ECE 302: Probability and Applications[1]

Week 10 Topics

- Vector Random Variables
  - pmf, pdf, independence, Markov
- Jointly Gaussian Vector Random Variables
- ML, MAP, and Minimum MSE Estimation

1 Vector Random Variables

Vector random variables arise naturally in random experiments where several quantities are of interest. For example, we may be interested in the age, weight, and height of an individual drawn from some population. In another example, the temperatures at a given time and location for a week forms an 7-dimensional random vector.

The definition for a pair of random variables is readily generalized: An n-dimensional **vector random variable $\mathbf{X}$** is a function that assigns a vector of n real numbers to each outcome ζ in S, the sample space of the random experiment. The probability that $\mathbf{X}$ is in some event R^n is then found by finding the equivalent event A in S that maps outcomes to B, as shown in Figure 1. The same development that was used for pairs of random variables can then be used to define for $\mathbf{X}$: the joint cumulative distribution functions and, where applicable, joint probability mass function and/or joint probability density function.

Vector notation greatly simplifies our expressions. We denote vector random variables using uppercase boldface notation. By convention $\mathbf{X}$ is a column vector (n rows by 1 column), so the vector random variable with components $X_1, X_2, \ldots, X_n$ corresponds to

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$$

where "T" denotes the transpose of a matrix or vector. We will sometimes write $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ to save space and omit the transpose unless dealing with matrices where it makes a difference. Possible values of the vector random variable are denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where x_i corresponds to the value of X_i.

Example 1: Packet Arrivals

Suppose that packets arrive at each of 3 input ports according to a Bernoulli random variable with probability 1/2. A packet arrival is equally likely to be destined to one of the three output ports. Let $X = (X_1, X_2, X_3)$ be the number of packets arriving at the output ports.

Example 2: Joint Poisson Counts

A random experiment consists of finding the number of defects in a semiconductor chip and identifying their locations. The outcome of this experiment consists of the vector $\zeta = (n, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, where the first component specifies the total number of defects and the remaining components specify the coordinates of their location. Suppose that the chip consists of M regions. Let $N_1(\zeta), N_2(\zeta), \ldots, N_M(\zeta)$ be the number of defects in each of these regions, that is, $N_k(\zeta)$ is the number of $\mathbf{y}$'s that fall in region k. The vector $\mathbf{N}(\zeta) = (N_1, N_2, \ldots, N_M)$ is then a vector random variable.

Example 3: Samples of an Audio Signal

Let the outcome ζ of a random experiment be an audio signal $X(t)$. Let the random variable $X_k = X(kT)$ be the sample of the signal taken at time kT. An MP3 codec processes the audio in blocks of n samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. $\mathbf{X}$ is a vector random variable.

1.1 Joint Distribution Functions

The **joint cumulative distribution function** of $X_1, X_2, \ldots, X_n$ is defined as the probability of an n-dimensional semi-infinite rectangle associated with the point $(x_1, \ldots, x_n)$:

$$F_{\mathrm{X}}(\mathbf{x}) \triangleq F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P\left[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\right] \tag{1}$$

The joint cdf is defined for discrete, continuous, and random variables of mixed type. The probability of product-form (rectangular-shape) events can be expressed in terms of the joint cdf.

Every subset of variables of a random is itself a random vector of lower dimension, and so the joint cdf generates a family of **marginal cdf's** for subcollections of the random variables $X_1, \ldots, X_n$. These marginal cdf's are obtained by setting the appropriate entries to $+\infty$ in the joint cdf in Eq. (1). For example:

Joint cdf for $X_1, \ldots, X_{n-1}$ is given by $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_{n-1}, \infty)$

Joint cdf for X_1 and X_2 is given by $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \infty, \ldots, \infty)$.

Example 4

A radio transmitter sends a signal to a receiver using three paths. Let X_1, X_2, and X_3 be the signals that arrive at the receiver along each path. Find $P[\max(X_1, X_2, X_3) \leq 5]$.

The maximum of three numbers is less than 5 if and only if each of the three numbers is less than 5; therefore

$$\begin{aligned} P[A] &= P[\{X_1 \leq 5\} \cap \{X_2 \leq 5\} \cap \{X_3 \leq 5\}] \\ &= F_{X_1,X_2,X_3}(5,5,5). \end{aligned}$$

The **joint probability mass function** of n discrete random variables is defined by

$$p_X(\mathbf{x}) \triangleq p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]$$

The probability of any n-dimensional event A is found by summing the pmf over the points in the event

$$P[\mathbf{X} \text{ in } A] = \sum \underset{\mathbf{x} \text{ in } A}{\cdots} \sum p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

The joint pmf generates a family of **marginal pmf's** that specifies the joint probabilities for subcollections of the n random variables. For example, the one-dimensional pmf of X_j is found by adding the joint pmf over all variables other than x_j:

$$p_{X_j}(x_j) = P[X_j = x_j] = \sum_{x_1} \cdots \sum_{x_{j-1} x_{j+1}} \cdots \sum_{x_n} p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

The two-dimensional joint pmf of any pair X_j and X_k is found by adding the joint pmf over all $n-2$ other variables, and so on. Thus, the marginal pmf for $X_1, \ldots, X_{n-1}$ is given by

$$p_{X_1,\ldots,X_{n-1}}(x_1, x_2, \ldots, x_{n-1}) = \sum_{x_n} p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n)$$

A family of **conditional pmf's** is obtained from the joint pmf by conditioning on different subcollections of the random variables. For example, if $p_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1}) > 0$:

$$p_{X_n}(x_n \mid x_1, \ldots, x_{n-1}) = \frac{p_{X_1}, \ldots, X_n(x_1, \ldots, x_n)}{p_{X_1}, \ldots, X_{n-1}(x_1, \ldots, x_{n-1})} \tag{2}$$

Repeated applications of Eq. (2) yield the following very useful expression:

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \\ p_{X_n}(x_n \mid x_1,\ldots,x_{n-1})\, p_{X_{n-1}}(x_{n-1} \mid x_1,\ldots,x_{n-2}) \ldots p_{X_2}(x_2 \mid x_1)\, p_{X_1}(x_1) \quad (3)$$

Example 5: Arrivals at a Packet Switch

Find the joint pmf of $\mathbf{X} = (X_1, X_2, X_3)$ in Example 1. Find $\mathrm{P}[X_1 > X_3]$.

