CHAPTER 3

Random Variables, Vectors, and Sequences

So far we have dealt with deterministic signals, that is, signals whose amplitude is uniquely specified by a mathematical formula or rule. However, there are many important examples of signals whose precise description (i.e., as deterministic signals) is extremely difficult, if not impossible. As mentioned in Section 2.1, such signals are called *random signals*. Although random signals are evolving in time in an unpredictable manner, their average properties can be often assumed to be deterministic; that is, they can be specified by explicit mathematical formulas. This is the key for the modeling of a random signal as a stochastic process.

Our aim in the subsequent discussions is to present some basic results from the theory of random variables, random vectors, and discrete-time stochastic processes that will be useful in the chapters that follow. We assume that most readers have some basic knowledge of these topics, and so parts of this chapter may be treated as a review exercise. However, some specific topics are developed in greater depth with a viewpoint that will serve as a foundation for the rest of the book. A more complete treatment can be found in Papoulis (1991), Helstrom (1992), and Stark and Woods (1994).

3.1 RANDOM VARIABLES

The concept of random variables begins with the definition of probability. Consider an experiment with a finite or infinite number of unpredictable outcomes from a *universal set*, denoted by $\mathcal{S} = \{\zeta_1, \zeta_2, \ldots\}$. A collection of subsets of $\mathcal{S}$ containing $\mathcal{S}$ itself and that is closed under countable set operations is called a σ *field* and denoted by $\mathcal{F}$. Elements of $\mathcal{F}$ are called *events*. The unpredictability of these events is measured by a nonnegative set function $\Pr\{\zeta_k\}$, $k = 1, 2, \ldots$, called the *probability* of event ζ_k. This set function satisfies three well-known and intuitive axioms (Papoulis 1991) such that the probability of any event produced by set-theoretic operations on the events of $\mathcal{S}$ can be uniquely determined. Thus, any situation of random nature, abstract or otherwise, can be studied using the axiomatic definition of probability by defining an appropriate probability space $(\mathcal{S}, \mathcal{F}, \Pr)$.

In practice it is often difficult, if not impossible, to work with this probability space for two reasons. First, the basic space contains abstract events and outcomes that are difficult to manipulate. In engineering applications, we want random outcomes that can be measured and manipulated in a meaningful way by using numerical operations. Second, the probability function $\Pr\{\cdot\}$ is a set function that again is difficult, if not impossible, to manipulate by using calculus. These two problems are addressed through the concept of the random variable.

DEFINITION 3.1 (RANDOM VARIABLE). A random variable $x(\zeta)$ is a mapping that assigns a real number x to every outcome ζ from an abstract probability space. This mapping should satisfy the following two conditions: (1) the interval $\{x(\zeta) \leq x\}$ is an event in the abstract probability space for every x; (2) $\Pr\{x(\zeta) = \infty\} = 0$ and $\Pr\{x(\zeta) = -\infty\} = 0$.

A complex-valued random variable is defined by $x(\zeta) = x_{\mathrm{R}}(\zeta) + jx_{\mathrm{I}}(\zeta)$ where $x_{\mathrm{R}}(\zeta)$ and $x_{\mathrm{I}}(\zeta)$ are real-valued random variables. We will discuss complex-valued random variables in Section 3.2. Strictly speaking, a random variable is neither random nor a variable but is a function or a mapping. As shown in Figure 3.1, the domain of a random variable is the universal set $\mathcal{S}$, and its range is the real line $\mathbb{R}$. Since random variables are numbers, they can be added, subtracted, or manipulated otherwise.

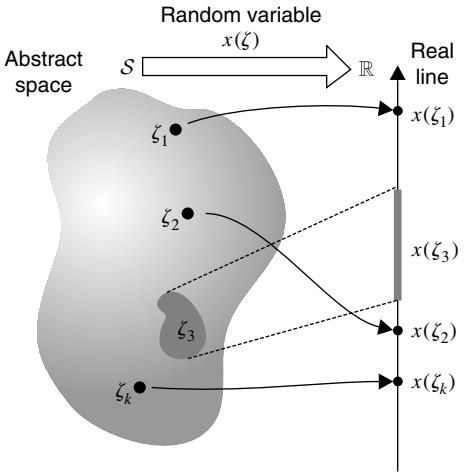


FIGURE 3.1
Graphical illustration of random variable mapping.

An important comment on notation. We will use $x(\zeta), y(\zeta), \ldots,$ to denote random variables and the corresponding lowercase alphabet without parentheses to denote their values; for example, $x(\zeta) = x$ means that the random variable $x(\zeta)$ takes value equal to x. We believe that this notation will not cause any confusion because the meaning of the lowercase variable will be clear from the context.[†] A specific value of the random variable realization will be denoted by $x(\zeta_0) = x_0$ (corresponding to a particular event ζ_0 in the original space).

A random variable is called *discrete-valued* if x takes a discrete set of values $\{x_k\}$; otherwise, it is termed a *continuous-valued* random variable. A *mixed* random variable takes both discrete and continuous values.

3.1.1 Distribution and Density Functions

The probability set function $\Pr\{x(\zeta) \leq x\}$ is a function of the set $\{x(\zeta) \leq x\}$, but it is also a number that varies with x. Hence it is also a function of a point x on the real line $\mathbb{R}$. This point function is the well-known *cumulative distribution function* (cdf) $F_x(x)$ of a random variable $x(\zeta)$ and is defined by

$$F_x(x) \triangleq \Pr\{x(\zeta) \leq x\} \tag{3.1.1}$$

The second important probability function is the *probability density function* (pdf) $f_x(x)$,

[†]Traditionally, the uppercase alphabet is used to denote random variables. We have reserved the use of uppercase alphabet for transform-domain quantities.

which is defined as a formal derivative

$$f_x(x) \triangleq \frac{\mathrm{d}F_x(x)}{\mathrm{d}x} \tag{3.1.2}$$

Note that the pdf $f_x(x)$ is not the probability, but must be multiplied by a certain interval Δx to obtain a probability, that is,

$$f_x(x)\Delta x \approx \Delta F_x(x) \triangleq F_x(x + \Delta x) - F_x(x) = \Pr\{x < x(\zeta) \leq x + \Delta x\} \tag{3.1.3}$$

Integrating both sides of (3.1.2), we obtain

$$F_x(x) = \int_{-\infty}^{x} f_x(\nu)\,\mathrm{d}\nu \tag{3.1.4}$$

For discrete-valued random variables, we use the *probability mass function* (pmf) p_k, defined as the probability that random variable $x(\zeta)$ takes a value equal to x_k, or

$$p_k \triangleq \Pr\{x(\zeta) = x_k\} \tag{3.1.5}$$

These probability functions satisfy several important properties (Papoulis 1991), such as

$$0 \leq F_x(x) \leq 1 \qquad F_x(-\infty) = 0 \qquad F_x(\infty) = 1 \tag{3.1.6}$$

$$f_x(x) \geq 0 \qquad \int_{-\infty}^{\infty} f_x(x)\,\mathrm{d}x = 1 \tag{3.1.7}$$

Using these functions and their properties, we can compute the probabilities of any event (or interval) on $\mathbb{R}$. For example,

$$\Pr\{x_1 < x(\zeta) \leq x_2\} = F_x(x_2) - F_x(x_1) = \int_{x_1}^{x_2} f_x(x)\,\mathrm{d}x \tag{3.1.8}$$

3.1.2 Statistical Averages

To completely characterize a random variable, we have to know its probability density function. In practice, it is desirable to summarize some of the key aspects of a density function by using a few numbers rather than to specify the entire density function. These numbers, which are called *statistical averages* or *moments*, are evaluated by using the mathematical expectation operation. Although density functions are needed to theoretically compute moments, in practice, moments are easily estimated without the explicit knowledge of density functions.

Mathematical expectation

This is one of the most important operations in the theory of random variables. It is generally used to describe various statistical averages, and it is also needed in estimation theory. The *expected* or *mean value* of a random variable $x(\zeta)$ is given by

$$E\{x(\zeta)\} \triangleq \mu_x = \begin{cases} \displaystyle\sum_k x_k p_k & x(\zeta) \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} x f_x(x)\,\mathrm{d}x & x(\zeta) \text{ continuous} \end{cases} \tag{3.1.9}$$

Although, strictly speaking, to compute $E\{x(\zeta)\}$ we need the definitions for both the discrete and continuous random variables, we will follow the engineering practice of using the expression for the continuous random variable (which can also describe a discrete random variable if we allow impulse functions in its pdf). The expectation operation computes a statistical average by using the density $f_x(x)$ as a weighting function. Hence, the mean μ_x can be regarded as the "location" (or the "center of gravity") of the density $f_x(x)$, as shown in Figure 3.2(*a*). If $f_x(x)$ is symmetric about $x = a$, then $\mu_x = a$ and, in particular, if $f_x(x)$

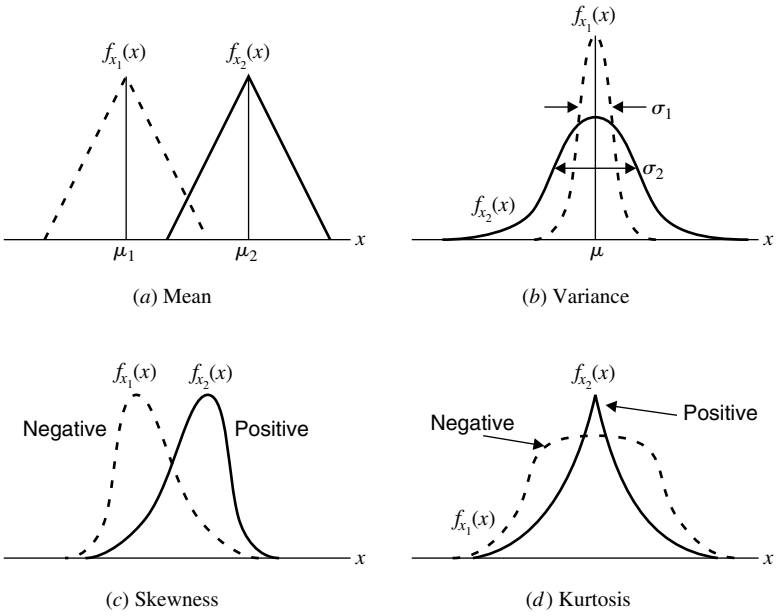


FIGURE 3.2
Illustration of mean, standard deviation, skewness, and kurtosis.

is an even function, then $\mu_x = 0$. One important property of expectation is that it is a linear operation, that is,

$$E\{\alpha x(\zeta) + \beta\} = \alpha\mu_x + \beta \tag{3.1.10}$$

Let $y(\zeta) = g[x(\zeta)]$ be a random variable obtained by transforming $x(\zeta)$ through a suitable function.[†] Then the expectation of $y(\zeta)$ is given by

$$E\{y(\zeta)\} \triangleq E\{g[x(\zeta)]\} = \int_{-\infty}^{\infty} g(x) f_x(x)\,\mathrm{d}x \tag{3.1.11}$$

Moments

Using the expectation operations (3.1.9) and (3.1.11), we can define various moments of the random variable $x(\zeta)$ that describe certain useful aspects of the density function. Let $g[x(\zeta)] = x^m(\zeta)$. Then

$$r_x^{(m)} \triangleq E\{x^m(\zeta)\} = \int_{-\infty}^{\infty} x^m f_x(x)\,\mathrm{d}x \tag{3.1.12}$$

is called the mth-order *moment* of $x(\zeta)$. In particular, $r_x^{(0)} = 1$, and the first-order moment $r_x^{(1)} = \mu_x$. The second-order moment $r_x^{(2)} = E\{x^2(\zeta)\}$ is called the *mean-squared value*, and it plays an important role in estimation theory. Note that

$$E\{x^2(\zeta)\} \neq E^2\{x(\zeta)\} \tag{3.1.13}$$

Corresponding to these moments we also have central moments. Let $g[x(\zeta)] = [x(\zeta) - \mu_x]^m$, then

$$\gamma_x^{(m)} \triangleq E\{[x(\zeta) - \mu_x]^m\} = \int_{-\infty}^{\infty} (x - \mu_x)^m f_x(x)\,\mathrm{d}x \tag{3.1.14}$$

is called the mth-order *central moment* of $x(\zeta)$. In particular, $\gamma_x^{(0)} = 1$ and $\gamma_x^{(1)} = 0$, which is obvious. Clearly, a random variable's moments and central moments are identical if its

[†]Such a function $g(\cdot)$ is called a *Baire function* (Papoulis 1991).

mean value is zero. The second central moment is of considerable importance and is called the *variance* of $x(\zeta)$, denoted by σ_x^2. Thus

$$\text{var}[x(\zeta)] \triangleq \sigma_x^2 \triangleq \gamma_x^{(2)} = E\{[x(\zeta) - \mu_x]^2\} \tag{3.1.15}$$

The quantity $\sigma_x = \sqrt{\gamma_x^{(2)}}$ is called the *standard deviation* of $x(\zeta)$ and is a measure of the spread (or dispersion) of the observed values of $x(\zeta)$ around its mean μ_x [see Figure 3.2(*b*)]. The relation between a random variable's moments and central moments is given by (see Problem 3.3)

$$\gamma_x^{(m)} = \sum_{k=0}^{m} \binom{m}{k} (-1)^k \mu_x^k r_x^{(m-k)} \tag{3.1.16}$$

In particular, and also from (3.1.15), we have

$$\sigma_x^2 = r_x^{(2)} - \mu_x^2 = E\{x^2(\zeta)\} - E^2\{x(\zeta)\} \tag{3.1.17}$$

The quantity *skewness* is related to the third-order central moment and characterizes the degree of asymmetry of a distribution around its mean, as shown in Figure 3.2(*c*). It is defined as a *normalized* third-order central moment, that is,

$$\text{Skew} \triangleq \tilde{\kappa}_x^{(3)} \triangleq E\left\{\left[\frac{x(\zeta) - \mu_x}{\sigma_x}\right]^3\right\} = \frac{1}{\sigma_x^3}\gamma_x^{(3)} \tag{3.1.18}$$

and is a *dimensionless* quantity. It is a pure number that attempts to describe leaning of the shape of the distribution. The skewness is zero if the density function is symmetric about its mean value, is positive if the shape leans towards the right, or is negative if it leans towards the left.

The quantity related to the fourth-order central moment is called *kurtosis*, which is also a dimensionless quantity. It measures the relative flatness or peakedness of a distribution about its mean as shown in Figure 3.2(*d*). This relative measure is with respect to a normal distribution, which will be introduced in the next section. The kurtosis is defined as

$$\text{Kurtosis} \triangleq \tilde{\kappa}_x^{(4)} \triangleq E\left\{\left[\frac{x(\zeta) - \mu_x}{\sigma_x}\right]^4\right\} - 3 = \frac{1}{\sigma_x^4}\gamma_x^{(4)} - 3 \tag{3.1.19}$$

where the term -3 makes the kurtosis $\tilde{\kappa}_x^{(4)} = 0$ for the normal distribution [see (3.1.40) for explanation].

Chebyshev's inequality. A useful result in the interpretation and use of the mean μ and the variance σ^2 of a random variable is given by Chebyshev's inequality. Given a random variable $x(\zeta)$ with its mean μ_x and variance σ_x^2, we have the inequality

$$\Pr\{|x(\zeta) - \mu_x| \geq k\sigma_x\} \leq \frac{1}{k^2} \qquad k > 0 \tag{3.1.20}$$

The interpretation of the above inequality is that regardless of the shape of $f_x(x)$, the random variable $x(\zeta)$ deviates from its mean by k times its standard deviation with probability less than or equal to $1/k^2$.

