Chapter 5 Multivariate Probability Distributions

Abstract: In this chapter we study the relationships among random variables, which will be characterized by the joint probability distribution of random variables. Most insights into multivariate distributions can be gained by focusing on bivariate distributions. We first introduce the joint probability distribution of a bivariate random vector (X, Y) via the characterization of the joint cumulative distribution function, the joint probability mass function (when (X, Y) are discrete), and the joint probability density function (when (X, Y) are continuous) respectively. We then characterize various aspects of the relationship between X and Y using the conditional distributions, correlation, and conditional expectations. The concept of independence and its implications on the joint distributions, conditional distributions and correlation are also discussed. We also introduce a class of bivariate normal distributions.

Key words: Bivariate normal distribution, Bivariate transformation, Conditional distribution, Conditional mean, Conditional variance, Correlation, Independence, Joint moment generating function, Joint probability distribution, Law of iterated expectations, Marginal distribution.

5.1 Random Vectors and Joint Probability Distributions

Any economy is a system which consists of different units (e.g., households, assets, sectors, markets, etc). These units are generally related to each other. As a consequence, economic variables are interrelated. The most important goal of economic analysis and econometric analysis is to identify the relationships between economic events or economic variables. As discussed in Chapter 2, for two events A and B, the joint probability $P(A \cap B)$ and conditional probability $P(A|B)$ describe the joint and predictive relationships between them. Such a relationship can be exploited to predict one using the other.

Definition 1 (5.1)**. [Random Vector]:** An n-dimensional random vector, denoted as $Z = (Z_1, \cdots, Z_n)'$, is a function from a sample space S into $\mathbb{R}^n$, the n-dimensional Euclidean space. For each outcome $s \in S$, $Z(s)$ is an n-dimensional real-valued vector and is called a realization of the random vector Z.

In this chapter, we will mainly focus on bivariate probability distributions, which can illustrate most (but not all) essentials of multivariate probability distributions. We now consider two random variables (X, Y) in most of the subsequent discussion, where both X and Y are defined on the same probability space $(S, \mathbb{B}, P)$. A realization of (X, Y) will be a pair $(x, y) \in \mathbb{R}^2$.

Having defined a bivariate random vector (X, Y), we can now discuss probabilities of events that are defined in terms of (X, Y).

Like in the univariate case, we can use the CDF, now called the joint CDF, of X and Y, to characterize their joint distribution.

Definition 2 (5.2)**. [Joint CDF]:** The joint CDF of X and Y is defined as follows:

$$\begin{aligned} F_{XY}(x,y) &= P(X \le x, Y \le y) \\ &= P(X \le x \cap Y \le y) \end{aligned}$$

for any pair $(x,y) \in \mathbb{R}^2$.

We first examine the properties of the joint CDF.

Lemma 1 (5.1)**. [Properties of $F_{XY}(x,y)$] :**
(1) $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0; F_{XY}(\infty, \infty) = 1$.
(2) $F_{XY}(x,y)$ is non-decreasing in both x and y.
(3) $F_{XY}(x,y)$ is right continuous in both x and y.

Below we state a useful result.

Theorem 1 (5.2)**.** $F_X(x) = F_{XY}(x, +\infty)$ and $F_Y(y) = F_{XY}(+\infty, y)$.

Proof: Define two events: $A = \{X \le x\}$, $B = \{Y \le \infty\}$. Since B always holds, we have $A \cap B = A$. It follows that $P(A) = P(A \cap B)$, that is, $F_X(x) = F_{XY}(x, \infty)$.

Theorem 5.2 implies that individual CDF's of X and Y can be obtained from the joint CDF $F_{XY}(x,y)$. These individual CDF's are called the marginal CDF's of X and Y respectively.

The joint and marginal CDF's are widely used in statistics and econometrics. Below we provide an important example called copula, which can be used to model the dependence between random variables.

Example 1 (5.1)**. [Bivariate Copula]:** Put $U = F_X(X), V = F_Y(Y)$. Then both the probability integral transforms U and V are $U[0,1]$ random variables. The joint CDF of (U,V)

$$C(u,v) = P(U \le u, V \le v)$$

is called the copula associated with the joint probability distribution of (X,Y). The copula $C(u,v)$ is closely related to the joint CDF $F_{XY}(x,y)$. Suppose $F_X(\cdot)$ and $F_Y(\cdot)$ are strictly increasing functions. Then

$$\begin{aligned} F_{XY}(x,y) &= P(X \le x, Y \le y) \\ &= P\left[F_X(X) \le F_X(x), F_Y(Y) \le F_Y(y)\right] \\ &= P\left[U \le F_X(x), V \le F_Y(y)\right] \\ &= C\left[F_X(x), F_Y(y)\right]. \end{aligned}$$

This suggests that the joint distribution $F_{XY}(x,y)$ can be decomposed into two separate components: the marginal distributions $F_X(\cdot)$ and $F_Y(\cdot)$ on one hand, and the "pure" dependence function $C(\cdot,\cdot)$ between X and Y on the other hand. In other words, the copula $C(\cdot,\cdot)$ only specifies the dependence or association between X and Y regardless of their marginal distributions. It separates the marginal behaviors from the association between X and Y. Given the functional form $C(\cdot,\cdot)$, different marginal distributions will yield different joint distributions of (X,Y). The copula has been widely used in financial econometrics and financial industries, to model comovements among markets or among assets (Cherubini, Luciano and Vecchiato (2004)).

5.1.1 The Discrete Case

We now consider the case when both X and Y are DRV's, where we can introduce a joint PMF to characterize their joint probability distribution.

Definition 3 (5.3)**. [Joint PMF]** Let X and Y be two DRV's. Then their joint PMF is defined as

$$f_{XY}(x, y) = P(X = x \ \cap \ Y = y) = P(X = x, Y = y)$$

for any point $(x, y) \in \mathbb{R}^2$.

We now examine the properties of the joint PMF.

Lemma 2 (5.3)**. [Properties of $f_{XY}(x, y)$]** :
(1) $f_{XY}(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$.
(2) $\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f_{XY}(x, y) = 1$, where Ω_X and Ω_Y are the supports of X and Y respectively.

Definition 4 (5.4)**. [Support]** The support of a bivariate random vector (X, Y) is defined as the set of all possible pairs of (x, y) which (X, Y) will take with strictly positive probability. That is,

$$\text{Support}(X, Y) = \Omega_{XY} = \{(x, y) \in \mathbb{R}^2 : f_{XY}(x, y) > 0\}.$$

It is convenient to work on the support of (X, Y) only.
Question: Suppose Ω_X and Ω_Y are the supports of X, Y respectively. Is it true that

$$\begin{aligned} \Omega_{XY} &= \Omega_X \times \Omega_Y \\ &= \big\{(x, y) \in \mathbb{R}^2 : f_X(x) > 0, f_Y(y) > 0\big\}? \end{aligned}$$

The answer is generally no. Consider an example of bivariate distribution for which both X and Y takes nonnegative integers but with the restriction that $X \leq Y$. Then we have

$$\Omega_X = \Omega_Y = \{0, 1, 2, \cdots\}$$

while

$$\Omega_{XY} = \{(x, y) : 0 \leq x \leq y < \infty, \ x, y \text{ are integers}\}.$$

Obviously, Ω_{XY} is a subset of $\Omega_X \times \Omega_Y$.
Question: Is it true that

$$\sum_{(x,y) \in \Omega_{XY}} f_{XY}(x, y) = 1?$$

Note that in general, Ω_{XY} is a subset of $\Omega_X \times \Omega_Y$.

The joint PMF $f_{XY}(x, y)$ can be used to calculate the probability of any event defined in terms of (X, Y). For any subset $A \in \mathbb{R}^2$, we have

$$P\left[(X, Y) \in A\right] = \sum_{(x,y) \in A} f_{XY}(x, y).$$

Example 2 (5.2)**.** Suppose X and Y have the joint PMF

$$f_{XY}(x, y) = c|x + y| \text{ for } x = -1, 0, 1 \text{ and } y = 0, 1,$$

where c is an unknown constant.

Find (1) the supports of X, Y, and (X, Y) respectively; (2) the value of c; (3) $P(X = 0$ and $Y = 1)$; (4) $P(X = 1)$; and (5) $P(|X - Y| \leq 1)$.

Solution: (1) We have $\Omega_X = \{-1, 0, 1\}$, $\Omega_Y = \{0, 1\}$, and $\Omega_{XY} = \{(-1, 0), (0, 1), (1, 0), (1, 1)\}$. Note that Ω_{XY} is a subset of $\Omega_X \times \Omega_Y$ because $(0, 0) \in \Omega_X \times \Omega_Y$ but $(0, 0) \notin \Omega_{XY}$.

(2) Using the property that $\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f_{XY}(x, y) = 1$, we have

$$c\,[| - 1 + 0| + | - 1 + 1| + |0 + 0| + |0 + 1| + |1 + 0| + |1 + 1|] = 1.$$

Thus, $c = \frac{1}{5}$.

(3) By the definition of the joint PMF, we have

$$P(X = 0, Y = 1) = f_{XY}(0, 1) = \frac{1}{5}|0 + 1| = \frac{1}{5}.$$

(4) Noting that the event $\{X = 1\} = \{X = 1\} \cap \{Y \in \Omega_Y\}$, we have

$$\begin{aligned} P(X &= 1) \\ &= \sum_{y \in \Omega_Y} f_{XY}(1, y) \\ &= f_{XY}(1, 0) + f_{XY}(1, 1) \\ &= \frac{3}{5}. \end{aligned}$$

(5)

$$\begin{aligned} P(|X - Y| &\leq 1) = \sum_{(x,y) \in \Omega_{XY} : |x - y| \leq 1} f_{XY}(x, y) \\ &= f_{XY}(-1, 0) + f_{XY}(0, 1) + f_{XY}(1, 0) + f_{XY}(1, 1) \\ &= 1. \end{aligned}$$

We now investigate the relationship between $f_{XY}(x, y)$ and $F_{XY}(x, y)$, which is very similar to the relationship between $f_X(x)$ and $F_X(x)$ in the univariate case.

For DRV's X and Y, we can obtain the joint CDF

$$\begin{aligned} F_{XY}(x, y) &= P(X \leq x, Y \leq y) \\ &= \sum_{(u,v) \in \Omega_{XY}(x,y)} f_{XY}(u, v), \end{aligned}$$

where $\Omega_{XY}(x, y)$ is the set of all possible pairs of (u, v) in the support Ω_{XY} of (X, Y) such that $u \leq x, v \leq y$, namely,

$$\Omega_{XY}(x, y) = \{(u, v) \in \Omega_{XY}, u \leq x, v \leq y\}.$$

Thus, we can obtain $F_{XY}(x, y)$ from $f_{XY}(x, y)$.

On the other hand, by taking the differences of $F_{XY}(x, y)$ with respect to x and y, we can also recover $f_{XY}(x, y)$ from $F_{XY}(x, y)$. Without loss of generality, we assume that the possible values X can take are arranged in an increasing order: $x_1 < x_2 < x_3 < \cdots$, and the possible values Y can take are also arranged in an increasing order: $y_1 < y_2 < y_3 < \cdots$. Then, for $i > 1, j > 1$,

$$\begin{aligned} f_{XY}(x_i, y_j) &= \Delta_Y \Delta_X F_{XY}(x_i, y_j) \\ &= \Delta_Y [F_{XY}(x_i, y_j) - F_{XY}(x_{i-1}, y_j)] \\ &= [F_{XY}(x_i, y_j) - F_{XY}(x_i, y_{j-1})] - [F_{XY}(x_{i-1}, y_j) - F_{XY}(x_{i-1}, y_{j-1})] \\ &= F_{XY}(x_i, y_j) - F_{XY}(x_i, y_{j-1}) - F_{XY}(x_{i-1}, y_j) + F_{XY}(x_{i-1}, y_{j-1}), \end{aligned}$$

where Δ_X and Δ_Y are the difference operators with respect to x and y respectively. For example, for any given y, we have $\Delta_X F_{XY}(x_i, y) = F_{XY}(x_i, y) - F_{XY}(x_{i-1}, y)$.

The above formula does not cover the cases where $i = 1$ or $j = 1$. For these cases, we have

$$f_{XY}(x_i, y_j) = \begin{cases} F_{XY}(x_i, y_j) - F_{XY}(x_i, y_{j-1}), & \text{if } i = 1, j > 1, \\ F_{XY}(x_i, y_j) - F_{XY}(x_{i-1}, y_j), & \text{if } i > 1, j = 1, \\ F_{XY}(x_i, y_j), & \text{if } i = 1, j = 1. \end{cases}$$

The above bivariate concepts can be generalized to the multivariate case, where there are n random variables. For example, the joint PMF of n discrete random variables $X_1, \cdots, X_n$ is given by

$$f_{\mathbf{X}^n}(\mathbf{x}^n) = P(X_1 = x_1, \cdots, X_n = x_n)$$

for each n tuple $\mathbf{x}^n = (x_1, \cdots, x_n) \in \mathbb{R}^n$, and the joint CDF of $\mathbf{X}^n = (X_1, \cdots, X_n)$ is given by

$$F_{\mathbf{X}^n}(\mathbf{x}^n) = P\left(X_1 \le x_1, \cdots, X_n \le x_n\right)$$

for all $\mathbf{x}^n \in \mathbb{R}^n$.

5.1.2 The Continuous Case

Next, we consider the case of CRV's:

Definition 5 (5.5)**. [Joint PDF]:** Two random variables X and Y are said to have a continuous joint distribution if their joint CDF $F_{XY}(x, y)$ is absolutely continuous in both x and y. In this case, there exists a nonnegative function $f_{XY}(x, y)$ such that for any subset $A \in \mathbb{R}^2$,

$$P[(X, Y) \in A] = \int \int_{(x,y)\in A} f_{XY}(x, y) dx dy.$$

The function $f_{XY}(x, y)$ is called a joint PDF of (X, Y).

Lemma 3 (5.4). **[Properties of Joint PDF $f_{XY}(x,y)$]:**
(1) $f_{XY}(x,y) \geq 0$ for all (x,y) in the xy plane;
(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$.

Proof: (1) Denoting

$$A(x,y) = \{(u,v) : u \leq x, v \leq y\} \text{ for any given pair } (x,y) \in \mathbb{R}^2,$$

we have

$$\begin{aligned} P[(X,Y) &\in A(x,y)] = P(X \leq x, Y \leq y) \\ &= F_{XY}(x,y) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)dvdu. \end{aligned}$$

This formula is analogous to the double sum of the joint PMF in the discrete case. It indicates that $F_{XY}(x,y)$ can be obtained from the joint PDF $f_{XY}(x,y)$ by double integrations.

On the other hand, at the points of (x,y) where $F_{XY}(x,y)$ is differentiable, we have

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \geq 0,$$

where the equality follows from the fundamental theorem of calculus, and the inequality follows from the fact that $F_{XY}(x,y)$ is nondecreasing in (x,y). This formula is analogous to the double differences of the joint CDF $F_{XY}(x,y)$ with respect to (x,y) in the discrete case. It indicates that one can recover the joint PDF $f_{XY}(x,y)$ from differentiating the joint CDF $F_{XY}(x,y)$ with respect to (x,y).

(2) The integral $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$ follows immediately from the fact that $F_{XY}(\infty,\infty) = 1$. This completes the proof.

Question: What is the interpretation of the joint PDF $f_{XY}(x,y)$?

For any given pair (x,y) in the xy-plane, consider the event

$$A(x,y) = \left\{x - \frac{\epsilon}{2} < X \leq x + \frac{\epsilon}{2} \text{ and } y - \frac{\epsilon}{2} < Y \leq y + \frac{\epsilon}{2}\right\},$$

where $\epsilon > 0$ is a small constant. That is, $A(x,y)$ is the event that (X,Y) takes values in a small rectangular area centered at point (x,y) and with each size equal to ϵ. Suppose $f_{XY}(x,y)$ is continuous at the point (x,y). Then

$$\begin{aligned} P[A(x,y)] &= \int_{y-\epsilon/2}^{y+\epsilon/2}\int_{x-\epsilon/2}^{x+\epsilon/2} f_{XY}(u,v)dudv \\ &= f_{XY}(\bar{x},\bar{y}) \cdot \epsilon \cdot \epsilon \text{ for some } (\bar{x},\bar{y}) \\ &\approx f_{XY}(x,y)\epsilon^2 \text{ when } \epsilon \text{ is small.} \end{aligned}$$

Thus, although $f_{XY}(x,y)$ is not a probability measure, it is proportional to the probability that (X,Y) takes values in a small rectangular area centered at point (x,y). In other words,

$f_{XY}(x, y)$ is proportional to the probability that (X, Y) takes values in the area centered at point (x, y) in the xy plane.

The probability $P[(X, Y) \in A]$ has a 3-dimensional geometric interpretation. Recall that in the univariate case, when A is an interval on the real line (e.g., $A = \{x \in R : a < x \leq b\}$), the probability $P(X \in A)$ is equal to the area under the curve $f_X(x)$ over the interval A, as is shown in Figure 5.1(a). Now, for the bivariate case, suppose A is an area on the xy plane. Then the probability $P[(X, Y) \in A]$ is a volume under the surface of $f_{XY}(x, y)$ over the area A in the xy plane, as shown in Figure 5.1(b).

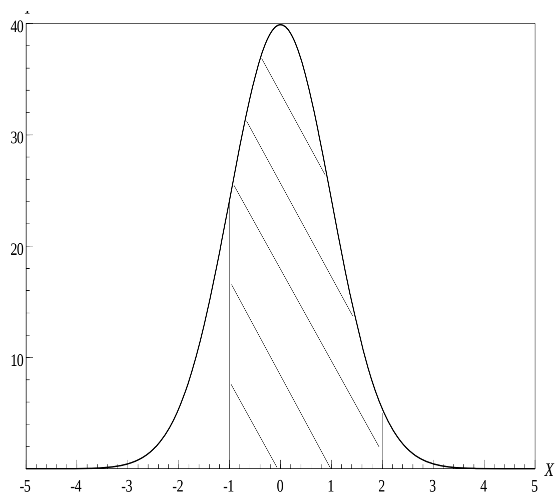

Figure 5.1 (a) Geometric interpretation for $P(X \in [-1, 2])$

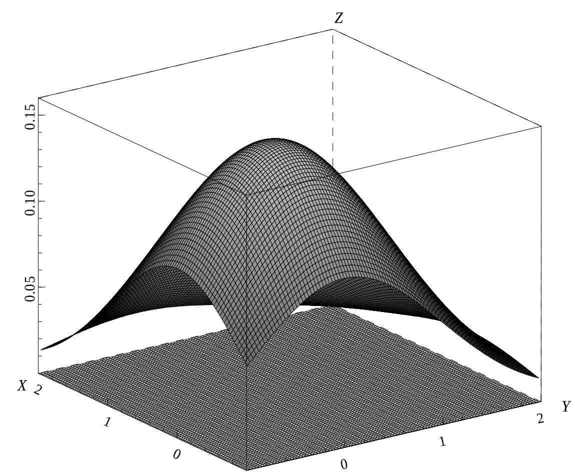

Figure 5.1(b) Geometric interpretation for $P((X, Y) \in [-1, 2]^2)$

The geometric interpretation for $P[(X, Y) \in A]$ has important implications: (1) an event that (X, Y) takes a value at any individual point (x, y), or values at any finite number of points in the xy plane, has probability zero; (2) an event that (X, Y) takes values on any one-dimensional curve in the xy plane has probability zero.

Because $F_{XY}(x, y)$ and $f_{XY}(x, y)$ can be recovered from each other, they are equivalent in the sense that they contain the same information about the joint distribution of (X, Y). However, it is often more convenient to use $f_{XY}(x, y)$ in practice. Also, like the univariate case, for each joint CDF $F_{XY}(x, y)$, there may exist some degree of arbitrariness in defining the joint PDF $f_{XY}(x, y)$ over a countable set of points (x, y) on the xy-plane, which will not alter the joint CDF $F_{XY}(x, y)$ in any way. However, there is always a most natural joint PDF which is smooth as much as possible over the xy plane.

It is important to note that the joint PDF $f_{XY}(x, y)$ is defined for all pairs of (x, y) in $\mathbb{R}^2$. The PDF $f_{XY}(x, y)$ may be zero for a large set A where $P[(X, Y) \in A] = 0$, although the PDF is defined for points in A. For this reason, it is convenient to focus on the set of pairs of (x, y) at which $f_{XY}(x, y)$ is strictly positive and such a set A is called the support of (X, Y) in the continuous case.

Definition 6 (5.6)**. [Support of** (X, Y)**]:** The support of the bivariate continuous random vector (X, Y) is defined as

$$\text{Support}(X, Y) = \{(x, y) \in \mathbb{R}^2 : f_{XY}(x, y) > 0\}.$$

Thus, Support(X, Y) is the set of all points (x, y) in the xy plane such that the probability that (X, Y) takes values in a small neighborhood of each of these points is always strictly positive. Since $f_{XY}(x, y) = 0$ for all points (x, y) outside the support of (X, Y), calculation of probabilities for any events related to (X, Y) will only involve integrals of $f_{XY}(x, y)$ on the support Ω_{XY}.

Example 3 (5.3)**.** Suppose X and Y have a joint PDF

$$f_{XY}(x, y) = cy^2 \text{ if } 0 \le x \le 2 \text{ and } 0 \le y \le 1,$$

where c is an unknown constant. Find (1) the value of c; (2) $P(X + Y > 2)$; (3) $P(X < 0.5)$; (4) $P(X = 3Y)$.

Solution: The support $\Omega_{XY} = \{(x, y) : 0 \le x \le 2, 0 \le y \le 1\}$ is a rectangular area in the xy plane. (1) Recalling $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)dxdy = 1$, we have

$$\begin{aligned}
\int_0^1 \left(\int_0^2 cy^2 dx \right) dy &= \int_0^1 cy^2 \left(\int_0^2 dx \right) dy \\
&= c \int_0^1 y^2 x\Big|_0^2 dy \\
&= 2c \int_0^1 y^2 dy \\
&= \frac{2c}{3} y^3|_0^1 \\
&= \frac{2c}{3} \\
&= 1.
\end{aligned}$$

Thus, $c = \frac{3}{2}$.
(2)

$$P(X + Y > 2) = \int_0^1 \left(\int_{2-y}^2 \frac{3}{2} y^2 dx \right) dy = \frac{3}{8};$$

(3)

$$P(X < 0.5) = \int_0^1 \left(\int_0^{\frac{1}{2}} \frac{3}{2} y^2 dx \right) dy = \frac{1}{4};$$

(4) Because $x = 3y$ is a line, the volume over a line is zero. Therefore, we have

$$P(X = 3Y) = 0.$$

Example 4 (5.4)**.** Suppose X and Y have a joint PDF

$$f_{XY}(x,y) = cx^2y \text{ for } x^2 \le y \le 1,$$

where c is an unknown constant. Find (1) the value of c; (2) $P(X \ge Y)$.

Solution: The support $\Omega_{XY} = \{(x,y) : x^2 \le y \le 1\}$ is plotted in Figure 5.2.

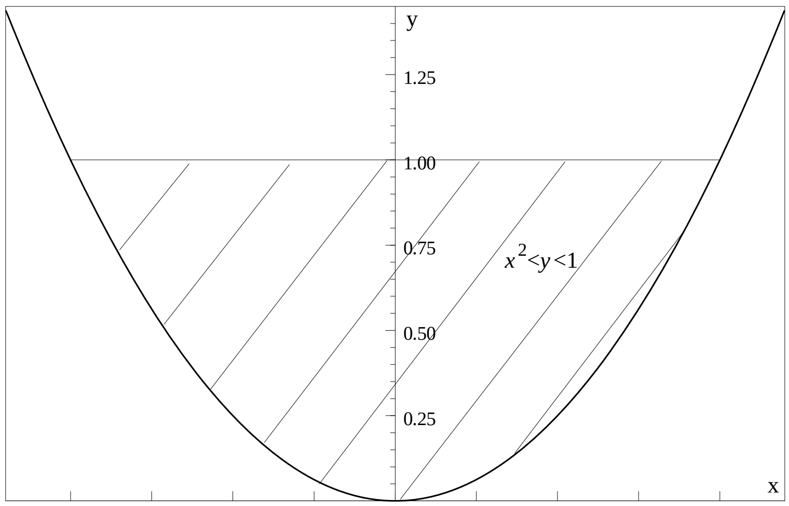

Figure 5.2: The Support of (X,Y)

(1) Using the property that

$$\int_0^1 \left(\int_{-\sqrt{y}}^{\sqrt{y}} cx^2 y dx \right) dy = 1,$$

we can solve for $c = \frac{21}{4}$.
(2)

$$P(X \ge Y) = \int_0^1 \left(\int_y^{\sqrt{y}} cx^2 y dx \right) dy = \frac{3}{20}.$$

Example 5 (5.5)**.** Suppose (X,Y) has a joint CDF $F_{XY}(x,y) = \frac{1}{16}xy(x+y)$ for $0 \le x \le 2$ and $0 \le y \le 2$. Find (1) the complete representation of the CDF $F_{XY}(x,y)$ in the entire xy-plane; (2) the joint PDF $f_{XY}(x,y)$; (3) $P(1 \le X \le 2, 1 \le Y \le 2)$.

Solution: (1) According to the properties of the joint CDF $F_{XY}(x,y)$ in Lemma 5.1, we obtain

$$F_{XY}(x,y) = \begin{cases} 0, & \text{if } x < 0 \text{ or } y < 0, \\ \frac{1}{16}xy(x+y), & \text{if } 0 \le x \le 2,\ 0 \le y \le 2, \\ \frac{1}{8}x(x+2), & \text{if } 0 \le x \le 2,\ y > 2, \\ \frac{1}{8}y(y+2), & \text{if } x > 2,\ 0 \le y \le 2. \end{cases}$$

(2) By partial differentiation, we have

$$\begin{aligned} f_{XY}(x,y) &= \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \\ &= \frac{1}{8}(x+y) \text{ for } 0 \le x \le 2, 0 \le y \le 2, \end{aligned}$$

Note that $f_{XY}(x,y) = 0$ for all (x,y) elsewhere.
(3)

$$\begin{aligned} P(1 &\le X \le 2, 1 \le Y \le 2) = \int_1^2 \int_1^2 \frac{1}{8}(x+y)dxdy \\ &= \frac{3}{8}. \end{aligned}$$

Before we conclude this section, we point out that the bivariate concepts in the continuous case can be generalized to the multivariate contexts. The joint CDF of n continuous random variables $\mathbf{X}^n = (X_1, \cdots, X_n)'$ is given by

$$F_{\mathbf{X}^n}(\mathbf{x}^n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}^n}(\mathbf{u}^n) d\mathbf{u}^n,$$

where $f_{\mathbf{X}^n}(\mathbf{x}^n)$ is the joint PDF of $\mathbf{X}^n$, $\mathbf{x}^n = (x_1, \cdots, x_n)$, and $\mathbf{u}^n = (u_1, \cdots, u_n)$. Partial differentiation yields

$$f_{\mathbf{X}^n}(\mathbf{x}^n) = \frac{\partial^n F_{\mathbf{X}^n}(\mathbf{x}^n)}{\partial x_1 \cdots \partial x_n}$$

wherever the partial derivative exists at point $\mathbf{x}^n$ in the $\mathbb{R}^n$ space.

5.2 Marginal Distributions

What information can be extracted from the joint PMF/PDF $f_{XY}(x,y)$? Intuitively, we can expect to extract the following information:

- Individual information of X, characterized by the PMF/PDF $f_X(x)$ of X;
- Individual information of Y, characterized by the PMF/PDF $f_Y(y)$ of Y;
- Predictive relationship between X and Y, characterized by suitable conditional distribution concepts.

5.2.1 The Discrete Case

We first investigate how to extract such information from the joint PMF/PDF $f_{XY}(x,y)$. We first consider the case of DRV's.