Let N be the total number of packets arriving in the three input ports. Each input port has an arrival with probability $p = 1/2$, so N is binomial with pmf:

$$p_N(n) = \binom{3}{n}\frac{1}{2^3} \text{ for } 0 \le n \le 3$$

Given $N = n$, the number of packets arriving for each output port has a multinomial distribution:

$$p_{X_1,X_2,X_3}(i,j,k \mid i{+}j{+}k = n) = \begin{cases} \frac{n!}{i!j!k!}\frac{1}{3^n} & \text{for } i+j+k=n, i \ge 0, j \ge 0, k \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

The joint pmf of $\mathbf{X}$ is then:

$$p_{\mathbf{x}}(i,j,k) = p_{\mathbf{x}}(i,j,k \mid n)\binom{3}{n}\frac{1}{2^3} \text{ for } i \ge 0, j \ge 0, k \ge 0, i+j+k = n \le 3$$

The explicit values of the joint pmf are:

$$p_{\mathbf{X}}(0,0,0) = \frac{0!}{0!0!0!}\frac{1}{3^0}\binom{3}{0}\frac{1}{2^3} = \frac{1}{8}$$

$$p_{\mathbf{X}}(1,0,0) = p_{\mathbf{X}}(0,1,0) = p_{\mathbf{X}}(0,0,1) = \frac{1!}{0!0!1!}\frac{1}{3^1}\binom{3}{1}\frac{1}{2^3} = \frac{3}{24}$$

$$p_{\mathbf{X}}(1,1,0) = p_{\mathbf{X}}(1,0,1) = p_{\mathbf{X}}(0,1,1) = \frac{2!}{0!1!1!}\frac{1}{3^2}\binom{3}{2}\frac{1}{2^3} = \frac{6}{72}$$

$$p_{\mathbf{X}}(2,0,0) = p_{\mathbf{X}}(0,2,0) = p_{\mathbf{X}}(0,0,2) = 3/72$$

$$p_{\mathbf{X}}(1,1,1) = 6/216$$

$$p_{\mathbf{X}}(0,1,2) = p_{\mathbf{X}}(0,2,1) = p_{\mathbf{X}}(1,0,2) = p_{\mathbf{X}}(1,2,0) = p_{\mathbf{X}}(2,0,1) = p_{\mathbf{x}}(2,1,0) = 3/216$$

$$p_{\mathbf{X}}(3,0,0) = p_{\mathbf{X}}(0,3,0) = p_{\mathbf{X}}(0,0,3) = 1/216$$

Finally:

$$\begin{aligned} P\left[X_1 > X_3\right] &= p_{\mathbf{X}}(1,0,0) + p_{\mathbf{X}}(1,1,0) + p_{\mathbf{X}}(2,0,0) + p_{\mathbf{X}}(1,2,0) \\ &+ p_{\mathbf{x}}(2,0,1) + p_{\mathbf{X}}(2,1,0) + p_{\mathbf{X}}(3,0,0) \\ &= 8/27. \end{aligned}$$

We say that the random variables $X_1, X_2, \ldots, X_n$ are **jointly continuous random variables** if the probability of any n-dimensional event A is given by an n-dimensional integral of a probability density function:

$$P[\mathbf{X} \text{ in } A] = \int_{\mathbf{x} \text{ in } A} \int f_{X_1,\ldots,X_n}\left(x_1', \ldots, x_n'\right) dx_1' \ldots dx_n'$$

where $f_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n\right)$ is the **joint probability density function**.

The joint cdf of $\mathbf{X}$ is obtained from the joint pdf by integration:

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= F_{X_1,X_2,\ldots,X_n}\left(x_1, x_2, \ldots, x_n\right) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}\left(x_1', \ldots, x_n'\right) dx_1' \ldots dx_n' \end{aligned}$$

The joint pdf (if the derivative exists) is given by

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq f_{X_1,X_2,\ldots,X_n}\left(x_1, x_2, \ldots, x_n\right) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n\right) \quad (4)$$

A family of **marginal pdf's** is associated with the joint pdf in Eq. (4). The marginal pdf for a subset of the random variables is obtained by integrating the other variables out. For example, the marginal pdf of X_1 is

$$f_{X_1}\left(x_1\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}\left(x_1, x_2', \ldots, x_n'\right) dx_2' \ldots dx_n'$$

As another example, the marginal pdf for $X_1, \ldots, X_{n-1}$ is given by

$$f_{X_1,\ldots,X_{n-1}}\left(x_1, \ldots, x_{n-1}\right) = \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}\left(x_1, \ldots, x_{n-1}, x_n'\right) dx_n'$$

A family of **conditional pdf's** is also associated with the joint pdf. For example, the pdf of X_n given the values of $X_1, \ldots, X_{n-1}$ is given by

$$f_{X_n}\left(x_n \mid x_1, \ldots, x_{n-1}\right) = \frac{f_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n\right)}{f_{X_1,\ldots,X_{n-1}}\left(x_1, \ldots, x_{n-1}\right)} \quad (5)$$

if $f_{X_1,\ldots,X_{n-1}}(x_1,\ldots,x_{n-1}) > 0$

Repeated applications of Eq. (5) yield an expression analogous to Eq. (3):

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \\ f_{X_n}(x_n \mid x_1,\ldots,x_{n-1})\, f_{X_{n-1}}(x_{n-1} \mid x_1,\ldots,x_{n-2}) \ldots f_{X_2}(x_2 \mid x_1)\, f_{X_1}(x_1)$$

Example 6

The random variables X_1, X_2, and X_3 have the joint Gaussian pdf

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \frac{e^{-\left(x_1^2+x_2^2-\sqrt{2}x_1x_2+1/2x_3^2\right)}}{2\pi\sqrt{\pi}}$$

Find the marginal pdf of X_1 and X_3. Find the conditional pdf of X_2 given X_1 and X_3.

The marginal pdf for the pair X_1 and X_3 is found by integrating the joint pdf over x_2:

$$f_{X_1,X_3}(x_1,x_3) = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-\left(x_1^2+x_2^2-\sqrt{2}x_1x_2\right)}}{2\pi/\sqrt{2}} dx_2$$

The above integral was carried out in a previous example by completing the square in the exponent. By substituting the result of the integration above, we obtain

$$f_{X_1,X_3}(x_1,x_3) = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} \frac{e^{-x_1^2/2}}{\sqrt{2\pi}}$$

Therefore X_1 and X_3 are independent zero-mean, unit-variance Gaussian random variables.

The conditional pdf of X_2 given X_1 and X_3 is:

$$\begin{aligned} f_{X_2}(x_2 \mid x_1,x_3) &= \frac{e^{-\left(x_1^2+x_2^2-\sqrt{2}x_1x_2+1/2x_3^2\right)}}{2\pi\sqrt{\pi}} \frac{\sqrt{2\pi}\sqrt{2\pi}}{e^{-x_3^2/2}e^{-x_1^2/2}} \\ &= \frac{e^{-\left(1/2x_1^2+x_2^2-\sqrt{2}x_1x_2\right)}}{\sqrt{\pi}} = \frac{e^{-\left(x_2-x_1/\sqrt{2}x_1\right)^2}}{\sqrt{\pi}}. \end{aligned}$$

We conclude that X_2 given X_1 and X_3 is a Gaussian random variable with mean $x_1/\sqrt{2}$ and variance $1/2$.