Characteristic functions

The Fourier and Laplace transforms find many uses in probability theory through the concepts of characteristic and moment generating functions. The *characteristic function* of a random variable $x(\zeta)$ is defined by the integral

$$\Phi_x(\xi) \triangleq E\{e^{j\xi x(\zeta)}\} = \int_{-\infty}^{\infty} f_x(x) e^{j\xi x}\, \mathrm{d}x \tag{3.1.21}$$

which can be interpreted as the Fourier transform of $f_x(x)$ with sign reversal in the complex exponential. To avoid confusion with the cdf, we do not use $F_x(\xi)$ to denote this Fourier transform. Furthermore, the variable ξ in $\Phi_x(\xi)$ is not and should not be interpreted as frequency. When $j\xi$ in (3.1.21) is replaced by a complex variable s, we obtain the *moment generating function* defined by

$$\bar{\Phi}_x(s) \triangleq E\{e^{sx(\zeta)}\} = \int_{-\infty}^{\infty} f_x(x)e^{sx}\,\mathrm{d}x \tag{3.1.22}$$

which again can be interpreted as the Laplace transform of $f_x(x)$ with sign reversal. Expanding e^{sx} in (3.1.22) in a Taylor series at $s = 0$, we obtain

$$\begin{aligned}\bar{\Phi}_x(s) = E\{e^{sx(\zeta)}\} &= E\left\{1 + sx(\zeta) + \frac{[sx(\zeta)]^2}{2!} + \cdots + \frac{[sx(\zeta)]^m}{m!} + \cdots\right\} \\ &= 1 + s\mu_x + \frac{s^2}{2!}r_x^{(2)} + \cdots + \frac{s^m}{m!}r_x^{(m)} + \cdots\end{aligned} \tag{3.1.23}$$

provided every moment $r_x^{(m)}$ exists. Thus from (3.1.23) we infer that if all moments of $x(\zeta)$ are known (and exist), then we can assemble $\bar{\Phi}_x(s)$ and upon inverse Laplace transformation, we can determine the density function $f_x(x)$. If we differentiate $\bar{\Phi}_x(s)$ with respect to s, we obtain

$$r_x^{(m)} = \left.\frac{\mathrm{d}^m[\bar{\Phi}_x(s)]}{\mathrm{d}s^m}\right|_{s=0} = (-j)^m \left.\frac{\mathrm{d}^m[\Phi_x(\xi)]}{\mathrm{d}\xi^m}\right|_{\xi=0} \qquad m = 1, 2, \ldots \tag{3.1.24}$$

which provides the mth-order moment of the random variable $x(\zeta)$.

The functions $\Phi_x(\xi)$ and $\bar{\Phi}_x(s)$ possess all the properties associated with the Fourier and Laplace transforms, respectively. Thus, since $f_x(x)$ is always a real-valued function, $\Phi_x(\xi)$ is conjugate symmetric; and if $f_x(x)$ is also an even function, then $\Phi_x(\xi)$ is a real-valued even function. In addition, they possess several properties due to the basic nature of the pdf. Therefore, the characteristic function $\Phi_x(\xi)$ always exists[†] since

$$\int |f_x(x)|\,\mathrm{d}x = \int f_x(x)\,\mathrm{d}x = 1$$

and $\Phi_x(\xi)$ is maximum at the origin, that is,

$$|\Phi_x(\xi)| \le \Phi_x(0) = 1 \tag{3.1.25}$$

since $f_x(x) \ge 0$.

Cumulants

These statistical descriptors are similar to the moments, but provide better information for higher-order moment analysis, which we will consider in detail in Chapter 12. The cumulants are derived by considering the moment generating function's natural logarithm. This logarithm is commonly referred to as the *cumulant generating function* and is given by

$$\bar{\Psi}_x(s) \triangleq \ln \bar{\Phi}_x(s) = \ln E\{e^{sx(\zeta)}\} \tag{3.1.26}$$

When s is replaced by $j\xi$ in (3.1.26), the resulting function is known as the *second characteristic function* and is denoted by $\Psi_x(\xi)$.

The *cumulants* $\kappa_x^{(m)}$ of a random variable $x(\zeta)$ are defined as the derivatives of the cumulant generating function, that is,

$$\kappa_x^{(m)} \triangleq \left.\frac{\mathrm{d}^m[\bar{\Psi}_x(s)]}{\mathrm{d}s^m}\right|_{s=0} = (-j)^m \left.\frac{\mathrm{d}^m[\Psi_x(\xi)]}{\mathrm{d}\xi^m}\right|_{\xi=0} \qquad m = 1, 2, \ldots \tag{3.1.27}$$

[†]We will generally choose the characteristic function over the moment generating function.

Clearly, $\kappa_x^{(0)} = 0$. It can be shown that (see Problem 3.4) for a zero-mean random variable, the first five cumulants as functions of the central moments are given by

$$\kappa_x^{(1)} = r_{(x)}^{(1)} = \mu_x = 0 \tag{3.1.28}$$

$$\kappa_x^{(2)} = \gamma_x^{(2)} = \sigma_x^2 \tag{3.1.29}$$

$$\kappa_x^{(3)} = \gamma_x^{(3)} \tag{3.1.30}$$

$$\kappa_x^{(4)} = \gamma_x^{(4)} - 3\sigma_x^4 \tag{3.1.31}$$

$$\kappa_x^{(5)} = \gamma_x^{(5)} - 10\gamma_x^{(3)}\sigma_x^2 \tag{3.1.32}$$

which show that the first two cumulants are identical to the first two central moments. Clearly due to the logarithmic function in (3.1.26), cumulants are useful for dealing with products of characteristic functions (see Section 3.2.4).

3.1.3 Some Useful Random Variables

Random variable models are needed to describe (or approximate) complex physical phenomena using simple parameters. For example, the random phase of a sinusoidal carrier can be described by a uniformly distributed random variable so that we can study its statistical properties. This approximation allows us to investigate random signals in a sound mathematical way. We will describe three continuous random variable models although there are several other known continuous as well as discrete models available in the literature.

Uniformly distributed random variable. This is an appropriate model in situations in which random outcomes are "equally likely." Here $x(\zeta)$ assumes values on $\mathbb{R}$ according to the pdf

$$f_x(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{elsewhere} \end{cases} \tag{3.1.33}$$

where $a < b$ are specified parameters. This pdf is shown in Figure 3.3. The corresponding

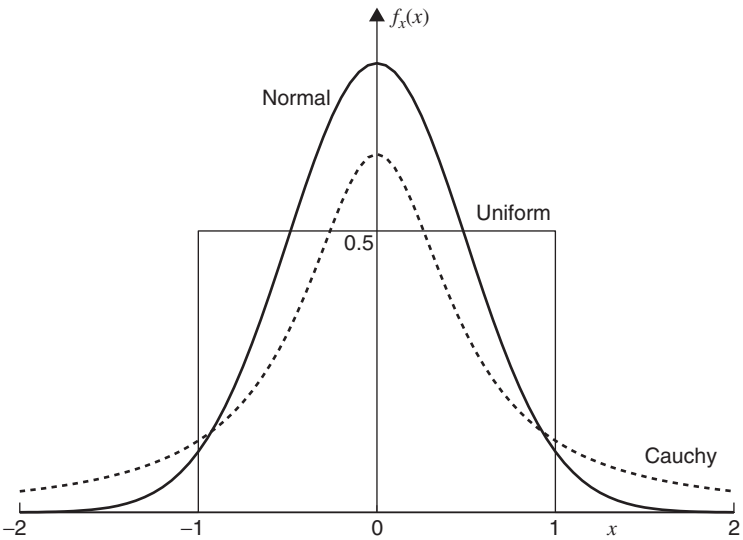


FIGURE 3.3
Probability density functions of useful random variables.

cdf is given by

$$F_x(x) = \int_{-\infty}^{x} f_x(v)\,dv = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > a \end{cases} \tag{3.1.34}$$

and the characteristic function is given by

$$\Phi_x(\xi) = \frac{e^{j\xi b} - e^{j\xi a}}{j\xi(b-a)} \tag{3.1.35}$$

The mean and the variance of this random variable are given by, respectively,

$$\mu_x = \frac{a+b}{2} \qquad \text{and} \qquad \sigma_x^2 = \frac{(b-a)^2}{12} \tag{3.1.36}$$

Normal random variable. This is the most useful and convenient model in many applications, as we shall see later. It is also known as a *Gaussian* random variable, and we will use both terms interchangeably. The pdf of a normally distributed random variable $x(\zeta)$ with mean μ_x and standard deviation σ_x is given by

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right] \tag{3.1.37}$$

where $-\infty < \mu < \infty$ and $\sigma \ge 0$ (see Figure 3.3). The characteristic function of the normal random variable is given by

$$\Phi_x(\xi) = \exp(j\mu_x\xi - \tfrac{1}{2}\sigma_x^2\xi^2) \tag{3.1.38}$$

Clearly, the pdf of a normal random variable is completely described by its mean μ_x and standard deviation σ_x and is denoted by $\mathcal{N}(\mu_x, \sigma_x^2)$. We note that *all higher-order moments of a normal random variable can be determined in terms of the first two moments*, that is,

$$\gamma_x^{(m)} = E\{[x(\zeta) - \mu_x]^m\} = \begin{cases} 1 \cdot 3 \cdot 5 \cdots (m-1)\sigma_x^m & \text{if } m \text{ even} \\ 0 & \text{if } m \text{ odd} \end{cases} \tag{3.1.39}$$

In particular, we obtain the fourth moment as

$$\gamma_x^{(4)} = 3\sigma_x^4 \tag{3.1.40}$$

or from (3.1.19), kurtosis $= 0$, which explains the term -3 in (3.1.19).

From (3.1.37), we observe that the Gaussian random variable is completely determined by its first two moments (mean μ_x and variance σ_x^2), which means that the higher moments do not provide any additional information about the Gaussian density function. In fact, all higher-order moments can be obtained in terms of the first two moments [see Equation (3.1.39)]. Thus for a non-Gaussian random variable, we would like to know how different that random variable is from a Gaussian random variable (this is also known as a departure from the Gaussian-ness). This measurement of the deviation from being Gaussian is given by the cumulants that were defined in (3.1.27). Roughly speaking, the cumulants are like central moments (which measure deviations from the mean) of non-Gaussian random variables for Gaussian departure. Also from (3.1.30) and (3.1.31), we see that all higher-order (that is, $m > 2$) cumulants of a Gaussian random variable are zero. This fact is used in the analysis and estimation of non-Gaussian random variables (and later for non-Gaussian random processes).

Cauchy random variable. This is an appropriate model in which a random variable takes large values with significant probability (heavy-tailed distribution). The Cauchy pdf with parameters μ and β is given by

$$f_x(x) = \frac{\beta}{\pi} \frac{1}{(x-\mu)^2 + \beta^2} \tag{3.1.41}$$

and is shown in Figure 3.3. The corresponding cdf is given by

$$F_x(x) = 0.5 + \frac{1}{\pi} \arctan \frac{x - \mu}{\beta} \tag{3.1.42}$$

and the characteristic function is given by

$$\Phi_x(\xi) = \exp(j\mu\xi - \beta|\xi|) \tag{3.1.43}$$

The Cauchy random variable has mean $\mu_x = \mu$. However, its variance does not exist because $E\{x^2\}$ fails to exist in any sense, and hence the moment generating function does not exist, in general. It has the property that the sum of M independent Cauchy random variables is also Cauchy (see Example 3.2.3). Thus a Cauchy random variable is an example of an infinite-variance random variable.

Random number generators. Random numbers, by definition, are truly unpredictable, and hence it is not possible to generate them by using a well-defined algorithm on a computer. However, in many simulation studies, we need to use sequences of numbers that appear to be random and that possess required properties, for example, Gaussian random numbers in a Monte Carlo analysis. These numbers are called *pseudo random numbers*, and many excellent algorithms are available to generate them on a computer (Park and Miller 1988). In MATLAB, the function `rand` generates numbers that are uniformly distributed over (0, 1) while the function `randn` generates $\mathcal{N}(0, 1)$ pseudo random numbers.

3.2 RANDOM VECTORS

In many applications, a group of signal observations can be modeled as a collection of random variables that can be grouped to form a *random vector*. This is an extension of the concept of random variable and generalizes many scalar quantities to vectors and matrices. One example of a random vector is the case of a complex-valued random variable $x(\zeta) = x_{\mathrm{R}}(\zeta) + jx_{\mathrm{I}}(\zeta)$, which can be considered as a group of $x_{\mathrm{R}}(\zeta)$ and $x_{\mathrm{I}}(\zeta)$. In this section, we provide a review of the basic properties of random vectors and related results from linear algebra. We first begin with real-valued random vectors and then extend their concepts to complex-valued random vectors.

3.2.1 Definitions and Second-Order Moments

A real-valued vector containing M random variables

$$\mathbf{x}(\zeta) = [x_1(\zeta), x_2(\zeta), \ldots, x_M(\zeta)]^T \tag{3.2.1}$$

is called a *random M vector* or a random vector when dimensionality is unimportant. As usual, superscript T denotes the transpose of the vector. We can think of a real-valued random vector as a mapping from an abstract probability space to a vector-valued, real space $\mathbb{R}^M$. Thus the range of this mapping is an M-dimensional space.

Distribution and density functions

A random vector is completely characterized by its *joint* cumulative distribution function, which is defined by

$$F_{\mathbf{x}}(x_1, \ldots, x_M) \triangleq \Pr\{x_1(\zeta) \le x_1, \ldots, x_M(\zeta) \le x_M\} \tag{3.2.2}$$

and is often written as

$$F_{\mathbf{x}}(\mathbf{x}) = \Pr\{\mathbf{x}(\zeta) \le \mathbf{x}\} \tag{3.2.3}$$

for convenience. A random vector can be also characterized by its *joint* probability density function, which is defined by

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \lim_{\substack{\Delta x_1 \to 0 \\ \vdots \\ \Delta x_M \to 0}} \frac{\Pr\{x_1 < x_1(\zeta) \le x_1 + \Delta x_1, \ldots, x_M < x_M(\zeta) \le x_M + \Delta x_M\}}{\Delta x_1 \cdots \Delta x_M} \\ &\triangleq \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_M} F_{\mathbf{x}}(\mathbf{x}) \end{aligned} \tag{3.2.4}$$

The function

$$f_{x_j}(x_j) = \int_{(M-1)} \cdots \int f_{\mathbf{x}}(\mathbf{x})\, \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1}\, \mathrm{d}x_{j+1} \cdots \mathrm{d}x_M \tag{3.2.5}$$

is known as a *marginal* density function and describes individual random variables. Thus the probability functions defined for a random variable in the previous section are more appropriately called marginal functions. The joint pdf $f_{\mathbf{x}}(\mathbf{x})$ must be multiplied by a certain M-dimensional region $\Delta\mathbf{x}$ to obtain a probability. From (3.2.4) we obtain

$$F_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_M} f_{\mathbf{x}}(\boldsymbol{\nu})\, \mathrm{d}\nu_1 \cdots \mathrm{d}\nu_M = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{x}}(\boldsymbol{\nu})\, \mathrm{d}\boldsymbol{\nu} \tag{3.2.6}$$

These joint probability functions also satisfy several important properties that are similar to (3.1.6) through (3.1.8) for random variables. In particular, note that both $f_{\mathbf{x}}(\mathbf{x})$ and $F_{\mathbf{x}}(\mathbf{x})$ are positive multidimensional functions.

The joint [and conditional probability (see Papoulis 1991)] functions can also be used to define the concept of independent random variables. Two random variables $x_1(\zeta)$ and $x_2(\zeta)$ are *independent* if the events $\{x_1(\zeta) \le x_1\}$ and $\{x_2(\zeta) \le x_2\}$ are jointly independent, that is, if

$$\Pr\{x_1(\zeta) \le x_1, x_2(\zeta) \le x_2\} = \Pr\{x_1(\zeta) \le x_1\} \Pr\{x_2(\zeta) \le x_2\}$$

which implies that

$$F_{x_1,x_2}(x_1, x_2) = F_{x_1}(x_1) F_{x_2}(x_2) \qquad \text{and} \qquad f_{x_1,x_2}(x_1, x_2) = f_{x_1}(x_1) f_{x_2}(x_2) \tag{3.2.7}$$

Complex-valued random variables and vectors

As we shall see in later chapters, in applications such as channel equalization, array processing, etc., we encounter complex signal and noise models. To formulate these models, we need to describe complex random variables and vectors, and then extend our standard definitions and results to the complex case. A complex random variable is defined as[†] $x(\zeta) = x_{\mathrm{R}}(\zeta) + jx_{\mathrm{I}}(\zeta)$, where $x_{\mathrm{R}}(\zeta)$ and $x_{\mathrm{I}}(\zeta)$ are real-valued random variables. Thus we can think of $x(\zeta)$ as a mapping from an abstract probability space $\mathcal{S}$ to a complex space $\mathbb{C}$. Alternatively, $x(\zeta)$ can be thought of as a real-valued random vector $[x_{\mathrm{R}}(\zeta), x_{\mathrm{I}}(\zeta)]^T$ with a joint cdf $F_{x_{\mathrm{R}},x_{\mathrm{I}}}(x_{\mathrm{R}}, x_{\mathrm{I}})$ or a joint pdf $f_{x_1,x_2}(x_1, x_2)$ that will allow us to define its statistical averages. The mean of $x(\zeta)$ is defined as

$$E\{x(\zeta)\} = \mu_x = E\{x_{\mathrm{R}}(\zeta) + jx_{\mathrm{I}}(\zeta)\} = \mu_{x_{\mathrm{R}}} + j\mu_{x_{\mathrm{I}}} \tag{3.2.8}$$

and the variance is defined as

$$\sigma_x^2 = E\{|x(\zeta) - \mu_x|^2\} \tag{3.2.9}$$

which can be shown to be equal to

$$\sigma_x^2 = E\{|x(\zeta)|^2\} - |\mu_x|^2 \tag{3.2.10}$$

[†]We will not make any distinction in *notation* between a real-valued and a complex-valued random variable. The actual type should be evident from the context.