Definition 7 (5.7)**. [Discrete Marginal PMF's]:** Suppose X and Y have a joint discrete distribution with joint PMF $f_{XY}(x,y)$. Then the marginal PMF's of X and Y are defined as

$$\begin{aligned} f_X(x) &= P(X=x) = \sum_{y \in \Omega_Y} f_{XY}(x,y), \text{ where } -\infty < x < \infty, \\ f_Y(y) &= P(Y=y) = \sum_{x \in \Omega_X} f_{XY}(x,y), \text{ where } -\infty < y < \infty. \end{aligned}$$

To understand the definition of marginal PMF's, we define the event $\{X = x\}$ in a bivariate context. Observe that the event

$$\begin{aligned}
\{X = x\} &= \{X = x\} \cap \{Y \in \Omega_Y\} \\
&= \{X = x\} \bigcap \left[\bigcup_{y \in \Omega_Y} \{Y = y\} \right] \\
&= \bigcup_{y \in \Omega_Y} \left[\{X = x\} \cap \{Y = y\} \right],
\end{aligned}$$

where the last equality follows by the distributive laws in Theorem 2.1. This is analogous to the rule of total probability that $P(A) = \sum_{i=1}^{\infty} P(A \cap A_i)$ for a sequence of mutually exclusive and collectively exhaustive events $\{A_1, A_2, \cdots\}$ and any event A. It follows that

$$\begin{aligned}
P(X &= x) = P(\{X = x\} \cap \{Y \in \Omega_Y\}) \\
&= \sum_{y \in \Omega_Y} f_{XY}(x, y).
\end{aligned}$$

Intuitively, the marginal PMF of X is the probability that X takes a given value x regardless of the values taken by Y. By taking into account of all possibilities of Y, we get rid of all the information of Y, and only the information of X remains. We call $f_X(x)$ the marginal PMF of X to emphasize the fact that it is the PMF of X but in the bivariate context that gives the joint distribution of the vector (X, Y). Technically, the adjective "marginal" is redundant.

The marginal PMF's have the following properties.

Lemma 4 (5.5)**. [Properties of $f_X(x)$ and $f_Y(y)$]:**
(1) $f_X(x) \geq 0$ for all $x \in (-\infty, \infty)$;
(2) $\sum_{x \in \Omega_X} f_X(x) = 1$, where Ω_X is the support of X.
Similar results for $f_Y(y)$ also hold.

Example 6 (5.6)**.** Suppose X and Y have the joint PMF

$$f_{XY}(x, y) = \frac{1}{5}|x + y|, \qquad x = -1, 0, 1, \quad y = 0, 1.$$

Find (1) $f_X(x)$; and (2) $f_Y(y)$.

Solution: (1) For $x = -1$, the event $\{X = -1\}$ contains two basic outcomes: $\{X = -1, Y = 0\}$ and $\{X = -1, Y = 1\}$. These basic outcomes are mutually exclusive. Thus, it follows that

$$\begin{aligned}
f_X(-1) &= P(X = -1) \\
&= f_{XY}(-1, 0) + f_{XY}(-1, 1) \\
&= \frac{1}{5}.
\end{aligned}$$

Similarly,

$$\begin{aligned} f_X(0) &= f_{XY}(0,0) + f_{XY}(0,1) = \frac{1}{5}, \\ f_X(1) &= f_{XY}(1,0) + f_{XY}(1,1) = \frac{3}{5}. \end{aligned}$$

Then the marginal PMF of X is given by

$$f_X(x) = \begin{cases} \frac{1}{5}, & \text{if } x = -1, \\ \frac{1}{5}, & \text{if } x = 0, \\ \frac{3}{5}, & \text{if } x = 1. \end{cases}$$

(2) Similarly, we can obtain the PMF of Y

$$f_Y(y) = \begin{cases} \frac{2}{5}, & \text{if } y = 0, \\ \frac{3}{5}, & \text{if } y = 1. \end{cases}$$

We can compactly represent the joint PMF $f_{XY}(x, y)$ and the marginal PMF's of (X, Y) as a matrix form, where the marginal PMFs by summing rows and columns respectively:

Y/X	-1	0	1
0	$\frac{1}{5}$	0	$\frac{1}{5}$
1	0	$\frac{1}{5}$	$\frac{2}{5}$

It is possible to have the same marginal PMF's, but different joint PMF's. There are many joint distributions that have the same marginal distributions. The joint PMF not only tells the marginal information but also the relationship between X and Y that is not available in the marginals. If the relationship between X and Y is changed, a different joint distribution will arise, although the marginal distributions may remain the same.

We consider a simple example.

Example 7 (5.7)**.** Suppose both X and Y are binary random variables, taking only value 1 or 0. Consider the following two joint PMF's for (X, Y) :

Case I:

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

Case II:

$$f_{XY}(x, y) = \begin{cases} p^{x+y}(1-p)^{2-(x+y)}, & \text{if } x = 1, 0;\ y = 1, 0, \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, these are two different joint distributions of (X, Y). However, it can be shown that in both cases, we have $X \sim \text{Bernoulli}(p), Y \sim \text{Bernoulli}(p)$.

5.2.2 The Continuous Case

Suppose the bivariate continuous random vector (X, Y) have a joint PDF $f_{XY}(x, y)$. We first consider how to obtain the CDF of X. Observe that the event $\{X \leq x\} = \{X \leq x\} \cap \{-\infty < Y < \infty\}$, we have

$$\begin{aligned} F_X(x) &\equiv P(X \leq x) \\ &= P(X \leq x, -\infty < Y < \infty) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(u, y) du dy \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{XY}(u, y) dy \right] du \\ &= \int_{-\infty}^{x} f_X(u) du. \end{aligned}$$

Differentiating both sides of the above equation, we obtain

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy.$$

Thus by integrating out y, we can obtain the marginal PDF $f_X(x)$ from the joint PDF $f_{XY}(x, y)$.

Definition 8 (5.8)**. [Marginal PDF's]:** Suppose X and Y have a joint continuous distribution with joint PDF $f_{XY}(x, y)$. Then the marginal PDF's of X and Y are defined as follows:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y) dy, \quad -\infty < x < \infty, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y) dx, \quad -\infty < y < \infty. \end{aligned}$$

Unlike such marginal concepts as marginal utility and marginal productivity, the marginal PDF's are obtained by integrating the other variable, rather than by taking derivatives. Intuitively, by integrating out y in the joint PDF $f_{XY}(x, y)$, what remains is the information of Y.

The marginal PDF's have the following properties.

Lemma 5 (5.6)**. [Properties of $f_X(x)$ and $f_Y(y)$]:**
(1) $f_X(x) \geq 0$ for all $x \in (-\infty, \infty)$;
(2) $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
Similar results hold for $f_Y(y)$.

Example 8 (5.8)**.** Suppose (X, Y) have a joint PDF $f_{XY}(x, y) = 4xy$ if $0 < x < 1$ and $0 < y < 1$. Find the marginal PDF's $f_X(x)$ and $f_Y(y)$.

Solution: The support $\Omega_{XY} = \{(x, y) : 0 < x < 1, 0 < y < 1\}$ is a unit rectangular area. For $0 < x < 1$, the marginal PDF of X is

$$f_X(x) = \int_0^1 4xydy = 2x.$$

For $x \le 0$ or $x \ge 1$, we have $f_X(x) = 0$.

For $0 < y < 1$, the marginal PDF of Y is

$$f_Y(y) = \int_0^1 4xydx = 2y.$$

For $y \le 0$ or $y \ge 1$, we have $f_Y(y) = 0$.

Example 9 (5.9)**.** Suppose $f_{XY}(x, y) = cy^2$ for $x^2 < y < 1$. Find (1) $f_X(x)$; (2) $f_Y(y)$.

Solution: The support of (X, Y) is the same as the support of (X, Y) in Example 5.4, which was given in Figure 5.2, from which we can obtain the support of X as $\Omega_X = \{x \in \mathbb{R} : -1 < x < 1\}$ and the support of Y as $\Omega_Y = \{y \in \mathbb{R} : 0 < y < 1\}$.

We first determine the value of c by using the property that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)dxdy = 1.$$

Given the support of (X, Y) as depicted in Figure 5.2, we have

$$\begin{aligned} c\int_{-1}^{1} \left(\int_{x^2}^{1} y^2 dy\right) dx &= c\int_{-1}^{1} \frac{1}{3}\left(1 - x^6\right) dx \\ &= \frac{2c}{3}\int_0^1 \left(1 - x^6\right) dx \\ &= \frac{4c}{7} \\ &= 1. \end{aligned}$$

It follows that $c = \frac{7}{4}$.

(1) For $-1 < x < 1$, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy \\ &= \int_{x^2}^{1} cy^2 dy \\ &= c\frac{1}{3}y^3|_{x^2}^{1} \\ &= \frac{7}{12}(1 - x^6). \end{aligned}$$

Then

$$f_X(x) = \begin{cases} \frac{7}{12}(1 - x^6), & -1 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

(2) For $0 < y < 1$, we have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dx \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} cy^2 dx \\ &= cy^2 \cdot 2\sqrt{y} \\ &= \frac{7}{2} y^{\frac{5}{2}}. \end{aligned}$$

Then the marginal PDF if Y is

$$f_Y(y) = \begin{cases} \frac{7}{2} y^{\frac{5}{2}}, & 0 \le y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

When there are more than two random variables, we can define not only the marginal distributions of individual random variables but also the joint distributions of any subset of random variables. For example, in the case of n discrete random variables $X_1, \cdots, X_n$, we have the marginal PMF of X_1

$$f_{X_1}(x_1) = \sum_{x_2 \in \Omega_2} \cdots \sum_{x_n \in \Omega_n} f_{\mathbf{X}^n}(\mathbf{x}^n), \qquad -\infty < x_1 < \infty,$$

and the joint PMF of the subset (X_1, X_2, X_3)

$$f_{X_1 X_2 X_3}(x_1, x_2, x_3) = \sum_{x_4 \in \Omega_4} \cdots \sum_{x_n \in \Omega_n} f_{\mathbf{X}^n}(\mathbf{x}^n), \qquad -\infty < x_1, x_2, x_3 < \infty,$$

where $\mathbf{X}^n = (X_1, \cdots, X_n), \mathbf{x}^n = (x_1, \cdots, x_n)$, and Ω_i is the support of $X_i, i = 1, \cdots, n$.

5.3 Conditional Distributions

Oftentimes when two random variables, (X, Y), are observed, the values of the two variables are related. Knowledge about the value of X gives us some information about the value of Y even if it does not tell us the value of Y exactly.

Question: How to characterize the predictive relationship between X and Y?

We can use the concept of conditional distribution of Y given X. Again, we consider the cases of discrete and continuous random variables respectively.

5.3.1 The Discrete Case

We first consider the case of DRV's.

Definition 9 (5.9)**. [Conditional PMF's]:** Let X and Y have a joint discrete distribution with joint PMF $f_{XY}(x, y)$ and marginal PMF's $f_X(x)$ and $f_Y(y)$. Then the conditional PMF of Y given $X = x$ is defined as

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= \frac{f_{XY}(x, y)}{f_X(x)} \end{aligned}$$

provided $f_X(x) > 0$.

Similarly, the conditional PMF of X given $Y = y$ is defined as

$$\begin{aligned} f_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= \frac{f_{XY}(x, y)}{f_Y(y)} \end{aligned}$$

provided $f_Y(y) > 0$.

Intuitively, the conditional PMF $f_{Y|X}(y|x)$ is the probability that the random variable Y will take any arbitrary value y given that the value x of the random variable X has been observed. Recall the conditional probability formula $P(A|B) = P(A \cap B)/P(B)$ for any two events in Chapter 2. Define events $A = \{X = x\}$ and $B = \{Y = y\}$. Then

$$f_{Y|X}(y|x) = P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Example 10 (5.10)**.** Suppose two random variables X and Y are independent, and $X \sim$Poisson(λ_1), $Y \sim$Poisson(λ_2). We can verify that $X + Y \sim$Poisson($\lambda_1 + \lambda_2$). Find the conditional distribution of X given $X + Y = n$.

Solution: By the definition of conditional PMF and independence between X and Y, we have

$$\begin{aligned} P(X = k|X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{\frac{\lambda_1^k}{k!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^n}{n!} e^{-(\lambda_1+\lambda_2)}} \\ &= \frac{n!}{k!(n-k)!} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{(n-k)} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(1 - \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{(n-k)}, \qquad k = 0, 1, \cdots, n. \end{aligned}$$

Therefore, given $X + Y = n$, X follows a binomial distribution $B(n, p)$ with $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

Questions: What happens to the conditional PMF $f_{Y|X}(y|x)$ if $f_X(x) = 0$?

In this case, $f_{Y|X}(y|x)$, is not well-defined, because it does not make any sense in practice to condition on something that is unlikely to occur.

It is important to emphasize that given any x with $f_X(x) > 0$, the conditional PMF $f_{Y|X}(y|x)$ is a PMF of Y. That is, given each x with $f_X(x) > 0$, we have the following:

(1) $f_{Y|X}(y|x) \geq 0$ for all $y \in (-\infty, \infty)$;

(2) $\sum_{y \in \Omega_{Y|X}(x)} f_{Y|X}(y|x) = 1$, where $\Omega_{Y|X}(x) = \{y \in \Omega_Y : f_{Y|X}(y|x) > 0\}$ is the set of all possible values of $y \in \Omega_Y$ with $f_{Y|X}(y|x) > 0$.

Different values of x can be associated with different conditional PMF's for Y. For example, X can be a state variable taking two possible values: 0 and 1. When $X = 0$ (representing a bear market), the distribution for the stock return Y may have a large dispersion; when $X = 1$ (representing a bull market), the distribution for the stock return Y may have a small dispersion.

With the definition of conditional PMF, we have the following multiplication rules:

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

where the first equality always holds provided $f_X(x) > 0$ and the second equality holds provided $f_Y(y) > 0$. This multiplication rule suggests that joint PMFs can be equivalently characterized by specifying the conditional PMF $f_{Y|X}(y|x)$ and the marginal PMF $f_X(x)$.

Since the conditional distribution of Y given $X = x$ is possibly a different probability distribution for each value of x, we have a family of probability distributions for Y, one for each x. When we wish to describe this entire family, we will use the phrase "the distribution of $Y|X$." If, for example, X is a positive integer-valued random variable and the conditional distribution of Y given $X = x$ is Binomial(x, p), then we might say the distribution of $Y|X$ is Binomial(X, p) or write $Y|X \sim$ Binomial(X, p). Whenever we use the symbol $Y|X$ or have a random variable X as the parameter of a probability distribution of Y, we are describing the family of conditional probability distributions.

The conditional PMF $f_{Y|X}(y|x)$ is very useful because it tells how the information of X can be used to predict the probability of Y. It should be emphasized that the conditional PMF is a predictive relationship. It is not the causal relationship from X to Y. For example, it is possible that both X and Y are caused by an unobservable variable Z. In this case, X and Y are generally related to each other, and one can use the information of X to predict the probability distribution of Y, although X does not cause Y.

5.3.2 The Continuous Case

Next, we consider the case of CRV's. Suppose we have observed a value x for random variable X. It is desired to specify probabilities for $Y \in A$ given $X = x$, for various sets A that are of interest.

However, if X and Y are CRV's, then $P(X = x) = 0$ for every value of x. Therefore, to compute a conditional probability such as $P(Y > 5|X = 10)$, the definition $P(B|A) =$

$P(A \cap B)/P(A)$, where $A = \{X = 10\}$, cannot be used, since the denominator $P(X = 10) = 0$. In the other words, the concept of conditional PMF cannot be used for the continuous case. Yet, in reality, a value of $X = 10$ may be well observed. If, to the limit of our measurement, we see $X = 10$, this knowledge might give us information about Y. It turns out that when X and Y are continuous, the appropriate way to define a conditional probability distribution for Y given $X = x$ is analogous to the discrete case but with PDF's replacing PMF's. In other words, the concept of the conditional PMF for the discrete case should be extended when we move to the continuous case.

Definition 10 (5.10)**. [Conditional PDF's]**: Let X and Y have a joint continuous distribution with joint PDF $f_{XY}(x, y)$ and marginal PDF's $f_X(x)$ and $f_Y(y)$. Then the conditional PDF of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

if $f_X(x) > 0$.

Similarly, the conditional PDF of X given $Y = y$ can be defined as

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

if $f_Y(y) > 0$.

It should be noted that for DRV's (X, Y), the conditional PMF $f_{Y|X}(y|x)$ is derived as a conditional probability for the event $\{Y = y\}$ given the event $\{X = x\}$, whereas for CRV's (X, Y), the conditional PDF $f_{Y|X}(y|x)$ is defined as the ratio of the joint PDF $f_{XY}(x, y)$ to the marginal PDF $f_X(x)$.

Example 11 (5.11)**.** Suppose two dimensional random variable (X, Y) follows a uniform distribution on $\{(x, y) : x^2 + y^2 \leq 1\}$. Find (1) the conditional probability of Y given $X = x$; (2) $P(Y > 0|X = 0)$.

Solution: The joint PDF of (X, Y) is

$$f_{XY}(x, y) = \begin{cases} \frac{1}{\pi}, & x^2 + y^2 \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

and the support $\Omega_{XY} = \{(x, y) : x^2 + y^2 \leq 1\}$ is a unit circle centered at the origin $(0, 0)$ in the xy-plane. For $-1 \leq x \leq 1$, we have:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)dy \\ &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy \\ &= \frac{2}{\pi}\sqrt{1 - x^2} \end{aligned}$$

Then, for $-1 < x < 1$, the conditional PDF of Y given $X = x$ is:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \begin{cases} \frac{1}{2\sqrt{1-x^2}}, & \text{if } -\sqrt{1-x^2} \le y \le \sqrt{1-x^2}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

That is, for any given $x \in (-1, 1)$, the conditional PDF of Y given $X = x$ follows a uniform distribution on the interval $[-\sqrt{1-x^2}, \sqrt{1-x^2}]$.

(2) For $x = 0$, we have

$$f_{Y|X}(y|0) = \begin{cases} \frac{1}{2}, & \text{if } -1 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} P(Y &> 0|X = 0) = \int_0^\infty f_{Y|X}(y|0)dx \\ &= \int_0^1 \frac{1}{2}dy \\ &= \frac{1}{2}. \end{aligned}$$

We now examine the properties and provide an interpretation for the conditional PDF $f_{Y|X}(y|x)$.

Lemma 6 (5.7)**. [Properties of Conditional PDF's]:** For any given $x \in \mathbb{R}$ with $f_X(x) > 0$, $f_{Y|X}(y|x)$ is a PDF of Y. That is,

(1) $f_{Y|X}(y|x) \ge 0$ for all $y \in (-\infty, \infty)$;

(2) $\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = 1$.

The same properties hold for $f_{X|Y}(x|y)$.

Proof: Suppose $f_X(x) > 0$. Then $f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x)$ is well-defined and is nonnegative for all $y \in (-\infty, \infty)$. Moreover,

$$\begin{aligned} \int_{-\infty}^{\infty} f_{Y|X}(y|x)dy &= \int_{-\infty}^{\infty} \frac{f_{XY}(x,y)}{f_X(x)}dy \\ &= \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\ &= \frac{1}{f_X(x)} f_X(x) \\ &= 1. \end{aligned}$$

Thus, given any value of x with $f_X(x) > 0$, $f_{Y|X}(y|x)$ is a PDF of Y. Different values of x can be associated with different distributions for Y. This implies that one can then use the information of X to predict the distribution of Y. Figure 5.3 plots a continuous family of conditional PDF's $f_{Y|X}(y|x)$ for $x \in (-1, 1)$ of Example 5.11:

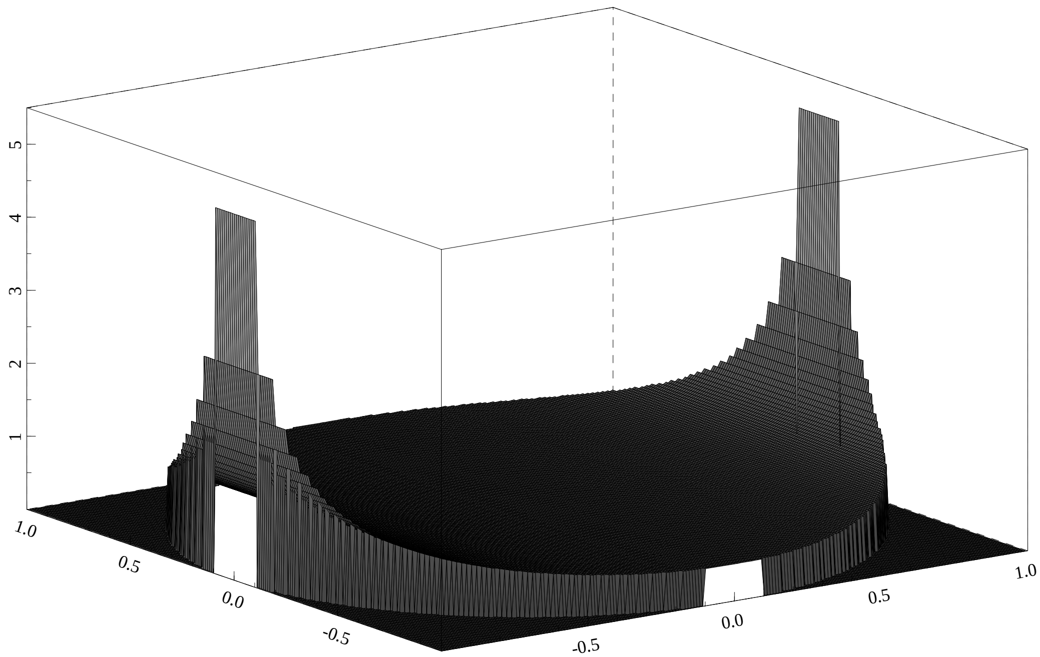

Figure 5.3: Conditional PDF's $f_{Y|X}(y|x)$ for $x \in (-1, 1)$ in Example 5.11

Question: What is the interpretation for the conditional PDF $f_{Y|X}(y|x)$?

For any given pair of (x, y) with $f_{XY}(x, y) > 0$, consider two events $A(x) = \{x - \frac{\epsilon}{2} < X \leq x + \frac{\epsilon}{2}\}$ and $B(y) = \{y - \frac{\epsilon}{2} < Y \leq y + \frac{\epsilon}{2}\}$, where ϵ is a small positive constant. Then by the mean value theorem, we can obtain

$$
\begin{aligned}
P[A(x) \cap B(y))] &= \int_{y-\epsilon/2}^{y+\epsilon/2} \int_{x-\epsilon/2}^{x+\epsilon/2} f_{XY}(u, v) du dv \\
&\approx f_{XY}(x, y)\epsilon^2, \\
P[A(x)] &= \int_{x-\epsilon/2}^{x+\epsilon/2} f_X(u) du \\
&\approx f_X(x)\epsilon.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
P[B(y)|A(x)] &= \frac{P[A(x) \cap B(y)]}{P[A(x)]} \\
&\approx \frac{f_{XY}(x, y)\epsilon^2}{f_X(x)\epsilon} = f_{Y|X}(y|x)\epsilon.
\end{aligned}
$$

This implies that the conditional PDF $f_{Y|X}(y|x)$ is proportional to

$$
P[B(y)|A(x)] = P\left(y - \frac{\epsilon}{2} < Y \leq y + \frac{\epsilon}{2} \Big| x - \frac{\epsilon}{2} < X \leq x + \frac{\epsilon}{2}\right),
$$

the conditional probability that Y takes values in the small interval $(y-\frac{\epsilon}{2}, y+\frac{\epsilon}{2}]$ given that X takes values in the small interval $(x-\frac{\epsilon}{2}, x+\frac{\epsilon}{2}]$. In other words, $f_{Y|X}(y|x)$ is proportional to the probability that Y takes values near y given that X has taken values near x.

Like in the discrete case, we also have the multiplication rules:

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

where the first equality holds whenever $f_X(x) > 0$, and the second equality holds whenever $f_Y(y) > 0$.

When we are dealing with more than two random variables, we can consider various kinds of conditional distributions. For example, we can define

$$f_{X_i|\mathbf{X}^{i-1}}(x_i|\mathbf{x}^{i-1}) = \frac{f_{X^i}(\mathbf{x}^i)}{f_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1})}$$

if $f_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}) > 0$, where $\mathbf{X}^{i-1} = (X_1, \cdots, X_{i-1})'$ and $\mathbf{x}^{i-1} = (x_1, \cdots, x_{i-1})'$. We can also define

$$f_{(X_1,X_2)|(X_3,X_4)}(x_1,x_2|x_3,x_4) = \frac{f_{X_1X_2X_3X_4}(x_1,x_2,x_3,x_4)}{f_{X_3X_4}(x_3,x_4)}$$

if $f_{X_3X_4}(x_3,x_4) > 0$.

5.4 Independence

The marginal distributions of X and Y, described by the marginal PMF's/PDF's $f_X(x)$ and $f_Y(y)$, do not completely describe the joint distribution of X and Y. Indeed, there are many different joint distributions that have the same marginal distributions. Thus, it is hopeless to try to determine the joint PMF's/PDF's, $f_{XY}(x,y)$, from knowledge of only the marginal PMF's/PDF's, $f_X(x)$ and $f_Y(y)$. However, there is a special but important case in which we can use the marginal distributions to determine the joint distributions. This occurs when knowledge about X gives us no information about Y. This is the case of so-called independence between X and Y. It is a characterization of no association between X and Y.

Definition 11 (5.11)**. [Independence]**: Two random variables X and Y are independent if

$$F_{XY}(x,y) = F_X(x)F_Y(y) \text{ for all } -\infty < x, y < \infty.$$

where $F_{XY}(\cdot), F_X(\cdot), F_Y(\cdot)$ are the joint and marginal CDF's.

The above definition of independence is equivalent to the definition that two random variables X and Y defined on the same sample space are independent if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all subsets of $A \in \Omega_X$ and $B \in \Omega_Y$.

The definition of independence allows that X and Y are discrete or continuous. In the discrete case, we can use the joint and marginal PMF's to characterize independence.

Theorem 2 (5.8)**.** Two discrete random variables (X, Y) are independent if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \text{ for all pairs of } (x, y) \in \mathbb{R}^2,$$

where $f_{XY}(x, y), f_X(x), f_Y(y)$ are the joint and marginal PMF's.

Proof: (1) [*Necessity*] By the definition of independence, we have

$$F_{XY}(x, y) = F_X(x) F_Y(y) \text{ for all } -\infty < x, y < \infty.$$

Without loss of generality, we assume that the possible values that X can take are arranged in an increasing order: $x_1 < x_2 < x_3 < \cdots$ and the possible values that Y can take are also arranged in an increasing order: $y_1 < y_2 < y_3 < \cdots$.
For $i > 1$, taking a difference of the above equation with respect to x, i.e., from x_{i-1} to x_i, we obtain

$$\begin{aligned} \Delta_X F_{XY}(x_i, y) &= F_{XY}(x_i, y) - F_{XY}(x_{i-1}, y) \\ &= [F_X(x_i) - F_X(x_{i-1})] F_Y(y). \end{aligned}$$

If in addition $j > 1$, we can further take a difference of the above equation with respect to y and obtain

$$\Delta_Y \Delta_X F_{XY}(x_i, y_j) = [F_X(x_i) - F_X(x_{i-1})][F_Y(y_j) - F_Y(y_{j-1})].$$

This yields the following relationship among the joint and marginal PMF's:

$$f_{XY}(x_i, y_j) = f_X(x_i) f_Y(y_j), \qquad i, j > 1.$$

We can also obtain the same relationship when $i = 1$ or $j = 1$. (Please verify it!)
(2) [*Sufficiency*] Suppose now

$$f_{XY}(x_i, y_j) = f_X(x_i) f_Y(y_j) \text{ for all } i, j = 1, 2, \cdots.$$

Assuming $x_i \le x < x_{i+1}, y_j \le y < y_{j+1}$, we have

$$\begin{aligned} F_{XY}(x, y) &= P(X \le x, Y \le y) \\ &= \sum_{i'=1}^{i} \sum_{j'=1}^{j} f_{XY}(x_{i'}, y_{j'}) \\ &= \sum_{i'=1}^{i} f_X(x_{i'}) \sum_{j'=1}^{j} f_Y(y_{j'}) \\ &= F_X(x) F_Y(y). \end{aligned}$$

Since i, j are arbitrary, so are x and y. Thus, $F_{XY}(x, y) = F_X(x) F_Y(y)$ for all pairs of (x, y) on the xy-plane. The proof is completed.

Next, we show that for the continuous case, we can use the joint and marginal PDF's to characterize independence.

Theorem 3 (5.9)**.** Suppose X and Y are two CRV's. Then X and Y are independent if and only if

$$f_{XY}(x,y) = f_X(x)f_Y(y) \text{ for all } (x,y) \in \mathbb{R}^2,$$

where $f_{XY}(x,y)$, $f_X(x)$, and $f_Y(y)$ are the joint and marginal PDF's.

Proof: (1) [*Necessity*] We first show if (X,Y) are independent, then $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all pairs of $(x,y) \in \mathbb{R}^2$.