Example 7: Multiplicative Sequence

Let X_1 be uniform in $[0,1]$, X_2 be uniform in $[0, X_1]$, and X_3 be uniform in $[0, X_2]$. (Note that X_3 is also the product of three uniform random variables.) Find the joint pdf of $\mathbf{X}$ and the marginal pdf of X_3.

For $0 < z < y < x < 1$, the joint pdf is nonzero and given by:

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = f_{X_3}(z \mid x, y) f_{X_2}(y \mid x) f_{X_1}(x) = \frac{1}{y}\frac{1}{x}1 = \frac{1}{xy}.$$

The joint pdf of X_2 and X_3 is nonzero for $0 < z < y < 1$ and is obtained by integrating x between y and 1:

$$f_{X_2,X_3}(x_2, x_3) = \int_y^1 \frac{1}{xy} dx = \frac{1}{y} \ln x \bigg|_y^1 = \frac{1}{y} \ln \frac{1}{y}$$

We obtain the pdf of X_3 by integrating y between z and 1:

$$f_{X_3}(x_3) = -\int_z^1 \frac{1}{y} \ln y dy = -\frac{1}{2}(\ln y)^2 \bigg|_z^1 = \frac{1}{2}(\ln z)^2$$

Note that the pdf of X_3 is concentrated at the values close to $x = 0$.

1.2 Independent Random Variables

We say that a collection of random variables $X_1, \ldots, X_n$ is **independent** if

$$P[X_1 \text{ in } A_1, X_2 \text{ in } A_2, \ldots, X_n \text{ in } A_n] = P[X_1 \text{ in } A_1] P[X_2 \text{ in } A_2] \ldots P[X_n \text{ in } A_n]$$

for *any* one-dimensional events $A_1, \ldots, A_n$. It can be shown that $X_1, \ldots, X_n$ *are independent if and only if*

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \tag{6}$$

for all $x_1, \ldots, x_n$. If the random variables are discrete, Eq. (6) is equivalent to

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n) \qquad \text{for all } x_1, \ldots, x_n$$

If the random variables are jointly continuous, Eq. (6) is equivalent to

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

for all $x_1,\ldots,x_n$.

Example 8

The n samples $X_1, X_2, \ldots, X_n$ of a noise signal have joint pdf given by

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{e^{-\left(x_1^2+\ldots+x_n^2\right)/2}}{(2\pi)^{n/2}} \quad \text{for all } x_1,\ldots,x_n$$

It is clear that the above is the product of n one-dimensional Gaussian pdf's. Thus $X_1,\ldots,X_n$ are independent Gaussian random variables.

2 Jointly Gaussian Random Vectors

Jointly Gaussian random variables appear in numerous applications. They are frequently used to model signals in signal processing applications, and they are the most important model used in communication systems that involve dealing with signals in the presence of noise. They also play a central role in many statistical methods. Jointly Gaussian random variables have many nice properties. For example, the marginals of jointly Gaussian vectors are also jointly Gaussian. Also, the linear transformation of a jointly Gaussian vector is also jointly Gaussian. In this section we will see that jointly Gaussian vectors are best studied using vector and matrix notation. We will first consider the two-dimensional case.

2.1 Pairs of Jointly Gaussian Random Variables

The random variables X and Y are said to be **jointly Gaussian** if their joint pdf has the form

$$f_{X,Y}(x,y) = \frac{\exp\left\{\frac{-1}{2\left(1-\rho_{X,Y}^2\right)}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{X,Y}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{X,Y}^2}} \tag{7}$$

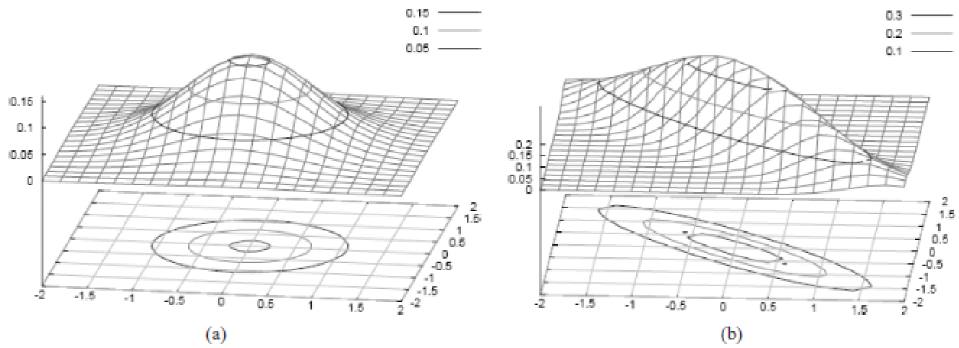

Figure 1: Jointly Gaussian pdf (a) $\rho = 0$ (b) $\rho = -0.9$.

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

The pdf is centered at the point (m_1, m_2), and it has a bell shape that depends on the values of σ_1, σ_2, and $\rho_{X,Y}$ as shown in Fig. 1. The pdf is constant for values x and y for which the argument of the exponent is constant:

$$\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{X,Y}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right] = \text{ constant.}$$

Figure 2 shows the orientation of these elliptical contours for various values of σ_1, σ_2, and $\rho_{X,Y}$. When $\rho_{X,Y} = 0$, that is, when X and Y are independent, the equal-pdf contour is an ellipse with principal axes aligned with the x - and y-axes. When $\rho_{X,Y} \neq 0$, the major axis of the ellipse is oriented along the angle [Edwards and Penney, pp. 570-571]

$$\theta = \frac{1}{2}\arctan^{-1}\tan\left(\frac{2\rho_{X,Y}\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right) \tag{8}$$

Note that the angle is 45° when the variances are equal.

The marginal pdf of X is found by integrating $f_{X,Y}(x,y)$ over all y. The integration is carried out by completing the square in the exponent as was done in Example **??**. The result is that the marginal pdf of X is

$$f_X(x) = \frac{e^{-(x-m_1)^2/2\sigma_1^2}}{\sqrt{2\pi}\sigma_1} \tag{9}$$

that is, X is a Gaussian random variable with mean m_1 and variance σ_1^2. Similarly, the marginal pdf for Y is found to be Gaussian with pdf mean m_2 and variance σ_2^2.