A complex-valued random vector is given by

$$\mathbf{x}(\zeta) = \mathbf{x}_{\mathrm{R}}(\zeta) + j\mathbf{x}_{\mathrm{I}}(\zeta) = \begin{bmatrix} x_{\mathrm{R}1}(\zeta) \\ \vdots \\ x_{\mathrm{R}M}(\zeta) \end{bmatrix} + j \begin{bmatrix} x_{\mathrm{I}1}(\zeta) \\ \vdots \\ x_{\mathrm{I}M}(\zeta) \end{bmatrix} \tag{3.2.11}$$

and we can think of a complex-valued random vector as a mapping from an abstract probability space to a vector-valued, complex space $\mathbb{C}^M$. The cdf for the complex-valued random vector $\mathbf{x}(\zeta)$ is then defined as

$$F_{\mathbf{x}}(\mathbf{x}) \triangleq \Pr\{\mathbf{x}(\zeta) \leq \mathbf{x}\} \triangleq \Pr\{\mathbf{x}_{\mathrm{R}}(\zeta) \leq \mathbf{x}_{\mathrm{R}}, \mathbf{x}_{\mathrm{I}}(\zeta) \leq \mathbf{x}_{\mathrm{I}}\} \tag{3.2.12}$$

while its joint pdf is defined as

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \lim_{\substack{\Delta x_{\mathrm{R}1}\to 0 \\ \vdots \\ \Delta x_{\mathrm{I}M}\to 0}} \frac{\Pr\{\mathbf{x}_{\mathrm{R}} < \mathbf{x}_{\mathrm{R}}(\zeta) \leq \mathbf{x}_{\mathrm{R}} + \Delta\mathbf{x}_{\mathrm{R}}, \mathbf{x}_{\mathrm{I}} < \mathbf{x}_{\mathrm{I}}(\zeta) \leq \mathbf{x}_{\mathrm{I}} + \Delta\mathbf{x}_{\mathrm{I}}\}}{\Delta x_{\mathrm{R}1}\Delta x_{\mathrm{I}1}\cdots\Delta x_{\mathrm{R}M}\Delta x_{\mathrm{I}M}} \\ &\triangleq \frac{\partial}{\partial x_{\mathrm{R}1}}\frac{\partial}{\partial x_{\mathrm{I}1}}\cdots\frac{\partial}{\partial x_{\mathrm{R}M}}\frac{\partial}{\partial x_{\mathrm{I}M}} F_{\mathbf{x}}(\mathbf{x}) \end{aligned} \tag{3.2.13}$$

From (3.2.13), the cdf is obtained by integrating the pdf over all real and imaginary parts, that is

$$F_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{x_{\mathrm{R}1}} \cdots \int_{-\infty}^{x_{\mathrm{I}M}} f_{\mathbf{x}}(\boldsymbol{\nu})\, \mathrm{d}\nu_{\mathrm{R}1} \cdots \mathrm{d}\nu_{\mathrm{I}M} = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{x}}(\boldsymbol{\nu})\, \mathrm{d}\boldsymbol{\nu} \tag{3.2.14}$$

where the single integral in the last expression is used as a compact notation for multidimensional integrals and should not be confused with a complex-contour integral. These probability functions for a complex-valued random vector possess properties similar to those of the real-valued random vectors. In particular,

$$\int_{-\infty}^{\infty} f_{\mathbf{x}}(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1 \tag{3.2.15}$$

Statistical description

Clearly the above probability functions require an enormous amount of information that is not easy to obtain or is too complex mathematically for practical use. In practical applications, random vectors are described by less complete but more manageable statistical averages.

Mean vector. As we have seen before, the most important statistical operation is the expectation operation. The marginal expectation of a random vector $\mathbf{x}(\zeta)$ is called the *mean vector* and is defined by

$$\boldsymbol{\mu}_{\mathbf{x}} = E\{\mathbf{x}(\zeta)\} = \begin{bmatrix} E\{x_1(\zeta)\} \\ \vdots \\ E\{x_M(\zeta)\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_M \end{bmatrix} \tag{3.2.16}$$

where the integral is taken over the entire $\mathbb{C}^M$ space. The components of $\boldsymbol{\mu}$ are the means of the individual random variables.

Correlation and covariance matrices. The second-order moments of a random vector $\mathbf{x}(\zeta)$ are given as matrices and describe the spread of its distribution. The *autocorrelation matrix* is defined by

$$\mathbf{R}_{\mathbf{x}} \triangleq E\{\mathbf{x}(\zeta)\mathbf{x}^H(\zeta)\} = \begin{bmatrix} r_{11} & \cdots & r_{1M} \\ \vdots & \ddots & \vdots \\ r_{M1} & \cdots & r_{MM} \end{bmatrix} \tag{3.2.17}$$

where superscript H denotes the conjugate transpose operation, the diagonal terms

$$r_{ii} \triangleq E\{|x_i(\zeta)|^2\} \qquad i = 1, \ldots, M \tag{3.2.18}$$

are the second-order moments, denoted earlier as $r_{x_i}^{(2)}$, of random variables $x_i(\zeta)$, and the off-diagonal terms

$$r_{ij} \triangleq E\{x_i(\zeta)x_j^*(\zeta)\} = r_{ji}^* \qquad i \neq j \tag{3.2.19}$$

measure the *correlation*, that is, the statistical similarity between the random variables $x_i(\zeta)$ and $x_j(\zeta)$. From (3.2.19) we note that the correlation matrix $\mathbf{R_x}$ is conjugate symmetric or *Hermitian*, that is, $\mathbf{R_x} = \mathbf{R_x}^H$.

The *autocovariance matrix* is defined by

$$\mathbf{\Gamma_x} \triangleq E\{[\mathbf{x}(\zeta) - \boldsymbol{\mu}_\mathbf{x}][\mathbf{x}(\zeta) - \boldsymbol{\mu}_\mathbf{x}]^H\} \triangleq \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \cdots & \gamma_{MM} \end{bmatrix} \tag{3.2.20}$$

where the diagonal terms

$$\gamma_{ii} = E\{|x_i(\zeta) - \mu_i|^2\} \qquad i = 1, \ldots, M \tag{3.2.21}$$

are the (self-)variances of $x_i(\zeta)$ denoted earlier as $\sigma_{x_i}^2$ while the off-diagonal terms

$$\gamma_{ij} = E\{[x_i(\zeta) - \mu_i][x_j(\zeta) - \mu_j]^*\} = E\{x_i(\zeta)x_j^*(\zeta)\} - \mu_i\mu_j^* = \gamma_{ji}^* \qquad i \neq j \tag{3.2.22}$$

are the values of the *covariance* between $x_i(\zeta)$ and $x_j(\zeta)$. The covariance matrix $\mathbf{\Gamma_x}$ is also a Hermitian matrix. The covariance γ_{ij} can also be expressed in terms of standard deviations of $x_i(\zeta)$ and $x_j(\zeta)$ as $\gamma_{ij} = \rho_{ij}\sigma_i\sigma_j$, where

$$\rho_{ij} \triangleq \frac{\gamma_{ij}}{\sigma_i\sigma_j} = \rho_{ji} \tag{3.2.23}$$

is called the *correlation coefficient* between $x_i(\zeta)$ and $x_j(\zeta)$. Note that

$$|\rho_{ij}| \leq 1 \qquad i \neq j \qquad \rho_{ii} = 1 \tag{3.2.24}$$

The correlation coefficient measures the degree of statistical similarity between two random variables. If $|\rho_{ij}| = 1$, then random variables are said to be *perfectly correlated*; but if $\rho_{ij} = 0$ (that is, when the covariance $\gamma_{ij} = 0$), then $x_i(\zeta)$ and $x_j(\zeta)$ are said to *uncorrelated*.

The autocorrelation and autocovariance matrices are related. Indeed, we can easily see that

$$\mathbf{\Gamma_x} \triangleq E\{[\mathbf{x}(\zeta) - \boldsymbol{\mu}_\mathbf{x}][\mathbf{x}(\zeta) - \boldsymbol{\mu}_\mathbf{x}]^H\} = \mathbf{R_x} - \boldsymbol{\mu}_\mathbf{x}\boldsymbol{\mu}_\mathbf{x}^H \tag{3.2.25}$$

which shows that these two moments have essentially the same amount of information. In fact, if $\boldsymbol{\mu}_\mathbf{x} = \mathbf{0}$, then $\mathbf{\Gamma_x} = \mathbf{R_x}$. The autocovariance measures a *weaker* form of interaction between random variables called *correlatedness* that should be contrasted with the *stronger* form of independence that we described in (3.2.7). If random variables $x_i(\zeta)$ and $x_j(\zeta)$ are independent, then they are also uncorrelated since (3.2.7) implies that

$$E\{x_i(\zeta)x_j^*(\zeta)\} = E\{x_i(\zeta)\}E\{x_j^*(\zeta)\} \qquad \text{or} \qquad \gamma_{ij} = 0 \tag{3.2.26}$$

but uncorrelatedness does not imply independence unless random variables are jointly Gaussian (see Problem 3.15). The autocorrelation also measures another weaker form of interaction called *orthogonality*. Random variables $x_i(\zeta)$ and $x_j(\zeta)$ are *orthogonal* if their correlation

$$r_{ij} = E\{x_i(\zeta)x_j^*(\zeta)\} = 0 \qquad i \neq j \tag{3.2.27}$$

Clearly, from (3.2.26) if one or both random variables have zero means, then uncorrelatedness also implies orthogonality.

We can also define correlation and covariance functions between two random vectors. Let $\mathbf{x}(\zeta)$ and $\mathbf{y}(\zeta)$ be random M- and L-vectors, respectively. Then the $M \times L$ matrix

$$\mathbf{R}_{\mathbf{xy}} \triangleq E\{\mathbf{x}\mathbf{y}^H\} = \begin{bmatrix} E\{x_1(\zeta)y_1^*(\zeta)\} & \cdots & E\{x_1(\zeta)y_L^*(\zeta)\} \\ \vdots & \ddots & \vdots \\ E\{x_M(\zeta)y_1^*(\zeta)\} & \cdots & E\{x_M(\zeta)y_L^*(\zeta)\} \end{bmatrix} \tag{3.2.28}$$

is called a *cross-correlation* matrix whose elements r_{ij} are the correlations between random variables $x_i(\zeta)$ and $y_j(\zeta)$. Similarly the $M \times L$ matrix

$$\mathbf{\Gamma}_{\mathbf{xy}} \triangleq E\{[\mathbf{x}(\zeta)-\boldsymbol{\mu}_{\mathbf{x}}][\mathbf{y}(\zeta)-\boldsymbol{\mu}_{\mathbf{y}}]^H\} = \mathbf{R}_{\mathbf{xy}} - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{y}}^H \tag{3.2.29}$$

is called a *cross-covariance* matrix whose elements c_{ij} are the covariances between $x_i(\zeta)$ and $y_j(\zeta)$. In general the cross-matrices are not square matrices, and even if $M = L$, they are not necessarily symmetric. Two random vectors $\mathbf{x}(\zeta)$ and $\mathbf{y}(\zeta)$ are said to be

- Uncorrelated if

$$\mathbf{\Gamma}_{\mathbf{xy}} = \mathbf{0} \Rightarrow \mathbf{R}_{\mathbf{xy}} = \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{y}}^H \tag{3.2.30}$$

- Orthogonal if

$$\mathbf{R}_{\mathbf{xy}} = \mathbf{0} \tag{3.2.31}$$

Again, if $\boldsymbol{\mu}_{\mathbf{x}}$ or $\boldsymbol{\mu}_{\mathbf{y}}$ or both are zero vectors, then (3.2.30) implies (3.2.31).

3.2.2 Linear Transformations of Random Vectors

Many signal processing applications involve linear operations on random vectors. Linear transformations are relatively simple mappings and are given by the matrix operation

$$\mathbf{y}(\zeta) = g[\mathbf{x}(\zeta)] = \mathbf{A}\mathbf{x}(\zeta) \tag{3.2.32}$$

where $\mathbf{A}$ is an $L \times M$ (not necessarily square) matrix. The random vector $\mathbf{y}(\zeta)$ is completely described by the density function $f_{\mathbf{y}}(\mathbf{y})$. If $L > M$, then only M $y_i(\zeta)$ random variables can be independently determined from $\mathbf{x}(\zeta)$. The remaining $(L - M)$ $y_i(\zeta)$ random variables can be obtained from the first $y_i(\zeta)$ random variables. Thus we need to determine $f_{\mathbf{y}}(\mathbf{y})$ for M random variables from which we can determine $f_{\mathbf{y}}(\mathbf{y})$ for all L random variables. If $M > L$, then we can augment $\mathbf{y}$ into an M-vector by introducing auxiliary random variables

$$y_{L+1}(\zeta) = x_{L+1}(\zeta), \ldots, y_M(\zeta) = x_M(\zeta) \tag{3.2.33}$$

to determine $f_{\mathbf{y}}(\mathbf{y})$ for M random variables from which we can determine $f_{\mathbf{y}}(\mathbf{y})$ for the original L random variables. Therefore, for the determination of the pdf $f_{\mathbf{y}}(\mathbf{y})$, we will assume that $L = M$ and that $\mathbf{A}$ is nonsingular.

Furthermore, we will first consider the case in which both $\mathbf{x}(\zeta)$ and $\mathbf{y}(\zeta)$ are real-valued random vectors, which also implies that $\mathbf{A}$ is a real-valued matrix. This approach is necessary because the complex case leads to a slightly different result. Then the pdf $f_{\mathbf{y}}(\mathbf{y})$ is given by

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(g^{-1}(\mathbf{y}))}{|\mathbf{J}|} \tag{3.2.34}$$

where $\mathbf{J}$ is called the *Jacobian* of the transformation (3.2.32), given by

$$\mathbf{J} = \det \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots & \dfrac{\partial y_M}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y_1}{\partial x_M} & \cdots & \dfrac{\partial y_M}{\partial x_M} \end{bmatrix} = \det \mathbf{A} \tag{3.2.35}$$

From (3.2.34) and (3.2.35), the pdf of $\mathbf{y}(\zeta)$ is given by

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y})}{|\det \mathbf{A}|} \qquad \text{real-valued random vector} \tag{3.2.36}$$

from which moment computations of any order of $\mathbf{y}(\zeta)$ can be performed. Now we consider the case of the complex-valued random vectors. Then by applying the above approach to both real and imaginary parts, the result (3.2.36) becomes

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y})}{|\det \mathbf{A}|^2} \qquad \text{complex-valued random vector} \tag{3.2.37}$$

This shows that sometimes we can get different results depending upon whether we assume real- or complex-valued random vectors in our analysis.