Suppose (X,Y) are independent. Then by definition, we have

$$F_{XY}(x,y) = F_X(x)F_Y(y) \text{ for all } x, y.$$

Differentiating the both sides of this equation with respect to x and y respectively, we obtain

$$\begin{aligned} \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} &= \frac{\partial^2 F_X(x)F_Y(y)}{\partial x \partial y} \\ &= \frac{\partial F_X(x)}{\partial x}\frac{\partial F_Y(y)}{\partial y}. \end{aligned}$$

This implies

$$f(x,y) = f_X(x)f_Y(y) \text{ for all } (x,y) \in \mathbb{R}^2.$$

(2) [*Sufficiency*] Next, we show that if $f_{XY}(x,y) = f_X(x)f_Y(y)$, then (X,Y) are independent. Suppose $f_{XY}(u,v) = f_X(u)f_Y(v)$ for all $(u,v) \in \mathbb{R}^2$. Then by integration, we have

$$\begin{aligned} \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)dudv &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)dudv \\ &= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv. \end{aligned}$$

That is,

$$F_{XY}(x,y) = F_X(x)F_Y(y), \text{ for all } -\infty < x, y < \infty,$$

which implies that X and Y are independent. The proof is completed.

Example 12 (5.12)**.** Suppose $f_{XY}(x,y) = 4xy$ if $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Are X and Y independent?

Solution: In Example 5.8, we have obtained $f_X(x) = 2x$ for $0 \leq x \leq 1$ and $f_Y(y) = 2y$ for $0 \leq y \leq 1$. Then

$$f_X(x)f_Y(y) = 4xy = f_{XY}(x,y) \text{ for } 0 \leq x \leq, 0 \leq y \leq 1.$$

Also, we have $f_{XY}(x,y) = f_X(x)f_Y(y) = 0$ for all (x,y) outside the rectangular area defined by $0 \leq x \leq 1$ and $0 \leq y \leq 1$. It follows that $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all (x,y) in the xy plane. Therefore, X and Y are independent.

Example 13 (5.13)**.** Suppose $f_{XY}(x,y) = 8xy$, $0 \leq x \leq y \leq 1$. Are X and Y independent?

Solution: The support $\Omega_{XY} = \{(x, y) : 0 \le x \le y \le 1\}$ is an upper triangular area, as shown in Figure 5.4 below.

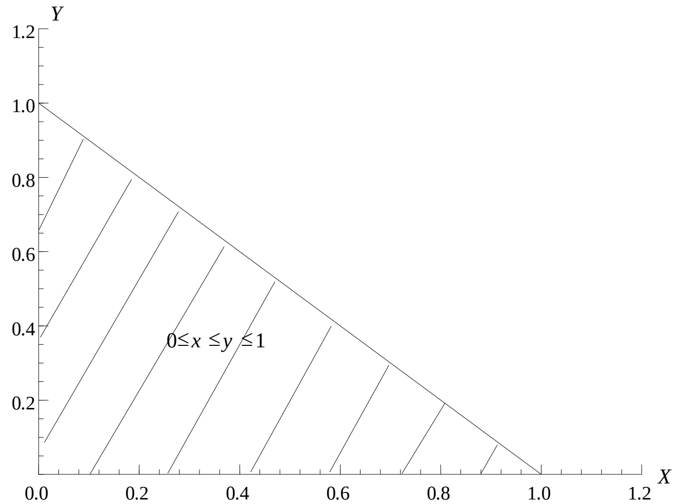

Figure 5.4 The Support $\Omega_{XY} = \{(x, y) : 0 \le x \le y \le 1\}$

By integrating out y and x respectively, we can obtain the marginal PDF's:

$$f_X(x) = 4x(1 - x^2), \qquad 0 \le x \le 1,$$

and

$$f_Y(y) = 4y^3, \qquad 0 \le y \le 1.$$

Thus, X and Y are not independent, since $f_X(x)f_Y(y) \ne f_{XY}(x, y)$ when $0 \le x \le y \le 1$.

We now explore some implications of independence.

Suppose X and Y have a joint PMF or PDF $f_{XY}(x, y)$ that can be decomposed as the production of two functions $h(x)g(y)$ for all (x, y) on the xy-plane, where $h(x)$ and $g(y)$ are not necessarily a PMF or PDF of X and Y. Then the theorem below shows that X and Y are independent.

Theorem 4 (5.10)**. [Factorization Theorem]** The two random variables X and Y are independent if and only if the joint PMF/PDF can be written as

$$f_{XY}(x, y) = g(x)h(y), \text{ for all } -\infty < x, y < \infty.$$

Proof: We shall show the continuous case only; the proof for the discrete case is very similar.

(1) If X and Y are independent, then

$$f_{XY}(x, y) = f_X(x)f_Y(y) \; = g(x)h(y) \text{ for all } -\infty < x, y < \infty,$$

where we set $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.

(2) Now suppose

$$f_{XY}(x, y) = g(x)h(y) \text{ for all } -\infty < x, y < \infty$$

for some functions $g(\cdot)$ and $h(\cdot)$. Then we have

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\
&= \int_{-\infty}^{\infty} g(x)h(y)dy \\
&= g(x)\int_{-\infty}^{\infty} h(y)dy, \\
f_Y(y) &= h(y)\int_{-\infty}^{\infty} g(x)dx.
\end{aligned}$$

It follows that

$$\begin{aligned}
f_X(x)f_Y(y) &= \left[g(x)\int_{-\infty}^{\infty} h(v)dv\right]\left[h(y)\int_{-\infty}^{\infty} g(u)du\right] \\
&= g(x)h(y)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)h(v)dudv \\
&= f_{XY}(x,y) \text{ for all } -\infty < x, y < \infty,
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(u,v)dudv &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u)h(v)dudv \\
&= 1.
\end{aligned}$$

This completes the proof.

The factorization theorem provides a very convenient way to check independence. The key is to check whether the joint PMF/PDF can be partitioned into a product of two separate functions of x and y respectively. We note that it is important to check whether the partition holds for all points (x, y) in the whole xy-plane, rather than on a subregion of the xy-plane only.

On the other hand, for CRV's (X, Y), it is possible that $f_{XY}(x, y) \neq f_X(x)f_Y(y)$ on a set A of (x, y) values for which $\int_A dxdy = 0$. In such cases, X and Y are still called independent. This is because two PDF's that differ only on a set like A still define the same probability distribution for (X, Y).

Next, we explore the implication of independence on the conditional probability distributions.

Theorem 5 (5.11)**.** Suppose two random variables X and Y are independent. Then for the conditional PMF/PDF

$$f_{Y|X}(y|x) = f_Y(y) \text{ for all } (x, y) \in \mathbb{R}^2$$

where $f_X(x) > 0$, and

$$f_{X|Y}(x|y) = f_X(x) \text{ for all } (x, y) \in \mathbb{R}^2$$

where $f_Y(y) > 0$.

Proof: When X and Y are independent, we have $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all (x,y) in the xy-plane. It follows that

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= f_Y(y) \text{ for all } (x,y) \end{aligned}$$

provided $f_X(x) > 0$.

Similarly, we have

$$f_{X|Y}(x|y) = f_X(x) \text{ for all } (x,y)$$

provided $f_Y(y) > 0$. The proof is completed.

This theorem implies that when X and Y are independent, the information of X has no predictive power for the probability distribution of Y, and vice versa.

Example 14 (5.14)**.** Suppose two random variables X and Y have a joint PDF

$$f_{XY}(x,y) = \begin{cases} e^{-y}, & 0 < x < y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

(1) Find $f_X(x)$ and $f_Y(y)$; (2) find $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$; (3) check if X and Y are independent.

Solution: The support $\Omega_{XY} = \{(x,y) : 0 < x < y < \infty\}$ is an upper triangular area as shown in Figure 5.5.

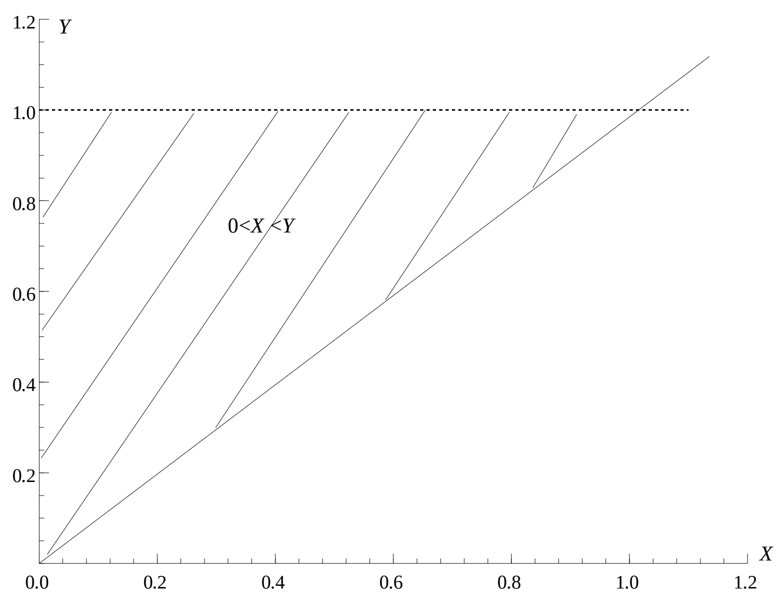

Figure 5.5: The Support $\Omega_{X,Y}$ of (X,Y) in Example 5.14

(1) The support of X is $0 < x < \infty$. By definition,

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy.$$

For $x \leq 0$, $f_{XY}(x, y) = 0$. So, $f_X(x) = 0$ for $x \leq 0$.
For $x > 0$,

$$\begin{aligned} f_X(x) &= \int_x^{\infty} e^{-y} dy \\ &= e^{-x}. \end{aligned}$$

Therefore,

$$f_X(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

It is important to note that no matter whether X and Y are independent, the marginal PDF $f_X(x)$ has nothing to do with y.

Next, we calculate $f_Y(y)$. The support of Y is $0 < y < \infty$. For $y \leq 0, f_{XY}(x, y) = 0$. So $f_Y(y) = 0$. For $y > 0$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y) dx \\ &= \int_0^y e^{-y} dx \\ &= ye^{-y}. \end{aligned}$$

It follows that

$$f_Y(y) = \begin{cases} ye^{-y}, & y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

(2) First, we calculate $f_{Y|X}(y|x)$. Since $f_X(x) = 0$ for $x \leq 0$, the conditional PDF $f_{Y|X}(y|x)$ is defined for any given $x > 0$. Given any $x > 0$, we have

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x, y)}{f_X(x)} \\ &= \frac{e^{-y}}{e^{-x}} \text{ for } 0 < x < y < \infty. \end{aligned}$$

It follows that

$$f_{Y|X}(y|x) = \begin{cases} e^{-(y-x)}, & y \in (x, \infty), \\ 0, & y \in (-\infty, x]. \end{cases}$$

This implies that, conditional on $X = x$, Y follows an exponential distribution with support $y \in (x, \infty)$.

Next, we calculate $f_{X|Y}(y|x)$. Since $f_Y(y) = 0$ for $y \leq 0$, the conditional PDF $f_{Y|X}(x|y)$ is defined for any given $y > 0$. By definition,

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x, y)}{f_Y(y)} \\ &= \frac{e^{-y}}{ye^{-y}} \\ &= \frac{1}{y} \text{ for all } x \in (0, y). \end{aligned}$$

Thus, for any given $y > 0$,

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

This implies that conditional on $Y = y$, X follows a uniform distribution on the interval $(0, y)$.
(3) Because $f_{Y|X}(y|x) \neq f_Y(y)$ for $0 < x < y < \infty$, X and Y are not independent.

If X and Y are independent, then we have $f_{XY}(x, y) = f_X(x)f_Y(y) > 0$ on the set $\Omega_X \times \Omega_Y = \{(x, y) : x \in \Omega_X \text{ and } y \in \Omega_Y\}$, where $\Omega_X = \{x : f_X(x) > 0\}$ and $\Omega_Y = \{y : f_Y(y) > 0\}$ are the supports of X and Y respectively. Memberships in the cross-product $\Omega_X \times \Omega_Y$ can be checked by considering the x and y values separately. If $f_{XY}(x, y)$ is a joint PDF or PMF and the set where $f_{XY}(x, y) > 0$ is not a cross-product, then the random variables X and Y are not independent. An example of not being a cross-product is the set $0 < x < y < \infty$.

We now extend the definition of independence when there are more than two random variables.

Definition 12 (5.12)**.** The random variables $X_1, \cdots, X_n$ are mutually independent if the joint CDF is equal to the product of their marginal CDF's, namely,

$$F_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} F_{X_i}(x_i) \text{ for all } -\infty < x_1, \cdots, x_n < \infty.$$

where $\mathbf{X}^n = (X_1, \cdots, X_n)'$ and $\mathbf{x}^n = (x_1, \cdots, x_n)'$.

For more than two random variables, it is possible that two random variables of any pair are independent, but all of them together are not independent. This is illustrated by the example below.

Example 15 (5.15)**.** Suppose random variables X_1, X_2, X_3 have the following joint PDF

$$f_{X_1X_2X_3}(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)e^{-x_3}, & 0 < x_1 < 1, 0 < x_2 < 1, x_3 > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

It can be verified that X_1, X_2 and X_3 are pairwise independent but they are not jointly independent. This is analogous to Example 2.55 in Chapter 2 where three events are not jointly independent but any two events are independent of each other.

When can we have the results that pairwise independence also implies joint independence? If this can be the case, then this will greatly simplify the verification of independence in the multivariate contexts. One example is the multivariate normal distribution, in which pairwise independence implies joint independence. See Sections 5.5 and 5.8 for discussion.

The concept of independence has important applications in economics and finance. As discussed in Example 2.50, Chapter 2, a geometric random walk is often used to model asset prices, where the increments in the log-price, which are approximately equal to the relative price changes, are independent across different time periods. Thus, one can test the random walk hypothesis by checking whether relative price changes of different time periods are mutually independent. In practice, the definition of independence and various characterizations provide powerful operational ways to verify independence using observed data. For example, one could construct estimators for the joint and marginal PDF, and then check whether the joint PDF is equal to the product of the marginal PDF's. See Hong and White (2005) and Robinson (1994) for more discussion.

5.5 Bivariate Transformation

In Chapter 3, we have discussed how to find the PMF/PDF $f_Y(y)$ of the univariate transformation $Y = g(X)$ when the PMF/PDF $f_X(x)$ of X is given. Suppose now we have a bivariate transformation

$$\begin{aligned} U &= g_1(X, Y), \\ V &= g_2(X, Y), \end{aligned}$$

where the joint PMF/PDF $f_{XY}(x, y)$ of (X, Y) is given. How can one find $f_{UV}(u, v)$, the joint PMF/PDF of (U, V)?

We first provide some examples to show why a bivariate transformation is useful in economics.

Example 16 (5.16)**.** Suppose X and Y are two common factors that drive the dynamics of the Eurodollar U and Japanese Yen V. Then both U and V are the functions of X and Y.

Example 17 (5.17)**.** Suppose both demand U and supply V of some commodity are the functions of the commodity price X and the consumer income Y.

Sometimes, our interest is in finding the probability distribution of $U = g_1(X, Y)$ given the joint probability distribution of (X, Y). In this case, we can first derive the joint PMF/PDF $f_{UV}(u, v)$ of $U = g_1(X, Y)$ and $V = X$ (say), and then integrate out v to obtain the PMF/PDF $f_U(u)$.

To find the joint PMF/PDF $f_{UV}(u, v)$ of (U, V), we consider the cases of DRV's and CRV's (X, Y) respectively. For the discrete case, the new random variables (U, V) are also discrete as well. Their joint PMF can be obtained via the formula

$$\begin{aligned} f_{UV}(u, v) &= P(U = u, V = v) \\ &= \sum_{(x,y) \in A(u,v)} f_{XY}(x, y), \end{aligned}$$

where

$$A(u, v) = \{(x, y) \in \Omega_{XY} : g_1(x, y) = u, g_2(x, y) = v\}.$$

This is the set of all possible values of (x, y) in the support of (X, Y) that satisfy the restrictions that $u = g_1(x, y)$ and $v = g_2(x, y)$.

Example 18 (5.18)**.** Let X and Y be independent Poisson random variables with parameters θ and λ, respectively. Thus the joint PMF of (X, Y) is:

$$f_{XY}(x, y) = \frac{\theta^x e^{-\theta}}{x!} \frac{\lambda^y e^{-\lambda}}{y!}, \qquad x = 0, 1, 2, \cdots, y = 0, 1, 2, \cdots$$

Find the PMF of $X + Y$.

Solution: Define $U = X + Y, V = Y$. The support of (U, V) is given by $\Omega_{UV} = \{(u, v) : u = v, v + 1, v + 2, \cdots; v = 0, 1, \cdots\}$. For any $(u, v) \in \Omega_{UV}$, the only (x, y) value that satisfies $x + y = u, y = v$ is the pair $(x, y) = (u - v, v)$. Thus, $A(u, v) = \{(u - v, v)\}$. Then the joint PMF of (U, V) is:

$$\begin{aligned} f_{UV}(u, v) &= f_{XY}(u - v, v) \\ &= f_X(u - v) f_V(v) \\ &= \frac{\theta^{u-v} e^{-\theta}}{(u - v)!} \frac{\lambda^v e^{-\lambda}}{v!} \quad u = v, v + 1, v + 2 \cdots; v = 0, 1, \cdots. \end{aligned}$$

Since for any given integer $u \geq 0$, $f_{UV}(u, v) > 0$ only for $v = 0, 1, \cdots, u$. It follows that the marginal PMF of U

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} f_{UV}(u, v) \\ &= \sum_{v=0}^{u} \frac{\theta^{u-v} e^{-\theta}}{(u - v)!} \frac{\lambda^v e^{-\lambda}}{v!} \\ &= e^{-(\theta+\lambda)} \sum_{v=0}^{u} \frac{\theta^{u-v}}{(u - v)!} \frac{\lambda^v}{v!} \\ &= \frac{e^{-(\theta+\lambda)}}{u!} \sum_{v=0}^{u} \binom{u}{v} \lambda^v \theta^{u-v} \\ &= \frac{e^{-(\theta+\lambda)}}{u!} (\theta + \lambda)^u, \quad u = 0, 1, 2, \cdots. \end{aligned}$$

That is, $U = X + Y$ follows a Poisson$(\theta + \lambda)$ distribution. This example is related to Example 5.10.

Next, we now consider the continuous case. For this purpose, we first review the concepts of the Jacobian matrix and Jacobian.

Definition 13 (5.13)**. [Jacobian Matrix and Jacobian]:** Consider the bivariate transformation:

$$\begin{aligned} U &= g_1(X, Y), \\ V &= g_2(X, Y), \end{aligned}$$

where functions $g_1(\cdot, \cdot)$ and $g_2(\cdot, \cdot)$ are continuously differentiable with respect to (x, y). Then the 2×2 matrix

$$J_{UV}(x, y) = \begin{bmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_1(x,y)}{\partial y} \\ \\ \frac{\partial g_2(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial y} \end{bmatrix}$$

is called the Jacobian matrix of (U, V), and its determinant is called the Jacobian of (U, V).

Note that the Jacobian matrix $J_{UV}(x, y)$ is not necessarily a symmetric matrix.

Definition 14 (5.14)**. [Inverse Function]:** Suppose $\mathbb{A}$ and $\mathbb{B}$ are subsets of $\mathbb{R}^2$, and the functions $g_1 : \mathbb{A} \to \mathbb{B}$ and $g_2 : \mathbb{A} \to \mathbb{B}$ are continuously differentiable with the determinant of $J_{UV}(x, y)$ not zero for all $(x, y) \in \mathbb{A}$. Then the following functions exist:

$$\begin{aligned} X &= h_1(U, V), \\ Y &= h_2(U, V), \end{aligned}$$

where $h_1 : \mathbb{B} \to \mathbb{A}$ and $h_2 : \mathbb{A} \to \mathbb{B}$ are continuously differentiable on $\mathbb{B}$ and they satisfy the conditions

$$\begin{aligned} h_1[g_1(x, y), g_2(x, y)] &= x, \\ h_2[g_1(x, y), g_2(x, y)] &= y. \end{aligned}$$

The functions $\{h_1(\cdot), h_2(\cdot)\}$ are called the inverse functions of $\{g_1(\cdot), g_2(\cdot)\}$.

Intuitively, the inverse functions $h_1(U, V)$ and $h_2(U, V)$ can be obtained by representing (X, Y) in terms of (U, V) via solving the system of equations $U = g_1(X, Y)$ and $V = g_2(X, Y)$.

The following theorem shows that the Jacobian matrix of the inverse functions is equal to the inverse of the Jacobian matrix of the original functions.

Theorem 6 (5.12)**.** The Jacobian matrix of (X, Y)

$$\begin{aligned} J_{XY}(u, v) &= \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \\ &= J_{UV}(x, y)^{-1} \end{aligned}$$

where $x = h_1(u, v), y = h_2(u, v)$.

Proof: Recall the identities

$$\begin{aligned} h_1[g_1(x,y),g_2(x,y)] &= x, \\ h_2[g_1(x,y),g_2(x,y)] &= y, \end{aligned}$$

Put $u = g_1(x,y), v = g_2(x,y)$. Differentiating the first identity with respect to x, y respectively, we obtain the following results:

$$\begin{aligned} \frac{\partial h_1(u,v)}{\partial u}\frac{\partial g_1(x,y)}{\partial x} + \frac{\partial h_1(u,v)}{\partial v}\frac{\partial g_2(x,y)}{\partial x} &= 1, \\ \frac{\partial h_1(u,v)}{\partial u}\frac{\partial g_1(x,y)}{\partial y} + \frac{\partial h_1(u,v)}{\partial v}\frac{\partial g_2(x,y)}{\partial y} &= 0, \end{aligned}$$

Similarly, differentiating the second identity with respect to x, y respectively, we obtain the following results:

$$\begin{aligned} \frac{\partial h_2(u,v)}{\partial u}\frac{\partial g_1(x,y)}{\partial x} + \frac{\partial h_2(u,v)}{\partial v}\frac{\partial g_2(x,y)}{\partial x} &= 0, \\ \frac{\partial h_2(u,v)}{\partial u}\frac{\partial g_1(x,y)}{\partial y} + \frac{\partial h_2(u,v)}{\partial v}\frac{\partial g_2(x,y)}{\partial y} &= 1. \end{aligned}$$

Representing these four derivative equations in a matrix form, we obtain

$$\begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_1(x,y)}{\partial y} \\ \frac{\partial g_2(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

or

$$J_{XY}(u,v)J_{UV}(x,y) = I_2,$$

where I_2 is a 2×2 identity matrix. Because $J_{UV}(x,y)$ is nonsingular for all $(x,y) \in \mathbb{A}$, by multiplying the inverse of $J_{UV}(x,y)$ from the right hand side, we obtain

$$J_{XY}(u,v) = J_{UV}^{-1}(x,y).$$

This completes the proof.

Example 19 (5.19)**.** Two random variables $U, V : \mathbb{R}^2 \to \mathbb{R}^2$ are defined by $U = XY$ and $V = X$. Then $g : \mathbb{R}^2 \to \mathbb{R}^2$ is 1-1 and has an inverse function $h : \mathbb{R}^2 \to \mathbb{R}^2$. (1) Find the inverse function; (2) verify that $J_{XY}(u,v) = J_{UV}^{-1}(x,y)$, where $u = xy$ and $v = x$.

Solution: (1) Given

$$\begin{aligned} U &= XY = g_1(X,Y), \\ V &= X = g_2(X,Y), \end{aligned}$$

we have the inverse functions:

$$\begin{aligned} X &= h_1(U,V) = V, \\ Y &= h_2(U,V) = \frac{U}{V}. \end{aligned}$$

(2) By definition, the Jacobian matrix of the inverse function is given by

$$\begin{aligned} J_{XY}(u,v) &= \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{bmatrix}. \end{aligned}$$

Furthermore, the Jacobian matrix of the original functions $U = g_1(X,Y) = XY$, $V = g_2(X,Y) = X$ is given by

$$\begin{aligned} J_{UV}(x,y) &= \begin{bmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_1(x,y)}{\partial y} \\ \\ \frac{\partial g_2(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial y} \end{bmatrix} \\ &= \begin{bmatrix} y & x \\ 1 & 0 \end{bmatrix} \end{aligned}$$

The inverse

$$\begin{aligned} J_{UV}(x,y)^{-1} &= -\frac{1}{x}\begin{bmatrix} 0 & -x \\ -1 & y \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ \frac{1}{x} & -\frac{y}{x} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{bmatrix} \\ &= J_{XY}(u,v), \end{aligned}$$

where $u = xy$ and $v = x$.

Now, we consider the following general problem: suppose (X,Y) have a joint PDF $f_{XY}(x,y)$, and

$$\begin{aligned} U &= g_1(X,Y) \\ V &= g_2(X,Y). \end{aligned}$$

Then how can we find the joint PDF $f_{UV}(u,v)$ of (U,V)?

The following bivariate transformation theorem provides a powerful method to calculate the joint PDF $f_{UV}(u,v)$.

Theorem 7 (5.13)**. [Bivariate Transformation]**: Let (X,Y) be a bivariate continuous random vector with joint PDF $f_{XY}(x,y)$, and let $\Omega_{XY} = \{(x,y) \in \mathbb{R}^2 : f_{XY}(x,y) > 0\}$ be the support of (X,Y). Define

$$\begin{aligned} U &= g_1(X,Y), \\ V &= g_2(X,Y), \end{aligned}$$

where $g : \Omega_{XY} \to \mathbb{R}^2$ is a 1−1 and continuously differentiable function on Ω_{XY}, with $\det[J_{UV}(x,y)] \neq 0$ for all $(x,y) \in \Omega_{XY}$. Then the joint PDF of (U,V) is given by

$$f_{UV}(u,v) = f_{XY}(x,y) \left|\det\left[J_{XY}(u,v)\right]\right| \text{ for all } (u,v) \in \Omega_{UV},$$

where $x = h_1(u,v)$ and $y = h_2(u,v)$, and

$$\Omega_{UV} = \{(u,v) \in \mathbb{R}^2 : u = g_1(x,y), v = g_2(x,y) \text{ for all } (x,y) \in \Omega_{XY}\}$$

is the the support of (U,V).

Proof: For (u,v) in the support Ω_{UV} of (U,V), we have

$$\begin{aligned} F_{UV}(u,v) &= P(U \leq u, V \leq v) \\ &= P\left[g_1(X,Y) \leq u, g_2(X,Y) \leq v\right] \\ &= \int\int_{\mathbb{A}(u,v)} f_{XY}(x',y')dx'dy', \end{aligned}$$

where the double integration is taken over the set $\mathbb{A}(u,v) = \{(x,y) \in \mathbb{R}^2 : g_1(x,y) \leq u$ and $g_2(x,y) \leq v\}$. Making the transformation $s = g_1(x',y')$ and $t = g_2(x',y')$ and applying from calculus on the change of variable formula for double integrals, we obtain

$$F_{UV}(u,v) = \int_{-\infty}^{u}\int_{-\infty}^{v} \frac{1}{|\det[J_{UV}(x',y')]|} f_{XY}(x',y')dsdt,$$

where (x',y') satisfies the restrictions that $g_1(x',y') = s$ and $g_2(x',y') = t$. By taking the partial derivatives of both sides with respect to (u,v), we can obtain

$$\begin{aligned} f_{UV}(u,v) &= \frac{\partial^2 F_{UV}(u,v)}{\partial u \partial v} \\ &= f_{XY}(x,y)\frac{1}{|\det[J_{UV}(x,y)]|}, \end{aligned}$$

where $x = h_1(u,v), y = h_2(u,v)$. Since $J_{UV}(x,y) = J_{XY}^{-1}(u,v)$, we also have

$$f_{UV}(u,v) = f_{XY}(x,y) \left|\det[J_{XY}(u,v)]\right|.$$

This completes the proof.

As discussed in Chapter 3, for the univariate transformation $Y = g(X)$, where $g(\cdot)$ is a continuously differentiable monotonic function, the PDF

$$f_Y(y) = f_X(x)|h'(y)| \; = f_X(x)|g'(x)|^{-1},$$

where $x = h(y)$ is the inverse function of $y = g(x)$. See Theorem 3.12 in Chapter 3 for details. Thus, we can view the bivariate transformation theorem as a generalization of the univariate transformation theorem. In fact, we can also derive a similar multivariate transformation involving more than two random variables.

It is difficult to overemphasize the importance of the bivariate transformation theorem. One can use it to efficiently derive results, obtain univariate distributions, and, of course, determine bivariate distributions. We note that the 1–1 mapping is a crucial condition in applying the bivariate transformation. This is similar to the univariate transformation where it is crucial that $g(\cdot)$ is monotonic.

We first apply the bivariate transformation to show that independence between X and Y is equivalent to independence between $g_1(X)$ and $g_2(Y)$ for any continuously differentiable 1–1 transformations $g_1(\cdot)$ and $g_2(\cdot)$.

