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2} \tag{12.21}$$

Sketch it......

Exercise 1 (☕☕): Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $N \sim \mathcal{N}(\mu_N, \sigma_N^2)$ be independent and

$$Y = X - 2N.$$

We measure $Y = y$. Determine $p_{X|Y}(x \mid y)$.

Exercise 2: (☕☕) (i) Let X be a real-valued continuous random variable with pdf p_X and $Y = aX$, for $a \in \mathbb{R}$. Show that

$$p_{X,Y}(x,y) = \begin{cases} p_X(x), & \text{if } y = ax \\ 0, & \text{otherwise} \end{cases} \tag{12.22}$$

(ii) Let $Z = f(X)$. Determine $p_{X,Z}$, (iii) Let $Z_0, Z_1, \ldots, Z_N$ be a sequence of mutually independent (vector-valued) random variables with pdfs $p_{Z_0}, \ldots, p_{Z_N}$ and let $Y_0 = Z_0$ and $Y_i = Y_{i-1} + Z_i$ for $i \in \mathbb{N}_{[1,N]}$. Determine the joint distribution

$$p_{Z_0,\ldots,Z_N,Y_0,\ldots,Y_N}(z_0, \ldots, z_N, y_0, \ldots, y_N).$$

Exercise 3 (☕)[2]**:** Alice has two coins in her pocket, a fair coin (head on one side and tail on the other side) and a two-headed coin. She picks one at random from her pocket, tosses it and obtains head. What is the probability that she flipped the fair coin?

Exercise 4: Taxi-cab problem (☕)[3]**:** Suppose you were told that a taxi-cab was involved in a hit-and-run accident one night. Of the taxi-cabs in the city, 85% belonged to the Green company and 15% to the Blue company. You are then told that an eyewitness had identified the cab as a blue cab. But when her ability to identify cabs under appropriate visibility conditions was tested, she was wrong 20% of the time. (i) What is the probability that the cab is blue? (ii) How would you solve this problem if you didn't know the proportion of blue and green taxis in the city?

Exercise 5 (☕)[4]**:** In a casino in Blackpool there are two slot machines: one that pays out 10% of the time, and one that pays out 20% of the time. Obviously, you would like to play on the machine that pays out 20% of the time but you do not know which of the two machines is the more generous. You thus adopt the following strategy: you assume initially that the two machines are equally likely to be the generous machine. You then select one of the two machines at random and put a coin into it. Given that you loose that first bet estimate the probability that the machine you selected is the more generous of the two machines.

Exercise 6 (☕☕☕) We say that a random variable X follows the Gamma (Γ) distribution, $\Gamma(\alpha, \beta)$, with parameters $\alpha, \beta > 0$ if $p_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, defined for $x > 0$. Suppose that $Y \sim \Gamma(\alpha_Y, \beta_Y)$, where α_Y is known, but β_Y is not. Suppose that $\beta_Y \sim \Gamma(\alpha, \beta)$. We measure $Y = y$. What is the posterior distribution, $p_{\beta_Y|Y}(\beta_Y \mid y)$?[5]

---

[2]Credit: https://www.statlect.com/fundamentals-of-probability/Bayes-rule

[3]Credit: https://jrnold.github.io/bayesian_notes/bayes-theorem.html

[4]Credit: http://gtribello.github.io/mathNET/bayes-theorem-problems.html

[5]It is $\beta_Y \mid Y = y \sim \Gamma\left(\alpha_Y + \alpha, \beta + y\right)$.