Determining $f_{\mathbf{y}}(\mathbf{y})$ is, in general, tedious except in the case of Gaussian random vectors, as we shall see later. In practice, the knowledge of $\boldsymbol{\mu}_{\mathbf{y}}$, $\boldsymbol{\Gamma}_{\mathbf{y}}$, $\boldsymbol{\Gamma}_{\mathbf{xy}}$, or $\boldsymbol{\Gamma}_{\mathbf{yx}}$ is sufficient in many applications. If we take the expectation of both sides of (3.2.32), we find that the mean vector is given by

$$\boldsymbol{\mu}_{\mathbf{y}} = E\{\mathbf{y}(\zeta)\} = E\{\mathbf{A}\mathbf{x}(\zeta)\} = \mathbf{A}E\{\mathbf{x}(\zeta)\} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} \tag{3.2.38}$$

The autocorrelation matrix of $\mathbf{y}(\zeta)$ is given by

$$\mathbf{R}_{\mathbf{y}} = E\{\mathbf{y}\mathbf{y}^H\} = E\{\mathbf{A}\mathbf{x}\mathbf{x}^H\mathbf{A}^H\} = \mathbf{A}E\{\mathbf{x}\mathbf{x}^H\}\mathbf{A}^H = \mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^H \tag{3.2.39}$$

Similarly, the autocovariance matrix of $\mathbf{y}(\zeta)$ is given by

$$\boldsymbol{\Gamma}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\Gamma}_{\mathbf{x}}\mathbf{A}^H \tag{3.2.40}$$

Consider the cross-correlation matrix

$$\mathbf{R}_{\mathbf{xy}} = E\{\mathbf{x}(\zeta)\mathbf{y}^H(\zeta)\} = E\{\mathbf{x}(\zeta)\mathbf{x}^H(\zeta)\mathbf{A}^H\} \tag{3.2.41}$$

$$= E\{\mathbf{x}(\zeta)\mathbf{x}^H(\zeta)\}\mathbf{A}^H = \mathbf{R}_{\mathbf{x}}\mathbf{A}^H \tag{3.2.42}$$

and hence $\mathbf{R}_{\mathbf{yx}} = \mathbf{A}\mathbf{R}_{\mathbf{x}}$. Similarly, the cross-covariance matrices are

$$\boldsymbol{\Gamma}_{\mathbf{xy}} = \boldsymbol{\Gamma}_{\mathbf{x}}\mathbf{A}^H \qquad \text{and} \qquad \boldsymbol{\Gamma}_{\mathbf{yx}} = \mathbf{A}\boldsymbol{\Gamma}_{\mathbf{x}} \tag{3.2.43}$$

3.2.3 Normal Random Vectors

If the components of the random vector $\mathbf{x}(\zeta)$ are jointly normal, then $\mathbf{x}(\zeta)$ is a normal random M-vector. Again, the pdf expressions for the real- and complex-valued cases are slightly different, and hence we consider these cases separately. The real-valued normal random vector has the pdf

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Gamma}_{\mathbf{x}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\right] \qquad \text{real} \tag{3.2.44}$$

with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance $\boldsymbol{\Gamma}_{\mathbf{x}}$. It will be denoted by $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Gamma}_{\mathbf{x}})$. The term in the exponent $(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ is a positive definite quadratic function of x_i and is also given by

$$(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}) = \sum_{i=1}^{M}\sum_{j=1}^{M}\langle\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}\rangle_{ij}(x_i-\mu_i)(x_j-\mu_j) \tag{3.2.45}$$

where $\langle\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}\rangle_{ij}$ denotes the (i, j)th element of $\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}$. The characteristic function of the normal random vector is given by

$$\Phi_{\mathbf{x}}(\boldsymbol{\xi}) = \exp(j\boldsymbol{\xi}^T\boldsymbol{\mu}_x - \tfrac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Gamma}_{\mathbf{x}}\boldsymbol{\xi}) \tag{3.2.46}$$

where $\boldsymbol{\xi}^T = [\xi_1, \ldots, \xi_M]$.

The complex-valued normal random vector has the pdf

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\pi^M |\mathbf{\Gamma}_{\mathbf{x}}|} \exp[-(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^H \mathbf{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})] \qquad \text{complex} \tag{3.2.47}$$

with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance $\mathbf{\Gamma}_{\mathbf{x}}$. This pdf will be denoted by $\mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{\Gamma}_{\mathbf{x}})$. If $\mathbf{x}(\zeta)$ is a scalar complex-valued random variable $x(\zeta)$ with mean μ_x and variance σ_x^2, then (3.2.47) reduces to

$$f_x(x) = \frac{1}{\pi\sigma_x^2} \exp\left(-\frac{|x-\mu|^2}{\sigma_x^2}\right) \tag{3.2.48}$$

which should be compared with the pdf given in (3.1.37). Note that the pdf in (3.1.37) is not obtained by setting the imaginary part of $x(\zeta)$ in (3.2.48) equal to zero. For a more detailed discussion on this aspect, see Therrien (1992) or Kay (1993). The term $(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^H \mathbf{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ in the exponent of (3.2.47) is also a positive definite quadratic function and is given by

$$(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^H \mathbf{\Gamma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}) = \sum_{i=1}^{M}\sum_{j=1}^{M} \langle \mathbf{\Gamma}_{\mathbf{x}}^{-1}\rangle_{ij}(x_i-\mu_i)^*(x_j-\mu_j) \tag{3.2.49}$$

The characteristic function for the complex-valued normal random vector is given by

$$\Phi_{\mathbf{x}}(\boldsymbol{\xi}) = \exp[j\mathrm{Re}(\boldsymbol{\xi}^H \boldsymbol{\mu}_{\mathbf{x}}) - \tfrac{1}{4}\boldsymbol{\xi}^H \mathbf{\Gamma}_{\mathbf{x}} \boldsymbol{\xi}] \tag{3.2.50}$$

The normal distribution is a useful model of a random vector because of its many important properties:

1. The pdf is completely specified by the mean vector and the covariance matrix, which are relatively easy to estimate in practice. All other higher-order moments can be obtained from these parameters.
2. If the components of $\mathbf{x}(\zeta)$ are mutually uncorrelated, then they are also independent. (See Problem 3.15.) This is useful in many derivations.
3. A linear transformation of a normal random vector is also normal. This can be easily seen by using (3.2.38), (3.2.40), and (3.2.44) in (3.2.36); that is, for the real-valued case we obtain

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{(2\pi)^{M/2}|\mathbf{\Gamma}_{\mathbf{y}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})^T \mathbf{\Gamma}_{\mathbf{y}}^{-1}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})\right] \qquad \text{real} \tag{3.2.51}$$

   This result can also be proved by using the moment generating function in (3.2.46) (see Problem 3.6). Similarly for the complex-valued case, from (3.2.37) and (3.2.47) we obtain

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\pi^M |\mathbf{\Gamma}_{\mathbf{y}}|} \exp[-(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})^H (\mathbf{A}^{-1})^H \mathbf{\Gamma}_{\mathbf{x}}^{-1} \mathbf{A}^{-1}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})] \qquad \text{complex} \tag{3.2.52}$$

4. The fourth-order moment of a normal random vector

$$\mathbf{x}(\zeta) = [x_1(\zeta)\ x_2(\zeta)\ x_3(\zeta)\ x_4(\zeta)]^T$$

   can be expressed in terms of its second-order moments. For the real case, that is, when $\mathbf{x}(\zeta) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_{\mathbf{x}})$, we have

$$\begin{aligned} E\{x_1(\zeta)x_2(\zeta)x_3(\zeta)x_4(\zeta)\} = {} & E\{x_1(\zeta)x_2(\zeta)\}E\{x_3(\zeta)x_4(\zeta)\} \\ & + E\{x_1(\zeta)x_3(\zeta)\}E\{x_2(\zeta)x_4(\zeta)\} \\ & + E\{x_1(\zeta)x_4(\zeta)\}E\{x_2(\zeta)x_3(\zeta)\} \end{aligned} \tag{3.2.53}$$

For the complex case, that is, when $\mathbf{x}(\zeta) \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Gamma}_{\mathbf{x}})$, we have

$$E\{x_1^*(\zeta)x_2(\zeta)x_3^*(\zeta)x_4(\zeta)\} = E\{x_1^*(\zeta)x_2(\zeta)\}E\{x_3^*(\zeta)x_4(\zeta)\} + E\{x_1^*(\zeta)x_4(\zeta)\}E\{x_2(\zeta)x_3^*(\zeta)\} \tag{3.2.54}$$

The proof of (3.2.53) is tedious but straightforward. However, the proof of (3.2.54) is complicated and is discussed in Kay (1993).

3.2.4 Sums of Independent Random Variables

In many applications, a random variable $y(\zeta)$ can be expressed as a linear combination of M statistically independent random variables $\{x_k(\zeta)\}_1^M$, that is,

$$y(\zeta) = c_1x_1(\zeta) + c_2x_2(\zeta) + \cdots + c_Mx_M(\zeta) = \sum_{k=1}^{M} c_kx_k(\zeta) \tag{3.2.55}$$

where $\{c_k\}_1^M$ is a set of fixed coefficients. In these situations, we would like to compute the first two moments and the pdf of $y(\zeta)$. The moment computation is straightforward, but the pdf computation requires the use of characteristic functions. When these results are extended to the sum of an infinite number of statistically independent random variables, we obtain a powerful theorem called the *central limit theorem* (*CLT*). Another interesting concept develops when the sum of IID random variables preserves their distribution, which results in stable distributions.

Mean. Using the linearity of the expectation operator and taking the expectation of both sides of (3.2.55), we obtain

$$\mu_y = \sum_{k=1}^{M} c_k\mu_{x_k} \tag{3.2.56}$$

Variance. Again by using independence, the variance of $y(\zeta)$ is given by

$$\sigma_y^2 = E\left\{\left|\sum_{k=1}^{M} c_k[x_k(\zeta) - \mu_{x_k}]\right|^2\right\} = \sum_{k=1}^{M} |c_k|^2\sigma_{x_k}^2 \tag{3.2.57}$$

where we have used the statistical independence between random variables.

Probability density function. Before we derive the pdf of $y(\zeta)$ in (3.2.55), we consider two special cases. First, let

$$y(\zeta) = x_1(\zeta) + x_2(\zeta) \tag{3.2.58}$$

where $x_1(\zeta)$ and $x_2(\zeta)$ are statistically independent. Then its characteristic function is given by

$$\Phi_y(\xi) = E\{e^{j\xi y(\zeta)}\} = E\{e^{j\xi[x_1(\zeta)+x_2(\zeta)]}\} = E\{e^{j\xi x_1(\zeta)}\}E\{e^{j\xi x_2(\zeta)}\} \tag{3.2.59}$$

where the last equality follows from the independence. Hence

$$\Phi_y(\xi) = \Phi_{x_1}(\xi)\Phi_{x_2}(\xi) \tag{3.2.60}$$

or from the convolution property of the Fourier transform

$$f_y(y) = f_{x_1}(y) * f_{x_2}(y) \tag{3.2.61}$$

From (3.2.60) the second characteristic function of $y(\zeta)$ is given by

$$\Psi_y(\xi) = \Psi_{x_1}(\xi) + \Psi_{x_2}(\xi) \tag{3.2.62}$$

or the mth-order cumulant of $y(\zeta)$ is given by

$$\kappa_y^{(m)} = \kappa_{x_1}^{(m)} + \kappa_{x_2}^{(m)} \tag{3.2.63}$$

These results can be easily generalized to the sum of M independent random variables.

EXAMPLE 3.2.1. Let $\{x_k(\zeta)\}_{k=1}^4$ be four IID random variables uniformly distributed over $[-0.5, 0.5]$. Compute and plot the pdfs of $y_M(\zeta) \triangleq \sum_{k=1}^M x_k$ for $M = 2$, 3, and 4. Compare these pdfs with that of a zero-mean Gaussian random variable.

Solution. Let $f(x)$ be the pdf of a uniform random variable over $[-0.5, 0.5]$, that is,

$$f(x) = \begin{cases} 1 & -0.5 \le x \le 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.64}$$

Then from (3.2.61)

$$f_{y_2}(y) = f(y) * f(y) = \begin{cases} 1+y & -1 \le y \le 0 \\ 1-y & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.65}$$

Similarly, we have

$$f_{y_3}(y) = f_{y_2}(y) * f(y) = \begin{cases} \frac{1}{2}(y+\frac{3}{2})^2 & -\frac{3}{2} \le y \le -\frac{1}{2} \\ \frac{3}{4} - y^2 & -\frac{1}{2} \le y \le \frac{1}{2} \\ \frac{1}{2}(y-\frac{3}{2})^2 & \frac{1}{2} \le y \le \frac{3}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.2.66}$$

and

$$f_{y_4}(y) = f_{y_3}(y) * f(y) = \begin{cases} \frac{1}{6}(y+2)^3 & -2 \le y \le -1 \\ -\frac{1}{2}y^3 - y^2 + \frac{2}{3} & -1 \le y \le 0 \\ \frac{2}{3} + \frac{1}{2}y^3 - y^2 & 0 \le y \le 1 \\ -\frac{1}{6}(-2+y)^3 & 1 \le y \le 2 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.67}$$

The plots of $f_{y_2}(y)$, $f_{y_3}(y)$, and $f_{y_4}(y)$ are shown in Figure 3.4 along with the zero-mean Gaussian pdf. The variance of the Gaussian random variable is chosen so that 99.92 percent of the pdf area is over $[-2, 2]$. We observe that as M increases, the pdf plots appear to get closer to the shape of the Gaussian pdf. This observation will be explored in detail in the CLT.

Next, let $y(\zeta) = ax(\zeta) + b$; then the characteristic function of $y(\zeta)$ is

$$\Phi_y(\xi) = E\{e^{j[ax(\zeta)+b]\xi}\} = E\{e^{ja\xi x(\zeta)} e^{jb\xi}\} = \Phi_x(a\xi) e^{jb\xi} \tag{3.2.68}$$

and by using the properties of the Fourier transform, the pdf of $y(\zeta)$ is given by

$$f_y(y) = \frac{1}{|a|} f_x\left(\frac{y-b}{a}\right) \tag{3.2.69}$$

From (3.2.68), the second characteristic function is given by

$$\Psi_y(\xi) = \Psi_x(a\xi) + jb\xi \tag{3.2.70}$$

and the cumulants are given by

$$\begin{aligned} \kappa_y^{(m)} &= (-j)^m \left.\frac{\mathrm{d}^m \Psi_y(\xi)}{\mathrm{d}\xi^m}\right|_{\xi=0} = a^m (-j)^m \left.\frac{\mathrm{d}^m \Psi_x(a\xi)}{\mathrm{d}\xi^m}\right|_{\xi=0} \\ &= a^m \kappa_x^{(m)} \qquad m > 1 \end{aligned} \tag{3.2.71}$$

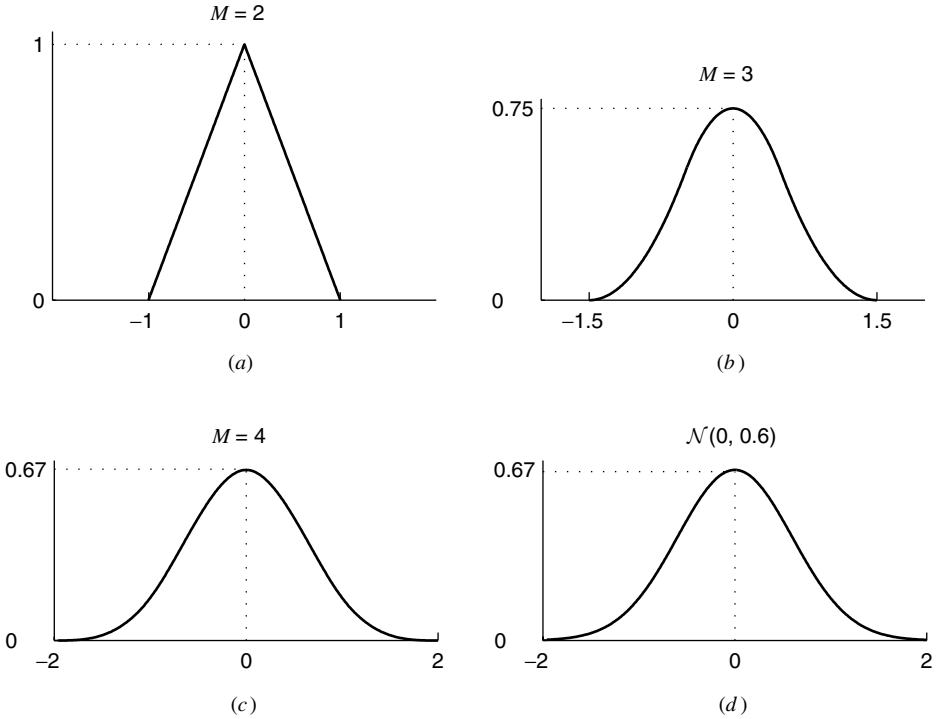


FIGURE 3.4
The pdf plots of (*a*) sum of two, (*b*) sum of three, (*c*) sum of four, and (*d*) Gaussian random variables in Example 3.2.1.

Finally, consider $y(\zeta)$ in (3.2.55). Using the results in (3.2.60) and (3.2.68), we have

$$\Phi_y(\xi) = \prod_{k=1}^{M} \Phi_{x_k}(c_k \xi) \tag{3.2.72}$$

from which the pdf of $y(\zeta)$ is given by

$$f_y(y) = \frac{1}{|c_1|} f_{x_1}\left(\frac{y}{c_1}\right) * \frac{1}{|c_2|} f_{x_2}\left(\frac{y}{c_2}\right) * \cdots * \frac{1}{|c_M|} f_{x_M}\left(\frac{y}{c_M}\right) \tag{3.2.73}$$

From (3.2.62) and (3.2.70), the second characteristic function is given by

$$\Psi_y(\xi) = \sum_{k=1}^{M} \Psi_{x_k}(c_k \xi) \tag{3.2.74}$$

and hence from (3.2.63) and (3.2.71), the cumulants of $y(\zeta)$ are

$$\kappa_y^{(m)} = \sum_{k=1}^{M} c_k^m \kappa_{x_k}^{(m)} \tag{3.2.75}$$

where c_k^m is the mth power of c_k.