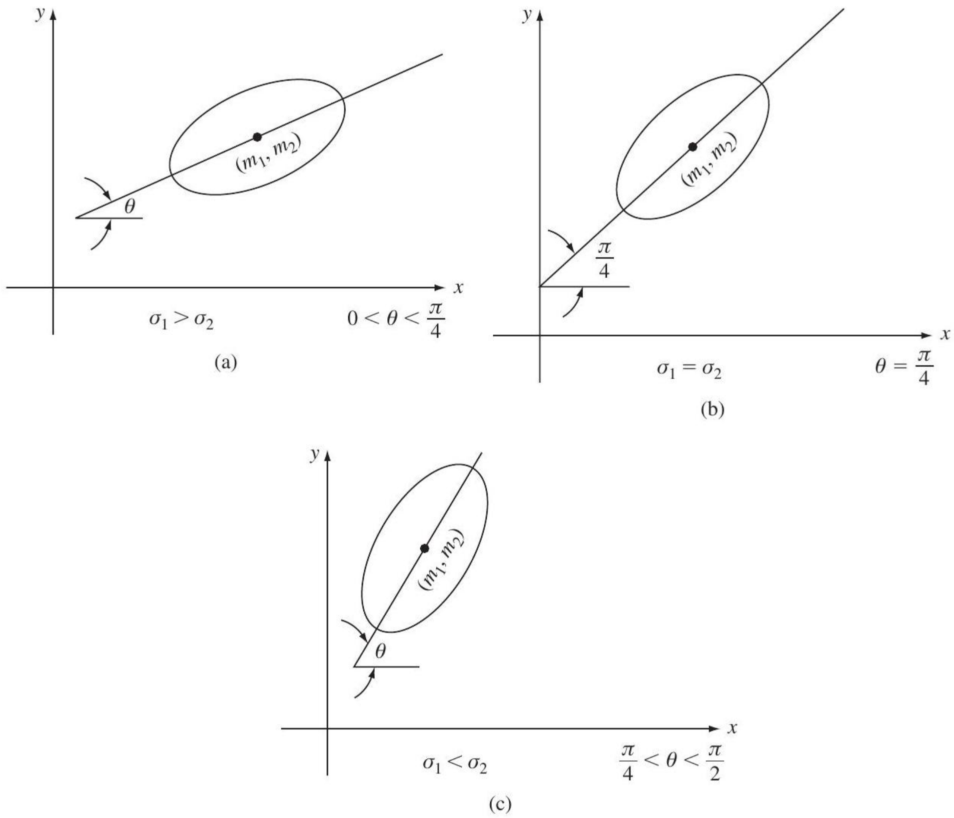

Figure 2: Orientation of contours of equal value of joint Gaussian pdf for $\rho_{X,Y} > 0$.

The conditional pdf's $f_X(x \mid y)$ and $f_Y(y \mid x)$ give us information about the interrelation between X and Y. The conditional pdf of X given $Y = y$ is

$$f_X(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\exp\left\{\frac{-1}{2\left(1-\rho_{X,Y}^2\right)\sigma_1^2}\left[x - \rho_{X,Y}\frac{\sigma_1}{\sigma_2}(y-m_2) - m_1\right]^2\right\}}{\sqrt{2\pi\sigma_1^2\left(1-\rho_{X,Y}^2\right)}} \tag{10}$$

Equation (10) shows that the conditional pdf of X given $Y = y$ is also Gaussian but with conditional mean $m_1 + \rho_{X,Y}(\sigma_1/\sigma_2)(y - m_2)$ and conditional variance $\sigma_1^2\left(1-\rho_{X,Y}^2\right)$. Note that when $\rho_{X,Y}=0$, the conditional pdf of X given $Y=y$ equals the marginal pdf of X. This is consistent with the fact that X and Y are independent when $\rho_{X,Y}=0$. On the other hand, as $|\rho_{X,Y}|\to 1$ the variance of X about the conditional mean approaches zero, so the conditional pdf approaches a delta function at the conditional mean. Thus when $|\rho_{X,Y}|=1$, the conditional variance is zero and X is equal to the conditional mean with probability one. We note that similarly $f_Y(y \mid x)$ is Gaussian with conditional mean $m_2+\rho_{X,Y}\left(\sigma_2/\sigma_1\right)(x-m_1)$ and conditional variance $\sigma_2^2\left(1-\rho_{X,Y}^2\right)$.

We now show that the $\rho_{X,Y}$ in Eq. (7) is indeed the correlation coefficient between X and Y. The covariance between X and Y is defined by

$$\begin{aligned}\mathrm{COV}(X,Y) &= E\left[(X-m_1)(Y-m_2)\right]\\ &= E\left[E\left[(X-m_1)(Y-m_2)\mid Y\right]\right]\end{aligned}$$

Now the conditional expectation of $(X-m_1)(Y-m_2)$ given $Y=y$ is

$$\begin{aligned}E\left[(X-m_1)(Y-m_2)\mid Y=y\right] &= (y-m_2)\,E\left[X-m_1\mid Y=y\right]\\ &= (y-m_2)\left(E[X\mid Y=y]-m_1\right)\\ &= (y-m_2)\left(\rho_{X,Y}\frac{\sigma_1}{\sigma_2}(y-m_2)\right)\end{aligned}$$

where we have used the fact that the conditional mean of X given $Y=y$ is $m_1+\rho_{X,Y}\left(\sigma_1/\sigma_2\right)(y-m_2)$. Therefore

$$E\left[(X-m_1)(Y-m_2)\mid Y\right]=\rho_{X,Y}\frac{\sigma_1}{\sigma_2}(Y-m_2)^2$$

and

$$\begin{aligned}\mathrm{COV}(X,Y) &= E\left[E\left[(X-m_1)(Y-m_2)\mid Y\right]\right]=\rho_{X,Y}\frac{\sigma_1}{\sigma_2}E\left[(Y-m_2)^2\right]\\ &= \rho_{X,Y}\sigma_1\sigma_2\end{aligned}$$

The above equation is consistent with the definition of the correlation coefficient, $\rho_{X,Y}=\mathrm{COV}(X,Y)/\sigma_1\sigma_2$. Thus the $\rho_{X,Y}$ in Eq. (7) is indeed the correlation coefficient between X and Y.

Example 9

The amount of yearly rainfall in city 1 and in city 2 is modeled by a pair of jointly Gaussian random variables, X and Y, with pdf given by Eq.(7). Find the most likely value of X given that we know $Y = y$.

The most likely value of X given $Y = y$ is the value of x for which $f_X(x \mid y)$ is maximum. The conditional pdf of X given $Y = y$ is given by Eq. (10), which is maximum at the conditional mean

$$E[X \mid y] = m_1 + \rho_{X,Y} \frac{\sigma_1}{\sigma_2} (y - m_2)$$

Note that this "maximum likelihood" estimate is a linear function of the observation y.

Example 10: Estimation of Signal in Noise

Let $Y = X + N$ where X (the "signal") and N (the "noise") are independent zero-mean Gaussian random variables with different variances. Find the correlation coefficient between the observed signal Y and the desired signal X. Find the value of x that maximizes $f_X(x \mid y)$.