12.3 Estimates, Estimators and Minimum Variance Estimation of Jointly Distributed RVs

Problem statement. Suppose X and Y are jointly distributed random variables. We measure Y. What is a good estimate of X given Y? We are looking for a function $\widehat{X}$ that takes an *observation* (a sample of Y) and returns an estimate of X

$$y \longrightarrow \boxed{\widehat{X}(\,\cdot\,)} \longrightarrow \hat{x} = \widehat{X}(y)$$

The *value* $\hat{x}(y)$ is an **estimate** of X given $Y = y$. The *random variable* $X(Y)$ is an **estimator**. The *estimator error* is the random variable

$$e = X - \widehat{X}(Y). \tag{12.23}$$

The expectation of the error is the *estimator bias*, bias $= \mathbb{E}[e] = \mathbb{E}[X - \widehat{X}(Y)]$. The variance-covariance matrix of the error is called the *variance-covariance matrix of the estimator*,

$$\Sigma := \mathbb{E}[ee^\intercal]. \tag{12.24}$$

Theorem 12.1 (Minimum variance estimation) *Suppose X and Y are jointly distributed continuous real-valued random variables and we measure $Y = y$. Let $\hat{x} = \hat{x}(y)$ be a minimiser of the problem*

$$\operatorname*{minimise}_{z} \mathbb{E}[\|X - z\|^2 \mid Y = y],$$

that is, $\hat{x}$ is a minimum variance estimate of X given $Y = y$. Then, $\hat{x}$ is unique and

$$\hat{x} = \mathbb{E}[X \mid Y = y].$$

Proof:

$\blacksquare$

The estimator

$$\widehat{X} = \mathbb{E}[X \mid Y], \tag{12.25}$$

is an unbiased minimum variance estimator, that is[6],

$$\text{bias} = \mathbb{E}\left[X - \widehat{X}\right] = 0, \tag{12.26}$$

and for any estimator Z,

$$\mathbb{E}\left[\|X - \widehat{X}(Y)\|^2\right] \leq \mathbb{E}\left[\|X - Z(Y)\|^2\right]. \tag{12.27}$$

The *risk* of $\widehat{X}$ is

$$\mathbb{E}\left[\|X - \hat{x}\|^2 \mid Y\right] = \mathbb{E}\left[\|X\|^2 \mid Y\right] - \|\hat{x}\|^2. \tag{12.28}$$

Remark. Here X and Y are treated as jointly distributed random variables and the objective is to determine an estimate of X, which is a deterministic value, given a measurement $Y = y$.

[6] Here we use the definition of bias given in the book of Anderson and Moore, *Optimal Filtering*, Prentice Hall, 1979.

As discussed above, the conditional expectation of X given $Y = y$ has some favourable properties (unbiased, minimum variance; see Theorem 12.1). Note that this conditional expectation can be computed as follows:

$$\mathbb{E}[X \mid Y = y] = \int p_{X|Y}(x \mid y)\mathrm{d}x. \tag{12.29}$$

In the following example we will see that the conditional pdf $p_{X|Y}(x \mid y)$ is more informative compared to the above point-estimate. If we have $p_{X|Y}(x \mid y)$, we can determine the conditional expectation (which comes with the nice properties stated in Theorem 12.1), we can determine the corresponding variance, or we may decide to use the mode or median of this pdf as an estimator.

12.3.1 Example

Consider a pair $(X, Y) \sim \mathcal{N}\left(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 0.6\\0.6 & 0.7\end{smallmatrix}\right]\right)$. We define the estimator $\widehat{X}(Y) = \mathbb{E}[X \mid Y]$, which is a minimum variance estimator.

Exercise 7: Determine the estimator $\widehat{X}(Y)$ and its variance.

The estimation of X given $Y = y$ is illustrated in Figure 12.1. Note that the conditional variance of $\widehat{X}(Y)$ given $Y = y$ is independent of y — in other words, no particular measurement is more informative than any other.