In the following two examples, we consider two interesting cases in which the sum of IID random variables retains their original distribution. The first case concerns Gaussian random variables that have finite variances while the second case involves Cauchy random variables that possess infinite variance.

EXAMPLE 3.2.2. Let $x_k(\zeta) \sim \mathcal{N}(\mu_k, \sigma_k^2), k = 1, \ldots, M$ and let $y(k) = \sum_1^M x_k(\zeta)$. The characteristic function of $x_k(\zeta)$ is

$$\Phi_{x_k}(\xi) = \exp\left(j\mu_k\xi - \frac{\xi^2\sigma_k^2}{2}\right)$$

and hence from (3.2.72), we have

$$\Phi_y(\xi) = \exp\left(j\xi \sum_{k=1}^{M} \mu_k - \frac{\xi^2 \sum_{k=1}^{M} \sigma_k^2}{2}\right)$$

which means that $y(\zeta)$ is also a Gaussian random variable with mean $\sum_{k=1}^M \mu_k$ and variance $\sum_{k=1}^M \sigma_k^2$, that is, $y(\zeta) \sim \mathcal{N}(\sum_{k=1}^M \mu_k, \sum_{k=1}^M \sigma_k^2)$. In particular, if the $x_k(\zeta)$ are IID with a pdf $\mathcal{N}(\mu, \sigma^2)$, then

$$\Phi_y(\xi) = \exp\left(jM\mu\xi - \frac{\xi^2 M\sigma^2}{2}\right) = \exp\left[M\left(j\xi\mu - \frac{\xi^2\sigma^2}{2}\right)\right] \tag{3.2.76}$$

This behavior of $y(\zeta)$ is in contrast with that of the sum of the IID random variables in Example 3.2.1 in which the uniform pdf changed its form after M-fold convolutions.

EXAMPLE 3.2.3. As a second case, consider M IID random variables $\{x_k(\zeta)\}_{k=1}^M$ with Cauchy distribution

$$f_{x_k}(x) = \frac{\beta}{\pi}\frac{1}{(x-\alpha)^2 + \beta^2}$$

and let $y(k) = \sum_1^M x_k(\zeta)$. Then from (3.1.43), we have

$$\Phi_x(\xi) = \exp(j\alpha\xi - \beta|\xi|)$$

and hence

$$\Phi_y(\xi) = \exp(jM\alpha\xi - M\beta|\xi|) = \exp[M(j\alpha\xi - \beta|\xi|)] \tag{3.2.77}$$

This once again shows that the sum random variable has the same distribution (up to a scale factor) as that of the individual random variables, which in this case is the Cauchy distribution.

From these examples, we note that the Gaussian and the Cauchy random variables are *invariant*, or that they have a "self-reproducing" property under linear transformations. These two examples also raise some interesting questions. Are there any other random variables that possess this *invariance* property? If such random variables exist, what is the form of their pdfs or, alternatively, of their characteristic functions, and what can we say about their means and variances? From (3.2.76) and (3.2.77), observe that if the characteristic function has a general form

$$\Phi_{x_k}(\xi) = a^{\theta(\xi)} \tag{3.2.78}$$

where a is some constant and $\theta(\xi)$ is some function of ξ, then we have

$$\Phi_y(s) = a^{M\theta(s)} \tag{3.2.79}$$

that is, the characteristic function of the sum has the same *functional form* except for a change in scale. Are Gaussian and Cauchy both special cases of some general situation? These questions are answered by the concept of *stable* (more appropriately, linearly invariant or self-reproducing) *distributions*.

Stable distributions. These distributions satisfy the "stability" property, which in simple terms means that the distributions are preserved (or that they self-reproduce) under convolution. The only stable distribution that has finite variance is the Gaussian distribution, which has been well understood and is used extensively in the literature and in

practice. The remaining stable distributions have infinite variances (and in some cases, infinite means) which means that the corresponding random variables exhibit large fluctuations. These distributions can then be used to model signals with large variability and hence are finding increasing use in many diverse applications such as the gravitational fields of stars, temperature distributions in a nuclear reaction, or stock market fluctuations (Lamperti 1996; Samorodnitsky and Taqqu 1994; Feller 1966).

Before we formally define stable distributions, we introduce the following notation for convenience

$$y(\zeta) \stackrel{\mathrm{d}}{=} x(\zeta) \tag{3.2.80}$$

to indicate that the random variables $x(\zeta)$ and $y(\zeta)$ have the same distribution. For example, if $y(\zeta) = ax(\zeta) + b$, we have

$$F_y(y) = F_x\left(\frac{y-b}{a}\right) \tag{3.2.81}$$

and therefore $x(\zeta) \stackrel{\mathrm{d}}{=} ax(\zeta) + b$.

DEFINITION 3.2. Let $x_1(\zeta), x_2(\zeta), \ldots, x_M(\zeta)$ be IID random variables with a common distribution $F_x(x)$ and let $s_M(\zeta) = x_1(\zeta) + \cdots + x_M(\zeta)$ be their sum. The distribution $F_x(x)$ is said to be *stable* if for each M there exist constants $a_M > 0$ and b_M such that

$$s_M(\zeta) \stackrel{\mathrm{d}}{=} a_M x(\zeta) + b_M \tag{3.2.82}$$

and that $F_x(x)$ is not concentrated at one point.

If (3.2.82) holds for $b_M = 0$, we say that $F_x(x)$ is stable in the *strict sense*. The condition that $F_x(x)$ is not concentrated at one point is necessary because such a distribution is always stable. Thus it is a degenerate case that is of no practical interest. A stable distribution is called *symmetric stable* if the distribution is symmetric, which also implies that it is strictly stable.

It can be shown that for any stable random variable $x(\zeta)$ there is a number α, $0 < \alpha \leq 2$, such that the constant a_M in (3.2.82) is $a_M = M^{1/\alpha}$. The number α is known as the *index of stability* or *characteristic exponent*. A stable random variable $x(\zeta)$ with index α is called *α stable*.

Since there is no closed-form expression for the probability density function of stable random variables, except in special cases, they are specified by their characteristic function $\Phi(\xi)$. This characteristic function is given by

$$\Phi(\xi) = \begin{cases} \exp\{j\mu\xi - |\sigma\xi|^\alpha \cdot [1 - j\beta \operatorname{sign}(\xi) \tan\left(\frac{\pi\alpha}{2}\right)]\} & \alpha \neq 1 \\ \exp\{j\mu\xi - |\sigma\xi|^\alpha \cdot [1 - j\beta\left(\frac{2}{\pi}\right) \operatorname{sign}(\xi) \ln|\xi|]\} & \alpha = 1 \end{cases} \tag{3.2.83}$$

where $\operatorname{sign}(\xi) = \xi/|\xi|$ if $\xi \neq 0$ and zero otherwise. We shall use the notation $S_\alpha(\sigma, \beta, \mu)$ to denote the stable random variable defined by (3.2.83). The parameters in (3.2.83) have the following meaning:

1. The characteristic exponent α, $0 < \alpha \leq 2$, determines the shape of the distribution and hence the flatness of the tails.
2. The skewness (or alternatively, symmetry) parameter β, $-1 < \beta < 1$, determines the symmetry of the distribution: $\beta = 0$ specifies a symmetric distribution, $\beta < 0$ a left-skewed distribution, and $\beta > 0$ a right-skewed distribution.
3. The scale parameter σ, $0 \leq \sigma < \infty$, determines the range or dispersion of the stable distribution.
4. The location parameter μ, $-\infty < \mu < \infty$, determines the center of the distribution.

We next list some useful properties of stable random variables.

1. For $0 < \alpha < 2$, the tails of a stable distribution decay as a power law, that is,

$$\Pr[|x(\zeta) - \mu| \geq x] \simeq \frac{C}{x^{\alpha}} \qquad \text{as } x \to \infty \tag{3.2.84}$$

where C is a constant that depends on the scale parameter σ. As a result of this behavior, α-stable random variables have infinite second-order moments. In particular,

$$\begin{aligned} E\{|x(\zeta)|^p\} &< \infty \qquad \text{for any } 0 < p \leq \alpha \\ E\{|x(\zeta)|^p\} &= \infty \qquad \text{for any } p > \alpha \end{aligned} \tag{3.2.85}$$

Also $\text{var}[x(\zeta)] = \infty$ for $0 < \alpha < 2$, and $E\{|x(\zeta)|\} = \infty$ if $0 < \alpha < 1$.

2. A stable distribution is symmetric about μ iff $\beta = 0$. A symmetric α-stable distribution is denoted as $S\alpha S$, and its characteristic function is given by

$$\Phi(\xi) = \exp(j\mu\xi - |\sigma\xi|^{\alpha}) \tag{3.2.86}$$

3. If $x(\zeta)$ is $S\alpha S$ with $\alpha = 2$ in (3.2.83), we have a Gaussian distribution with variance equal to $2\sigma^2$, that is, $\mathcal{N}(\mu, 2\sigma^2)$, whose tails decay exponentially and not as a power law. Thus, the Gaussian is the only stable distribution with finite variance.
4. If $x(\zeta)$ is $S\alpha S$ with $\alpha = 1$, we have a Cauchy distribution with density

$$f_x(x) = \frac{\sigma/\pi}{(x-\mu)^2 + \sigma^2} \tag{3.2.87}$$

A standard ($\mu = 0, \sigma = 1$) Cauchy random variable $x(\zeta)$ can be generated from a $[0, 1]$ uniform random variable $u(\zeta)$, by using the transformation $x = \tan[\pi(u - \frac{1}{2})]$.

5. If $x(\zeta)$ is $S\alpha S$ with $\alpha = \frac{1}{2}$, we have a *Levy distribution*, which has both infinite variance and infinite mean. The pdf of this distribution does not have a functional form and hence must be computed numerically.

In Figure 3.5, we display characteristic and density functions of Gaussian, Cauchy, and Levy random variables. The density plots were computed numerically using the MATLAB function `stablepdf`.

Infinitely divisible distributions. A distribution $F_x(x)$ is infinitely divisible if and only if for each M there exists a distribution $F_M(x)$ such that

$$f_x(x) = f_M(x) * f_M(x) * \cdots * f_M(x) \tag{3.2.88}$$

or by using the convolution theorem,

$$\Phi_x(\xi) = \Phi_M(\xi)\ \Phi_M(\xi)\ \cdots\ \Phi_M(\xi) = \Phi_M^M(\xi) \tag{3.2.89}$$

that is, for each M the random variable $x(\zeta)$ can be represented as the sum $x(\zeta) = x_1(\zeta) + \cdots + x_M(\zeta)$ of M IID random variables with a *common* distribution $F_M(x)$. Clearly, all stable distributions are infinitely divisible. An example of an infinitely divisible pdf is shown in Figure 3.6 for $M = 4$, $\alpha = 1.5$, $\mu = 0$, and $\beta = 0$.

Central limit theorem. Consider the random variable $y(\zeta)$ defined in (3.2.55). We would like to know about the convergence of its distribution as $M \to \infty$. If $y(\zeta)$ is a sum of IID random variables with a stable distribution, the distribution of $y(\zeta)$ also converges to a stable distribution. What result should we expect if the individual distributions are not stable and, in particular, are of finite variance? As we observed in Example 3.2.1, the sum of uniformly distributed independent random variables appears to converge to a Gaussian distribution. Is this result valid for any other distribution? The following version of the CLT answers these questions.

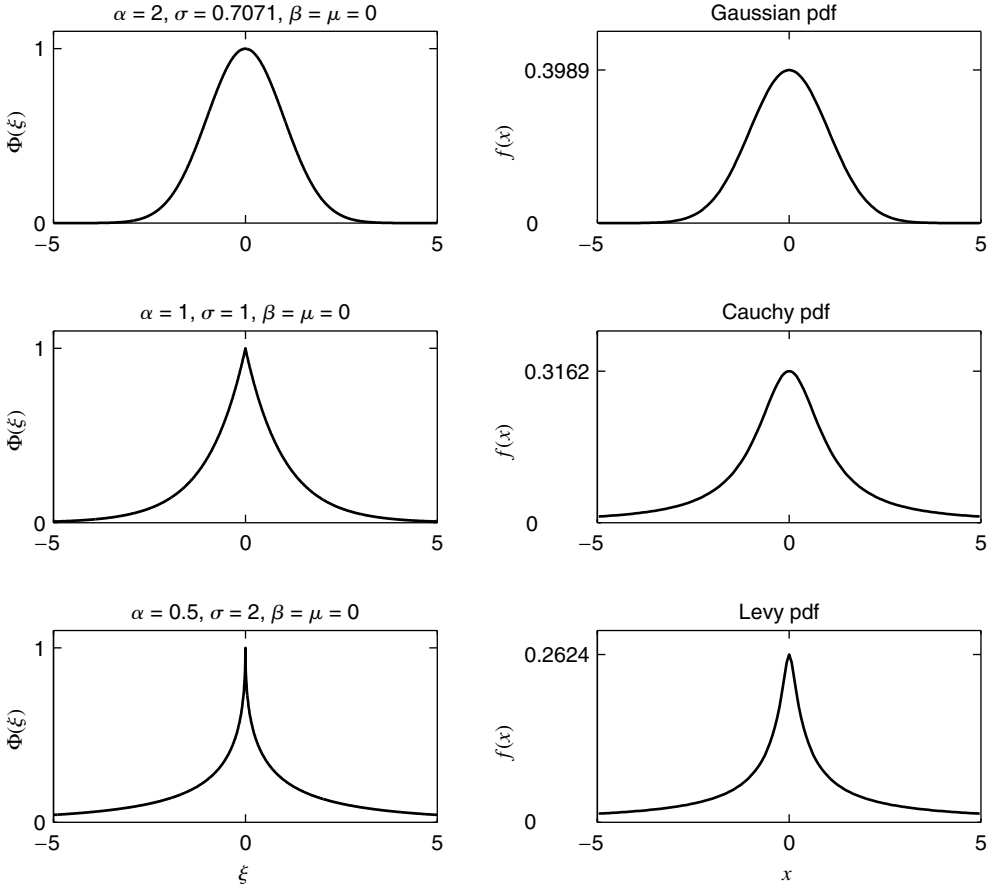


FIGURE 3.5
The characteristic and density function plots of Gaussian, Cauchy, and Levy random variables.

THEOREM 3.1 (CENTRAL LIMIT THEOREM). Let $\{x_k(\zeta)\}_{k=1}^{M}$ be a collection of random variables such that $x_1(\zeta), x_2(\zeta), \ldots, x_M(\zeta)$ (a) are mutually independent and (b) have the same distribution, and (c) the mean and variance of each random variable exist and are finite, that is, $\mu_{x_k} < \infty$ and $\sigma_{x_k}^2 < \infty$ for all $k = 1, 2, \ldots, M$. Then, the distribution of the normalized sum

$$y_M(\zeta) = \frac{\sum_{k=1}^{M} x_k(\zeta) - \mu_{y_M}}{\sigma_{y_M}}$$

approaches that of a normal random variable with zero mean and unit standard deviation as $M \to \infty$.

Proof. See Borkar (1995).

Comments. The following important comments are in order regarding the CLT.

1. Since we are assuming IID components in the normalized sum, the above theorem is known as the *equal-component case of the CLT*.
2. It should be emphasized that the convergence in the above theorem is in *distribution* (cdf) and not necessarily in density (pdf). Suppose we have M discrete and IID random variables. Then their normalized sum will always remain discrete no matter how large M is, but the distribution of the sum will converge to the the integral of the Gaussian pdf.