The mean and variance of Y and the covariance of X and Y are:

$$\begin{aligned}
E[Y] &= E[X] + E[N] = 0 \\
\sigma_Y^2 &= E\left[Y^2\right] = E\left[(X+N)^2\right] = E\left[X^2 + 2XN + N^2\right] \\
&= E\left[X^2\right] + E\left[N^2\right] = \sigma_X^2 + \sigma_N^2 \\
\mathrm{COV}(X,Y) &= E[(X - E[X])(E(Y - E[Y])] = E[XY] \\
&= E[X(X+N)] = \sigma_X^2.
\end{aligned}$$

Therefore, the correlation coefficient is:

$$\rho_{X,Y} = \frac{\mathrm{COV}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_X}{\left(\sigma_X^2 + \sigma_N^2\right)^{1/2}} = \frac{1}{\left(1 + \frac{\sigma_N^2}{\sigma_X^2}\right)^{1/2}}$$

Note that $\rho_{X,Y}^2 = \sigma_X^2/\sigma_Y^2 = 1 - \sigma_N^2/\sigma_Y^2$.

To find the joint pdf of X and Y consider the following linear transformation:

$$\begin{aligned}
X &= X \qquad \text{which has inverse} \qquad & X &= X \\
Y &= X + N & N &= -X + Y.
\end{aligned}$$

From Eq. (**??**) we have:

$$
\begin{aligned}
f_{X,Y}(x,y) = \left.\frac{f_{X,N}(x,y)}{\det A}\right|_{x=x,n=y-x} &= \left.\frac{e^{-x^2/2\sigma_X^2}}{\sqrt{2\pi}\sigma_X}\frac{e^{-n^2 12\sigma_N^2}}{\sqrt{2\pi}\sigma_N}\right|_{x=x,n=y-x} \\
&= \frac{e^{-x^2/2\sigma_X^2}}{\sqrt{2\pi}\sigma_X}\frac{e^{-(y-x)^2/2\sigma_N^2}}{\sqrt{2\pi}\sigma_N}
\end{aligned}
$$

The conditional pdf of the signal X given the observation Y is then:

$$
\begin{aligned}
f_X(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} &= \frac{e^{-x^2/2\sigma_X^2}}{\sqrt{2\pi}\sigma_X}\frac{e^{-(y-x)^2/2\sigma_N^2}}{\sqrt{2\pi}\sigma_N}\frac{\sqrt{2\pi}\sigma_Y}{e^{-y^2(2\sigma_X^2}} \\
&= \frac{\exp\left\{-\frac{1}{2}\left(\left(\frac{x}{\sigma_X}\right)^2 + \left(\frac{y-x}{\sigma_N}\right)^2 - \left(\frac{y}{\sigma_Y}\right)^2\right)\right\}}{\sqrt{2\pi}\sigma_N\sigma_X/\sigma_Y} \\
&= \frac{\exp\left\{-\frac{1}{2}\frac{\sigma_Y^2}{\sigma_X^2}\left(x-\frac{\sigma_X^2}{\sigma_Y^2}y\right)^2\right\}}{\sqrt{2\pi}\sigma_N\sigma_X/\sigma_Y} \\
&= \frac{\exp\left\{-\frac{1}{2\left(1-\rho_{X},y\right)\sigma_X^2}\left(x-\left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_X^2}\right)y\right)^2\right\}}{\sqrt{1-\rho_{X,Y}^2\sigma_X}}
\end{aligned}
$$

This pdf has its maximum value, when the argument of the exponent is zero, that is,

$$
x = \left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2}\right)y = \left(\frac{1}{1+\frac{\sigma_N^2}{\sigma_X^2}}\right)y
$$

The signal-to-noise ratio (SNR) is defined as the ratio of the variance of X and the variance of N. At high SNRs this estimator gives $x \approx y$, and at very low signal-to-noise ratios, it gives $x \approx 0$.

Example 11: Rotation of Jointly Gaussian Random Variables

The ellipse corresponding to an arbitrary two-dimensional Gaussian vector forms an angle

$$
\theta = \frac{1}{2}\arctan\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2-\sigma_2^2}\right)
$$

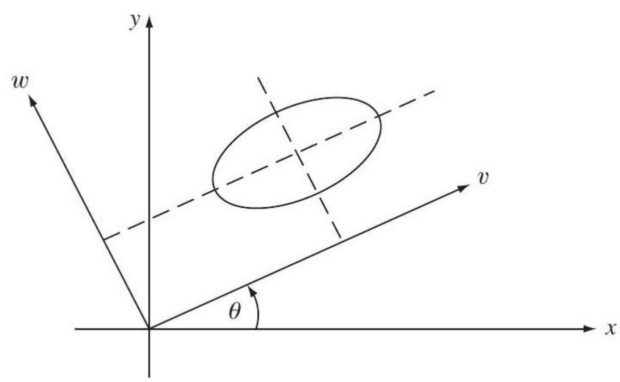

Figure 3: A rotation of the coordinate system transforms a pair of dependent Gaussian random variables into a pair of independent Gaussian random variables.

relative to the x-axis. Suppose we define a new coordinate system whose axes are aligned with those of the ellipse as shown in Fig. 3. This is accomplished by using the following rotation matrix:

$$\begin{bmatrix} V \\ W \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

To show that the new random variables are independent it suffices to show that they have covariance zero:

$$\begin{aligned} \text{COV}(V, W) =& E[(V - E[V])(W - E[W])] \\ =& E\left[\{(X - m_1)\cos\theta + (Y - m_2)\sin\theta\}\right. \\ & \left.\times \{-(X - m_1)\sin\theta + (Y - m_2)\cos\theta\}\right] \\ =& -\sigma_1^2 \sin\theta\cos\theta + \text{COV}(X, Y)\cos^2\theta \\ & - \text{COV}(X, Y)\sin^2\theta + \sigma_2^2 \sin\theta\cos\theta \\ =& \frac{(\sigma_2^2 - \sigma_1^2)\sin 2\theta + 2\,\text{COV}(X, Y)\cos 2\theta}{2} \\ =& \frac{\cos 2\theta\left[(\sigma_2^2 - \sigma_1^2)\tan 2\theta + 2\,\text{COV}(X, Y)\right]}{2} \end{aligned}$$

If we let the angle of rotation θ be such that

$$\tan 2\theta = \frac{2\,\text{COV}(X, Y)}{\sigma_1^2 - \sigma_2^2}$$

then the covariance of V and W is zero as required.

2.2 Jointly Gaussian Random Vectors

The random variables $X_1, X_2, \ldots, X_n$ are said to be jointly Gaussian if their joint pdf is given by

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq f_{X_1,X_2,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}K^{-1}(\mathbf{x}-\mathbf{m})\right\}}{(2\pi)^{n/2}|K|^{1/2}} \tag{11}$$

where $\mathbf{x}$ and $\mathbf{m}$ are column vectors defined by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix}$$

and K is the covariance matrix that is defined by

$$K = \begin{bmatrix} \mathrm{VAR}(X_1) & \mathrm{COV}(X_1,X_2) & \ldots & \mathrm{COV}(X_1,X_n) \\ \mathrm{COV}(X_2,X_1) & \mathrm{VAR}(X_2) & \ldots & \mathrm{COV}(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{COV}(X_n,X_1) & \ldots & & \mathrm{VAR}(X_n) \end{bmatrix} \tag{12}$$

The $(.\ )^{\mathrm{T}}$ in Eq. (11) denotes the transpose of a matrix or vector. Note that the covariance matrix is a symmetric matrix since $\mathrm{COV}(X_i,X_j) = \mathrm{COV}(X_j,X_i)$.