Theorem 8 (5.14)**.** Suppose $U = g_1(X)$ and $V = g_2(Y)$ are some continuously differentiable 1-1 measurable functions. Then X and Y are independent if and only if U and V are independent.

Proof: (1) [*Necessity*] We first show that independence between X and Y implies independence between U and V. By definition, the Jacobian matrix

$$J_{UV}(x,y) = \begin{bmatrix} g_1'(x) & 0 \\ \\ 0 & g_2'(y) \end{bmatrix}$$

and the Jacobian

$$\det\left[J_{UV}(x,y)\right] = g_1'(x)g_2'(y).$$

It follows from the bivariate transformation theorem that

$$\begin{aligned} f_{UV}(u,v) &= f_{XY}(x,y)\left|\det\left[J_{UV}(x,y)\right]\right|^{-1} \\ &= f_X(x)f_Y(y)|g_1'(x)g_2'(y)|^{-1} \\ &= \left[f_X(x)|g_1'(x)|^{-1}\right]\cdot\left[f_Y(y)|g_2'(y)|^{-1}\right] \\ &= f_U(u)f_V(v), \end{aligned}$$

where $x = g_1^{-1}(u)$ and $y = g_2^{-1}(v)$, which are the inverse functions of $g_1(u)$ and $g_2(v)$. Because (u,v) is arbitrary, we have that U and V are independent.

(2) [*Sufficiency*] Next, we show that independence between U and V implies independence between X and Y. The proof is analogous to the reasoning in (1) by applying the bivariate transformation theorem to the bivariate transformation $X = g_1^{-1}(U)$ and $Y = g_2^{-1}(V)$. This completes the proof.

Now we consider a variety of examples on application of the bivariate transformation.

Example 20 (5.20)**.** Suppose the random variable $X \sim Beta(\alpha,\beta)$, namely its PDF

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1,$$

$Y \sim Beta(\alpha+\beta,\gamma)$, namely its PDF

$$f_Y(y) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha+\beta)\Gamma(\gamma)}y^{\alpha+\beta-1}(1-y)^{\gamma-1}, \qquad 0 < y < 1,$$

and X and Y are independent. Define $U = XY$, $V = X$. Find the joint PDF $f_{UV}(u,v)$.

Solution: (1) We first note that the support of (X, Y)

$$\begin{aligned}\Omega_{XY} &= \{(x,y) \in \mathbb{R}^2 : f_{XY}(x,y) = f_X(x)f_Y(y) > 0\} \\ &= \{(x,y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < 1\}.\end{aligned}$$

This is a rectangular area.

Given $U = XY$ and $V = X$, we obtain the support of (U, V)

$$\begin{aligned}\Omega_{UV} &= \{(u,v) \in \mathbb{R}^2 : f_{UV}(u,v) > 0\} \\ &= \{(u,v) \in \mathbb{R}^2 : 0 < u < v, 0 < v < 1\} \\ &= \{(u,v) \in \mathbb{R}^2 : 0 < u < v < 1\}.\end{aligned}$$

(2) We now solve for the inverse function $x = h_1(u,v)$ and $y = h_2(u,v)$. Given $u = xy, v = x$, we have

$$\begin{aligned}x &= h_1(u,v) = v, \\ y &= h_2(u,v) = \frac{u}{v}.\end{aligned}$$

(3) The Jacobian matrix of (X, Y)

$$\begin{aligned}J_{XY}(u,v) &= \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{bmatrix}.\end{aligned}$$

Therefore, the Jacobian

$$\det[J_{XY}(u,v)] = \det\left(\begin{bmatrix} 0 & 1 \\ v^{-1} & -u/v^2 \end{bmatrix}\right) = -\frac{1}{v}.$$

It follows from the bivariate transformation theorem that

$$\begin{aligned}f_{UV}(u,v) &= f_{XY}(x,y)|\det J_{XY}(u,v)| \\ &= f_X(x)f_Y(y)\left|-\frac{1}{v}\right| \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha+\beta)\Gamma(\gamma)}y^{\alpha+\beta-1}(1-y)^{\gamma-1}\frac{1}{v} \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}v^{\alpha-1}(1-v)^{\beta-1}\left(\frac{u}{v}\right)^{\alpha+\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{1}{v}\end{aligned}$$

for $0 < u < v < 1$, where we have $x = v, y = u/v$.

Example 21 (5.21)**.** Suppose X and Y are independent $N(0, \sigma^2)$. Define $U = X^2 + Y^2, V = X/\sqrt{U} = X/\sqrt{X^2+Y^2}$. (1) Find $f_{UV}(u, v)$; (2) show that U and V are independent.

Solution: (1) From $U = X^2 + Y^2, V = X/\sqrt{X^2+Y^2}$, and $X \sim N(0, \sigma^2), Y \sim N(0, \sigma^2)$, we have the support of (U, V)

$$\Omega_{UV} = \{(u, v) \in \mathbb{R}^2 : 0 < u < \infty, -1 < v < 1\}.$$

We first note that (U, V) is not a 1–1 mapping, because there exist two pairs, (x, y) and $(x, -y)$, that correspond to the same value of (u, v). As a result, the bivariate transformation theorem cannot be applied directly. Putting $Z = Y^2$, we will consider a transformation from (X, Z) to (U, V), which is a 1–1 mapping. Then the bivariate transformation theory is applicable. For the distribution of $Z = Y^2$, we have for any $z \geq 0$,

$$\begin{aligned}
F_Z(z) &= P(Y^2 \leq z) \\
&= P(-\sqrt{z} \leq Y \leq \sqrt{z}) \\
&= F_Y(\sqrt{z}) - F_Y(-\sqrt{z}), \\
f_Z(z) &= F'_Z(z) \\
&= f_Y(\sqrt{z})\frac{1}{2\sqrt{z}} + f_Y(-\sqrt{z})\frac{1}{2\sqrt{z}} \\
&= \frac{1}{\sqrt{2\pi z}\sigma} e^{-z/2\sigma^2}.
\end{aligned}$$

Since X is independent of Y, it is also independent of $Z = Y^2$. The joint PDF of (X, Z) is

$$f_{XZ}(x, z) = \frac{1}{2\pi\sigma^2\sqrt{z}} e^{-x^2/2\sigma^2} e^{-z/2\sigma^2}, -\infty < x < \infty, 0 \leq z < \infty.$$

Also, the support of (X, Z) is

$$\Omega_{XZ} = \{(x, z) \in \mathbb{R}^2 : -\infty < x < \infty, 0 \leq x < \infty\}$$

and the joint support of $U = X^2 + Z$ and $V = X/\sqrt{X^2+Z}$ is

$$\Omega_{UV} = \{(u, v) \in \mathbb{R}^2 : 0 < u < \infty, -1 < v < 1\}.$$

We now find the inverse functions $X = h_1(U, V)$ and $Z = h_2(U, V)$. Given

$$\begin{aligned}
U &= X^2 + Z, \\
V &= \frac{X}{\sqrt{X^2+Z}},
\end{aligned}$$

we have

$$\begin{aligned}
X &= h_1(U, V) = V\sqrt{U}, \\
Z &= h_2(U, V) = U(1 - V^2).
\end{aligned}$$

It follows that the Jacobian matrix of (X, Z),

$$\begin{aligned} J_{XZ}(u,v) &= \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \\ &= \begin{bmatrix} \frac{v}{2\sqrt{u}} & \sqrt{u} \\ 1-v^2 & -2uv \end{bmatrix}. \end{aligned}$$

By the bivariate transformation theorem, we have

$$\begin{aligned} f_{UV}(u,v) &= f_{XZ}(x,z)|\det J_{XZ}(u,v)| \\ &= \frac{1}{2\pi\sigma^2\sqrt{z}} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{z}{2\sigma^2}} |\det J_{XZ}(u,v)| \\ &= \frac{1}{2\pi\sigma^2\sqrt{u\,(1-v^2)}} e^{-\frac{u}{2\sigma^2}} \sqrt{u} \\ &= \frac{1}{2\pi\sigma^2\sqrt{1-v^2}} e^{-\frac{u}{2\sigma^2}}, \text{ for } u>0, -1<v<1. \end{aligned}$$

It follows that

$$f_{UV}(u,v) = \begin{cases} \frac{1}{2\pi\sigma^2\sqrt{1-v^2}} e^{-\frac{u}{2\sigma^2}}, & 0<u<\infty, -1<v<1, \\ 0, & \text{otherwise.} \end{cases}$$

(2) Although both U, V are functions of (X, Y), U and V are independent because their joint PDF $f_{UV}(u, v)$ can be partitioned into the product of two separate functions of u and v respectively for all $(u, v) \in \mathbb{R}^2$, with

$$g(u) = \begin{cases} e^{-\frac{u^2}{2\sigma^2}u,} & \text{if } u>0, \\ 0, & \text{if } u \le 0, \end{cases}$$

and

$$h(v) = \begin{cases} \frac{1}{2\pi\sigma^2\sqrt{1-v^2}}, & \text{if } -1<v<1, \\ 0, & \text{otherwise.} \end{cases}$$

The independence between U and V follows from the factorization theorem (Theorem 5.10).

Example 22 (5.22)**.** Suppose $X \sim N(\mu, \sigma^2), Y \sim N(\mu, \sigma^2)$, and X, Y are independent. Put $U = X+Y, V = X-Y$. (1) Find the joint PDF of (U, V); (2) show that U and V are independent.

Solution: The support of (U, V) is the entire xy plane. From $U = X + Y, V = X - Y$, we have

$$\begin{aligned} X &= h_1(U,V) = \frac{1}{2}(U+V), \\ Y &= h_2(U,V) = \frac{1}{2}(U-V). \end{aligned}$$

The Jacobian matrix of (X, Y) is

$$J_{XY}(u,v) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

It follows from the bivariate transformation theorem (Theorem 5.13) and $x = \frac{1}{2}(u+v), y = \frac{1}{2}(u-v)$ that

$$\begin{aligned} f_{UV}(u,v) &= f_{XY}(x,y)|\det J_{XY}(u,v)| \\ &= \frac{1}{2\pi\sigma^2}e^{-\frac{1}{2\sigma^2}[(x-\mu)^2+(y-\mu)^2]}\cdot\frac{1}{2} \\ &= \frac{1}{4\pi\sigma^2}e^{-\frac{1}{8\sigma^2}(u+v-2\mu)^2}e^{-\frac{1}{8\sigma^2}(u-v-2\mu)^2} \\ &= \frac{1}{4\pi\sigma^2}e^{-\frac{1}{4\sigma^2}(u-2\mu)^2}e^{-\frac{1}{4\sigma^2}v^2} \\ &= \frac{1}{\sqrt{2\pi 2\sigma^2}}e^{-\frac{1}{4\sigma^2}(u-2\mu)^2}\frac{1}{\sqrt{2\pi 2\sigma^2}}e^{-\frac{1}{4\sigma^2}v^2}, \quad -\infty < u, v < \infty. \end{aligned}$$

Hence, $U \sim N(2\mu, 2\sigma^2)$, $V \sim N(0, 2\sigma^2)$, and U and V are independent by the factorization theorem (Theorem 5.10), because we can partition the joint PDF

$$f_{UV}(u,v) = g(u)h(v), \quad -\infty < u, v < \infty,$$

where

$$g(u) = \frac{1}{\sqrt{2\pi 2\sigma^2}}e^{-\frac{1}{4\sigma^2}(u-2\mu)^2}, -\infty < u < \infty$$

is the PDF of a $N(2\mu, 2\sigma^2)$ distribution, and

$$h(v) = \frac{1}{\sqrt{2\pi 2\sigma^2}}e^{-\frac{1}{4\sigma^2}v^2}, -\infty < v < \infty$$

is the PDF of a $N(0, 2\sigma^2)$ distribution.

Example 23 (5.23)**.** Find the PDF of $X+Y$, where $X \sim U[0,1]$ and $Y \sim U[0,1]$, and X and Y are independent.

Solution: Put $U = X+Y$ and $V = X$. We shall first find the joint PDF $f_{UV}(u,v)$ of (U,V) and then integrate out v to obtain the PDF $f_U(u)$ of U.
(1)From $U = X+Y, V = X$, we find the support of (U,V) as

$$\Omega_{UV} = \{(u,v) \in \mathbb{R}^2 : v \le u \le 1+v, 0 \le v \le 1\}.$$

Figure 5.6 gives the plot of the support Ω_{UV}.

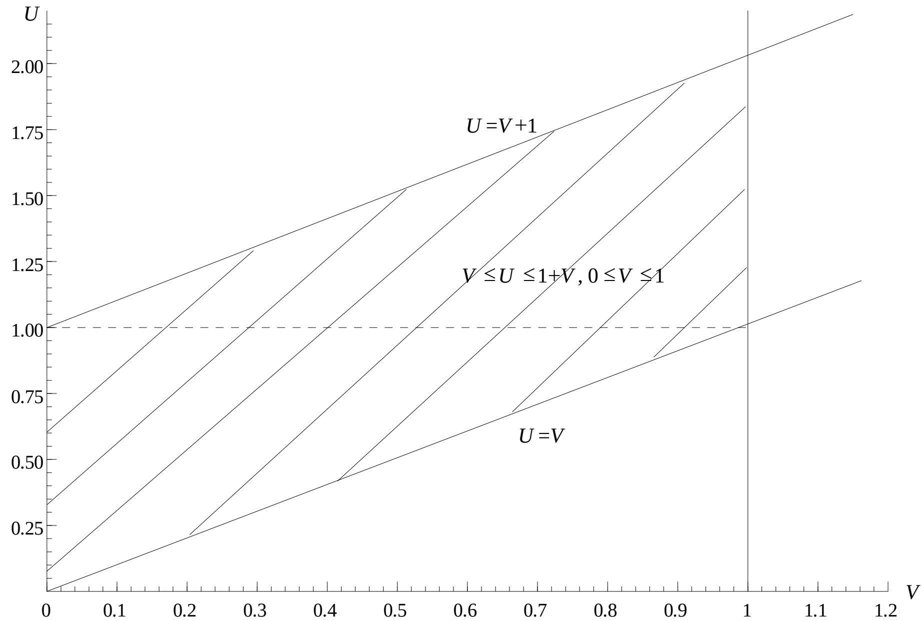

Figure 5.6: The Support of (U, V)

(2) The Jacobian matrix $J_{UV}(x, y)$ of (U, V) is

$$\begin{aligned} J_{UV}(x,y) &= \begin{bmatrix} \frac{\partial g_1(x,y)}{\partial x} & \frac{\partial g_1(x,y)}{\partial y} \\ \\ \frac{\partial g_2(x,y)}{\partial x} & \frac{\partial g_2(x,y)}{\partial y} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}. \end{aligned}$$

Hence, the Jacobian of (U, V) is

$$\det\left[J_{UV}(x,y)\right] = -1.$$

(3) By the bivariate transformation theorem, we obtain

$$\begin{aligned} f_{UV}(u,v) &= f_{XY}(x,y)|\det\left[J_{UV}(x,y)\right]|^{-1} \\ &= f_X(x)f_Y(y)|-1|^{-1} \\ &= 1, \quad 0 \le v \le 1, v \le u \le 1+v. \end{aligned}$$

We now integrate out v to obtain the PDF $f_U(u)$ of U. The support of U is $0 \le u \le 2$. We consider two subintervals: $0 \le u \le 1$ and $1 \le u \le 2$:

Case (1): For $0 \le u \le 1$, we have

$$f_U(u) = \int_0^u dv = u.$$

Case (2): For $1 \le u \le 2$, we have

$$\begin{aligned} f_U(u) &= \int_{u-1}^1 dv \\ &= 2-u \end{aligned}$$

It follows that

$$f_U(u) = \begin{cases} u, & 0 \le u \le 1, \\ 2-u, & 1 \le u < 2, \\ 0, & \text{otherwise.} \end{cases}$$

This is a triangular density function over the support $u \in [0,2]$ with peak at $u = 1$.

As pointed out earlier, we can extend the bivariate transformation further to a multivariate transformation involving more than two random variables. Many important statistics (e.g., econometric estimators and test statistics) are functions of more than two random variables. Thus, the multivariate transformation is very useful to obtain or understand the distribution of these important statistics.

Nevertheless, the multivariate transformation is often very tedious to use in practice, especially when the number of random variables involved is large. Fortunately, there exist other methods, such as those based on the moment generating function, that are convenient to use to obtain the distribution of interest. See examples in the subsequent sections of this chapter.

5.6 Bivariate Normal Distribution

We now consider a very important bivariate joint distribution called bivariate normal distribution.

Definition 15 (5.15)**. [Bivariate Normal Distribution]:** (X,Y) are jointly normally distributed, denoted as $BN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$, where $|\rho| \le 1$, if their joint PDF

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right]}.$$

When $(\mu_1,\mu_2,\sigma_1,\sigma_2) = (0,0,1,1)$, $BN(0,0,1,1,\rho)$ is called a standard bivariate normal distribution.

An alternative representation of $f_{XY}(x,y)$ is

$$f_{XY}(x,y) = \frac{1}{\sqrt{(2\pi)^2\det(\Sigma)}} e^{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)},$$

where $z = (x,y)', \mu = (\mu_1,\mu_2)'$, and

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}. \end{aligned}$$

Figure 5.7 below plots the joint PDF $f_{XY}(x,y)$ of $(X,Y) \sim BN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$, with various combinations of parameter values:

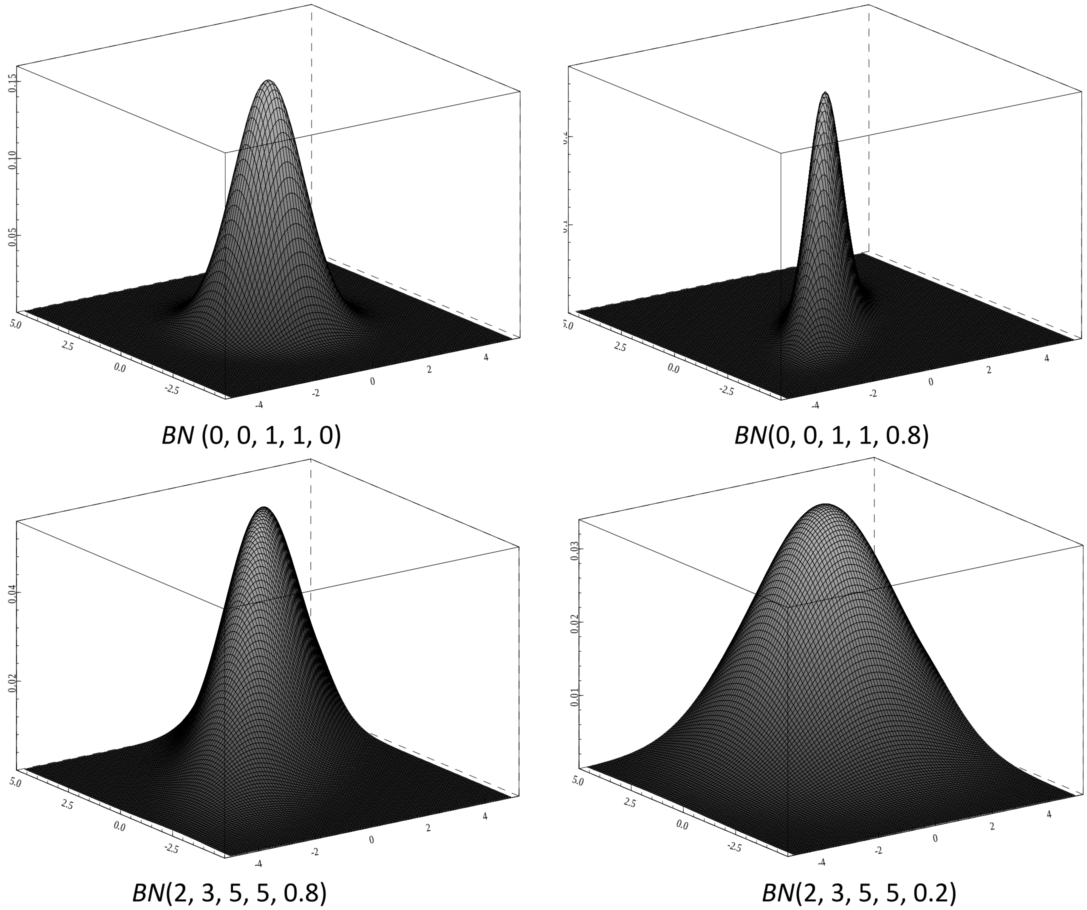

Figure 5.7: Joint PDF $f_{XY}(x, y)$ of $(X, Y) \sim BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

We now calculate the marginal PDF's of X and Y and the conditional PDF's of Y given X and of X given Y when $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Put

$$q(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Then we can write the joint PDF of a bivariate normal distribution as

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}q(x,y)}.$$

By straightforward algebra, we have

$$\begin{aligned}(1-\rho^2)q(x, y) &= (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left[\left(\frac{y-\mu_2}{\sigma_2}\right) - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right]^2 \\ &= (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu}{\sigma_2}\right)^2,\end{aligned}$$

where

$$\mu = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1).$$

Thus,

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} e^{-\frac{(y-\mu)^2}{2\sigma_2^2(1-\rho^2)}} dy \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \qquad -\infty < x < \infty, \end{aligned}$$

where the integral is equal to unity because it is the integral of the PDF of a $N[\mu, \sigma_2^2(1-\rho^2)]$ distribution.

By symmetry, we can also obtain

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}, \qquad -\infty < y < \infty.$$

These results imply that when X and Y are jointly normally distributed, then X and Y have marginal normal distributions respectively. In particular, $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim N(\mu_2, \sigma_2^2)$.

With $f_X(x) > 0$, the conditional PDF of Y given $X = x$ is

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} e^{-\frac{(y-\mu)^2}{2\sigma_2^2(1-\rho^2)}}, \qquad -\infty < y < \infty, \end{aligned}$$

where, as before, $\mu = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1)$. Therefore, the conditional distribution of Y given $X = x$ is also a normal distribution, $N[\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)]$. Similarly, the conditional distribution of X given $Y = y$ is a normal distribution, $N[\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2)]$. Figure 5.8 below plots the conditional PDF $f_{Y|X}(y|x)$ when $(X, Y) \sim BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$:

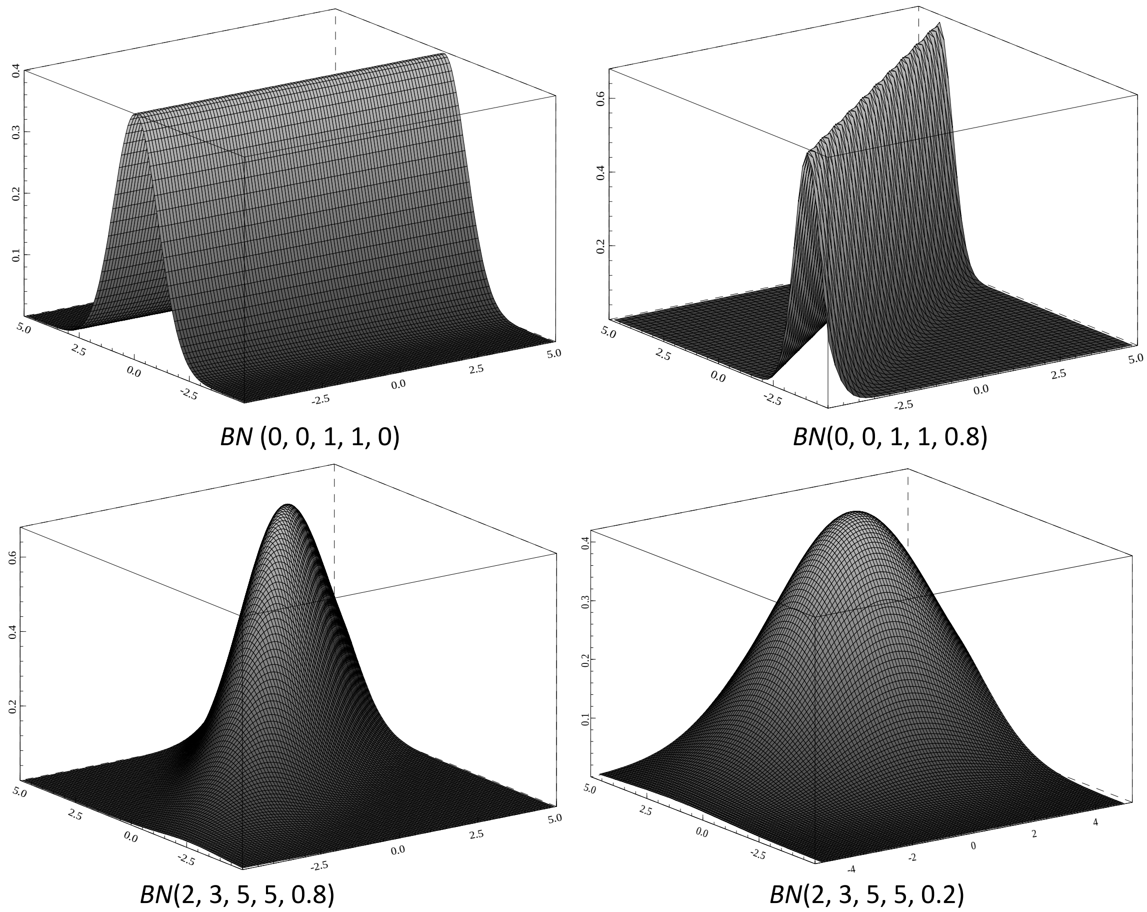

Figure 5.8 The Conditional PDF $f_{Y|X}(y|x)$ When $(X, Y) \sim BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

The constant ρ characterizes dependence between X and Y. When $\rho = 0$, the joint PDF of (X, Y) becomes

$$\begin{aligned} f_{XY}(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]} \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2} \\ &= f_X(x) f_Y(y) \text{ for all } (x, y) \in \mathbb{R}^2. \end{aligned}$$

Therefore, two jointly normally distributed random variables X and Y are independent if and only if $\rho = 0$.

We have shown that two jointly normally distributed random variables have marginal normal distributions. Is the converse true? That is, suppose the marginal PDFs of X and Y are a normal PDF respectively, are X and Y jointly normal? The following example shows that this is not necessarily the case.

Example 24 (5.24)**.** Let X and Y are random variables with joint PDF

$$f_{XY}(x,y)=\begin{cases}2f_X(x)f_Y(y), & \text{if } xy>0,\\ 0, & \text{if } xy\le 0,\end{cases}$$

where

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},\\ f_Y(y) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}. \end{aligned}$$

Then it can be shown that X and Y are normal random variables but obviously they are not jointly normally distributed.

In a multivariate setup, we can also define a multivariate normal distribution. The random variables $X_1,\cdots,X_n$ follow a joint normal distribution, denoted as $N(\mu,\Sigma)$ if their joint PDF

$$f_{\mathbf{X}^n}(\mathbf{x}^n)=\frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\left[-\frac{1}{2}(\mathbf{x}^n-\mu)'\Sigma^{-1}(\mathbf{x}^n-\mu)\right],$$

where $\mathbf{X}^n=(X_1,\cdots,X_n)', \mathbf{x}^n=(x_1,\cdots,x_n)', \mu=(\mu_1,\cdots,\mu_n)'$, and Σ is a $n\times n$ symmetric and positive definite matrix. The matrix Σ is called the variance-covariance matrix of the random vector $\mathbf{X}^n$ because its diagonal elements are the variances of the X_i, and its off-diagonal elements are the covariances between X_i and X_j, for all $i\neq j$. Obviously, Σ is a diagonal matrix if all covariances between X_i and X_j for $i\neq j$ are zero.

The assumption of multivariate normality offers a great deal of convenience in practice when calculating probabilities. For example, consider the return on a portfolio, which is a weighted average of n assets. Suppose the asset returns, $X_1,\cdots,X_n$, follow a multivariate normal distribution, it can be shown via a multivariate transformation that any linear combination of them, namely $X\equiv\sum_{i=1}^n c_iX_i$, will follow a normal distribution. It follows that the portfolio return also follows a normal distribution. Thus, calculation of the probabilities (e.g., $P(X<-V_{0.01})=0.01$) of the portfolio return will be rather convenient, under the assumption of the joint normal distribution for asset returns. Here, $V_{0.01}$ is the so-called called Value at Risk at the 1% significance level, which is widely used is financial risk management (see Section 3.7 in Chapter 3 for more discussion on value at risk).

5.7 Expectations and Covariance

Question: What information can we extract from a bivariate distribution?

We first define the mathematical expectation under a bivariate joint distribution.