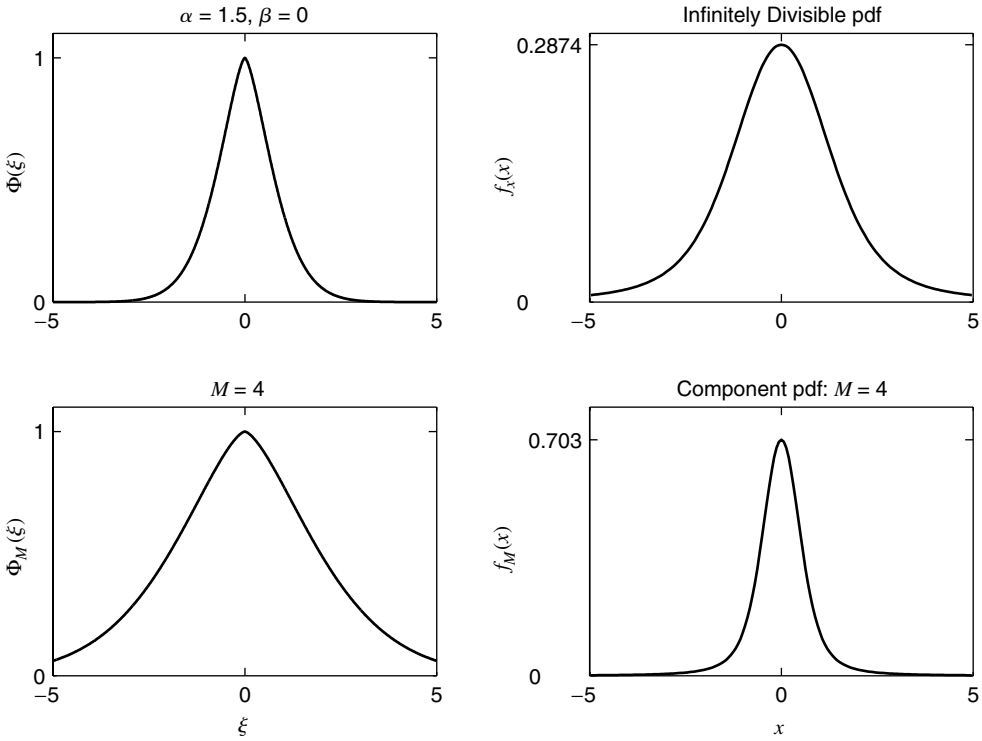


FIGURE 3.6
The characteristic and density function plots of an infinitely divisible distribution.

3. The word *central* in the CLT is a reminder that the distribution converges to the Gaussian distribution *around the center*, that is, around the mean. Note that while the limit distribution is found to be Gaussian, frequently the Gaussian limit gives a poor approximation for the tails of the actual distribution function of the sum when M is finite, even though the actual value under consideration might seem to be quite large.
4. As a final point, we note that in the above theorem the assumption of finite variance is critical to obtain a Gaussian limit. This implies that *all* distributions with finite variance will converge to the Gaussian when independent copies of their random variables are added. What happens if the variance is infinite? Then in this case the sum converges to one of the stable distributions. For example, as shown in Example 3.2.3, the sum of Cauchy random variables converges to a Cauchy distribution.

3.3 DISCRETE-TIME STOCHASTIC PROCESSES

After studying random variables and vectors, we can now extend these concepts to discrete-time signals (or sequences). Many natural sequences can be characterized as random signals because we cannot determine their values precisely, that is, they are unpredictable. A natural mathematical framework for the description of these discrete-time random signals is provided by discrete-time stochastic processes.

To obtain a formal definition, consider an experiment with a finite or infinite number of unpredictable outcomes from a sample space $\mathcal{S} = \{\zeta_1, \zeta_2, \ldots\}$, each occurring with a probability $\Pr\{\zeta_k\}, k = 1, 2, \ldots$. By some rule we assign to each element ζ_k of $\mathcal{S}$ a deterministic sequence $x(n, \zeta_k), -\infty < n < \infty$. The sample space S, the probabilities $\Pr\{\zeta_k\}$, and the sequences $x(n, \zeta_k), -\infty < n < \infty$, constitute a *discrete-time stochastic*

process or random sequence. Formally,

> $x(n, \zeta)$, $-\infty < n < \infty$, is a random sequence if for a fixed value n_0 of n, $x(n_0, \zeta)$ is a random variable.

The set of all possible sequences $\{x(n, \zeta)\}$ is called an *ensemble*, and each individual sequence $x(n, \zeta_k)$, corresponding to a specific value of $\zeta = \zeta_k$, is called a *realization* or a *sample sequence* of the ensemble.

There are four possible interpretations of $x(n, \zeta)$, depending on the character of n and ζ, as illustrated in Figure 3.7:

- $x(n, \zeta)$ is a random variable if n is *fixed* and ζ is a variable.
- $x(n, \zeta)$ is a sample sequence if ζ is *fixed* and n is a variable.
- $x(n, \zeta)$ is a number if both n and ζ are *fixed*.
- $x(n, \zeta)$ is a stochastic process if both n and ζ are variables.

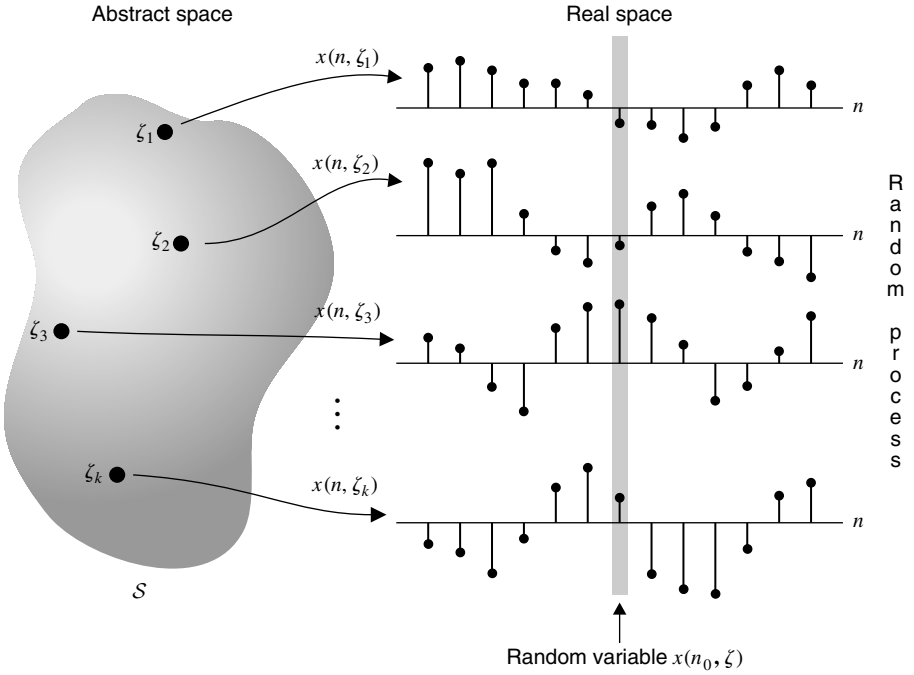


FIGURE 3.7
Graphical description of random sequences.

A random sequence is also called a *time series* in the statistics literature. It is a sequence of random variables, or it can be thought of as an *infinite-dimensional* random vector. As with any collection of infinite objects, one has to be careful with the asymptotic (or convergence) properties of a random sequence. If n is a continuous variable taking values in $\mathbb{R}$, then $x(n, \zeta)$ is an *uncountable* collection of random variables or an ensemble of waveforms. This ensemble is called a continuous-time stochastic process or a *random process*. Although these processes can be handled similarly to sequences, they are more difficult to deal with in a rigorous mathematical manner than sequences are. Furthermore, practical signal processing requires discrete-time signals. Hence in this book we consider random sequences rather than random waveforms.

Finally, in passing we note that the word *stochastic* is derived from the Greek word *stochasticos*, which means skillful in aiming or guessing. Hence, the terms *random process* and *stochastic process* will be used interchangeably throughout this book.

As mentioned before, a deterministic signal is by definition exactly predictable. This assumes that there exists a certain functional relationship that completely describes the signal, even if this relationship is not available. The unpredictability of a random process is, in general, the combined result of two things. First, the selection of a single realization is based on the outcome of a random experiment. Second, no functional description is available for all realizations of the ensemble. However, in some special cases, such a functional relationship is available. This means that after the occurrence of a specific realization, its future values can be predicted exactly from its past ones. If the future samples of any realization of a stochastic process can be predicted from the past ones, the process is called *predictable* or *deterministic*; otherwise, it is said to be a *regular process*. For example, the process $x(n, \zeta) = c$, where c is a random variable, is a predictable stochastic process because every realization is a discrete-time signal with constant amplitude. In practice, we most often deal with regular stochastic processes.

The simplest description of any random signal is provided by an amplitude-versus-time plot. Inspection of this plot provides qualitative information about some significant features of the signal that are useful in many applications. These features include, among others, the following:

1. The frequency of occurrence of various signal amplitudes, described by the probability distribution of samples.
2. The degree of dependence between two signal samples, described by the correlation between them.
3. The existence of "cycles" or quasi-periodic patterns, obtained from the signal power spectrum (which will be described in Section 3.3.6).
4. Indications of variability in the mean, variance, probability density, or spectral content.

The first feature above, the amplitude distribution, is obtained by plotting the histogram, which is an estimate of the first-order probability density of the underlying stochastic process. The probability density indicates waveform features such as "spikiness" and boundedness. Its form is crucial in the design of reliable estimators, quantizers, and event detectors.

The dependence between two signal samples (which are random variables) is given theoretically by the autocorrelation sequence and is quantified in practice by the empirical correlation (see Chapter 1), which is an estimate of the autocorrelation sequence of the underlying process. It affects the rate of amplitude change from sample to sample.

Cycles in the data are related to sharp peaks in the power spectrum or periodicity in the autocorrelation. Although the power spectrum and the autocorrelation contain the same information, they present it in different fashions.

Variability in a given quantity (e.g., variance) can be studied by evaluating this quantity for segments that can be assumed locally stationary and then analyzing the segment-to-segment variation. Such short-term descriptions should be distinguished from long-term ones, where the whole signal is analyzed as a single segment.

All the above features, to a lesser or greater extent, are interrelated. Therefore, it is impossible to point out exactly the effect of each one upon the visual appearance of the signal. However, a lot of insight can be gained by introducing the concepts of signal variability and signal memory, which are discussed in Sections 3.3.5 and 3.4.3 respectively.

3.3.1 Description Using Probability Functions

From Figure 3.7, it is clear that at $n = n_0$, $x(n_0, \zeta)$ is a random variable that requires a first-order probability function, say cdf $F_x(x; n_0)$, for its description. Similarly, $x(n_1, \zeta)$ and $x(n_2, \zeta)$ are joint random variables at instances n_1 and n_2, respectively, requiring a joint cdf $F_x(x_1, x_2; n_1, n_2)$. Stochastic processes contain infinitely many such random variables. Hence they are completely described, in a statistical sense, if their kth-order distribution

function

$$F_x(x_1, \ldots, x_k; n_1, \ldots, n_k) = \Pr\{x(n_1) \le x_1, \ldots, x(n_k) \le x_k\} \tag{3.3.1}$$

is known for every value of $k \ge 1$ and for all instances $n_1, n_2, \ldots, n_k$. The kth-order pdf is given by

$$f_x(x_1, \ldots, x_k; n_1, \ldots, n_k) \triangleq \frac{\partial^{2k} F_x(x_1, \ldots, x_k; n_1, \ldots, n_k)}{\partial x_{\mathrm{R}1} \cdots \partial x_{\mathrm{I}k}} \qquad k \ge 1 \tag{3.3.2}$$

Clearly, the probabilistic description requires a lot of information that is difficult to obtain in practice except for simple stochastic processes. However, many (but not all) properties of a stochastic process can be described in terms of *averages* associated with its first- and second-order densities.

For simplicity, in the rest of the book, we will use a compact notation $x(n)$ to represent either a random process $x(n, \zeta)$ or a single realization $x(n)$, which is a member of the ensemble. Thus we will drop the variable ζ from all notations involving random variables, vectors, or processes. We believe that this will not cause any confusion and that the exact meaning will be clear from the context. Also the random process $x(n)$ is assumed to be complex-valued unless explicitly specified as real-valued.

3.3.2 Second-Order Statistical Description

The second-order statistic of $x(n)$ at time n is specified by its *mean value* $\mu_x(n)$ and its *variance* $\sigma_x^2(n)$, defined by

$$\mu_x(n) = E\{x(n)\} = E\{x_{\mathrm{R}}(n) + jx_{\mathrm{I}}(n)\} \tag{3.3.3}$$

and

$$\sigma_x^2(n) = E\{|x(n) - \mu_x(n)|^2\} = E\{|x(n)|^2\} - |\mu_x(n)|^2 \tag{3.3.4}$$

respectively. Note that both $\mu_x(n)$ and $\sigma_x(n)$ are, in general, deterministic sequences.

The second-order statistics of $x(n)$ at two different times n_1 and n_2 are given by the two-dimensional autocorrelation (or autocovariance) sequences. The *autocorrelation sequence* of a discrete-time random process is defined as the joint moment of the random variables $x(n_1)$ and $x(n_2)$, that is,

$$r_{xx}(n_1, n_2) = E\{x(n_1)x^*(n_2)\} \tag{3.3.5}$$

It provides a measure of the dependence between values of the process at two different times. In this sense, it also provides information about the time variation of the process. The *autocovariance* sequence of $x(n)$ is defined by

$$\begin{aligned} \gamma_{xx}(n_1, n_2) &= E\{[x(n_1) - \mu_x(n_1)][x(n_2) - \mu_x(n_2)]^*\} \\ &= r_{xx}(n_1, n_2) - \mu_x(n_1)\mu_x^*(n_2) \end{aligned} \tag{3.3.6}$$

We will use notations such as $\gamma_x(n_1, n_2)$, $r_x(n_1, n_2)$, $\gamma(n_1, n_2)$, or $r(n_1, n_2)$ when there is no confusion as to which signal we are referring. Note that, in general, the second-order statistics are defined on a two-dimensional grid of integers.

The statistical relation between two stochastic processes $x(n)$ and $y(n)$ that are jointly distributed (i.e., they are defined on the same sample space $\mathcal{S}$) can be described by their *cross-correlation* and *cross-covariance* functions, defined by

$$r_{xy}(n_1, n_2) = E\{x(n_1)y^*(n_2)\} \tag{3.3.7}$$

and

$$\begin{aligned} \gamma_{xy}(n_1, n_2) &= E\{[x(n_1) - \mu_x(n_1)][y(n_2) - \mu_y(n_2)]^*\} \\ &= r_{xy}(n_1, n_2) - \mu_x(n_1)\mu_y^*(n_2) \end{aligned} \tag{3.3.8}$$

The *normalized cross-correlation* of two random processes $x(n)$ and $y(n)$ is defined by

$$\rho_{xy}(n_1, n_2) = \frac{\gamma_{xy}(n_1, n_2)}{\sigma_x(n_1)\sigma_y(n_2)} \tag{3.3.9}$$

Some definitions

We now describe some useful types of stochastic processes based on their statistical properties. A random process is said to be

- An *independent* process if

$$f_x(x_1, \ldots, x_k; n_1, \ldots, n_k) = f_1(x_1; n_1) \cdots f_k(x_k; n_k) \qquad \forall k, n_i, i = 1, \ldots, k \quad (3.3.10)$$

  that is, $x(n)$ is a sequence of independent random variables. If all random variables have the same pdf $f(x)$ for all k, then $x(n)$ is called an IID (independent and identically distributed) random sequence.
- An *uncorrelated* process if $x(n)$ is a sequence of uncorrelated random variables, that is,

$$\gamma_x(n_1, n_2) = \begin{Bmatrix} \sigma_x^2(n_1) & n_1 = n_2 \\ 0 & n_1 \neq n_2 \end{Bmatrix} = \sigma_x^2(n_1)\delta(n_1 - n_2) \quad (3.3.11)$$

  Alternatively, we have

$$r_x(n_1, n_2) = \begin{cases} \sigma_x^2(n_1) + |\mu_x(n_1)|^2 & n_1 = n_2 \\ \mu_x(n_1)\mu_x^*(n_2) & n_1 \neq n_2 \end{cases} \quad (3.3.12)$$

- An *orthogonal* process if it is a sequence of orthogonal random variables, that is,

$$r_x(n_1, n_2) = \begin{Bmatrix} \sigma_x^2(n_1) + |\mu_x(n_1)|^2 & n_1 = n_2 \\ 0 & n_1 \neq n_2 \end{Bmatrix} = E\{|x(n_1)|^2\}\delta(n_1 - n_2) \quad (3.3.13)$$

- An *independent increment* process if $\forall k > 1$ and $\forall n_1 < n_2 < \cdots < n_k$, the *increments*

$$\{x(n_1)\}, \{x(n_2) - x(n_1)\}, \ldots, \{x(n_k) - x(n_{k-1})\}$$

  are jointly independent. For such sequences, the kth-order probability function can be constructed as products of the probability functions of its increments.
- A *wide-sense periodic* (*WSP*) process with period N if

$$\mu_x(n) = \mu_x(n + N) \qquad \forall n \quad (3.3.14)$$

  and $$r_x(n_1, n_2) = r_x(n_1 + N, n_2) = r_x(n_1, n_2 + N) = r_x(n_1 + N, n_2 + N) \quad (3.3.15)$$

  Note that in the above definition, $\mu_x(n)$ is periodic in one dimension while $r_x(n_1, n_2)$ is periodic in two dimensions.
- A *wise-sense cyclostationary* process if there exists an integer N such that

$$\mu_x(n) = \mu_x(n + N) \qquad \forall n \quad (3.3.16)$$

  and $$r_x(n_1, n_2) = r_x(n_1 + N, n_2 + N) \quad (3.3.17)$$

  Note that in the above definition, $r_x(n_1, n_2)$ is *not* periodic in a two-dimensional sense. The correlation sequence is invariant to shift by N in *both* of its arguments.
- If all kth-order distributions of a stochastic process are jointly Gaussian, then it is called a Gaussian random sequence.