Equation (11) shows that the pdf of jointly Gaussian random variables is completely specified by the individual means and variances and the pairwise covariances. It can be shown using the joint characteristic function that all the marginal pdf's associated with Eq. (11) are also Gaussian and that these too are completely specified by the same set of means, variances, and covariances.

Example 12

Verify that the two-dimensional Gaussian pdf has the form of Eq. (11).

The covariance matrix for the two-dimensional case is given by

$$K = \begin{bmatrix} \sigma_1^2 & \rho_{X,Y}\sigma_1\sigma_2 \\ \rho_{X,Y}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where we have used the fact the $\text{COV}(X_1, X_2) = \rho_{X,Y}\sigma_1\sigma_2$. The determinant of K is $\sigma_1^2\ \sigma_{\frac{2}{2}}^2 \left(1 - \rho_{X,Y}^2\right)$ so the denominator of the pdf has the correct form. The inverse of the covariance matrix is also a real symmetric matrix:

$$K^{-1} = \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho_{X,Y}^2\right)}\begin{bmatrix} \sigma_2^2 & -\rho_{X,Y}\sigma_1\sigma_2 \\ -\rho_{X,Y}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

The term in the exponent is therefore

$$\begin{aligned}
&\frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho_{X,Y}^2\right)}\left(x-m_1, y-m_2\right)\begin{bmatrix} \sigma_2^2 & -\rho_{X,Y}\sigma_{\sigma_1}\sigma_2 \\ -\rho_{X,Y}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}\begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix} \\
&= \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho_{X,Y}^2\right)}\left(x-m_1, y-m_2\right)\begin{bmatrix} \sigma_2^2\left(x-m_1\right) - \rho_{X,Y}\sigma_1\sigma_2\left(y-m_2\right) \\ -\rho_{X,Y}\sigma_1\sigma_2\left(x-m_1\right) + \sigma_1^2\left(y-m_2\right) \end{bmatrix} \\
&= \frac{\left(\left(x-m_1\right)/\sigma_1\right)^2 - 2\rho_{X,Y}\left(\left(x-m_1\right)/\sigma_1\right)\left(\left(y-m_2\right)/\sigma_2\right) + \left(\left(y-m_2\right)/\sigma_2\right)^2}{\left(1-\rho_{X,Y}^2\right)}
\end{aligned}$$

Thus the two-dimensional pdf has the form of Eq. (11).

Example 13

The vector of random variables (X, Y, Z) is jointly Gaussian with zero means and covariance matrix:

$$K = \begin{bmatrix} \text{VAR}(X) & \text{COV}(X,Y) & \text{COV}(X,Z) \\ \text{COV}(Y,X) & \text{VAR}(Y) & \text{COV}(Y,Z) \\ \text{COV}(Z,X) & \text{COV}(Z,Y) & \text{VAR}(Z) \end{bmatrix} = \begin{bmatrix} 1.0 & 0.2 & 0.3 \\ 0.2 & 1.0 & 0.4 \\ 0.3 & 0.4 & 1.0 \end{bmatrix}$$

Find the marginal pdf of X and Z.

We can solve this problem two ways. The first involves integrating the pdf directly to obtain the marginal pdf. The second involves using the fact that the marginal pdf for X and Z is also Gaussian and has the same set of means, variances, and covariances. We will use the second approach.

The pair (X, Z) has zero-mean vector and covariance matrix:

$$K' = \begin{bmatrix} \text{VAR}(X) & \text{COV}(X,Z) \\ \text{COV}(Z,X) & \text{VAR}(Z) \end{bmatrix} = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$

The joint pdf of X and Z is found by substituting a zero-mean vector and this covariance matrix into Eq. (11).

Example 14: Independence of Uncorrelated Jointly Gaussian Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are jointly Gaussian random variables with $\mathrm{COV}\left(X_i, X_j\right) = 0$ for $i \neq j$. Show that $X_1, X_2, \ldots, X_n$ are independent random variables.

From Eq. (12) we see that the covariance matrix is a diagonal matrix:

$$K = \operatorname{diag}\left[\mathrm{VAR}\left(X_i\right)\right] = \operatorname{diag}\left[\sigma_i^2\right]$$

Therefore

$$K^{-1} = \operatorname{diag}\left[\frac{1}{\sigma_i^2}\right]$$

and

$$(\mathbf{x}-\mathbf{m})^{\mathrm{T}} K^{-1}(\mathbf{x}-\mathbf{m}) = \sum_{i=1}^{n}\left(\frac{x_i - m_i}{\sigma_i}\right)^2$$

Thus from Eq. (11)

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\left(x_i-m_i\right)/\sigma_i\right]^2\right\}}{(2\pi)^{n/2}}|K|^{1/2} = \prod_{i=1}^{n}\frac{\exp\left\{-\frac{1}{2}\left[\left(x_i-m_i\right)/\sigma_i\right]^2\right\}}{\sqrt{2\pi\sigma_i^2}} = \prod_{i=1}^{n} f_{X_i}\left(x_i\right)$$

Thus $X_1, X_2, \ldots, X_n$ are independent Gaussian random variables.

3 ML and MAP Estimation

In this section, we are concerned with *estimating the value of a random variable X in terms of the observation of a random variable Y*. For example, X could be the input to a communication channel and Y could be the observed output. In a prediction application, X could be a future value of some quantity and Y its present value.

3.1 MAP and ML Estimators

There are two basic approaches to estimating a random variable X in terms of the random variable Y. For example, in estimating the output of a discrete communications channel we are interested in finding the most probable input given the observation $Y = y$, that is, the value of input x that maximizes $P[X = x \mid Y = y]$:

$$\max_x P[X = x \mid Y = y]$$

In general we refer to the above estimator for X in terms of Y as the **maximum a posteriori (MAP) estimator**. The a posteriori probability is given by:

$$P[X = x \mid Y = y] = \frac{P[Y = y \mid X = x]P[X = x]}{P[Y = y]}$$

and so the MAP estimator requires that we know the a priori probabilities $P[X = x]$. In some situations we know $P[Y = y \mid X = x]$ but we do not know the a priori probabilities, so we select the estimator value x as the value that maximizes the likelihood of the observed value $Y = y$:

$$\max_x P[Y = y \mid X = x]$$

We refer to this estimator of X in terms of Y as the **maximum likelihood (ML) estimator**.