Definition 16 (5.16)**. [Expectation under Bivariate Joint Distribution]**: Let $g:\Omega_{XY}\to\mathbb{R}$ be a real-valued measurable function, where Ω_{XY} is the support of (X,Y). Then the expectation of function $g(X,Y)$ is defined as

$$\begin{aligned} E[g(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)dF_{XY}(x,y) \\ &= \begin{cases} \sum\sum_{(x,y)\in\Omega_{XY}} g(x,y)f_{X,Y}(x,y) & \text{if } (X,Y) \text{ are DRV's,} \\ \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{XY}(x,y)dxdy & \text{if } (X,Y) \text{ are CRV's,} \end{cases} \end{aligned}$$

provided the double sum or the double integral exists. Like the univariate case, we say that $E[g(X,Y)]$ exists if $E|g(X,Y)| < \infty$.

We now consider expectations of some important functions $g(X,Y)$. We first choose $g(X,Y) = X^rY^s$. This will yield various joint product moments of (X,Y).

Definition 17 (5.17)**. [Product Moments]** The r-th and s-th order product moment of (X,Y) about the origin is defined as

$$E(X^rY^s) = \begin{cases} \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} x^ry^sf_{XY}(x,y), & \text{if } (X,Y) \text{ are DRV's,} \\ \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^ry^sf_{XY}(x,y)dxdy, & \text{if } (X,Y) \text{ are CRV's.} \end{cases}$$

Similarly, the rth and sth central product moment is defined as

$$\begin{aligned} & E\left\{[X-E(X)]^r[Y-E(Y)]^s\right\} \\ = & \begin{cases} \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}(x-\mu_X)^r(y-\mu_Y)^sf_{XY}(x,y), & \text{if } (X,Y) \text{ are DRV's,} \\ \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)^r(y-\mu_Y)^sf_{XY}(x,y)dxdy, & \text{if } (X,Y) \text{ are CRV's.} \end{cases} \end{aligned}$$

When X and Y are not independent, we say that there exists a relationship or an association between them. However, if there is a relationship or an association, the relationship or association may be weak or strong. How to measure the strength of a relationship between X and Y? Among other things, the cross-products $E(X^rY^s)$ and $E[(X-\mu_X)^r(Y-\mu_Y)^s]$ provide a way to characterize the relationship or association between X and Y. Different choices of order (r,s) will capture different types of association between X and Y. To see this, we now examine a special case where $(r,s) = (1,1)$. This corresponds to the expectation of function $g(X,Y) = (X-\mu_X)(Y-\mu_Y)$. This function yields a positive value if X and Y tend to move above or below their means in the same direction, and yields a negative value if X and Y tend to move above or below their means in opposite directions. Thus, it can characterize the direction of the co-movement between X and Y.

Definition 18 (5.18)**. [Covariance]:** Suppose $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then the covariance between two variables X and Y is defined as

$$\begin{aligned} \text{cov}(X,Y) &= E\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)dF_{XY}(x,y). \end{aligned}$$

The covariance is a measure of the co-movement between X and Y. Suppose there is a high probability that large values of X tend to be observed with large values of Y, and small values of X with small values of Y, then $\operatorname{cov}(X,Y) > 0$. In this case, we say that X and Y are positively correlated. Like in the univariate case, the expectation operator in the definition of $\operatorname{cov}(X,Y)$ can be interpreted as the average value of the product $(X - \mu_X)(Y - \mu_Y)$ in infinitely many repeated independent experiments. On the other hand, suppose there is a high probability that large values of X tend to be observed with small values of Y, and small values of X tend to be observed with large values of Y, then $\operatorname{cov}(X,Y) < 0$. In this case, we say that X and Y are negatively correlated. If their changes are irrelevant, then $\operatorname{cov}(X,Y) = 0$. In this case, we call X and Y are uncorrelated. Thus, the signs of $\operatorname{cov}(X,Y)$ provide information regarding the relationship between X and Y. Note that $\operatorname{cov}(X,X) = \operatorname{var}(X)$.

Nonzero covariance implies that there exists an association between X and Y, but it does not necessarily imply causality between X and Y. If one finds a positive correlation between oil price increase and economic slowdown, for example, it does not necessarily mean that oil price increase causes economic slowdown. For another example, if one finds that there exists a positive correlation between smoking and cancer, it does not necessarily imply smoking causes cancer.

The following theorem provides a convenient formula to calculate covariance.

Theorem 9 (5.15)**.** Suppose (X, Y) have finite second moments. Then

$$\operatorname{cov}(X,Y) = E(XY) - \mu_X \mu_Y.$$

Proof: Noting that the expectation operator $E(\cdot)$ is a linear operator, we have

$$\begin{aligned}
\operatorname{cov}(X,Y) &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \\
&= E(XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y) \\
&= E(XY) - \mu_X \mu_Y.
\end{aligned}$$

The magnitude of $\operatorname{cov}(X,Y)$ does not provide useful information about the strength of the relationship between X and Y, because it depends on the scales of X and Y. We now provide a normalized measure which is robust to the scales of X and Y.

Definition 19 (5.19)**. [Correlation]:** The correlation coefficient between X and Y is defined as

$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

The correlation coefficient ρ_{XY} is the standardized covariance. This is analogous to the definitions of skewness and kurtosis. Now the magnitude of ρ_{XY} can indicate the strength of the association which ρ_{XY} is capturing. Before we examine the nature of the association which ρ_{XY} captures, we first show that its absolute value is always less than or at most equal to unity.

Theorem 10 (5.16)**.** $|\rho_{XY}| \leq 1$.

Proof: This follows immediately from the Cauchy-Schwartz inequality: for any measurable functions $g(X)$ and $h(Y)$, we have

$$E|g(X)h(Y)| \leq \{E[g^2(X)]E[h^2(Y)]\}^{1/2}.$$

Setting $g(X) = X - \mu_X$ and $h(Y) = Y - \mu_Y$, we obtain

$$\begin{aligned} |\mathrm{cov}(X,Y)| &\leq E\,|(X-\mu_X)(Y-\mu_Y)| \\ &\leq (\sigma_X^2\sigma_Y^2)^{1/2} \\ &= \sigma_X\sigma_Y. \end{aligned}$$

It follows that $|\rho_{XY}| \leq 1$. This completes the proof.

The correlation coefficient ρ_{XY} is a measure of linear association between X and Y. To gain an insight into the nature of linear association, we first state a result for ρ_{XY} when X is a linear function of Y.

Theorem 11 (5.17)**.** Suppose $Y = a + bX$, $b \neq 0$, where $\sigma_X^2 = \mathrm{var}(X)$ exists. Then $\rho_{XY} = 1$ if $b > 0$, and $\rho_{XY} = -1$ if $b < 0$.

Proof: Since $\mu_Y = a + b\mu_X$ and $\sigma_Y^2 = b^2\sigma_X^2$, the covariance

$$\begin{aligned} \mathrm{cov}(X,Y) &= E\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= E\left[(X-\mu_X)(a+bX-a-b\mu_X)\right] \\ &= bE(X-\mu_X)^2 \\ &= b\sigma_X^2. \end{aligned}$$

It follows that

$$\begin{aligned} \rho_{XY} &= \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{b\sigma_X^2}{|b|\sigma_X^2} \\ &= \frac{b}{|b|} \\ &= \begin{cases} 1, & \text{if } b > 0, \\ -1, & \text{if } b < 0. \end{cases} \end{aligned}$$

This completes the proof.

Thus, when there is a perfect linear relationship between X and Y, we always have $\rho_{XY} = \pm 1$, that is, ρ_{XY} achieves the maximum value of unity in absolute value. In this sense, it is a measure of linear association between X and Y.

For certain kinds of joint distributions of X and Y, the correlation coefficient ρ_{XY} proves to be a very useful characteristic of the joint distribution. Unfortunately, the formal definition of ρ_{XY}

does not reveal this fact. We have seen that if a joint distribution of two variables has a correlation coefficient, then $-1 \leq \rho_{XY} \leq 1$. If $\rho_{XY} = 1$, then there is a line with equation $Y = a + bX$, where $b > 0$, the graph of which contains all information of the probability distribution of X and Y. In this extreme case, we have $P(Y = a + bX) = 1$. If $\rho_{XY} = -1$, we have the same state of affairs except that $b < 0$. This suggests the following interesting question: When ρ_{XY} does not have the one of its extreme values, is there a line in the xy-plane such that the probability for X and Y tends to be concentrated in a band about this line? Under certain restrictive conditions this is in fact the case, and under those conditions we can look upon ρ_{XY} as a measure of the intensity of the concentration of the probability for X and Y around the straight line.

We now first illustrate this by an example.

Example 25 (5.25)**.** Suppose the joint PDF

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4b\varepsilon}, & -\varepsilon + a + bx < y < \varepsilon + a + bx, \quad -\varepsilon < x < \varepsilon, \\ 0, & \text{otherwise,} \end{cases}$$

where $\varepsilon > 0$. The support Ω_{XY} is depicted in Figure 5.9 below:

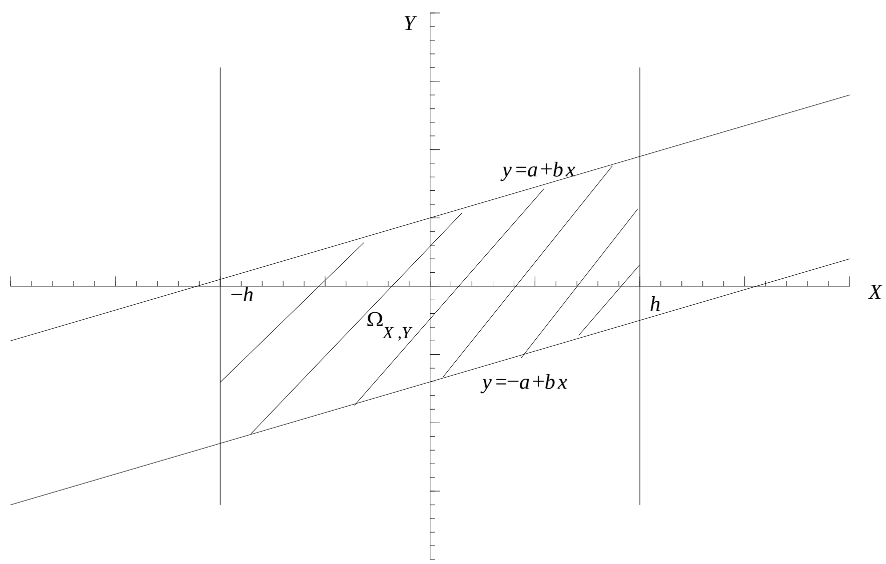

Figure 5.9: The Support Ω_{XY} of Example 5.25

Here, $|Y - (a + bX)|$ is bounded within a band of ε. If and only if $\varepsilon = 0$, there exists an exact linear relationship: $Y = a + bX$. It can be shown that

$$\rho_{XY} = \frac{b\varepsilon}{\sqrt{a^2 + b^2\varepsilon^2}}.$$

Obviously, suppose the band ε shrinks to zero, then ρ_{XY} will converge to 1 when $b > 0$.

In Example 5.25, the difference $|Y - (a + bX)|$ is bounded by a constant $\varepsilon > 0$. Below we provide an example where the difference $|Y - (a + bX)|$ is equal to a random variable ε which may have an unbounded support but a finite variance. In this case, ρ_{XY} can still measure the degree of linear association.

Example 26 (5.26)**. [Linear Regression Model]** Suppose we have

$$Y = a + bX + \varepsilon,$$

where ε is a random variable with $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma_\varepsilon^2 > 0$, and it is orthogonal to X in the sense $E(X\varepsilon) = 0$. This is usually called a linear regression model in statistics and econometrics. The random variable ε can be viewed as a disturbance to an otherwise perfect linear relationship $Y = a + bX$. Unlike Example 5.25, $\varepsilon = Y - (a + bX)$ may have unbounded support. When $\sigma_\varepsilon^2 > 0$, $|\rho_{XY}| < 1$ and we expect that $|\rho_{XY}|$ will shrink as σ_ε^2 increases. Given $E(X\varepsilon) = 0$, it can be shown that

$$\begin{aligned}
\rho_{XY} &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \\
&= \frac{b}{\sqrt{b^2 + \sigma_\varepsilon^2/\sigma_X^2}}.
\end{aligned}$$

Thus, the deviation of $|\rho_{XY}|$ from unity depends on the magnitude of the ratio $\sigma_\varepsilon^2/\sigma_X^2$, which is usually called a noise-to-signal ratio.

We now show how a linear regression model $Y = a + bX + \varepsilon$ can arise, where the disturbance ε is orthogonal to X, i.e., $E(X\varepsilon) = 0$.

Theorem 12 (5.18)**. [Best Linear Least Squares Prediction]:** Suppose X and Y are two random variables with finite second moments. If we use a linear function $\alpha + \beta X$ to predict Y, then the prediction error is $Y - (\alpha + \beta X)$. A popular criterion for prediction is the mean squared error, defined as

$$\text{MSE}(\alpha, \beta) = E[Y - (\alpha + \beta X)]^2.$$

Then the optimal coefficients (α^*, β^*) that minimizes $\text{MSE}(\alpha, \beta)$ are given by

$$\begin{aligned}
\alpha^* &= \mu_Y - \frac{\text{cov}(X, Y)}{\text{var}(X)} \mu_X, \\
\beta^* &= \frac{\text{cov}(X, Y)}{\text{var}(X)} = \rho_{XY} \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}.
\end{aligned}$$

Proof: We solve for the optimal coefficients (α^*, β^*) using the first order conditions(FOC):

$$\begin{aligned}
\frac{\partial \text{MSE}(\alpha, \beta)}{\partial \alpha}|_{(\alpha^*, \beta^*)} &= -2E[Y - (\alpha^* + \beta^* X)] = 0, \\
\frac{\partial \text{MSE}(\alpha, \beta)}{\partial \beta}|_{(\alpha^*, \beta^*)} &= -2E\{X[Y - (\alpha^* + \beta^* X)]\} = 0.
\end{aligned}$$

Solving these two equations, we can obtain the desired expressions for α^* and β^*.

Define the prediction error of the best linear least squares predictor $\alpha^* + \beta^* X$,

$$\varepsilon = Y - (\alpha^* + \beta^* X).$$

Then we can write

$$Y = \alpha^* + \beta^* X + \varepsilon,$$

where ε is orthogonal to X, namely,

$$E(X\varepsilon) = 0.$$

The FOC of the minimization of MSE(α, β) ensures $E(X\varepsilon) = 0$. That is, it is the very nature of the best linear least squares prediction that ensures that ε is orthogonal to X, implying that ε contains no linear component of X that is useful to predict Y. This provides a justification for Example 5.26. It is very important to note that the optimal slope coefficient β^* is proportional to covariance cov(X, Y). If cov$(X, Y) = 0$, then the linear predictor $\alpha^* + \beta^* X$ has no predictive power for Y in terms of the MSE criterion.

Covariance has been quite useful in economic analysis. For a most important empirical stylized fact in macroeconomics is called the Phillips curve, which states that unemployment and inflation tend to move in opposite directions, namely, they are negatively correlated. We now provide an important example in finance called capital asset pricing model (CAPM) to illustrate the usefulness of covariance.

Example 27 (5.27)**. [Capital Asset Pricing Model (CAPM)]:** Let R_{pt} be the return on a portfolio during a certain holding period t, r_{ft} is the risk-free interest rate, R_{mt} is the return on the market portfolio (often proxied by the return on the Standard & Poor 500 price index) in the same holding period. Then CAPM asserts that

$$R_{pt} - r_{ft} = \beta_p(R_{mt} - r_{ft}) + \varepsilon_{pt},$$

where $R_{pt} - r_{ft}$ is the excess return on the portfolio in the time period t, $R_{mt} - r_{ft}$ is the excess return on the market portfolio in the same period which represents the unavoidable systematic risk, and the random variable ε_{pt} represents the idiosyncratic risk peculiar to the portfolio, which could be eliminated by diversification (i.e, by forming a portfolio with a large number of assets; see Example 6.6 in Chapter 6). For simplicity, it is assumed here that $R_{mt} - r_{ft}$ and ε_{pt} are independent. This implies that in the equilibrium, any asset or portfolio should be paid only for its exposure to the market risk and not for its idiosyncratic risk. Hence, the expected excess return on the portfolio is β_p times the expected excess return on the market portfolio; that is, only the unavoidable systematic market risk should be compensated. It can be shown that the slope coefficient

$$\begin{aligned} \beta_p &= \frac{\text{cov}(R_{pt} - r_{ft}, R_{mt} - r_{ft})}{\text{var}(R_{mt} - r_{ft})} \\ &= \rho_{pm} \cdot \frac{\sqrt{\text{var}(R_{pt} - r_{ft})}}{\sqrt{\text{var}(R_{mt} - r_{ft})}}, \end{aligned}$$

where ρ_{pm} is the correlation between $R_{pt} - r_{ft}$ and $R_{mt} - r_{ft}$. The coefficient β_p is called the "investment beta" and has interesting economic interpretation: it measures both the riskiness of the portfolio and the reward for assuming that risk. The magnitude of β_p depends on the correlation strength between the portfolio and the market portfolio and the riskiness of the portfolio relative to the market portfolio. It indicates how aggressive the portfolio is. By definition, the beta for the market portfolio is unity, i.e., $\beta_m = 1$. It follows that the portfolio is more aggressive than the market portfolio if $\beta_p > 1$, and is less aggressive than the market portfolio if $\beta_p < 1$.

We note that under the above assumption for CAPM, we have

$$\text{var}\,(R_{pt} - r_{ft}) = \beta_p^2 \text{var}(R_{mt} - r_{ft}) + \text{var}(\varepsilon_{pt}),$$

where $\text{var}(R_{pt} - r_{ft})$ measures the total risk of the portfolio, $\text{var}(\varepsilon_{pt})$ measures the idiosyncratic risk peculiar to the portfolio, and $\beta_p^2 \text{var}(R_{mt} - r_{ft})$ is the unavoidable market risk tied to the portfolio. For another example, Z can be the total cost of production, where bX and cY are the costs of inputs (e.g., labor and capital), and a is a setup cost.

To further investigate the properties of $\text{cov}(X, Y)$, we now consider the mean and variance of the linear transformation

$$Z = a + bX + cY.$$

In economics, there are many examples of $Z = a + bX + cY$. As an example, Z is the return on a portfolio which consists of two risky assets and a risk-free asset, where X and Y are the return on the two risky assets respectively, and a, b, c are the portfolio weights on the risky assets and risk-free asset (assuming the return on the risk-free asset is unity). Then one may be interested in calculating the expected return and risk of such a portfolio.

Theorem 13 (5.19)**.** Suppose $Z = a + bX + cY$. Then
(1) $E(Z) = a + b\mu_X + c\mu_Y$.
(2) $\text{var}(Z) = b^2\sigma_X^2 + c^2\sigma_Y^2 + 2bc{\cdot}\text{cov}(X, Y)$.

For simplicity, we set $a = 0, b = c = 1$. Then Theorem 5.19 implies

$$\text{var}(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2\text{cov}(X, Y).$$

When $\text{cov}(X, Y) > 0$, we have $\text{var}(X + Y) > \sigma_X^2 + \sigma_Y^2$. That is, the variation of the sum is greater than the sum of individual variations. This follows because both X and Y move in the same direction with a high probability, and so it is more likely that $X + Y$ tends to be extremely large or small.

On the other hand, when $\text{cov}(X, Y) < 0$, the variation of the sum is less than the sum of individual variations. This follows because X and Y move in opposite directions with a high probability so that $X + Y$ does not vary much. These results have important implication on the risk of the portfolio investment: other things being equal, positive correlation between risky assets increases the risk of a portfolio, whereas negative correlation between risky assets decreases the risk of the portfolio.

More generally, we have the following multivariate result:

Theorem 14 (5.20)**.** Suppose $X_1, \cdot\cdot\cdot, X_n$ are a sequence of n random variables, and $Y = a_0 + \sum_{i=1}^n a_i X_i$, where the a_i are constants. Then we have
(1) $E(Y) = a_0 + \sum_{i=1}^n a_i E(X_i)$;
(2) $\text{var}(Y) = \sum_{i=1}^n a_i^2 \text{var}(X_i) + 2\sum_{i=2}^n \sum_{j=1}^{i-1} a_i a_j \text{cov}(X_i, X_j)$.

We now provide an interpretation for the constant ρ in a bivariate normal distribution $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ introduced in Section 5.6. That is, ρ is the correlation coefficient between X and Y.

Theorem 15 (5.21)**.** Suppose two random variables (X, Y) follow a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then the correlation coefficient $\rho_{XY} = \rho$.

Proof:

$$
\begin{aligned}
\rho_{XY} &= \frac{\operatorname{cov}(X,Y)}{\sigma_1\sigma_2} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) f(x,y)dxdy \\
&\quad \left(\text{setting } u = \frac{x-\mu_1}{\sigma_1}, v = \frac{y-\mu_2}{\sigma_2}\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv\frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(u^2-2u\rho v+(\rho v)^2-(\rho v)^2+v^2)}dudv \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uve^{-\frac{1}{2(1-\rho^2)}[(u-\rho v)^2+(1-\rho^2)v^2]}dudv \\
\text{(now setting } w &= u-\rho v, u = w+\rho v) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} wve^{-\frac{1}{2(1-\rho^2)}[w^2+(1-\rho^2)v^2]}dwdv \\
&\quad +\rho\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v^2e^{-\frac{1}{2(1-\rho^2)}[w^2+(1-\rho^2)v^2]}dwdv \\
&= 0+\rho \\
&= \rho.
\end{aligned}
$$

Thus, the constant ρ is the correlation coefficient between X and Y.

Example 28 (5.28)**.** How can one obtain a bivariate normal distribution $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ if two independent $N(0, 1)$ random variables Z_1 and Z_2 are given?

Solution: Define

$$
\begin{aligned}
X &= \mu_1 + aZ_1 + bZ_2, \\
Y &= \mu_2 + cZ_1 + dZ_2,
\end{aligned}
$$

where constants a, b, c, d satisfy the restrictions that

$$
\begin{aligned}
a^2 + b^2 &= \sigma_1^2, \\
c^2 + d^2 &= \sigma_2^2, \\
ac + bd &= \rho\sigma_1\sigma_2.
\end{aligned}
$$

Then using the bivariate transformation theorem, it can be shown that $(X, Y) \sim BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

In statistics, there is a so-called independent component analysis (ICA), where random variables $X_1, \cdots, X_n$ can be represented as a linear combination of a set of independent components $Z_1, \cdots, Z_m$, and n and m need not be equal to each other. Example 5.28 is a special case of independent component analysis. Here, independent components may be interpreted as independent random shocks to the system of variables $X_1, \cdots, X_n$. In this framework, it is straightforward to assess the impact of each shock to the system in terms of variance and the expected marginal effect.

Because ρ_{XY} is a measure of the linear association between X and Y, it is often referred to as the linear correlation coefficient. And, to avoid confusion, the word "linearly" is often used to describe the types of correlation with respect to ρ_{XY}. For instance, the term "positively linearly correlated" is often used in place of positively correlated.

Since the correlation coefficient ρ_{XY} is only a particular kind of linear association, two random variables X and Y may have a strong dependence but their correlation is zero, because their relationship is nonlinear. In other words, ρ_{XY} may not capture certain nonlinear associations. This is illustrated by the example below.

Example 29 (5.29)**.** Suppose $X \sim N(0, \sigma^2)$ and $Y = X^2$. Then

$$\begin{aligned} \text{cov}(X, Y) &= E(XY) - \mu_X \mu_Y \\ &= E(X^3) \\ &= \int_{-\infty}^{\infty} x^3 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= 0, \end{aligned}$$

where the integral is zero because the integrand is an odd function (i.e., $g(-x) = -g(x)$ for all x). Thus, although Y is a deterministic (nonlinear) function of X, they are uncorrelated. This indicates that $\text{cov}(X, Y)$ cannot capture some important nonlinear relationships between X and Y.

5.8 Joint Moment Generating Function

In this section, we first define the joint MGF of (X, Y) and then discuss its properties.

Definition 20 (5.20)**. [Joint MGF]:** The joint MGF of (X, Y) is defined as

$$M_{XY}(t_1, t_2) = E\left(e^{t_1 X + t_2 Y}\right), \qquad -\infty < t_1, t_2 < \infty,$$

provided the expectation exists for (t_1, t_2) in some neighborhood of $(0, 0)$.

The joint MGF $M_{XY}(t_1, t_2)$ is the bivariate generalization of the MGF $M_X(t)$ introduced in Section 3.8, Chapter 3. It contains rich information about the joint probability distribution of (X, Y). The marginal MGF's of X and Y can be obtained from the joint MGF $M_{XY}(t_1, t_2)$:

$$\begin{aligned} M_X(t_1) &= M_{XY}(t_1, 0), \\ M_Y(t_2) &= M_{XY}(0, t_2). \end{aligned}$$

When the joint MGF $M_{XY}(t_1, t_2)$ exists for (t_1, t_2) in some neighborhood of $(0, 0)$, the product moment $E(X^r Y^s)$ exists for all orders (r, s), and $M_{XY}(t_1, t_2)$ can be used to generate various product moments $E(X^r Y^s)$.

Theorem 16 (5.22)**.** Suppose the joint MGF $M_{XY}(t_1, t_2)$ exists for (t_1, t_2) in some neighborhood of $(0, 0)$. Then for all nonnegative integers $r, s \geq 0$,

$$E(X^r Y^s) = M_{XY}^{(r,s)}(0, 0),$$

and

$$\text{cov}(X^r, Y^s) = M_{XY}^{(r,s)}(0, 0) - M_X^{(r)}(0) M_Y^{(s)}(0).$$

In particular,

$$\text{cov}(X, Y) = M_{XY}^{(1,1)}(0, 0) - M_X^{(1)}(0) M_Y^{(1)}(0).$$

Proof: Given $M_{XY}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x + t_2 y} dF_{XY}(x, y)$, we have

$$\begin{aligned} M_{XY}^{(r,s)}(t_1, t_2) &= \frac{\partial^{r+s}}{\partial t_1^r \partial t_2^s} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x + t_2 y} dF_{XY}(x, y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial^{r+s} e^{t_1 x + t_2 y}}{\partial t_1^r \partial t_2^s} dF_{XY}(x, y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s e^{t_1 x + t_2 y} dF_{XY}(x, y). \end{aligned}$$

It follows that

$$M_{XY}^{(r,s)}(0, 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s dF_{XY}(x, y) = E(X^r Y^s).$$

Moreover,

$$\begin{aligned} M_{XY}^{(r,s)}(0, 0) - M_X^{(r)}(0) M_Y^{(s)}(0) &= E(X^r Y^s) - E(X^r) E(Y^s) \\ &= \text{cov}(X^r, Y^s). \end{aligned}$$

It is interesting to note that with the choice of $(r, s) = (1, 1)$, $\text{cov}(X, Y)$ can be obtained by differentiating the joint MGF $M_{XY}(t_1, t_2)$ and the marginal MGF's $M_X(t_1), M_Y(t_2)$. This completes the proof.

Like in the univariate case, the joint MGF $M_{XY}(t_1, t_2)$, when it exists for (t_1, t_2) in some neighborhood of $(0, 0)$, can be used to uniquely characterize the joint probability distribution of (X, Y).

Example 30 (5.30)**.** Suppose X, Y follow a bivariate normal distribution $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then their joint MGF

$$\begin{aligned} M_{XY}(t_1, t_2) &= e^{\mu_1 t + \mu_2 t + \frac{\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2}{2}} \\ &= e^{\mu' \mathbf{t} + \frac{1}{2}\mathbf{t}' \Sigma \mathbf{t}}, \end{aligned}$$

where $\mathbf{t} = (t_1, t_2)', \mu = (\mu_1, \mu_2)'$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

In a multivariate setup, we can define the joint MGF of random variables $X_1, \cdots, X_n$:

$$M_{\mathbf{X}^n}(\mathbf{t}) = E\left(e^{\mathbf{t}'\mathbf{X}^n}\right) = E\left(e^{\Sigma_{i=1}^n t_i X_i}\right),$$

provided the expectation exists for $\mathbf{t} = (t_1 \cdots, t_n)'$ in some neighborhood of the origin $(0, \cdots, 0)$, where $\mathbf{X}^n = (X_1, \cdots, X_n)', \mathbf{t} = (t_1, \cdots, t_n)'$.

The joint MGF may not exist for some joint distributions. However, we can always define a joint characteristic function, which exists for any joint distribution and has similar properties to the joint MGF. For space, we will not discuss the joint characteristic function. Interested readers are referred to Hong (1999) in a time series context.

5.9 Implications of Independence on Expectations

To examine the impact of independence on expectations under a joint probability distribution, we first state an important result of independence.