We can also extend some of these definitions to the case of two joint stochastic processes. The random processes $x(n)$ and $y(n)$ are said to be

- *Statistically independent* if for all values of n_1 and n_2

$$f_{xy}(x, y; n_1, n_2) = f_x(x; n_1) f_y(y; n_2) \quad (3.3.18)$$

- *Uncorrelated* if for every n_1 and n_2

$$\gamma_{xy}(n_1, n_2) = 0 \qquad \text{or} \qquad r_{xy}(n_1, n_2) = \mu_x(n_1)\mu_y^*(n_2) \quad (3.3.19)$$

- *Orthogonal* if for every n_1 and n_2

$$r_{xy}(n_1, n_2) = 0 \quad (3.3.20)$$

3.3.3 Stationarity

A random process $x(n)$ is called *stationary* if statistics determined for $x(n)$ are equal to those for $x(n+k)$, for every k. More specifically, we have the following definition.

DEFINITION 3.3 (STATIONARY OF ORDER N). A stochastic process $x(n)$ is called *stationary of order* N if

$$f_x(x_1, \ldots, x_N; n_1, \ldots, n_N) = f_x(x_1, \ldots, x_N; n_1 + k, \ldots, n_N + k) \tag{3.3.21}$$

for any value of k. If $x(n)$ is stationary for all orders $N = 1, 2, \ldots$, it is said to be *strict-sense stationary* (*SSS*).

An IID sequence is SSS. However, SSS is more restrictive than necessary for most practical applications. A more relaxed form of stationarity, which is sufficient for practical problems, occurs when a random process is *stationary up to order* 2, and it is also known as wide-sense stationarity.

DEFINITION 3.4 (WIDE-SENSE STATIONARITY). A random signal $x(n)$ is called *wide-sense stationary* (*WSS*) if

1. Its mean is a constant independent of n, that is,

$$E\{x(n)\} = \mu_x \tag{3.3.22}$$

2. Its variance is also a constant independent of n, that is,

$$\text{var}[x(n)] = \sigma_x^2 \tag{3.3.23}$$

and

3. Its autocorrelation depends only on the distance $l = n_1 - n_2$, called *lag*, that is,

$$r_x(n_1, n_2) = r_x(n_1 - n_2) = r_x(l) = E\{x(n+l)x^*(n)\} = E\{x(n)x^*(n-l)\} \tag{3.3.24}$$

From (3.3.22), (3.3.24), and (3.3.6) it follows that the autocovariance of a WSS signal also depends only on $l = n_1 - n_2$, that is,

$$\gamma_x(l) = r_x(l) - |\mu_x|^2 \tag{3.3.25}$$

EXAMPLE 3.3.1. Let $w(n)$ be a zero-mean, uncorrelated Gaussian random sequence with variance $\sigma^2(n) = 1$.

a. Characterize the random sequence $w(n)$.
b. Define $x(n) = w(n) + w(n-1)$, $-\infty < n < \infty$. Determine the mean and autocorrelation of $x(n)$. Also characterize $x(n)$.

Solution. Note that the variance of $w(n)$ is a constant.

a. Since uncorrelatedness implies independence for Gaussian random variables, $w(n)$ is an independent random sequence. Since its mean and variance are constants, it is at least stationary in the first order. Furthermore, from (3.3.12) or (3.3.13) we have

$$r_w(n_1, n_2) = \sigma^2 \delta(n_1 - n_2) = \delta(n_1 - n_2)$$

Hence $w(n)$ is also a WSS random process.

b. The mean of $x(n)$ is zero for all n since $w(n)$ is a zero-mean process. Consider

$$\begin{aligned} r_x(n_1, n_2) &= E\{x(n_1)x(n_2)\} \\ &= E\{[w(n_1) + w(n_1 - 1)][w(n_2) + w(n_2 - 1)]\} \\ &= r_w(n_1, n_2) + r_w(n_1, n_2 - 1) + r_w(n_1 - 1, n_2) \\ &\quad + r_w(n_1 - 1, n_2 - 1) \\ &= \sigma^2 \delta(n_1 - n_2) + \sigma^2 \delta(n_1 - n_2 + 1) \\ &\quad + \sigma^2 \delta(n_1 - 1 - n_2) + \sigma^2 \delta(n_1 - 1 - n_2 + 1) \\ &= 2\delta(n_1 - n_2) + \delta(n_1 - n_2 + 1) + \delta(n_1 - n_2 - 1) \end{aligned}$$

Clearly, $r_x(n_1, n_2)$ is a function of $n_1 - n_2$. Hence

$$r_x(l) = 2\delta(l) + \delta(l+1) + \delta(l-1)$$

Therefore, $x(n)$ is a WSS sequence. However, it is not an independent random sequence since both $x(n)$ and $x(n+1)$ depend on $w(n)$.

EXAMPLE 3.3.2 (WIENER PROCESS). Toss a fair coin at each n, $-\infty < n < \infty$. Let

$$w(n) = \begin{cases} +S & \text{if heads is outcome} \quad \Pr(\text{H}) = 0.5 \\ -S & \text{if tails is outcome} \quad \Pr(\text{T}) = 0.5 \end{cases}$$

where S is a step size. Clearly, $w(n)$ is an independent random process with

$$E\{w(n)\} = 0$$

and

$$E\{w^2(n)\} = \sigma_w^2 = S^2\left(\tfrac{1}{2}\right) + S^2\left(\tfrac{1}{2}\right) = S^2$$

Define a new random process $x(n)$, $n \geq 1$, as

$$\begin{aligned} x(1) &= w(1) \\ x(2) &= x(1) + w(2) = w(1) + w(2) \\ &\vdots \\ x(n) &= x(n-1) + w(n) = \sum_{i=1}^{n} w(i) \end{aligned}$$

Note that $x(n)$ is a running sum of independent steps or increments; thus it is an independent increment process. Such a sequence is called a *discrete Wiener process* or *random walk*. We can easily see that

$$E\{x(n)\} = E\left\{\sum_{i=1}^{n} w(i)\right\} = 0$$

and

$$E\{x^2(n)\} = E\left\{\sum_{i=1}^{n} w(i) \sum_{k=1}^{n} w(k)\right\} = E\left\{\sum_{i=1}^{n}\sum_{k=1}^{n} w(i)w(k)\right\}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{n} E\{w(i)w(k)\} = \sum_{i=1}^{n} E\{w^2(i)\} = nS^2$$

Therefore, random walk is a nonstationary (or *evolutionary*) process with zero mean and variance that grows with n, the number of steps taken.

It should be stressed at this point that although any strict-sense stationary signal is wide-sense stationary, the inverse is not always true, except if the signal is Gaussian. However in practice, it is very rare to encounter a signal that is stationary in the wide sense but not stationary in the strict sense (Papoulis 1991).

Two random signals $x(n)$ and $y(n)$ are called *jointly wide-sense stationary* if each is wide-sense stationary and their cross-correlation depends only on $l = n_1 - n_2$

$$r_{xy}(l) = E\{x(n)y^*(n-l)\};\ \gamma_{x,y}(l) = r_{xy}(l) - \mu_x\mu_y^* \tag{3.3.26}$$

Note that as a consequence of wide-sense stationarity the two-dimensional correlation and covariance sequences become one-dimensional sequences. This is a very important result that ultimately allows for a nice spectral description of stationary random processes.

Properties of autocorrelation sequences

The autocorrelation sequence of a stationary process has many important properties (which also apply to autocovariance sequences, but we will discuss mostly correlation sequences). Vector versions of these properties are discussed extensively in Section 3.4.4, and their proofs are explored in the problems.

PROPERTY 3.3.1. The average power of a WSS process $x(n)$ satisfies

$$r_x(0) = \sigma_x^2 + |\mu_x|^2 \geq 0 \tag{3.3.27}$$

and

$$r_x(0) \geq |r_x(l)| \qquad \text{for all } l \tag{3.3.28}$$

Proof. See Problem 3.21 and Property 3.3.6.

This property implies that the correlation attains its maximum value at zero lag and this value is nonnegative. The quantity $|\mu_x|^2$ is referred to as the *average dc power*, and the quantity $\sigma_x^2 = \gamma_x(0)$ is referred to as the *average ac power* of the random sequence. The quantity $r_x(0)$ then is the *total* average power of $x(n)$.

PROPERTY 3.3.2. The autocorrelation sequence $r_x(l)$ is a conjugate symmetric function of lag l, that is,

$$r_x^*(-l) = r_x(l) \tag{3.3.29}$$

Proof. It follows from Definition 3.4 and from (3.3.24).

PROPERTY 3.3.3. The autocorrelation sequence $r_x(l)$ is nonnegative definite; that is, for any $M > 0$ and any vector $\boldsymbol{\alpha} \in \mathbb{R}^M$

$$\sum_{k=1}^{M} \sum_{m=1}^{M} \alpha_k r_x(k-m) \alpha_m^* \geq 0 \tag{3.3.30}$$

This is a necessary and sufficient condition for a sequence $r_x(l)$ to be the autocorrelation sequence of a random sequence.

Proof. See Problem 3.22.

Since in this book we exclusively deal with wide-sense stationary processes, we will use the term *stationary* to mean wide-sense stationary. The properties of autocorrelation and cross-correlation sequences of jointly stationary processes, $x(n)$ and $y(n)$, are summarized in Table 3.1.

Although SSS and WSS forms are widely used in practice, there are processes with different forms of stationarity. Consider the following example.

EXAMPLE 3.3.3. Let $x(n)$ be a real-valued random process generated by the system

$$x(n) = \alpha x(n-1) + w(n) \qquad n \geq 0 \qquad x(-1) = 0 \tag{3.3.31}$$

where $w(n)$ is a stationary random process with mean μ_w and $r_w(l) = \sigma_w^2 \delta(l)$. The process $x(n)$ generated using (3.3.31) is known as a *first-order autoregressive*, or AR(1), process,[†] and the process $w(n)$ is known as a *white noise* process (defined in Section 3.3.6). Determine the mean $\mu_x(n)$ of $x(n)$ and comment on its stationarity.

Solution. To compute the mean of $x(n)$, we express it as a function of $\{w(n), w(n-1), \ldots, w(0)\}$ as follows

$$\begin{aligned}
x(0) &= \alpha x(-1) + w(0) = w(0) \\
x(1) &= \alpha x(0) + w(1) = \alpha w(0) + w(1) \\
&\vdots \\
x(n) &= \alpha^n w(0) + \alpha^{n-1} w(1) + \cdots + w(n) = \sum_{k=0}^{n} \alpha^k w(n-k)
\end{aligned}$$

[†]Note that from (3.3.31), $x(n-1)$ completely determines the distribution for $x(n)$, and $x(n)$ completely determines the distribution for $x(n+1)$, and so on. If

$$f_{x(n)|x(n-1)\ldots}(x_n | x_{n-1} \ldots) = f_{x(n)|x(n-1)}(x_n | x_{n-1})$$

then the process is termed a *Markov process*.

Hence the mean of $x(n)$ is given by

$$\mu_x(n) = E\left\{\sum_{k=0}^{n} \alpha^k w(n-k)\right\} = \mu_w\left(\sum_{k=0}^{n} \alpha^k\right) = \begin{cases} \dfrac{1-\alpha^{n+1}}{1-\alpha}\mu_w & \alpha \neq 1 \\ (n+1)\mu_w & \alpha = 1 \end{cases}$$

Clearly, the mean of $x(n)$ depends on n, and hence it is nonstationary. However, if we assume that $|\alpha| < 1$ (which implies that the system is BIBO stable), then as $n \to \infty$, we obtain

$$\mu_x(n) = \mu_w \frac{1-\alpha^{n+1}}{1-\alpha} \underset{n\to\infty}{\longrightarrow} \frac{\mu_w}{1-\alpha}$$

Thus $x(n)$ approaches first-order stationarity for large n. Similar analysis for the autocorrelation of $x(n)$ shows that $x(n)$ approaches wide-sense stationarity for large n (see Problem 3.23).

The above example illustrates a form of stationarity called *asymptotic* stationarity. A stochastic process $x(n)$ is *asymptotically stationary* if the statistics of random variables $x(n)$ and $x(n+k)$ become stationary as $k \to \infty$. When LTI systems are driven by zero-mean uncorrelated-component random processes, the output process becomes asymptotically stationary in the *steady state*. Another useful form of stationarity is given by stationary increments. If the increments $\{x(n) - x(n-k)\}$ of a process $x(n)$ form a stationary process for every k, we say that $x(n)$ is a process with *stationary increments*. Such processes can be used to model data in various practical applications (see Chapter 12).

The simplest way, to examine in practice if a real-world signal is stationary, is to investigate the physical mechanism that produces the signal. If this mechanism is time-invariant, then the signal is stationary. In case it is impossible to draw a conclusion based on physical considerations, we should rely on statistical methods (Bendat and Piersol 1986; Priestley 1981). Note that stationarity in practice means that a random signal has statistical properties that do not change over the time interval we observe the signal. For evolutionary signals the statistical properties change continuously with time. An example of a highly nonstationary random signal is the signals associated with the vibrations induced in space vehicles during launch and reentry. However, there is a kind of random signal whose statistical properties change slowly with time. Such signals, which are stationary over short periods, are called *locally stationary* signals. Many signals of great practical interest, such as speech, EEG, and ECG, belong to this family of signals.

Finally, we note that general techniques for the analysis of nonstationary signals do not exist. Thus only special methods that apply to specific types of nonstationary signals can be developed. Many such methods remove the nonstationary component of the signal, leaving behind another component that can be analyzed as stationary (Bendat and Piersol 1986; Priestley 1981).

3.3.4 Ergodicity

A stochastic process consists of the ensemble and a probability law. If this information is available, the statistical properties of the process can be determined in a quite straightforward manner. However, in the real world, we have access to only a limited number (usually one) of realizations of the process. The question that arises then is, Can we infer the statistical characteristics of the process from a single realization?

This is possible for the class of random processes that are called *ergodic* processes. Roughly speaking, ergodicity implies that all the statistical information can be obtained from any single representative member of the ensemble.

Time averages

All the statistical averages that we have defined up to this point are known as *ensemble averages* because they are obtained by "freezing" the time variable and averaging over the ensemble (see Fig. 3.7). Averages of this type are formally defined by using the expectation

operator $E\{\ \}$. Ensemble averaging is not used frequently in practice, because it is impractical to obtain the number of realizations needed for an accurate estimate. Thus the need for a different kind of average, based on only one realization, naturally arises. Obviously such an average can be obtained only by time averaging.

The *time average* of a quantity, related to a discrete-time random signal, is defined as

$$\langle(\cdot)\rangle \triangleq \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (\cdot) \tag{3.3.32}$$

Note that, owing to its dependence on a single realization, any time average is itself a random variable. The time average is taken over all time because all realizations of a stationary random process exist for all time; that is, they are power signals.

For every ensemble average we can define a corresponding time average. The following time averages are of special interest:

$$\begin{aligned}
\text{Mean value} &= \langle x(n)\rangle \\
\text{Mean square} &= \langle |x(n)|^2\rangle \\
\text{Variance} &= \langle |x(n) - \langle x(n)\rangle|^2\rangle \\
\text{Autocorrelation} &= \langle x(n)x^*(n-l)\rangle \\
\text{Autocovariance} &= \langle [x(n) - \langle x(n)\rangle][x(n-l) - \langle x(n)\rangle]^*\rangle \\
\text{Cross-correlation} &= \langle x(n)y^*(n-l)\rangle \\
\text{Cross-covariance} &= \langle [x(n) - \langle x(n)\rangle][y(n-l) - \langle y(n)\rangle]^*\rangle
\end{aligned} \tag{3.3.33}$$

It is necessary to mention at this point the remarkable similarity between time averages and the correlation sequences for deterministic power signals. Although this is just a formal similarity, due to the fact that random signals are power signals, both quantities have the same properties. However, we should always keep in mind that although time averages are random variables (because they are functions of ζ), the corresponding quantities for deterministic power signals are fixed numbers or deterministic sequences.