We can define MAP and ML estimators when X and Y are continuous random variables by replacing events of the form $\{Y = y\}$ by $\{y < Y < y + dy\}$. If X and Y are continuous, **the MAP estimator for $\boldsymbol{X}$ given the observation $\boldsymbol{Y}$** is given by:

$$\max_x f_X(X = x \mid Y = y)$$

and **the ML estimator for $\boldsymbol{X}$ given the observation $\boldsymbol{Y}$** is given by:

$$\max_x f_X(Y = y \mid X = x)$$

Example 15: Comparison of ML and MAP Estimators

Let X and Y be the random pair where the joint pdf is given by $e^{-(x+y)}$ $0 < x < y < \infty$, and zero elsewhere. Find the MAP and ML estimators for X in terms of Y.

In a previous example, we found the conditional pdf of X given Y is given by:

$$f_X(x \mid y) = e^{-(x-y)} \text{ for } y \leq x$$

which decreases as x increases beyond y. Therefore the MAP estimator is $\hat{X}_{\text{MAP}} = y$. On the other hand, the conditional pdf of Y given X is:

$$f_Y(y \mid x) = \frac{e^{-y}}{1 - e^{-x}} \text{ for } 0 < y \leq x$$

As x increases beyond y, the denominator becomes larger so the conditional pdf decreases. Therefore the ML estimator is $\hat{X}_{\text{ML}} = y$. In this example the ML and MAP estimators agree.

Example 16: Jointly Gaussian Random Variables

Find the MAP and ML estimator of X in terms of Y when X and Y are jointly Gaussian random variables.

The conditional pdf of X given Y is given by:

$$f_X(x \mid y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_X^2}\left(x - \rho\frac{\sigma_X}{\sigma_Y}(y - m_Y) - m_X\right)^2\right\}}{\sqrt{2\pi\sigma_X^2(1-\rho^2)}}$$

which is maximized by the value of x for which the exponent is zero. Therefore

$$\hat{X}_{\text{MAP}} = \rho\frac{\sigma_X}{\sigma_Y}(y - m_Y) + m_X$$

The conditional pdf of Y given X is:

$$f_Y(y \mid x) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_Y^2}\left(y - \rho\frac{\sigma_Y}{\sigma_X}(x - m_X) - m_Y\right)^2\right\}}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}$$

which is also maximized for the value of x for which the exponent is zero:

$$0 = y - \rho\frac{\sigma_Y}{\sigma_X}(x - m_X) - m_Y$$

The ML estimator for X given $Y = y$ is then:

$$\hat{X}_{\text{ML}} = \frac{\sigma_X}{\rho\sigma_Y}(y - m_Y) + m_X$$

Therefore we conclude that $\hat{X}_{\text{ML}} \neq \hat{X}_{\text{MAP}}$. In other words, knowledge of the a priori probabilities of X will affect the estimator.

3.2 Minimum MSE Linear Estimator

The estimate for X is given by a function of the observation $\hat{X} = g(Y)$. When X and Y are continuous random variables, we frequently use the

mean square error (MSE) to measure the goodness of the estimator. In this section we first consider the case where $g(Y)$ is constrained to be a linear function of Y, and then consider the case where $g(Y)$ can be any function, whether linear or nonlinear.

First, consider the problem of estimating a random variable X by a constant a so that the mean square error is minimized:

$$\min_a E\left[(X-a)^2\right] = E\left[X^2\right] - 2aE[X] + a^2 \tag{13}$$

The best a is found by taking the derivative with respect to a, setting the result to zero, and solving for a. The result is

$$a^* = E[X]$$

which makes sense since the expected value of X is the center of mass of the pdf. The mean square error for this estimator is equal to $E\left[(X-a^*)^2\right] = \mathrm{VAR}(X)$.

Now consider estimating X by a linear function $g(Y) = aY + b$:

$$\min_{a,b} E\left[(X - aY - b)^2\right] \tag{14}$$

Equation (14) can be viewed as the approximation of $X - aY$ by the constant b. This is the minimization posed in Eq. (13) and the best b is

$$b^* = E[X - aY] = E[X] - aE[Y] \tag{15}$$

Substitution into Eq. (14) implies that the best a is found by

$$\min_a E\left[\{(X - E[X]) - a(Y - E[Y])\}^2\right]$$

We once again differentiate with respect to a, set the result to zero, and solve for a:

$$0 = \frac{d}{da} E\left[(X - E[X]) - a(Y - E[Y])^2\right]$$

$$\begin{aligned} &= -2E[\{(X - E[X]) - a(Y - E[Y])\}(Y - E[Y])] \\ &= -2(\mathrm{COV}(X,Y) - a\,\mathrm{VAR}(Y)) \end{aligned} \tag{16}$$

The best coefficient a is found to be

$$a^* = \frac{\mathrm{COV}(X,Y)}{\mathrm{VAR}(Y)} = \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}$$

where $\sigma_Y = \sqrt{\mathrm{VAR}(Y)}$ and $\sigma_X = \sqrt{\mathrm{VAR}(X)}$. Therefore, ***the minimum mean square error (mmse) linear estimator*** for X in terms of Y is

$$\hat{X} = a^*Y + b^* \tag{17}$$

$$= \rho_{X,Y}\sigma_X\frac{Y - E[Y]}{\sigma_Y} + E[X] \tag{18}$$

The term $(Y - E[Y])/\sigma_Y$ is simply a zero-mean, unit-variance version of Y. Thus $\sigma_X(Y - E[Y])/\sigma_Y$ is a rescaled version of Y that has the variance of the random variable that is being estimated, namely σ_X^2. The term $E[X]$ simply ensures that the estimator has the correct mean. The key term in the above estimator is the correlation coefficient: $\rho_{X,Y}$ *specifies the sign and extent of the estimate of* Y *relative to* $\sigma_X(Y - E[Y])/\sigma_Y$. If X and Y are uncorrelated (i.e., $\rho_{X,Y} = 0$) then the best estimate for X is its mean, $E[X]$. On the other hand, if $\rho_{X,Y} = \pm 1$ then the best estimate is equal to $\pm\sigma_X(Y - E[Y])/\ \sigma_Y + E[X]$.

We draw our attention to the second equality in Eq. (16):

$$E\left[\{(X - E[X]) - a^*(Y - E[Y])\}(Y - E[Y])\right] = 0$$

This equation is called the **orthogonality condition** because it states that the error of the best linear estimator, the quantity inside the braces, is orthogonal to the observation $Y - E[Y]$. The orthogonality condition is a fundamental result in mean square estimation.