Theorem 17 (5.23)**.** Suppose (X, Y) are independent. Then for any measurable and integrable functions $h(X)$ and $q(Y)$,

$$E\left[h(X)q(Y)\right] = E[h(X)]E[q(Y)]$$

or equivalently,

$$\text{cov}[h(X), q(Y)] = 0.$$

Proof:

$$\begin{aligned} E[h(X)q(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)q(y)dF_{XY}(x,y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)q(y)dF_X(x)dF_Y(y) \quad \text{(by independence)} \\ &= \int_{-\infty}^{\infty} h(x)dF_X(x)\int_{-\infty}^{\infty} q(y)dF_Y(y) \\ &= E[h(X)]E[q(Y)]. \end{aligned}$$

Thus, when X and Y are independent, any measurable transformation of X and Y, no matter linear or nonlinear, will be uncorrelated.

5.9.1 Independence and Moment Generating Functions

To illustrate the usefulness of the above theorem, we first state a corollary about the impact of independence on the MGF.

Corollary 1 (5.24)**.** Suppose X and Y are independent and their marginal MGF's $M_X(t)$ and $M_Y(t)$ exist for t in a neighborhood of 0. Then $M_{X+Y}(t)$ exists for t in a small neighborhood of 0, and

$$M_{X+Y}(t) = M_X(t)M_Y(t) \text{ for all } t \text{ in a neighborhood of } (0,0).$$

Note that $M_{X+Y}(t)$ is the MGF of random variable $X+Y$. It is not the joint MGF $M_{XY}(t_1, t_2)$ of (X, Y).

Proof: Applying Theorem 5.23, we choose

$$g(X, Y) = e^{t(X+Y)} = e^{tX}e^{tY} = h(X)q(Y).$$

It follows that if X and Y are independent, then $M_{X+Y}(t)$ exists for all t in a small neighborhood of 0, and

$$E[e^{t(X+Y)}] = E(e^{tX})E(e^{tY}).$$

That is,

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

This property of the MGF for the sum of independent random variables is rather useful in characterizing the distributions for some random variables which themselves are the sum of other independent random variables. We now illustrate this via a variety of examples.

Example 31 (5.31)**. [Reproductive Property of the Normal Distribution]**: Suppose $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ and X, Y are independent. Then show

$$X \pm Y \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2).$$

Solution: The MGF's of X and Y are

$$\begin{aligned} M_X(t) &= Ee^{tX} = e^{\mu_1 t + \frac{\sigma_1^2}{2}t^2}, \\ M_Y(t) &= Ee^{tY} = e^{\mu_2 t + \frac{\sigma_2^2}{2}t^2}. \end{aligned}$$

By independence,

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} \cdot e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \end{aligned}$$

where $\mu = \mu_1 + \mu_2, \sigma^2 = \sigma_1^2 + \sigma_2^2$. It follows that

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

That is, the sum of two independent normal variables is also normal variable. In fact, this result can be extended to a linear combination of many independent normal random variables. This property is called the reproductive property of the normal distribution.

Example 32 (5.32). **[Reproductive Property of the Poisson Distribution]:** Suppose $X_1, \cdots, X_n$ are n independent random variables, with X_i having a Poisson(λ_i) distribution, $i = 1, \cdots, 2$. Show that $\Sigma_{i=1}^n X_i$ follows a Poisson(λ) distribution, where $\lambda = \Sigma_{i=1}^n \lambda_i$.

Solution: For a random variable X_i that follows a Poisson(λ_i) distribution, the MGF

$$M_i(t) = e^{\lambda_i(e^t-1)}, \qquad -\infty < t < \infty.$$

Put $X = \Sigma_{i=1}^n X_i$. Then its MGF

$$\begin{aligned} M_X(t) &= E\left(e^{t\Sigma_{i=1}^n X_i}\right) \\ &= \prod_{i=1}^n M_i(t) \\ &= \prod_{i=1}^n e^{\lambda_i(e^t-1)} \\ &= e^{\lambda(e^t-1)}, \end{aligned}$$

where $\lambda = \Sigma_{i=1}^n \lambda_i$. Therefore, $X \sim$ Poisson(λ). In other words, the sum of n independent Poisson(λ_i) random variables is a Poisson(λ) random variable with $\lambda = \sum_{i=1}^n \lambda_i$. This is called the reproductive property of the Poisson distribution.

Example 33 (5.33). **[Reproductive Property of the χ^2 Distribution]:** Suppose $X_1, X_2, \cdots$ are independent $\chi^2_{\nu_i}$ respectively. Show $\Sigma_{i=1}^n X_i$ follows a χ^2_ν, where $\nu = \Sigma_{i=1}^n \nu_i$.

Solution: From Chapter 4, we have that the MGF of a $\chi^2_{\nu_i}$ distribution

$$M_i(t) = (1-2t)^{-\nu_i/2}, \qquad t < \frac{1}{2}.$$

Put $X = \sum_{i=1}^n X_i$. Then its MGF

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \prod_{i=1}^n M_i(t) \\ &= \prod_{i=1}^n (1-2t)^{-\nu_i/2} \\ &= (1-2t)^{-\frac{1}{2}\sum_{i=1}^n \nu_i}. \end{aligned}$$

It follows that $X \sim \chi^2_\nu$, where $\nu = \sum_{i=1}^n \nu_i$. That is, the sum of independent χ^2 random variables is still a χ^2 random variable with the degree of freedom equal to the sum of individual degrees of freedom. This is called the reproductive property of the χ^2 distribution.

Dykstra and Hewett (1972) give two examples that shed light on some properties of the chi-square distribution. One example shows that the sum of two random variables can be distributed as a χ^2, where one of the variables is also χ^2-distributed but the other variable is positive but not necessarily χ^2-distributed. The other example shows the fact that the sum of two χ^2 random variables can have a χ^2-distribution, but these two χ^2 random variables need not be independent.

Example 34 (5.34)**.** Suppose $X_1, X_2, \cdots$ are independent random variables having exponential distributions with the same parameter $\beta > 0$. Show that $\Sigma_{i=1}^{n} X_i$ follows a Gamma(n, β) distribution.

Solution: For X_i having an EXP(β) distribution, we have

$$M_i(t) = (1 - \beta t)^{-1}, \qquad t < \frac{1}{\beta}.$$

Put $X = \sum_{i=1}^{n} X_i$. Then its MGF

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_i(t) \\ &= (1 - \beta t)^{-n}. \end{aligned}$$

This implies that $X \sim$ Gamma(n, β). That is, the sum of n independent exponential random variables with same parameter β follows a Gamma(n, β) distribution.

5.9.2 Independence and Uncorrelatedness

Applying Theorem 5.23, we immediately obtain an important result that independence implies uncorrelatedness.

Corollary 2 (5.25)**.** Suppose (X, Y) are independent. Then $\text{cov}(X, Y) = 0$.

Proof: Set $g(X) = X - \mu_X$ and $h(Y) = Y - \mu_Y$, and then apply Theorem (5.23).

Since independence rules out all possible types of association between X and Y, whereas uncorrelatedness only implies the absence of a linear association, independence is clearly a stronger condition than uncorrelatedness. Therefore, independence implies uncorrelatedness but uncorrelatedness does not necessarily imply independence. This is illustrated by the following examples.

Example 35 (5.35)**.** Put $Y = X^2$, where X follows a continuous probability distribution that is symmetric about 0 (i.e., $f_X(x) = f_X(-x)$ for all $-\infty < x < \infty$). Then

$$\text{cov}(X, Y) = 0.$$

Solution: Because the distribution of X is symmetric about 0, we have $E(X) = 0$ and $E(X^3) = 0$. It follows that

$$\begin{aligned} \text{cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(X^3) - 0 \cdot E(X^2) \\ &= E(X^3) \\ &= \int_{-\infty}^{\infty} x^3 f_X(x) dx \\ &= 0. \end{aligned}$$

Therefore, $\text{cov}(X, Y)$ cannot capture the quadratic association between X and $Y = X^2$ when X is symmetrically distributed about zero.

Example 36 (5.36)**. [Uncorrelated Bivariate Student's t-distribution]**: Random variables X and Y are called to follow the standard bivariate Student t-distribution if their joint PDF is given by

$$f_{XY}(x,y) = \frac{1}{2\pi}\left[1+\frac{1}{\nu}(x^2+y^2)\right]^{-\frac{\nu}{2}}, \qquad \nu \geq 3,$$

where the shape parameter ν is called the degree of freedom. It can be shown that $\mathrm{cov}(X,Y) = 0$ (verify it!). However, for any given $\nu < \infty$, X and Y are not independent because the joint PDF $f_{XY}(x,y)$ cannot be partitioned into the product of two separate functions of x and y respectively.

Note that when $\nu \to \infty$, we have

$$f_{XY}(x,y) \to \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x^2+y^2\right)\right]$$

given the fact that $(1+a/\nu)^{\nu} \to e^a$ as $\nu \to \infty$. In this case, X and Y are independent $N(0,1)$ random variables. Figure 5.10 below plots the joint PDF $f_{XY}(x,y)$ of the bivariate standard Student t-distribution with various choices of degree of freedom ν :

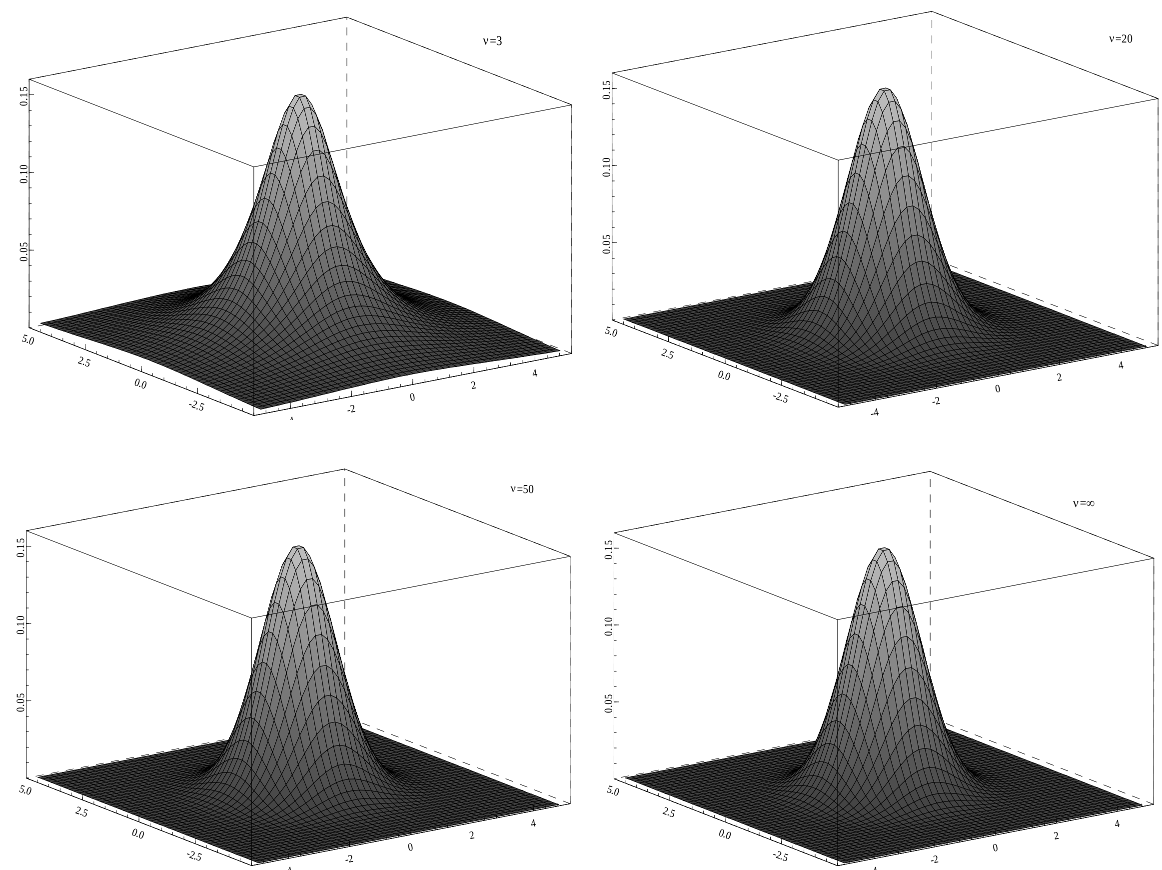

Figure 5.10: Joint PDF of the Bivariate Standard Student t-distribution with Different ν

Example 37 (5.37)**.** Suppose X has a PDF $f(x-\theta)$, where $f(\cdot)$ is symmetric about 0 and $E|X| < \infty$. Define $Y = \mathbf{1}(|X-\theta| < 2)$, where $\mathbf{1}(\cdot)$ is the indicator function, taking value 1 if $|X-\theta| < 2$ and 0 if $|X-\theta| \geq 2$. Check if $\text{cov}(X, Y) = 0$.

Solution: This is left as an exercise.

The above examples show that while independence implies uncorrelatedness, the converse is not true. However, there exist some special cases under which $\text{cov}(X, Y) = 0$ if and only if (X, Y) are independent. We consider two examples.

Theorem 18 (5.26)**.** Suppose (X, Y) are jointly normally distributed. Then $\text{cov}(X, Y) = 0$ if and only if X and Y are independent.

Proof: (1) [*Necessity*] By Corollary 5.25, if X, Y are independent, then we have $\text{cov}(X, Y) = 0$.

(2) [*Sufficiency*] Next, we show $\text{cov}(X, Y) = 0$ implies independence between X and Y. For a bivariate normal distribution $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, we have shown in Theorem 5.21 that the correlation coefficient $\rho_{XY} = \rho$. Thus, $\text{cov}(X, Y) = 0$ implies $\rho = 0$. It follows that the joint PDF

$$\begin{aligned} f(x,y) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(y-\mu_2)^2} \\ &= f_X(x) f_Y(y) \text{ for all } -\infty < x, y < \infty. \end{aligned}$$

Therefore, X and Y are independent.

The characterization of Theorem 5.26 provides a rather simple way to check independence between two jointly normally distributed random variables. That is, it suffices to check whether the correlation between X and Y is zero.

There also exists an example for discrete random variables for which uncorrelatedness implies independence.

Theorem 19 (5.27)**.** Suppose $X \sim \text{Bernoulli}(p_1)$ and $Y \sim \text{Bernoulli}(p_2)$. Then X and Y are independent if and only if $\text{cov}(X, Y) = 0$.

Proof: It suffices to show that if $\text{cov}(X, Y) = 0$, then X and Y are independent. We shall check whether the joint PMF $f_{XY}(x, y) = f_X(x) f_Y(y)$ for all $x = 1, 0$ and $y = 1, 0$. First, we observe $f_X(1) = p_1, f_X(0) = 1 - p_1, f_Y(1) = p_2, f_Y(0) = 1 - p_2, \mu_X = p_1$ and $\mu_Y = p_2$.

Now suppose $\text{cov}(X, Y) = 0$. Then

$$E(XY) = E(X)E(Y)$$

or equivalently

$$\sum_{x=0}^{1} \sum_{y=0}^{1} xy f_{XY}(x, y) = p_1 p_2.$$

Since $\sum_{x=0}^{1} \sum_{y=0}^{1} xy f_{XY}(x, y) = f_{XY}(1, 1)$. We have

$$f_{XY}(1, 1) = p_1 p_2 = f_X(1) f_Y(1).$$

Next, given $f_X(1) = \sum_{y=0}^{1} f_{XY}(1, y)$, we have

$$\begin{aligned} f_{XY}(1,0) &= f_X(1) - f_{XY}(1,1) \\ &= p_1(1-p_2) \\ &= f_X(1)f_Y(0). \end{aligned}$$

Similarly, we can show

$$f_{XY}(0,1) = (1-p_1)p_2 = f_X(0)f_Y(1)$$

and

$$f_{XY}(0,0) = (1-p_1)(1-p_2) = f_X(0)f_Y(0).$$

It follows that $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all $x = 1, 0$ and all $y = 1, 0$. Therefore, X and Y are independent. The proof is completed.

Question: Theorem 5.23 shows that if X and Y are independent, then

$$\mathrm{cov}[h(X), q(Y)] = 0$$

for any measurable functions $h(X)$ and $q(Y)$. Now, suppose $\mathrm{cov}[h(X), q(Y)] = 0$ for any measurable functions $h(X)$ and $q(Y)$. Are X and Y independent?

To answer this question, we consider $h_x(X) = \mathbf{1}(X \leq x), q_y(Y) = \mathbf{1}(Y \leq y)$ for any given $(x, y) \in \mathbb{R}^2$. Then

$$\begin{aligned} E[h_x(X)] &= E[\mathbf{1}(X \leq x)] \\ &= F_X(x), \\ E[q_y(Y)] &= E[\mathbf{1}(Y \leq y)] \\ &= F_Y(y), \\ E[h_x(X)q_y(Y)] &= E[\mathbf{1}(X \leq x)\mathbf{1}(Y \leq y)] \\ &= F_{XY}(x,y). \end{aligned}$$

Now, suppose we have

$$\mathrm{cov}[h_x(X), q_y(Y)] = 0 \text{ for all } -\infty < x, y < \infty$$

or equivalently

$$E[h_x(X)q_y(Y)] = E[h_x(X)]E[q_y(Y)] \text{ for all } -\infty < x, y < \infty.$$

Then

$$F_{XY}(x,y) = F_X(x)F_Y(y) \text{ for all } -\infty < x, y < \infty.$$

It follows that X and Y must be independent.

We can also use the joint MGF to characterize independence.

Theorem 20 (5.28)**.** Suppose the joint MGF $M_{XY}(t_1, t_2) = E(e^{t_1 X + t_2 Y})$ exists for (t_1, t_2) in some neighborhood of the origin $(0, 0)$. Then (X, Y) are independent if and only if

$$M_{XY}(t_1, t_2) = M_X(t_1) M_Y(t_2),$$

for all (t_1, t_2) in some neighborhood of $(0, 0)$.

Proof: (1) [*Necessity*] Suppose (X, Y) are independent. Then, by Theorem (5.23), we have

$$\begin{aligned} M_{XY}(t_1, t_2) &= E(e^{t_1 X + t_2 Y}) \\ &= E(e^{t_1 X} \cdot e^{t_2 Y}) \\ &= E\left(e^{t_1 X}\right) E\left(e^{t_2 Y}\right) \\ &= M_X(t_1) M_Y(t_2). \end{aligned}$$

(2) [*Sufficiency*] Next, we shall show that if $M_{XY}(t_1, t_2) = M_X(t_1) M_Y(t_2)$ for all (t_1, t_2) in some neighborhood of $(0, 0)$, then (X, Y) are independent. We apply the uniqueness theorem of the MGF. By the uniqueness theorem of the MGF, the bivariate distribution $F_{XY}(x, y)$ is the unique distribution that corresponds to the joint $M_{XY}(t_1, t_2)$, and the bivariate distribution $F_X(x) F_Y(y)$ is the unique distribution that corresponds to the joint MGF $M_X(t_1) M_Y(t_2)$. If $M_{XY}(t) = M_X(t_1) M_Y(t_2)$ for (t_1, t_2) in some neighborhood of $(0, 0)$, then $F_X(x) F_Y(y)$ is also the unique distribution associated with $M_{XY}(t)$. It follows that $F_{XY}(x, y) = F_X(x) F_Y(y)$ for all $-\infty < x, y < \infty$. This completes the proof.

Again, it is important to emphasize the difference between $M_{XY}(t_1, t_2)$ and $M_{X+Y}(t)$: $M_{XY}(t_1, t_2) = E(e^{t_1 X + t_2 Y})$ is the joint MGF of (X, Y), whereas $M_{X+Y}(t) = E(e^{t(X+Y)})$ is the MGF of the sum $X + Y$.

To sum up, we have learnt three basic approaches to characterizing independence, that is, checking whether the joint CDF, or the joint PMF/PDF, or the joint MGF, is equal to the product of their marginal counterparts, respectively.

There exists an alternative insightful representation of the MGF-based characterization of independence:

Theorem 21 (5.29)**.** Suppose $M_{XY}(t_1, t_2)$ exists for (t_1, t_2) in a neighborhood of $(0, 0)$. Then (X, Y) are independent if and only if

$$\text{cov}(e^{t_1 X}, e^{t_2 Y}) = 0$$

for all (t_1, t_2) in the neighborhood of $(0, 0)$.

Proof: Using the formula $\text{cov}(U, V) = E(UV) - E(U)E(V)$, it is straightforward to show that

$$\begin{aligned} \text{cov}(e^{t_1 X}, e^{t_2 Y}) &= E\left(e^{t_1 X} e^{t_2 Y}\right) - E\left(e^{t_1 X}\right) E\left(e^{t_2 Y}\right) \\ &= M_{XY}(t_1, t_2) - M_X(t_1) M_Y(t_2). \end{aligned}$$

Because $M_{XY}(t_1,t_2) = M_X(t_1)M_Y(t_2)$ for all (t_1,t_2) in a neighborhood of $(0,0)$ if and only if (X,Y) are independent, it follows immediately that $\text{cov}(e^{t_1X}, e^{t_2Y}) = 0$ for all t_1, t_2 in a neighborhood of 0 if and only if (X,Y) are independent. The proof is completed.

Since $\text{cov}(e^{t_1X}, e^{t_2Y})$ is the covariance between exponential transformations e^{t_1X} and e^{t_2Y}, it can be viewed as a generalized covariance. Theorem 5.29 shows that X and Y are independent if and only if this generalized covariance $\text{cov}(e^{t_1X}, e^{t_2Y})$ is zero for all pairs of (t_1,t_2) in some neighborhood of the origin $(0,0)$. The next theorem show that the generalized covariance $\text{cov}(e^{t_1X}, e^{t_2Y})$ can be viewed as a covariance generating function, that is, a function that can be used to generate various covariances $\text{cov}(X^r, Y^s)$.

Theorem 22 (5.30)**.** Suppose $M_{XY}(t_1,t_2)$ exists for (t_1,t_2) in a neighborhood of $(0,0)$. Then

$$\text{cov}(X,Y) = \left.\frac{\partial^2 \text{cov}(e^{t_1X}, e^{t_2Y})}{\partial t_1 \partial t_2}\right|_{(t_1,t_2)=(0,0)}.$$

Moreover, for any positive integers r, s,

$$\text{cov}(X^r,Y^s) = \left.\frac{\partial^{r+s} \text{cov}(e^{t_1X}, e^{t_2Y})}{\partial t_1^r \partial t_2^s}\right|_{(t_1,t_2)=(0,0)}.$$

When $M_{XY}(t_1,t_2)$ exists for (t_1,t_2) in a neighborhood of $(0,0)$, all derivatives of $\text{cov}(e^{t_1X}, e^{t_2Y})$ at the origin (0,0) exist. Therefore, using MacLaurin's series expansion $e^{tX} = \sum_{r=0}^{\infty} \frac{(tX)^r}{r!}$, we obtain

$$\begin{aligned}
\text{cov}(e^{t_1X}, e^{t_2Y}) &= \sum_{r=1}\sum_{s=1} \frac{t_1^r t_2^s}{r!s!} \frac{\partial^{r+s}\text{cov}(e^{t_1X}, e^{t_2Y})}{\partial t_1^r \partial t_2^s}|_{t=(0,0)} \\
&= \sum_{r=1}\sum_{s=1} \frac{t_1^r t_2^s}{r!s!} \text{cov}(X^r, Y^s).
\end{aligned}$$

Thus, $\text{cov}(e^{t_1X}, e^{t_2Y})$ contains information on all covariances $\{\text{cov}(X^r, Y^s)\}$ of various orders. If X and Y are independent, then $\text{cov}(X^r, Y^s) = 0$ for all $r, s > 0$, including $\text{cov}(X,Y) = 0$. In general, $\text{cov}(X,Y) = 0$ is only one of an infinite set of implications for independence. It is possible that $\text{cov}(X,Y) = 0$ but $\text{cov}(X^r, X^s) \neq 0$ for some choice of (r,s). In this case, X and Y are uncorrelated but they are not independent. An economic example is a sequence of high-frequency asset returns $\{X_t\}$ over time. It is often found that $\{X_t\}$ is serially uncorrelated over different time periods, that is, $\text{cov}(X_t, X_{t-j}) = 0$ for all $j > 0$. However, there exists persistent volatility clustering, that is, a large volatility today tends to be followed by another large volatility tomorrow, and a small volatility today tends to be followed by another small volatility, and these patterns alternate over times. This implies $\text{cov}(X_t^2, X_{t-j}^2) > 0$ at least for some $j > 0$.

Finally, we consider the following problem: Suppose all moments of X exist, and $\text{cov}(X^r, Y^s) = 0$ for all $r, s > 0$. Are X and Y independent?

This is analogous to the question in the univariate case whether the equality of all moments of two random variables X and Y imply identical distributions (see Theorem 3.25). The answer is yes, under the condition that both X and Y have bounded supports.

Theorem 23 (5.31)**.** Suppose X and Y have bounded supports. Then $\text{cov}(X^r, Y^s) = 0$ for all $r, s > 0$ if and only if X and Y are independent.

Proof: Given the bounded supports for both X and Y, there exists a constant $M < \infty$ such that $P(|X| < M) = 1$ and $P(|Y| < M) = 1$. Therefore, $E|X^r| \leq M^r$ and $E|Y^s| \leq M^s$. It follows that for any given $t_1, t_2 \in (-\infty, \infty)$, $\text{cov}(e^{t_1 X}, e^{t_2 Y})$ exists because

$$\begin{aligned}
\left|\text{cov}(e^{t_1 X}, e^{t_2 Y})\right| &= \left|\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \frac{t_1^r t_2^s}{r!s!}\text{cov}(X^r, Y^s)\right| \\
&\leq \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \frac{|t_1|^t |t_2|^s}{r!s!}|\text{cov}(X^r, Y^s)| \\
&\leq \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \frac{|t_1|^r |t_2|^s}{r!s!}(2M)^r (2M)^s \\
&= e^{2(|t_1|M + |t_2|M)} < \infty,
\end{aligned}$$

where we have made use of the fact that $|\text{cov}(X^r, Y^s)| \leq [\text{var}(X^{2r})\text{var}(Y^{2s})]^{1/2} \leq (2M)^{r+s}$.

(1) [*Necessity*] Suppose X and Y are independent. Then applying Theorem 5.23 with the choices of $h(X) = X^r, q(Y) = Y^s$, we have $\text{cov}(X^r, Y^s) = 0$ for all $r, s > 0$.

(2) [*Sufficiency*] Now, suppose $\text{cov}(X^r, Y^s) = 0$ for all $r, s > 0$. Then

$$\text{cov}(e^{t_1 X}, e^{t_2 Y}) = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \frac{t_1^r t_2^s}{r!s!}\text{cov}(X^r, Y^s) = 0$$

for all $t_1, t_2 \in (-\infty, \infty)$, which implies independence between X and Y by Theorem 5.29. This completes the proof.

Theorem 5.31 can be viewed as a generalization of Theorem 3.25 from the univariate context to the bivariate context. This implies that we can check all possible covariances between X^r and Y^s are zero in order to check independence between X and Y.

5.10 Conditional Expectations

Question: What information can we extract from a conditional distribution $f_{Y|X}(y|x)$?

We first define the expectation under a conditional distribution.

Definition 21 (5.21)**. [Conditional Expectation]:** The conditional expectation of $g(X, Y)$ given $X = x$ is defined as

$$\begin{aligned}
E[g(X,Y)|X &= x] = E[g(X,Y)|x] \\
&= \begin{cases} \sum_{y \in \Omega_Y(x)} g(x,y) f_{Y|X}(y|x), & \text{if } (X,Y) \text{ are DRV's,} \\ \\ \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x) dy, & \text{if } (X,Y) \text{ are CRV's,} \end{cases}
\end{aligned}$$

where $\Omega_Y(x) = \{y \in \Omega_Y : f_{Y|X}(y|x) > 0\}$ is the support of Y conditional on the event that $X = x$.

When taking a conditional expectation, one treats x as fixed, and conditional expectation is the expectation with respect to a conditional distribution of Y given $X = x$ instead of an unconditional distribution of Y.

Since y is integrated out, this conditional expectation $E[g(X,Y)|X = x]$ is a function of x only. Thus, $E[g(X,Y)|X]$ is a function of random variable X.

With the conditional expectation, $E[g(X,Y)|X]$, we can state the law of iterated expectations.