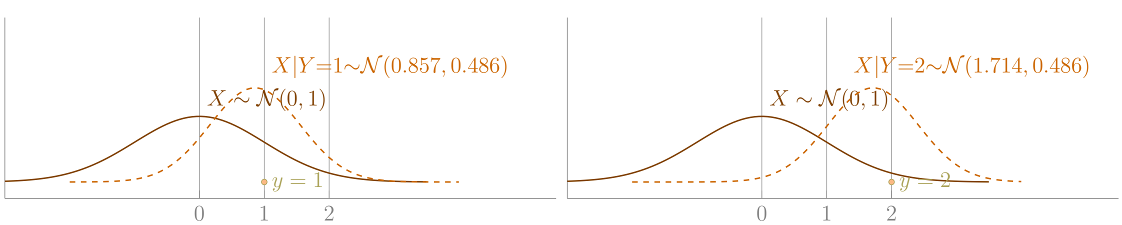

Figure 12.1: Estimation of X given Y by means of the estimator $\widehat{X}(Y) = \mathbb{E}[X \mid Y]$, which is a minimum variance estimator. (Left) Estimation of X given $Y = 1$ and (Right) given $Y = 2$

Figure 12.2 offers another interpretation of the conditional expectation: it is a *slice* of the joint probability density function of (X, Y).

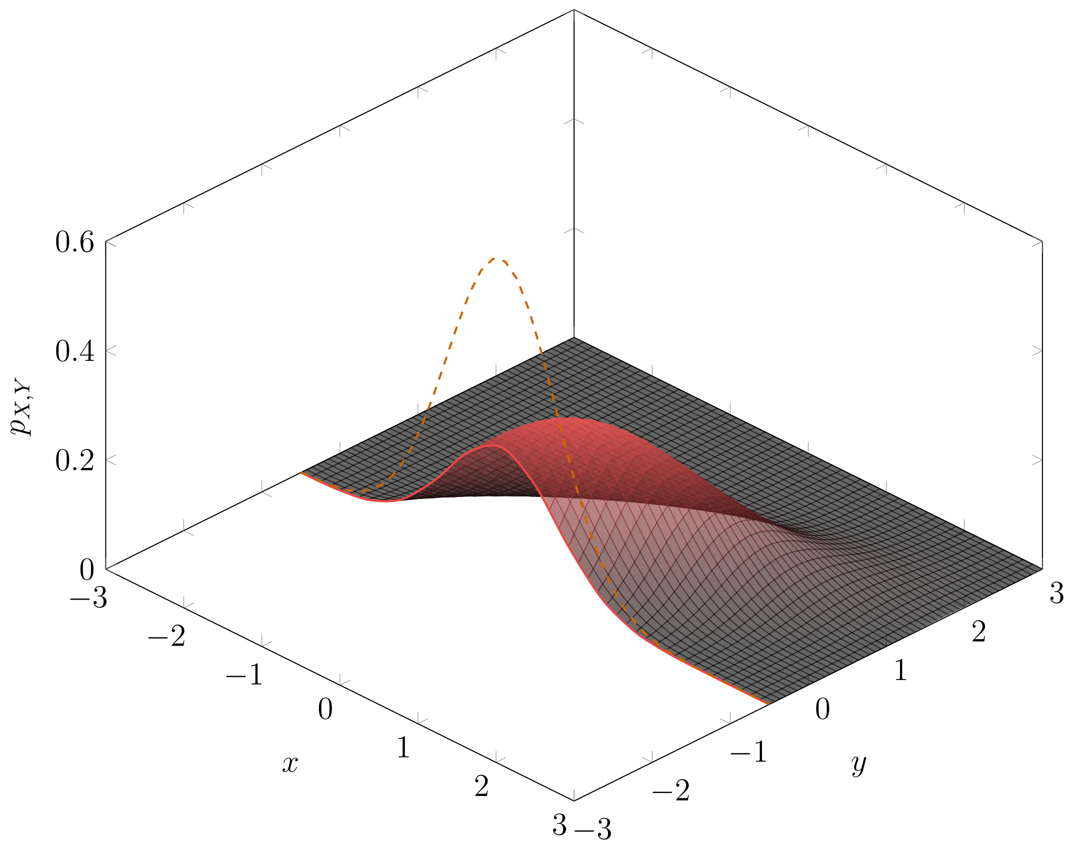

Figure 12.2: The estimator of X given Y can be interpreted as the "slicing" of the joint probability density function of (X, Y) along the line $Y = y$.