The mean square error of the best linear estimator is

$$\begin{aligned}
e_L^* =& E\left[((X - E[X]) - a^*(Y - E[Y]))^2\right] \\
=& E\left[((X - E[X]) - a^*(Y - E[Y]))(X - E[X])\right] \\
&\quad - a^*E\left[((X - E[X]) - a^*(Y - E[Y]))(Y - E[Y])\right] \\
=& E\left[((X - E[X]) - a^*(Y - E[Y]))(X - E[X])\right] \\
=& \mathrm{VAR}(X) - a^*\,\mathrm{COV}(X,Y) \\
=& \mathrm{VAR}(X)\left(1 - \rho_{X,Y}^2\right)
\end{aligned} \tag{19}$$

where the second equality follows from the orthogonality condition. Note that when $|\rho_{X,Y}| = 1$, the mean square error is zero. This implies that $P\left[|X - a^*Y - b^*| = 0\right] = P\left[X = a^*Y + b^*\right] = 1$, so that X is essentially a linear function of Y.

3.3 Minimum MSE Estimator

In general the estimator for X that minimizes the mean square error is a nonlinear function of Y. The estimator $g(Y)$ that best approximates X in the sense of minimizing mean square error must satisfy

$$\underset{g(.)}{\text{minimize}} E\left[(X - g(Y))^2\right]$$

The problem can be solved by using **conditional expectation**:

$$\begin{aligned} E\left[(X - g(Y))^2\right] &= E\left[E\left[(X - g(Y))^2 \mid Y\right]\right] \\ &= \int_{-\infty}^{\infty} E\left[(X - g(Y))^2 \mid Y = y\right] f_Y(y) dy \end{aligned}$$

The integrand above is positive for all y; therefore, the integral is minimized by minimizing $E\left[(X - g(Y))^2 \mid Y = y\right]$ for each y. But $g(y)$ is a constant as far as the conditional expectation is concerned, so the problem is equivalent to Eq. (13) and the "constant" that minimizes $E\left[(X - g(y))^2 \mid Y = y\right]$ is

$$g^*(y) = E[X \mid Y = y] \tag{20}$$

The function $g^*(y) = E[X \mid Y = y]$ is called the regression curve which simply traces the conditional expected value of X given the observation $Y = y$.

The mean square error of the best estimator is:

$$\begin{aligned} e^* = E\left[(X - g^*(Y))^2\right] &= \int_R E\left[(X - E[X \mid y])^2 \mid Y = y\right] f_Y(y) dy \\ &= \int_{R^n} \text{VAR}[X \mid Y = y] f_Y(y) dy \end{aligned}$$

Linear estimators in general are suboptimal and have larger mean square errors.

Example 17: Comparison of Linear and Minimum MSE Estimators

Let X and Y be the random pair in **Example 5.16**. Find the best linear and nonlinear estimators for X in terms of Y, and of Y in terms of X.

Example 5.28 provides the parameters needed for the linear estimator: $E[X] = 3/2$, $E[Y] = 1/2, \text{VAR}[X] = 5/4, \text{VAR}[Y] = 1/4$, and $\rho_{X,Y} = 1/\sqrt{5}$. **Example 5.32** provides the conditional pdf's needed to find the nonlinear estimator. The best linear and nonlinear estimators for X in terms of Y are:

$$\hat{X} = \frac{1}{\sqrt{5}}\frac{\sqrt{5}}{2}\frac{Y - 1/2}{1/2} + \frac{3}{2} = Y + 1$$

$$E[X \mid y] = \int_y^\infty xe^{-(x-y)}dx = y + 1 \text{ and so } E[X \mid Y] = Y + 1$$

Thus the optimum linear and nonlinear estimators are the same.

The best linear and nonlinear estimators for Y in terms of X are:

$$\hat{Y} = \frac{1}{\sqrt{5}}\frac{1}{2}\frac{X - 3/2}{\sqrt{5}/2} + \frac{1}{2} = (X+1)/5$$

$$E[Y \mid x] = \int_0^x y\frac{e^{-y}}{1-e^{-x}}dy = \frac{1 - e^{-x} - xe^{-x}}{1-e^{-x}} = 1 - \frac{xe^{-x}}{1-e^{-x}}$$

The optimum linear and nonlinear estimators are not the same in this case. Figure 4 compares the two estimators. It can be seen that the linear estimator is close to $E[Y \mid x]$ for lower values of x, where the joint pdf of X and Y are concentrated and that it diverges from $E[Y \mid x]$ for larger values of x.

Example 18

Let X be uniformly distributed in the interval $(-1, 1)$ and let $Y = X^2$. Find the best linear estimator for Y in terms of X. Compare its performance to the best estimator.

The mean of X is zero, and its correlation with Y is

$$E[XY] = E\left[XX^2\right] = \int_{-\frac{1}{2}}^{1} x^3/2dx = 0$$

Therefore $\text{COV}(X, Y) = 0$ and the best linear estimator for Y is $E[Y]$ by Eq. (18). The mean square error of this estimator is the $\text{VAR}(Y)$ by Eq. (19).

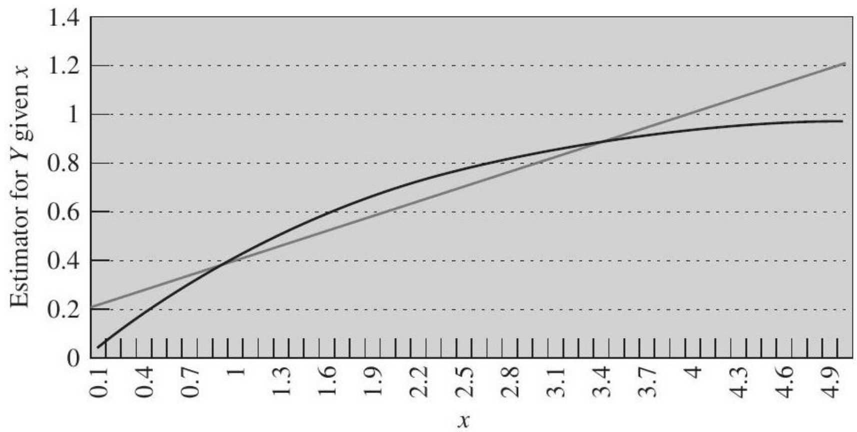

Figure 4: Comparison of linear and nonlinear estimators.

The best estimator is given by Eq. (20):

$$E[Y \mid X = x] = E\left[X^2 \mid X = x\right] = x^2$$

The mean square error of this estimator is

$$E\left[(Y - g(X))^2\right] = E\left[\left(X^2 - X^2\right)^2\right] = 0$$

Thus in this problem, the best linear estimator performs poorly while the nonlinear estimator gives the smallest possible mean square error, zero.

Example 19: Jointly Gaussian Random Variables

Find the minimum mean square error estimator of X in terms of Y when X and Y are jointly Gaussian random variables.

The minimum mean square error estimator is given by the conditional expectation of X given Y. From **Eq. (5.63)**, we see that the conditional expectation of X given $Y = y$ is given by

$$E[X \mid Y = y] = E[X] + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y}(Y - E[Y])$$

This is identical to the best linear estimator. Thus *for jointly Gaussian random variables the minimum mean square error estimator is linear.*