Theorem 24 (5.32)**. [Law of Iterated Expectations]:** Suppose $g(X,Y)$ is a measurable function and $E[g(X,Y)]$ exists. Then

$$\begin{aligned} E[g(X,Y)] &= E_X\{E[g(X,Y)|X]\} \\ &= E_Y\{E[g(X,Y)|Y]\}. \end{aligned}$$

Proof: We focus on the continuous case. Suppose (X,Y) have a joint PDF $f_{XY}(x,y)$. By the multiplication rule, we have $f_{XY}(x,y) = f_{Y|X}(y|x)f_X(x)$ if $f_X(x) > 0$. It follows that

$$\begin{aligned} E[g(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{Y|X}(y|x)f_X(x)dxdy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x,y)f_{Y|X}(y|x)dy\right] f_X(x)dx \\ &= \int_{-\infty}^{\infty} E[g(X,Y)|X = x]f_X(x)dx \\ &= E_X\{E\left[g(X,Y)|X\right]\}. \end{aligned}$$

Similarly it can be shown for the case of discrete random variables. This completes the proof.

The law of iterated expectations provides a two-stage procedure to compute an unconditional expectation. Thus, it is also called the law of total expectations. We shall provide an economic interpretation for this law of iterated expectations when we consider a special function of $g(X,Y)$ below.

In practice, we often write $E[g(X,Y)] = E\{E[g(X,Y)|X]\}$, by abusing the notation "E." The same notation "E" stands for different expectations in the same equation. The inside notation E is the expectation with respect to the conditional distribution of $Y|X$, and the outside notation E is the expectation with respect to the marginal distribution of X.

We now introduce various conditional moments of Y given X and examine their properties.

Definition 22 (5.22)**. [Conditional Mean]** The conditional mean of Y given $X = x$ is defined as

$$\begin{aligned} E(Y|x) &= E(Y|X=x) \\ &= \begin{cases} \sum_{y\in\Omega_Y(x)} y f_{Y|X}(y|x) & \text{if } (X,Y) \text{ are DRV's,} \\ \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy & \text{if } (X,Y) \text{ are CRV's.} \end{cases} \end{aligned}$$

The conditional mean $E(Y|X = x)$ is the average value of Y given that X has taken value x. We now provide an economic interpretation for $E(Y|X)$ and the law of iterated expectations via an example below.

Example 38 (5.38)**. [Average Wage and Law of Iterated Expectations]** Suppose Y is the wage of an employee, and X is a gender variable of the employee, taking value 0 if an employee is female, and taking value 1 if an employee is male. Then $E(Y|X = 0)$ is the average wage of female employees and $E(Y|X = 1)$ is the average wage of male employees.

Applying the law of iterated expectations, we obtain the overall average wage of employees

$$\begin{aligned} E(Y) &= E[E(Y|X)] \\ &= P(X=0)\cdot E(Y|X=0) + P(X=1)\cdot E(Y|X=1), \end{aligned}$$

where $P(X = 0)$ is the proportion of female employees in the labor force, and $P(X = 1)$ is the proportion of male employees in the labor force. The law of iterated expectations thus provides insight into the income distribution of the labor market.

The conditional mean $E(Y|X)$ is a function of X only. It is also called the regression function of Y on X to signify the predictive relationship of X for Y. It is the primary interest in econometrics and economics.

Why?

Suppose we use a function (or model) $g(X)$ of X to predict Y. Then the prediction error is $Y - g(X)$. A criterion to evaluate the predictive ability of model $g(X)$ is the mean squared error of $g(X)$, which is defined as

$$\text{MSE}(g) = E[Y - g(X)]^2.$$

MSE(g) is the average of the squared prediction errors. The smaller MSE(g), the better the predictor $g(X)$. The theorem below shows that the conditional mean is the optimal predictor for Y in terms of mean squared error.

Theorem 25 (5.33)**. [Mean Squared Error Criterion]** Let X and Y be random variables defined on the same sample space and suppose Y has a finite variance. Then the conditional mean $E(Y|X)$ is the optimal minimizer for the minimization problem of $E[Y - g(X)]^2$; that is,

$$E(Y|X) = \arg\min_{g(\cdot)} E[Y - g(X)]^2,$$

where the minimization is over all measurable and square-integrable functions.

Proof: Put $g_0(X) = E(Y|X)$. Then we have

$$\begin{aligned}
MSE(g) &= E[Y - g(X)]^2 \\
&= E\{[Y - g_0(X)] + [g_0(X) - g(X)]\}^2 \\
&= E\{[Y - g_0(X)]^2\} + E\{[g_0(X) - g(X)]^2\} \\
&\quad +2E\{[Y - g_0(X)][g_0(X) - g(X)]\} \\
&= E\{[Y - g_0(X)]^2\} + E\{[g_0(X) - g(X)]^2\},
\end{aligned}$$

where the last term, namely the cross product term, is identically zero:

$$\begin{aligned}
E\{[Y - g_0(X)]\,[g_0(X) - g(X)]\}\} &= E_X\{E[(Y - g_0(X))(g_0(X) - g(X))|X]\} \\
&= E_X\{(g_0(X) - g(X))E[(Y - g_0(X))|X]\} \\
&= E_X[(g_0(X) - g(X)) \cdot 0] \\
&= 0
\end{aligned}$$

by the law of iterated expectations and $E[(Y - g_0(X))|X] = 0$.

In the decomposition of MSE(g), the first term $E[Y - g_0(X)]^2$ does not depend on $g(\cdot)$. Thus, minimizing MSE(g) with respect to $g(\cdot)$ is equivalent to minimizing the second term $E[g_0(X) - g(X)]^2$, which achieves the minimum if and only if we choose $g(X) = g_0(X)$. This completes the proof.

Theorem 26 (5.34)**. [Regression Identify]:** Suppose $E(Y|X)$ exists. Then there is a random variable ε such that

$$Y = E(Y|X) + \varepsilon,$$

where ε satisfies the condition

$$E(\varepsilon|X) = 0.$$

Proof: Define $\varepsilon = Y - E(Y|X)$. Then we have $Y = E(Y|X) + \varepsilon$, where

$$\begin{aligned}
E(\varepsilon|X) &= E\{[Y - E(Y|X)]X\} \\
&= E(Y|X) - E[E(Y|X)|X] \\
&= E(Y|X) - E(Y|X) \\
&= 0.
\end{aligned}$$

The random variable ε is called a disturbance to the regression function $E(Y|X)$. It reflects the degree of uncertainty about the relationship between Y and X. When $\varepsilon = 0$, we have a perfect deterministic relationship between Y and X. An example is $Y = g_0(X)$ for some measurable function $g_0(\cdot)$.

What does it mean by the condition of $E(\varepsilon|X) = 0$? Put it simply, it implies that ε contains no systematic information of X that can be used to predict the expected value of Y with respect to stochastic factors rather than X. All systematic information of X that can be used to predict the expected value of Y has been incorporated in $E(Y|X)$.

The regression function $E(Y|X)$ can be a linear or nonlinear function of X. When $E(Y|X) = a + bX,$ we call it a linear regression function.

We now provide two examples for which the regression function is a linear function of X.

Example 39 (5.39)**.** Let the joint PDF $f_{XY}(x, y) = e^{-y}$ for $0 < x < y < \infty$ be given as in Example 5.14 of this Chapter. Then Find the regression function $E(Y|X = x)$.

Solution: In Example 5.14 of this chapter, we have shown that for any given $x > 0$,

$$f_{Y|X}(y|x) = e^{-(y-x)} \text{ for } y \in (x, \infty).$$

It follows that

$$\begin{aligned} E(Y|x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \\ &= \int_{x}^{\infty} y e^{-(y-x)} dy \\ &= e^{x} \int_{x}^{\infty} y e^{-y} dy \\ &= -e^{x} \int_{x}^{\infty} y de^{-y} \\ &= 1 + x. \end{aligned}$$

Example 40 (5.40)**.** [**Bivariate Normal Distribution**]: Suppose (X, Y) follows a bivariate normal distribution, $\text{BN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then the conditional mean

$$E(Y|X) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(X - \mu_1)$$

which is a linear function of X.

When $E(Y|X) = a + bX$, we have a linear regression equation

$$Y = a + bX + \varepsilon,$$

where $E(\varepsilon|X) = 0$.

Lemma 7 (5.35)**.** Suppose $Y = a + bX + \varepsilon$, where $E(\varepsilon|X) = 0$. Then $E(X\varepsilon) = 0$.

Proof: By the law of interacted expectations in Theorem 5.32, we have

$$\begin{aligned} E(X\varepsilon) &= E[E(X\varepsilon|X)] \\ &= E[XE(\varepsilon|X)] \\ &= E(X \cdot 0) \\ &= 0. \end{aligned}$$

This implies $\text{cov}(X, \varepsilon) = 0$. (Why?)

Since ε has no information on X that can be used to predict the expected value of Y, it should be orthogonal to X. By orthogonality, we mean that ε and X are uncorrelated, i.e., $\text{cov}(X, \varepsilon) = 0$. In fact, by the law of iterated expectations, $E(\varepsilon|X) = 0$ implies that ε is orthogonal to any measurable function $h(X)$ of X, that is, $E[\varepsilon h(X)] = 0$.

There is an important difference between $E(\varepsilon|X) = 0$ and $E(X\varepsilon) = 0$. Although $E(\varepsilon|X) = 0$ implies $E(X\varepsilon) = 0$, the converse is not true. This can be seen from the example below.

Example 41 (5.41)**.** Suppose $\varepsilon = (X^2 - 1) + u$, where X and u are independent $N(0, 1)$ random variables. Then

$$\begin{aligned} E(\varepsilon|X) &= X^2 - 1 + E(u|X) \\ &= X^2 - 1 \end{aligned}$$

where $E(u|X) = E(u) = 0$. On the other hand,

$$\begin{aligned} E(X\varepsilon) &= E(X^3 - X + Xu) \\ &= 0. \end{aligned}$$

In this example, ε contains a predictable (nonlinear) component $X^2 - 1$ so that $E(\varepsilon|X) \neq 0$ but ε is orthogonal to X or any linear function of X. This result has a profound implication. Suppose

$$Y = a + bX + \varepsilon,$$

where $E(X\varepsilon) = 0$. Then it does not necessarily imply that the linear model $g(X) = a + bX$ is the optimal predictor for Y in terms of the mean squared error criterion, because $E(X\varepsilon) = 0$ does not imply $E(\varepsilon|X) = 0$.

The concept of the conditional mean or regression function has wide applications in economics and finance. Below are a few examples in economics and finance.

Example 42 (5.42)**. [Consumption Function and Marginal Propensity to Consume]:** Suppose Y is consumption, X is income, and

$$Y = \alpha + \beta X + \varepsilon,$$

where ε represents other stochastic factors that affect consumption, with $E(\varepsilon|X) = 0$. The conditional expectation

$$E(Y|X) = \alpha + \beta X$$

is called a consumption function and the derivative

$$\frac{dE(Y|X)}{dX} = \beta$$

is the expected marginal propensity to consume; that is, the expected increase in consumption when income is increased by one unit. This is the most important concept in the Keynesian macroeconomic theory.

Example 43 (5.43)**. [Conditional Mean and Efficient Market Hypothesis (EMH)]:** Suppose Y_t is the asset return in time period t, and I_{t-1} is an information available at time $t-1$. Suppose one attempts to use the available information I_{t-1} to predict the asset return Y_t, but the expected return $E(Y_t|I_{t-1})$ using the information I_{t-1} is the same as the long-run market average return $E(Y_t)$. Then the information I_{t-1} has no predictive power for the future return Y_t, and we say that the asset market is efficient with respect to the information set I_{t-1}. Formally, such an hypothesis can be stated as follows:

$$E(Y_t|I_{t-1}) = E(Y_t).$$

There are three forms of the efficient market hypothesis:

- The weak form EMH, where I_{t-1} only contains the information of historical asset returns available at time $t-1$;
- The semi-strong form EMH, where I_{t-1} contains all publicly available information at time $t-1$;
- The strong form EMH, where I_{t-1} contains not only publicly available information but also some inside information at time $t-1$.

Example 44 (5.44)**. [Expected Shortfall and Financial Risk Management]:** In Example 3.33 of Chapter 3, we have introduced the concept of value at risk of a portfolio over a certain time period. The value of risk $V_t(\alpha)$ at level α is defined as

$$P[X_t < -V_t(\alpha)|I_{t-1}] = \alpha,$$

where X_t is the return on the portfolio in time period t, and I_{t-1} is the information available at time $t-1$. The value at risk $V_t(\alpha)$ is the threshold level which actual loss will exceed with probability α. This has been used to determine the level of risk capital in preventing extreme adverse market events.

In practice it has been known that the risk capital level based on the value at risk may not be prudent enough. Instead, some practitioners have advocated and used the so-called expected shortfall, defined as $ES_t(\alpha) = E[X_t|X_t < -V_t(\alpha)]$, to set the level of risk capital. The expected shortfall at level α is the expected loss given a crisis has occurred (a crisis occurs when actual loss exceeds the value at risk). When α is small, the expected shortfall is smaller than the value at risk, thus providing more prudent risk management. See Figure 5.11 below:

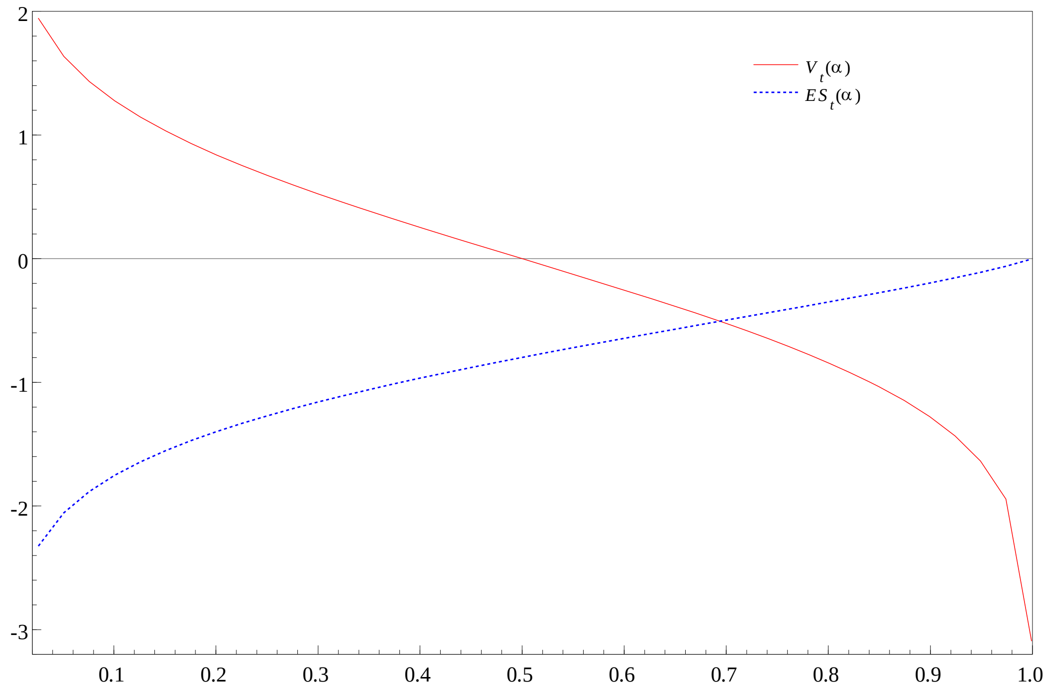

Figure 5.11: $ES_t(\alpha)$ and $V_t(\alpha)$ for $\alpha \in (0, 1)$

Example 45 (5.45)**. [Dynamic Asset Pricing Model and the Euler Equation]:** In the standard consumer based asset pricing model, the representative agent (investor) maximizes his expected lifetime utility

$$\max_{\{C_t\}} E \sum_{j=0}^{\infty} \beta^j u(C_{t+j})$$

by choosing an optimal path of consumptions $\{C_t, t = 1, 2, \cdots\}$ subject to an intertemporal budget constraint, where β is a time discount factor, and $u(\cdot)$ is the utility function of the economic agent in each time period. The first order condition for this intertemporal maximization problem is given by

$$E\left(M_{t+1} R_{t+1} | I_t\right) = 1,$$

where $M_{t+1} = \beta u'(C_{t+1})/u'(C_t)$ is called the stochastic discount factor which reflects the risk attitude of the representative economic agent, and R_{t+1} is the gross asset return from time t to time $t+1$. This FOC is called the Euler equation, which is a conditional mean characterization. Intuitively, the Euler equation says that the expected risk-compensated return in time period $t+1$ should be equal to the price of unity (the cost of investment in time period t). It characterizes the optimal path of asset investments over time or equivalently, the optimal path of consumptions over times under future uncertainty.

Next, we introduce the conditional variance of Y given X.

Definition 23 (5.23). **[Conditional Variance]:** The conditional variance of Y given X is defined as

$$\mathrm{var}(Y|X=x) = \int_{-\infty}^{\infty} [y - E(Y|X=x)]^2 dF_{Y|X}(y|x).$$

Intuitively, $\varepsilon = Y - E(Y|X)$ is the unexpected or surprise component of Y when one uses $E(Y|X)$ to predict Y. The conditional variance $\mathrm{var}(Y|X) = \mathrm{var}(\varepsilon|X)$ measures the magnitude of surprise or volatility of Y given the information X.

When $\mathrm{var}(Y|X) = \sigma^2$ is a constant, i.e., $\mathrm{var}(Y|X)$ does not depend on X, we say that there exists conditional homoskedasticity. This is an important assumption in the classical regression analysis (see, for example, Hong (2011, Chapter 3)).

In general, the conditional variance of Y given X is a function of X, i.e., it depends on X. In this case, $\mathrm{var}(Y|X) \neq \sigma^2$, and we say that there exists conditional heteroskedasticity. Conditional heteroskedasticity is a rule rather than exception. For example, large firms usually have large variations in output.

We now provide two important examples for conditional variance modeling.

Example 46 (5.46). **[Level Effect for Interest Rate Volatility]:** It is well known that there exists an important empirical stylized fact about the interest rate volatility: the interest rate volatility depends on the interest rate level. That is, the higher the interest rate level, the higher the interest rate volatility, as displayed in Figure 5.12 for the U.S. short-term interest rates. This phenomena is called the level effect of interest rate volatility and it is often modeled as

$$\mathrm{var}(r_t|I_{t-1}) = \alpha r_{t-1}^{\rho},$$

where r_t is the short-term interest rate in time t, and I_{t-1} is the information available at time $t-1$. See CKLS (1992) for more discussion.

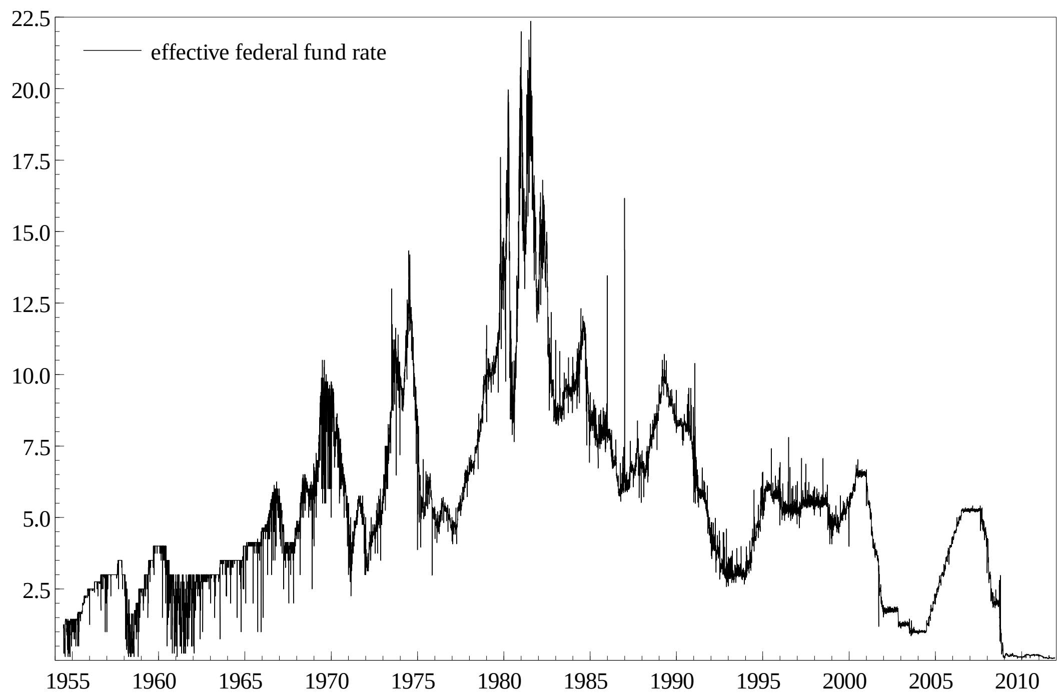

Figure 5.12: U.S. Short-term Interest Rates

Example 47 (5.47)**. [ARCH Models and Volatility Clustering]:** It is often observed in financial markets that a large volatility of an asset price today tends to be followed by another large volatility tomorrow, and a small volatility today tends to be followed by another small volatility tomorrow, as shown in Figure 1.2 for S&P500 daily returns. This phenomenon is called volatility clustering (Mandelbrot (1963)). To explain this stylized fact, Engle (1982) proposes a class of AutoRegressive Conditional Heteroskedasticity (ARCH) models to predict volatility. Suppose Y_t is the asset return in time t. Then the ARCH(1) model assumes

$$\mathrm{var}(Y_t|I_{t-1}) = \alpha + \beta Y_{t-1}^2,$$

where $\alpha, \beta > 0$, and I_{t-1} contains information on all past asset returns.

We now state a formula which provides a convenient way to calculate the conditional variance.

Lemma 8 (5.36)**.** :

$$\mathrm{var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2 .$$

Proof: This is left as an exercise.

This is a conditional version of the variance formula in Theorem 3.13.

Example 48 (5.48)**.** Let $f_{XY}(x, y) = e^{-y}$ for $0 < x < y < \infty$ be given as in Example 5.39, where we have shown $E(Y|x) = 1 + x$. Then

$$\begin{aligned}
\mathrm{var}(Y|x) &= E(Y^2|x) - [E(Y|x)]^2 \\
&= \int_x^{\infty} y^2 e^{-(y-x)} dy - (1+x)^2 \\
&= e^x \int_x^{\infty} y^2 e^{-y} dy - (1+x)^2 \\
&= -e^x \int_x^{\infty} y^2 de^{-y} - (1+x)^2 \text{ where } de^{-y} = -e^{-y} dy. \\
&= -e^x \left(y^2 e^{-y}|_x^{\infty} - \int_x^{\infty} e^{-y} dy^2 \right) - (1+x)^2 \\
&= -e^x \left(0 - x^2 e^{-x} - 2\int_x^{\infty} ye^{-y} dy \right) - (1+x)^2 \\
&= x^2 + 2e^x \int_x^{\infty} ye^{-y} dy - (1+x)^2 \\
&= x^2 + 2\int_x^{\infty} ye^{-(y-x)} dy - (1+x)^2 \\
&= x^2 + 2(1+x) - (1+x)^2 \\
&= 1.
\end{aligned}$$

There exists conditional homoskedasticity in this example. This is shown below.

Example 49 (5.49)**. [Bivariate Normal Distribution and Conditional Homoskedasticity]:** Suppose X and Y follow a bivariate normal distribution $BN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Show the conditional variance $\text{var}(Y|X) = \sigma_2^2(1-\rho^2)$.

Solution: In Section 5.6, we have shown that for the bivariate normal distribution, the conditional distribution of Y given X is a normal distribution with mean $\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X - \mu_1)$ and variance $\sigma_2^2(1-\rho^2)$. Therefore, $\text{var}(Y|X) = \sigma_2^2(1-\rho^2)$, which does not depend on X.

We now provide some examples of conditional heteroskedasticity.

Example 50 (5.50)**.** Suppose $Y = Z\sqrt{1+X^2}$, where Z is a random variable with mean 0 and variance 1, and is independent of X. Find (1) $E(Y|X)$; (2) $\text{var}(Y|X)$.

Solution: (1)

$$\begin{aligned} E(Y|X) &= E(Z\sqrt{1+X^2}|X) \\ &= \sqrt{1+X^2}E(Z|X) \\ &= \sqrt{1+X^2}E(Z) \\ &= 0 = E(Y). \end{aligned}$$

(2)

$$\begin{aligned} \text{var}(Y|X) &= E(Y^2|X) - [E(Y|X)]^2 \\ &= E(Y^2|X) \\ &= E[Z^2(1+X^2)|X] \\ &= (1+X^2)E(Z^2|X) \\ &= (1+X^2)E(Z^2) \\ &= 1+X^2. \end{aligned}$$

Example 51 (5.51)**. [Random Coefficient Model]**:Suppose

$$\begin{aligned} Y &= \alpha + \beta X + u_3 \\ &= (\alpha_0 + u_1) + (\beta_0 + u_2)X + u_3, \end{aligned}$$

where u_1, u_2, u_2, X are jointly independent, and $E(u_i) = 0$ for $i = 1, 2, 3$. This is called a random coefficient model. Find (1) $E(Y|X)$; (2) $\text{var}(Y|X)$.

Solution: (1) Wr first write

$$Y = \alpha_0 + \beta_0 X + \varepsilon,$$

where

$$\varepsilon = u_1 + u_2 X + u_3.$$

Since $E(\varepsilon|X) = 0$, we have

$$E(Y|X) = \alpha_0 + \beta_0 X.$$

(2)

$$
\begin{aligned}
\text{var}(Y|X) &= \text{var}(\varepsilon|X) \\
&= \sigma_{u_1}^2 + \sigma_{u_2}^2 X^2 + \sigma_{u_3}^2 \\
&= (\sigma_{u_1}^2 + \sigma_{u_3}^2) + \sigma_{u_2}^2 X^2.
\end{aligned}
$$

Theorem 27 (5.37)**. [Variance Decomposition]:** For any two random variables X and Y with finite second moments,

$$\text{var}(Y) = \text{var}[E(Y|X)] + E[\text{var}(Y|X)].$$

Proof: Putting $g_0(X) = E(Y|X)$, we have $Y = g_0(X) + \varepsilon$. It follows that

$$
\begin{aligned}
\text{var}(Y) &= \text{var}[g_0(X) + \varepsilon] \\
&= \text{var}[g_0(X)] + \text{var}(\varepsilon) + 2\text{cov}[g_0(X), \varepsilon] \\
&= \text{var}[g_0(X)] + \text{var}(\varepsilon),
\end{aligned}
$$

where

$$
\begin{aligned}
\text{cov}[g_0(X), \varepsilon] &= E[g_0(X)\varepsilon] - E[g_0(X)]E(\varepsilon) \\
&= 0
\end{aligned}
$$

by the law of iterated expectations in Theorem 5.32 and $E(\varepsilon|X) = 0$. Since the first term $\text{var}[g_0(X)] = \text{var}[E(Y|X)]$, it suffices to show $\text{var}(\varepsilon) = E[\text{var}(Y|X)]$. By $E(\varepsilon|X) = 0$ and the law of iterated expectations, we have $E(\varepsilon) = 0$ and

$$
\begin{aligned}
\text{var}(\varepsilon) &= E(\varepsilon^2) \\
&= E[E(\varepsilon^2|X)] \\
&= E[\text{var}(Y|X)],
\end{aligned}
$$

where $\text{var}(Y|X) = E(\varepsilon^2|X)$ by definition.

Intuitively, the variance decomposition theorem states that the total variation of Y, $\text{var}(Y)$, is equal to the sum of two components: the first is the variability of the best MSE predictor $E(Y|X)$, which measures how well $E(Y|X)$ can predict Y. The more $E(Y|X)$ can vary, the better it can predict Y. The second component, $E[\text{var}(Y|X)]$, is the averaged squared prediction error $Y - E(Y|X)$. Since $\text{var}(Y)$ is a constant, increasing the variability of the best predictor $E(Y|X)$ will decrease the average of squared prediction error. In the best case, when $Y = g(X)$ for some measurable function $g(\cdot)$, we have $E(Y|X) = Y$. In this case, $E(Y|X)$ can perfectly predict the variation of Y, and there is no prediction error. On the other hand, if X is independent of Y, then we have $E(Y|X) = E(Y)$, a constant. In this case, there is no variation in $E(Y|X)$, and the mean squared prediction error achieves its maximum value $\text{var}(Y)$.

In econometrics, the first two conditional moments are most important. However, there has been increasing interest in higher order conditional moments in economic and financial applications. Below are two examples:

- Conditional Skewness:

$$S(Y|X) = \frac{E(\varepsilon^3|X)}{[\text{var}(\varepsilon|X)]^{3/2}}.$$

- Conditional Kurtosis:

$$K(Y|X) = \frac{E(\varepsilon^4|X)}{[\text{var}(\varepsilon|X)]^2}.$$

Example 52 (5.52)**.** Suppose (X, Y) follow a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Find $E(Y|X)$, $\text{var}(Y|X)$, $S(Y|X)$ and $K(Y|X)$.

Solution: We have shown that the conditional distribution of Y given X is a normal distribution

$$Y|X \sim N\left[\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X - \mu_1), \sigma_2^2(1 - \rho^2)\right].$$

It follows that the conditional mean

$$E(Y|X) = \mu_2 + \frac{\sigma_2}{\sigma_1}\rho(X - \mu_1),$$

the conditional variance

$$\text{var}(Y|X) = \sigma_2^2(1 - \rho^2),$$

the conditional skewness

$$S(Y|X) = 0,$$

and the conditional kurtosis

$$K(Y|X) = 3.$$

This example shows that for the bivariate normal distribution, only the first conditional moment $E(Y|X)$ depends on X, all other conditional higher order comments are constant, i.e., do not depend on X.

On the other hand, it may be possible that the lower order conditional moments are constant, yet the higher order conditional moments will depend on X.

Example 53 (5.53)**.** Suppose the conditional distribution of Y given X is a $\text{Lognormal}(0, X^2)$ distribution. Then conditional on X, $\ln(Y)$ follows a $N(0, X^2)$ distribution. It follows that

$$E(Y^k|X) = e^{\frac{k^2}{2}X^2}, \qquad k = 1, 2, \cdots.$$

Put $\mu(X) = E(Y|X)$ and $\sigma^2(X) = \text{var}(Y|X)$. Define the standardized random variable

$$Z = \frac{Y - \mu(X)}{\sigma(X)}.$$

It can be shown that $E(Z|X) = 0$, $\text{var}(Z|X) = 1$, but $E(Z^3|X)$ is a function of X (please check it!). In this example, the first two conditional moments of Z do not depend on X, but all other higher order conditional moments of Z are functions of X.

In the practice of econometric modeling, there is an important question, namely, which conditional moment should be used in practice?

This depends on the economic application we have in mind. For some applications such as study on the market efficiency hypothesis and dynamic asset pricing, we need to model the conditional mean of economic variables. For applications such as studies on volatility spillover, we need to model the conditional variance. For applications such as financial risk management (e.g., managing financial crisis), hedging, and derivatives pricing, we need to model higher order conditional moments or even the entire conditional distribution.

More generally, we can consider the problem of decision under uncertainty. Suppose an economic agent has a loss function $l(e)$ and he is making an optimal decision a to minimize the expected loss conditional on the available information $X = x$ he has. The minimization problem is

$$\min_a E\left[l(Y-a)|X=x\right] = \min_a \int_{-\infty}^{\infty} l(y-a) f_{Y|X}(y|x)dx,$$

where Y is a random payoff which is unknown when the economic agent is making a decision, and $f_{Y|X}(y|x)$ is the conditional PDF of Y given $X = x$. In practice, the conditional PDF $f_{Y|X}(y|x)$ is usually unknown and need to be modeled. When the loss function is a quadratic loss, i.e.,

$$l(e) = e^2,$$

the optimal decision is the conditional mean:

$$a^*(X) = E(Y|X).$$

When the loss function is a so-called linexp function

$$l(e) = \frac{1}{\alpha^2}\left[e^{\alpha e} - (1+\alpha e)\right],$$

and the conditional distribution of Y given X is a normal distribution, the optimal decision is a linear combination of the conditional mean and conditional variance:

$$a^*(X) = E(Y|X) + \frac{\alpha}{2}\text{var}(Y|X).$$

The linexp loss function is asymmetric except for $\alpha = 0$, which yields the squared loss function. See Figure 5.13 for the shapes of the linexp loss function for various values of α.

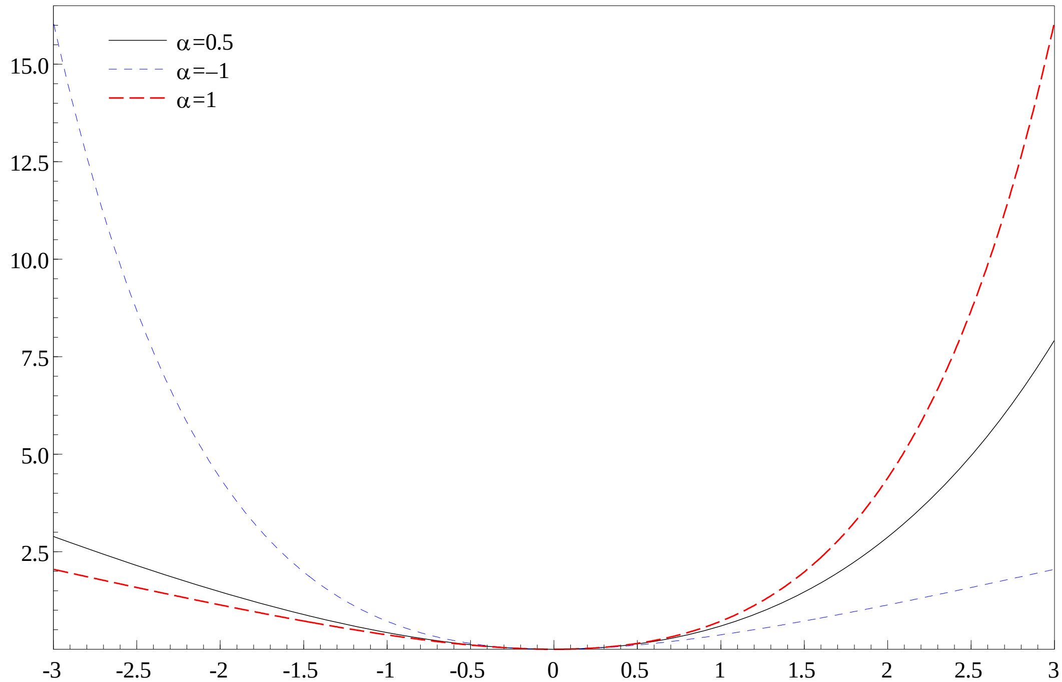

Figure 5.13: The Linexp Loss Function for Various Values of α

More generally, if the loss function $l(e)$ is a generic loss function, then the optimal decision $a^*(X)$ will depend on the entire conditional distribution of Y given X. In this case, one need to model the conditional PDF $f_{Y|X}(y|x)$; the first few conditional moments are not sufficient. For more discussion, see Granger (1999).

5.11 Conclusion

The most important goal in economic analysis and econometric analysis is to identify various economic relationships. In this chapter, we first characterize the joint probability distribution of two random variables X and Y by using the joint CDF, the joint PMF when (X, Y) are discrete, the joint PDF when (X, Y) are continuous, as well as their conditional counterparts. we also consider the methods of deriving the joint distribution of a bivariate transformation. We then examine the predictive relationships between X and Y using the conditional distributions, correlation, and conditional expectations. The concept of independence and its implications on the joint distributions, conditional distributions, correlation and joint moment generating functions are also fully discussed. We also introduce a class of bivariate normal distribution and examine its important properties. Throughout this chapter, economic interpretations and examples are provided for most of the probability concepts introduced.

EXERCISE 5

5.1. [# 4.4, p.192] A joint PDF is defined by

$$f_{XY}(x,y) = \begin{cases} c(x+2y), & \text{if } 0<y<1 \text{ and } 0<x<2, \\ 0, & \text{otherwise.} \end{cases}$$

(1) Find the value of c.
(2) Find the marginal PDF of X.
(3) Find the joint CDF of X and Y.
(4) Find the PDF of the random variable $Z = 9/(X+1)^2$.

5.2. Suppose (X, Y) has a joint pdf

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

where θ is a constant.
(1) Determine the possible value(s) of θ so that $f_{XY}(x,y)$ is a joint PDF. Give your reasoning.
(2) Let $\theta = 0$. Check if X and Y are independent. Give your reasoning.

5.3. For m random variables $X_i, i = 1, \cdots, m$, denote $F_{X_i}(x_i)$ as the marginal distribution of X_i and $F(x_1, \cdots, x_m)$ as the joint distribution of $(X_1, \cdots, X_m)$. Then the distribution function $C : [0,1]^m \to [0,1]$ that satisfies:

$$F(x_1, \cdots, x_m) = C[F_{X_1}(x_1), \cdots, F_{X_m}(x_m)], \quad x_j \in (-\infty, \infty),$$

is called the copula associate with $F(x_1, \cdots, x_m)$. Copula contains all information about dependence among components of a random vector but no information about the marginal distributions. Suppose $X = (X_1, \cdots, X_m)'$ with joint CDF $F_X(x) = P(X_1 \le x_1, \cdots, X_m \le x_m)'$ and marginal CDFs $F_{X_i}(x_i) = P(X_i \le x_i)$, where $x = (x_1, \cdots, x_m)'$. Assume that $F_{X_i}(\cdot)$ is strictly increasing for all $i = 1, \cdots, m$. Show:
(1) the copula of X is given by $C_X(u) = F_X[F_{X_1}^{-1}(u_1), \cdots, F_{X_m}^{-1}(u_m)]$;
(2) $F_X(x) = C_X[F_{X_1}(x_1), \cdots, F_{X_m}(x_m)]$.
(3) Suppose $Y_i = g_i(X_i), i = 1, \cdots, m$. Show $C_X(u) = C_Y(u)$ for all u. That is, the copula is invariant to strictly increasing transformation of the random variables.

5.4. [# 4.5, p.192] (1) Find $P(X > \sqrt{Y})$ if X and Y are jointly distributed with PDF $f_{XY}(x,y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$.
(2) Find $P(X^2 < Y < X)$ if X and Y are jointly distributed with PDF $f_{XY}(x,y) = 2x$ for $0 \le x \le 1, 0 \le y \le 1$.

5.5. [# 4.9, p.193] Prove that if the jointly CDF of X and Y satisfies $F_{XY}(x,y) = F_X(x)F_Y(y)$, that is, if X and Y are independent, then for any pair of intervals (a,b) and (c,d), $P(a \le X \le b, c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d)$.

5.6. [# 4.10, p.193] The random pair (X, Y) has the joint distribution

			X	
		1	2	3
	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
Y	3	$\frac{1}{6}$	0	$\frac{1}{6}$
	4	0	$\frac{1}{3}$	0

(1) Show that X and Y are dependent.
(2) Give a probability table for random variables U and V that have the same marginals as X and Y but are independent.

5.7. [# 4.14, p.194] Suppose X and Y are independent $N(0,1)$ random variables.
(1) Find $P(X^2 + Y^2 < 1)$.
(2) Find $P(X^2 < 1)$, after verifying that X^2 is distributed χ^2_1.

5.8. [# 4.17, p.194] Let X be an exponential(1) random variable, and define Y to be the integer part of $X + 1$, that is $Y = i + 1$ if and only if $i \le X < i + 1$, where $i = 0, 1, \cdots$.
(1) Find the distribution of Y. What well-known distribution does Y have?
(2) Find the conditional distribution of $X - 4$ given $Y \ge 5$.

5.9. [# 4.18, p.194] Suppose $g(x) \ge 0$ and $\int_0^\infty g(x)dx = 1$, show that $f(x, y) = \frac{2g(\sqrt{x^2+y^2})}{\pi\sqrt{x^2+y^2}}$, for $x, y > 0$, is a joint PDF.

5.10. Suppose (X, Y) has a joint PDF

$$f(x, y) = e^{-y} \text{ for } 0 < x < y < \infty.$$

Find (1) $f_X(x)$; (2) $f_Y(y)$; (3) $f_{X|Y}(x|y)$; (4) $f_{Y|X}(y|x)$; (5) are X and Y independent?

5.11. (X, Y) follows a bivariate normal distribution if their joint PDF

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]},$$

where $-\infty < \mu_1, \mu_2 < \infty, 0 < \sigma_1, \sigma_2 < \infty, -1 \le \rho \le 1$. Find (1) $f_X(x)$; (2) $f_Y(y)$; (3) $f_{Y|X}(y|x)$; (4) $f_{X|Y}(x|y)$; (5) under what conditions on parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, X and Y will be independent. [Hint: When finding $f_X(x)$, you can form a term with form

$$z^2 = \left[\left(\frac{y-\mu_2}{\sigma_2}\right) - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right]^2$$

and integrate it out first.]

5.12. Let X be $N(0, \sigma^2)$. Show that the CDF of the conditional distribution of X given $X > c$ is

$$F_{X|X>c}(x) = \frac{\Phi(x/\sigma) - \Phi(c/\sigma)}{1 - \Phi(c/\sigma)}, \ x > c$$

and the PDF of this distribution is

$$f_{X|X>c}(x) = \frac{\phi(x/\sigma)}{\sigma[1-\Phi(c/\sigma)]}, \qquad x > c,$$

where $\phi(x)$ and $\Phi(x)$ are the PDF and CDF of $N(0,1)$. Such a distribution is called a truncated distribution.

5.13. Suppose the random variables X and Y have the following joint pdf

$$f(x,y) = \begin{cases} 8xy & \text{for } 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $U = X/Y$ and $V = Y$. Determine the joint pdf of U and V.

5.14. [#4.19, p.194] (1) Let X_1 and X_2 be independent N(0,1) random variables. Find the PDF of $(X_1 - X_2)^2/2$.

(2) If $X_i, i = 1, 2$, are independent Gamma$(\alpha_i, 1)$ random variables, find the marginal distributions of $X_1/(X_1 + X_2)$ and $X_2/(X_1 + X_2)$.

5.15. Suppose X_1, X_2 are independent standard Gamma random variables, possibly with different parameters α_1, α_2. Show:

(1) The random variables

$$X_1 + X_2 \text{ and } \frac{X_1}{X_1 + X_2}$$

are mutually independent;

(2) The distribution of $X_1 + X_2$ is a standard Gamma with $\alpha = \alpha_1 + \alpha_2$.

(3) The distribution of $X_1/(X_1 + X_2)$ is a standard Beta with parameters α_1, α_2.

5.16. [#4.20, p.194] X_1 and X_2 are independent $N(0, \sigma^2)$ random variables.

(1) Find the joint distribution of Y_1 and Y_2, where $Y_1 = X_1^2 + X_2^2$ and $Y_2 = X_1/\sqrt{Y_1}$.

(2) Show that Y_1 and Y_2 are independent, and interpret this result geometrically.

5.17. [#4.23, p.195] For $X \sim Beta(\alpha, \beta)$, and $Y \sim Beta(\alpha + \beta, \gamma)$ be independent random variables, find the distribution XY by making the transformation given in (1) and (2) and integrating out V

(1) $U = XY, V = Y$

(2) $U = XY, V = X/Y$

5.18. [#4.27, p.195] Let $X \sim N(\mu, \sigma^2)$, and let $Y \sim N(\gamma, \sigma^2)$. Suppose X and Y are independent. Define $U = X + Y$ and $V = X - Y$. Show that U and V are independent normal random variables. Find the distribution of each of them.

5.19. Show that (1) if $X_1 \sim N(0, \sigma_1^2), X_2 \sim N(0, \sigma_2^2)$, and X_1 and X_2 are independent, then $X_1X_2/\sqrt{X_1^2 + X_2^2}$ is also normally distributed. (2) If in addition $\sigma_1^2 = \sigma_2^2$, then $(X_1^2 - X_2^2)/(X_1^2 + X_2^2)$ is also normally distributed.

5.20. Suppose $X_1 \sim N(0,1), X_2 \sim N(0,1)$ and X_1 and X_2 are independent. Find the distribution of (1) X_1/X_2 and (2) $X_1/|X_2|$ respectively.

5.21. Find the PDF of $X - Y$, where $X \sim U[0,1], Y \sim U[0,1]$, and X and Y are independent.

5.22. Suppose $X_1 \sim \text{Cauchy}(0,1), X_2 \sim \text{Cauchy}(0,1)$, and X_1 and X_2 are independent. Show $aX_1 + bX_2$ has Cauchy distribution.

5.23. Suppose $X_1 \sim \text{Gamma}(\alpha_1, 1), X_2 \sim \text{Gamma}(\alpha_2, 1)$, and X_1 and X_2 are independent. Show that $X_1 + X_2$ and $X_1/(X_1 + X_2)$ are independent. Also, find the marginal distributions of $X_1 + X_2$ and $X_1/(X_1 + X_2)$, respectively.

5.24. Suppose the PDF of X_i is

$$\frac{1}{\sigma_i} f\left(\frac{x - \theta_i}{\sigma_i}\right), \qquad i = 1, 2,$$

and X_1, X_2 are independent. Show that the PDF of $X_1 + X_2$ is of the form of

$$\frac{1}{\sigma} f\left(\frac{x - \theta}{\sigma}\right)$$

for some σ and θ.

5.25. Suppose X_1, X_2, X_3 have a continuous joint PDF $f(x, y, z)$. Define $Y_1 = F_1(X_1), Y_2 = F_2(X_1, X_2)$ and $Y_3 = F_3(X_1, X_2, X_3)$, where

$$\begin{aligned} F_1(x) &= P(X_1 \le x), \\ F_2(x_1, x_2) &= P(X_2 \le x_2 | X_1 = x_1), \\ F_3(x_1, x_2, x_3) &= P(X_3 \le x_3 | X_2 = x_2, X_1 = x_1). \end{aligned}$$

Show that Y_1, Y_2, Y_3 are mutually independent and each of them is uniformly distributed over [0,1]. [Hint: Firstly, define $f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_1}(x_1)}$,then $F_2(x_1, x_2) = \int_{-\infty}^{x_2} f_{X_2|X_1}(y|x_1) dy$].

5.26. [#4.30, p.195] Suppose the distribution of Y, conditional on $X = x$, is $N(x, x^2)$ and that the marginal distribution of X is uniform(0,1).

(1) Find $E(Y)$, var(Y), and cov(X, Y).

(2) Prove that Y/X and X are independent.

5.27. Consider two random variables (X, Y). Suppose X is uniformly distributed over $(-1, 1)$, that is, the pdf of X is

$$f_X(x) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Also, the conditional pdf of Y give $X = x$ is

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y - \alpha - \beta x)^2}{2}} \qquad \text{for } -\infty < y < \infty \text{ and } -1 < x < 1\ .$$

(1) Find $E(Y)$; (2) find $\text{cov}(X, Y)$.

5.28. [#4.40, p.198] A generalization of the beta distribution is the *Dirichlet* distribution. In its bivariate version, (X, Y) have a joint PDF $f_{XY}(x, y) = Cx^{a-1}y^{b-1}(1-x-y)^{c-1}$, $0 < x < 1$, $0 < y < 1$, $0 < y < 1 - x < 1$, where $a > 0$, $b > 0$, and $c > 0$ are constants.

(1) Show that $C = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}$.

(2) Show that, marginally, both X and Y are beta.

(3) Find the conditional distribution of $Y|X = x$, and show that $Y|(1 - X)$ is beta(b, c).

(4) Show that $E(XY) = \frac{ab}{(a+b+c+1)(a+b+c)}$, and find the covariance $\text{cov}(X, Y)$.

5.29. [#4.43, p.199] Let X_1, X_2 and X_3 be uncorrelated random variables, each with mean μ and variance σ^2. Find, in terms of μ and σ^2, $\text{cov}(X_1 + X_2, X_2 + X_3)$ and $\text{cov}(X_1 + X_2, X_1 - X_2)$

5.30. Suppose (X, Y) is a bivariate random vector with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2. Let $U = X + Y$ and $V = X - Y$. Show that U and V are uncorrelated if and only if $\sigma_X^2 = \sigma_Y^2$.

5.31. [#4.49(a,b,c), p.200] Behboodian (1990) illustrates how to construct bivariate random variables that are uncorrelated but dependent. Suppose that f_1, f_2, g_1, g_2 are univariate densities with means $\mu_1, \mu_2, \zeta_1, \zeta_2$, respectively, and the bivariate random variable (X, Y) has density

$$(X, Y) \sim af_1(x)g_1(y) + (1 - a)f_2(x)g_2(y)$$

where $0 < a < 1$ is known.

(1) Show that the marginal distributions are given by $f_X(x) = af_1(x) + (1 - a)f_2(x)$ and $f_X(y) = ag_1(x) + (1 - a)g_2(y)$.

(2) Show that X and Y are independent if and only if $[f_1(x) - f_2(x)][g_1(y) - g_2(y)] = 0$.

(3) Show that $\text{cov}(X, Y) = a(1 - a)(\mu_1 - \mu_2)(\zeta_1 - \zeta_2)$, and thus explain how to construct dependent uncorrelated random variables.

(4) Letting f_1, f_2, g_1, g_2 be binomial PMFs, give examples of combinations of parameters that lead to independent (X, Y) pairs, correlated (X, Y) pairs, and uncorrelated but dependent (X, Y) pairs.

5.32. [#4.50, p.201] Suppose (X, Y) has a bivariate normal PDF

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)}.$$

Show that $\text{corr}(X, Y) = \rho$, and $\text{corr}(X^2, Y^2) = \rho^2$. (Hint: Conditional expectation will simplify calculations.)

5.33. Suppose (X, Y) follows a standard bivariate normal distribution with correlation coefficient ρ. Define $U = (Y - \rho X)/\sqrt{1 - \rho^2}$. Show that U is normally distributed and independent of X.

5.34. Show $\text{var}(Y|X) = E(Y^2|X) - [E(Y|X)]^2$.

5.35. Suppose the joint PDF of X, Y is a uniform PDF on the circle $x^2 + y^2 \leq 1$. Find (1) $E(Y|X)$; (2) $\text{var}(Y|X)$.

5.36. Suppose (X, Y) have a joint Normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Find (1) $E(Y|X)$; (2) $\text{var}(Y|X)$.

5.37. Show that if $X = (X_1, \cdots, X_m)'$ follows a multivariate normal distribution with mean vector $\mu = E(X) = (\mu_1, \cdots, \mu_m)'$ and variance-covariance matrix $\Sigma = \text{cov}(X, X)$, then for any $\lambda = (\lambda_1, \cdots, \lambda_m)'$ with $\lambda'\lambda = 1$, $\lambda' X$, has a normal distribution with mean $\lambda'\mu$ and variance $\lambda'\Sigma\lambda$.

5.38. Suppose X and Y are random variables such that $E(Y|X) = 7 - \frac{1}{4}X$ and $E(X|Y) = 10 - Y$. Determine the correlation between X and Y.

5.39. Suppose $E(Y|X) = 1 + 2X$ and $\text{var}(X) = 2$. Find $\text{cov}(X, Y)$.

5.40. Suppose $Y = \alpha_0 + \alpha_1 X + \varepsilon\sqrt{\beta_0 + \beta_1 X^2}$, where ε and X are mutually independent random variables, with $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = 1$. Find (1) $E(Y|X)$; (2) $\text{var}(Y|X)$.

5.41. Suppose (X, Y) have a joint distribution such that $E(Y^2) < \infty$ and $0 < \text{var}(X) < \infty$. Let

$$\mathbb{A} = \{g : \mathbb{R} \to \mathbb{R} | g(x) = \alpha + \beta x, \qquad -\infty < \alpha, \beta < \infty\}$$

be a class of linear functions.

(1) Show that $g^*(X) = \alpha^* + \beta^* X$ is the optimal solution to $\min_{g \in \mathbb{A}} E\left[Y - g(x)\right]^2$ if and only if $E(u^*) = 0$ and $E(Xu^*) = 0$, where $u^* = Y - g^*(X)$.

(2) Find the expressions for α^* and β^* in terms of $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ and $\text{cov}(X, Y)$.

5.42. Let X and Y be two random variables and $0 < \sigma_X^2 < \infty$. Show that if $E(Y|X) = \alpha_o + \alpha_1 X$, then $\alpha_1 = \text{cov}(X, Y)/\sigma_X^2$.

5.43. Suppose $E(Y|X)$ is a linear function of X, i.e., $E(Y|X) = a + bX$ for some constants a, b. Find the expressions of a and b in terms of $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ and $\text{cov}(X, Y)$. Is it true that when $E(Y|X)$ is linear in X, then $E(Y|X)$ does not depend on X if and only if $\text{cov}(X, Y) = 0$? Give your reasoning.

5.44. Two random variables X and Y has a joint PDF

$$f_{XY}(x, y) = \frac{1}{\alpha + \beta x} e^{-\frac{y}{\alpha + \beta x}} \text{ for } 0 < y < \infty, 0 < x < 1,$$

where $0 < \alpha < \infty, 0 < \beta < \infty$ are two given constants.

(1) Find the conditional PDF $f_{Y|X}(y|x)$.

(2) Find the conditional mean $E(Y|X)$.

(3) Are (X, Y) independent? Give your reasoning.

5.45. Suppose (X, Y) have a joint PDF

$$f_{XY}(x, y) = \begin{cases} xe^{-y}, & \text{if } 0 < x < y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

(1) Find the conditional pdf $f_{Y|X}(y|x)$ of Y given $X = x$.
(2) Find the conditional mean $E(Y|x)$.
(3) Find the conditional variance $V(Y|x)$.
(4) Are X and Y independent? Give your reasoning.

5.46. Suppose (X, Y) is a bivariate random vector whose conditional mean $E(Y|X)$ has the form

$$E(Y|X) = \alpha_0 + \alpha_1 X + \alpha_2 X^2,$$

where $X \sim N(0, 1)$, and $\alpha_0, \alpha_1, \alpha_2$ are some given constants.
(1) Find the mean of Y.
(2) Suppose $\alpha_1 = \alpha_2 = 0$, is it always true that $\text{cov}(X, Y) = 0$? Give your reasoning.
(3) Suppose $\text{cov}(X, Y) = 0$, is it always true that $\alpha_1 = \alpha_2 = 0$? Give your reasoning.

5.47. Suppose (X, Y) is a bivariate vector whose conditional mean $E(Y|X)$ has the form

$$E(Y|x) = \alpha_o + \alpha_1 x + \alpha_2 x^2,$$

where $E(X) = 0$, $V(X) > 0$, and $\alpha_o, \alpha_1, \alpha_2$ are some constants.
(1) If $E(Y|x) = \alpha_o$ for all x, is it always true that $Cov(X, Y) = 0$? Give your reasoning.
(2) If $Cov(X, Y) = 0$, is it always true that $E(Y|x) = \alpha_o$ for all x? Give your reasoning.

5.48. [#4.58, page 202] For any two random variables X and Y with finite variances, prove that
(1) $\text{cov}(X, Y) = \text{cov}(X, E(Y|X))$.
(2) X and $Y - E(Y|X)$ are uncorrelated.
(3) $\text{var}[Y - E(Y|X)] = E[\text{var}(Y|X)]$.

5.49. (1) Suppose $E(Y|X) = E(Y)$. Show $\text{cov}(X, Y) = 0$. (2) Does $\text{cov}(X, Y) = 0$ imply $E(Y|X) = E(Y)$. If yes, prove it. If not, provide an example.

5.50. Suppose Y is a Bernoulli random variable, and X is a random variable. Show Y and X are independent if and only if $E(Y|X) = E(Y)$.

5.51. Suppose $X_1, \cdot\cdot\cdot, X_n$ are n IID random variables with variance σ^2. Denote $\bar{X}_n = n^{-1}\Sigma_{i=1}^n X_i$. Prove that for any integers i and j $(1 \le i < j \le n)$, the correlation coefficient between $X_i - \bar{X}_n$ and $X_j - \bar{X}_n$ is $-\frac{1}{n-1}$.

5.52. Suppose X has a probability density function

$$f(x) = \begin{cases} |x| & -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = X^2$.
(1) Find $Cov(X, Y)$;
(2) Are X and Y independent? Explain.

5.53. Suppose $\{X_i\}_{i=1}^n$ is an IID sequence from a $N(\mu, \sigma^2)$ distribution. Define $\bar{X}_n = n^{-1}\Sigma_{i=1}^n X_i$. Show that for all n, $\bar{X}_n$ and $g(X_1 - \bar{X}_n, \cdots, X_n - \bar{X}_n)$ are independent, where $g(\cdot, \cdots, \cdot)$ is any measurable function.

5.54. Suppose $X \sim \text{Poisson}(\lambda_1), Y \sim \text{Poisson}(\lambda_2)$, and X and Y are independent. Show:
(1) $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$;
(2) The conditional distribution of X given $X + Y = n$, is a binomial distribution $B(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

5.55. We have one unit of capital to invest in two bonds A and B. If we invest w_1 in bond A and the remaining $w_2 = 1 - w_1$ in bond B, then (w_1, w_2) constitutes a portfolio. Denote the return of bond A and B as random variable X with mean μ_X, variance σ_X^2 and random variable Y with mean μ_Y, variance σ_Y^2 respectively. The correlation coefficient between X and Y is ρ.
(1) Find the average return and risk of portfolio (w_1, w_2).
(2) Find the portfolio (w_1^*, w_2^*) which minimize the investment risk.