Ergodic random processes

As we have already mentioned, in many practical applications only one realization of a random signal is available instead of the entire ensemble. In general, a single member of the ensemble does not provide information about the statistics of the process. However, if the process is stationary and ergodic, then all statistical information can be derived from only one typical realization of the process.

A random signal $x(n)$ is called *ergodic*[†] if its ensemble averages equal appropriate time averages. There are several degrees of ergodicity (Papoulis 1991). We will discuss two of them: ergodicity in the mean and ergodicity in correlation.

DEFINITION 3.5 (ERGODIC IN THE MEAN). A random process $x(n)$ is ergodic *in the mean* if

$$\langle x(n)\rangle = E\{x(n)\} \tag{3.3.34}$$

DEFINITION 3.6 (ERGODIC IN CORRELATION). A random process $x(n)$ is *ergodic in correlation* if

$$\langle x(n)x^*(n-l)\rangle = E\{x(n)x^*(n-l)\} \tag{3.3.35}$$

Note that since $\langle x(n)\rangle$ is constant and $\langle x(n)x^*(n-l)\rangle$ is a function of l, if $x(n)$ is ergodic in both the mean and correlation, then it is also WSS. Thus only stationary signals can be ergodic. On the other hand, WSS does not imply ergodicity of any kind. Fortunately,

[†] Strictly speaking, the form of ergodicity that we will use is called *mean-square ergodicity* since the underlying convergence of random variables is in the mean-square sense (Stark and Woods 1994). Therefore, equalities in the definitions are in the mean-square sense.

in practice almost all stationary processes are also ergodic, which is very useful for the estimation of their statistical properties. From now on we will use the term *ergodic* to mean both ergodicity in the mean and ergodicity in correlation.

DEFINITION 3.7 (JOINT ERGODICITY). Two random signals are called *jointly ergodic* if they are individually ergodic and in addition

$$\langle x(n)y^*(n-l)\rangle = E\{x(n)y^*(n-l)\} \tag{3.3.36}$$

A physical interpretation of ergodicity is that one realization of the random signal $x(n)$, as time n tends to infinity, takes on values with the same statistics as the value $x(n_1)$, corresponding to all samples of the ensemble members at a given time $n = n_1$.

In practice, it is of course impossible to use the time-average formulas introduced above, because only finite records of data are available. In this case, it is common practice to replace the operator (3.3.32) by the operator

$$\langle(\cdot)\rangle_N = \frac{1}{2N+1}\sum_{n=-N}^{N}(\cdot) \tag{3.3.37}$$

to obtain *estimates* of the true quantities. Our desire in such problems is to find estimates that become increasingly accurate (in a sense to be defined in Section 3.6) as the length $2N + 1$ of the record of used data becomes larger.

Finally, to summarize, we note that whereas stationarity ensures the time invariance of the statistics of a random signal, ergodicity implies that any statistics can be calculated either by averaging over all members of the ensemble at a fixed time or by time-averaging over any single representative member of the ensemble.

3.3.5 Random Signal Variability

If we consider a stationary random sequence $w(n)$ that is IID with zero mean, its key characteristics depend on its first-order density. Figure 3.8 shows the probability density functions and sample realizations for IID processes with uniform, Gaussian, and Cauchy probability distributions. In the case of the uniform distribution, the amplitude of the random variable is limited to a range, with values occurring outside this interval with zero probability. On the other hand, the Gaussian distribution does not have a finite interval of support, allowing for the possibility of any value. The same is true of the Cauchy distribution, but its characteristics are dramatically different from those of the Gaussian distribution. The center lobe of the density is much narrower while the tails that extend out to infinity are significantly higher. As a result, the realization of the Cauchy random process contains numerous spikes or extreme values while the remainder of the process is more compact about the mean. Although the Gaussian random process allows for the possibility of large values, the probability of their occurrence is so small that they are not found in realizations of the process.

The major difference between the Gaussian and Cauchy distributions lies in the area found under the tails of the density as it extends out to infinity. This characteristic is related to the *variability* of the process. The heavy tails, as found in the Cauchy distribution, result in an abundance of spikes in the process, a characteristic referred to as *high variability*. On the other hand, a distribution such as the Gaussian does not allow for extreme values and indicates *low variability*. The extent of the variability of a given distribution is determined by the heaviness of the tails. Distributions with heavy tails are called *long-tailed* distributions and have been used extensively as models of impulsive random processes.

DEFINITION 3.8. A distribution is called *long-tailed* if its tails decay hyperbolically or algebraically as

$$Pr\{|x(n)| \geq x\} \sim Cx^{-\alpha} \qquad \text{as } x \to \infty \tag{3.3.38}$$

where C is a constant and the variable α determines the rate of decay of the distribution.

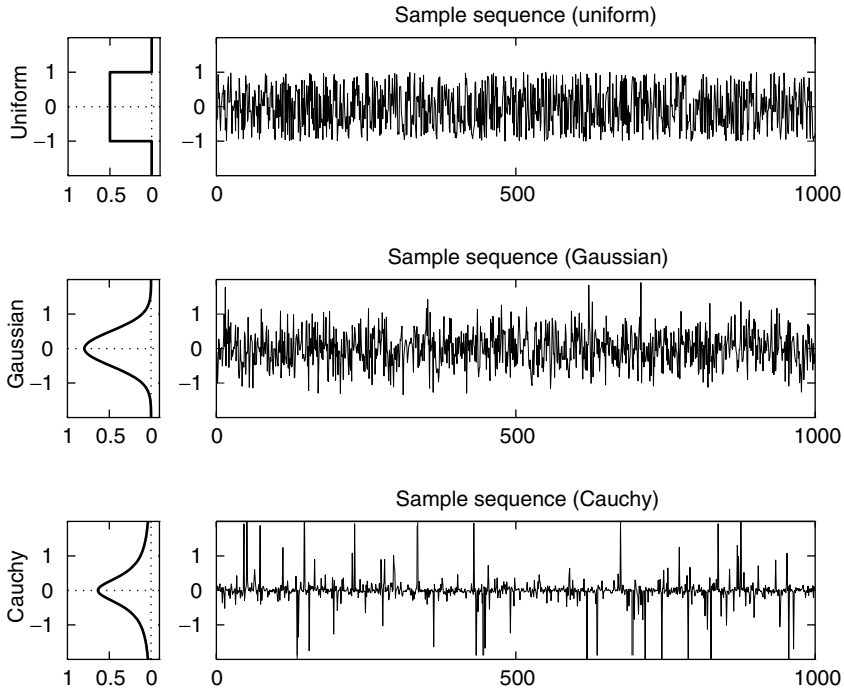


FIGURE 3.8
Probability density functions and sample realizations of an IID process with uniform, Gaussian, and Cauchy distributions.

By means of comparison, the Gaussian distribution has an exponential rate of decay. The implication of the algebraically decaying tail is that the process has infinite variance, that is,

$$\sigma_x^2 = E\{|x(n)|^2\} = \infty$$

and therefore lacks second-order moments. The lack of second-order moments means that, in addition to the variance, the correlation functions of these processes do not exist. Since most signal processing algorithms are based on second-order moment theory, infinite variance has some extreme implications for the way in which such processes are treated.

In this book, we shall model high variability, and hence infinite variance, using the family of symmetric stable distributions. The reason is twofold: First, a linear combination of stable random variables is stable. Second, stable distributions appear as limits in central limit theorems (see stable distributions in Section 3.2.4). Stable distributions are characterized by a parameter $\alpha, 0 < \alpha \leq 2$. They are Cauchy when $\alpha = 1$ and Gaussian when $\alpha = 2$. However, they have finite variance only when $\alpha = 2$.

In practice, the type of data under consideration governs the variability of the modeling distribution. Random signals restricted to a certain interval, such as the phase of complex random signals, are well suited for uniform distributions. On the other hand, signals allowing for any possible value but generally confined to a region are better suited for Gaussian models. However, if a process contains spikes and therefore has high variability, it is best characterized by a long-tailed distribution such as the Cauchy distribution. Impulsive signals have been found in a variety of applications, such as communication channels, radar signals, and electronic circuit noise. In all cases, the variability of the process dictates the appropriate model.

3.3.6 Frequency-Domain Description of Stationary Processes

Discrete-time stationary random processes have correlation sequences that are functions of a single index. This leads to nice and powerful representations in both the frequency and the z-transform domains.

Power spectral density

The *power spectral density* (PSD, or more appropriately autoPSD) of a stationary stochastic process $x(n)$ is a Fourier transformation of its autocorrelation sequence $r_x(l)$. If $r_x(l)$ is periodic (which corresponds to a wide-sense periodic stochastic process) in l, then the DTFS discussed in Section 2.2.1 can be used to obtain the PSD, which has the form of a *line spectrum*. If $r_x(l)$ is nonperiodic, the DTFT discussed in Section 2.2.1 can be used provided that $r_x(l)$ is absolutely summable. This means that the process $x(n)$ must be a zero-mean process. In general, a stochastic process can be a mixture of periodic and nonperiodic components.[†]

If we allow impulse functions in the DTFT to represent periodic (or almost periodic) sequences and non-zero-mean processes (see Section 2.2.1), then we can define the PSD as

$$R_x(e^{j\omega}) = \sum_{l=-\infty}^{\infty} r_x(l)e^{-j\omega l} \tag{3.3.39}$$

where ω is the frequency in radians per sample. If the process $x(n)$ is a zero-mean nonperiodic process, then (3.3.39) is enough to determine the PSD. If $x(n)$ is periodic (including nonzero mean) or almost periodic, then the PSD is given by

$$R_x(e^{j\omega}) = \sum_i 2\pi A_i \delta(\omega - \omega_i) \tag{3.3.40}$$

where the A_i are amplitudes of $r_x(l)$ at frequencies ω_i. For discussion purposes we will assume that $x(n)$ is a zero-mean nonperiodic process. The autocorrelation $r_x(l)$ can be recovered from the PSD by using the inverse DTFT as

$$r_x(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_x(e^{j\omega})e^{j\omega l}\,\mathrm{d}\omega \tag{3.3.41}$$

EXAMPLE 3.3.4. Determine the PSD of a zero-mean WSS process $x(n)$ with $r_x(l) = a^{|l|}$, $-1 < a < 1$.

Solution. From (3.3.39) we have

$$\begin{aligned} R_x(e^{j\omega}) &= \sum_{l=-\infty}^{\infty} a^{|l|}e^{-j\omega l} \qquad -1 < a < 1 \\ &= \frac{1}{1-ae^{j\omega}} + \frac{1}{1-ae^{-j\omega}} - 1 \\ &= \frac{1-a^2}{1+a^2-2a\cos\omega} \qquad -1 < a < 1 \end{aligned} \tag{3.3.42}$$

which is a real-valued, even, and nonnegative function of ω.

Properties of the autoPSD. The power spectral density $R_x(e^{j\omega})$ has three key properties that follow from corresponding properties of the autocorrelation sequence and the DTFT.

[†]Periodic components are predictable processes as discussed before. However, some nonperiodic components can also be predictable. Hence nonperiodic components are not always regular processes.

PROPERTY 3.3.4. The autoPSD $R_x(e^{j\omega})$ is a real-valued periodic function of frequency with period 2π for any (real- or complex-valued) process $x(n)$. If $x(n)$ is real-valued, then $R_x(e^{j\omega})$ is also an even function of ω, that is,

$$R_x(e^{j\omega}) = R_x(e^{-j\omega}) \tag{3.3.43}$$

Proof. It follows from autocorrelation and DTFT properties.

PROPERTY 3.3.5. The autoPSD is nonnegative definite, that is,

$$R_x(e^{j\omega}) \geq 0 \tag{3.3.44}$$

Proof. This follows from the nonnegative definiteness of the autocorrelation sequence [see also discussions leading to (3.4.27)].

PROPERTY 3.3.6. The area under $R_x(e^{j\omega})$ is *nonnegative* and it equals the average power of $x(n)$. Indeed, from (3.3.41) it follows with $l = 0$ that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} R_x(e^{j\omega})\,\mathrm{d}\omega = r_x(0) = E\{|x(n)|^2\} \geq 0 \tag{3.3.45}$$

Proof. It follows from Property 3.3.5.

White noise. A random sequence $w(n)$ is called a (second-order) *white noise process* with mean μ_w and variance σ_w^2, denoted by

$$w(n) \sim \mathrm{WN}(\mu_w, \sigma_w^2) \tag{3.3.46}$$

if and only if $E\{w(n)\} = \mu_w$ and

$$r_w(l) = E\{w(n)w^*(n-l)\} = \sigma_w^2\delta(l) \tag{3.3.47}$$

which implies that $$R_w(e^{j\omega}) = \sigma_w^2 \qquad -\pi \leq \omega \leq \pi \tag{3.3.48}$$

The term *white noise* is used to emphasize that all frequencies contribute the same amount of power, as in the case of white light, which is obtained by mixing all possible colors by the same amount. If, in addition, the pdf of $x(n)$ is Gaussian, then the process is called a (second-order) *white Gaussian noise* process, and it will be denoted by $\mathrm{WGN}(\mu_w, \sigma_w^2)$.

If the random variables $w(n)$ are independently and identically distributed with mean μ_w and variance σ_w^2, then we shall write

$$w(n) \sim \mathrm{IID}(\mu_w, \sigma_w^2) \tag{3.3.49}$$

This is sometimes referred to as a *strict* white noise.

We emphasize that the conditions of uncorrelatedness or independence do not put any restriction on the form of the probability density function of $w(n)$. Thus we can have an IID process with any type of probability distribution. Clearly, white noise is the simplest random process because it does not have any structure. However, we will see that it can be used as the basic building block for the construction of processes with more complicated dependence or correlation structures.

Harmonic processes. A *harmonic process* is defined by

$$x(n) = \sum_{k=1}^{M} A_k \cos(\omega_k n + \phi_k) \qquad \omega_k \neq 0 \tag{3.3.50}$$

where M, $\{A_k\}_1^M$, and $\{\omega_k\}_1^M$ are constants and $\{\phi_k\}_1^M$ are pairwise independent random variables uniformly distributed in the interval $[0, 2\pi]$. It can be shown (see Problem 3.9) that $x(n)$ is a stationary process with mean

$$E\{x(n)\} = 0 \qquad \text{for all } n \tag{3.3.51}$$

and autocorrelation

$$r_x(l) = \frac{1}{2}\sum_{k=1}^{M} A_k^2 \cos \omega_k l \qquad -\infty < l < \infty \tag{3.3.52}$$

We note that $r_x(l)$ consists of a sum of "in-phase" cosines with the same frequencies as in $x(n)$.

If $\omega_k/(2\pi)$ are rational numbers, $r_x(l)$ is periodic and can be expanded as a Fourier series. These series coefficients provide the power spectrum $R_x(k)$ of $x(n)$. However, because $r_x(l)$ is a linear superposition of cosines, it always has a line spectrum with $2M$ lines of strength $A_k^2/4$ at frequencies $\pm\omega_k$ in the interval $[-\pi, \pi]$. If $r_x(l)$ is periodic, then the lines are equidistant (i.e., harmonically related), hence the name *harmonic process*. If $\omega/(2\pi)$ is irrational, then $r_x(l)$ is almost periodic and can be treated in the frequency domain in almost the same fashion. Hence the power spectrum of a harmonic process is given by

$$R_x(e^{j\omega}) = \sum_{k=-M}^{M} 2\pi \left(\frac{A_k^2}{4}\right)\delta(\omega - \omega_k) = \sum_{k=-M}^{M} \frac{\pi}{2} A_k^2 \delta(\omega - \omega_k), \; -\pi < \omega \le \pi \tag{3.3.53}$$

EXAMPLE 3.3.5. Consider the following harmonic process

$$x(n) = \cos(0.1\pi n + \phi_1) + 2\sin(1.5n + \phi_2)$$

where ϕ_1 and ϕ_2 are IID random variables uniformly distributed in the interval $[0, 2\pi]$. The first component of $x(n)$ is periodic with $\omega_1 = 0.1\pi$ and period equal to 20 while the second component is almost periodic with $\omega_2 = 1.5$. Thus the sequence $x(n)$ is almost periodic. A sample function realization of $x(n)$ is shown in Figure 3.9(a). The mean of $x(n)$ is

$$\mu_x(n) = E\{x(n)\} = E\{\cos(0.1\pi n + \phi_1) + 2\sin(1.5n + \phi_2)\} = 0$$

and the autocorrelation sequence (using mutual independence between ϕ_1 and ϕ_2) is

$$\begin{aligned} r_x(n_1, n_2) &= E\{x(n_1)x_2^*(n_2)\} \\ &= E\{\cos(0.1\pi n_1 + \phi_1)\cos(0.1\pi n_2 + \phi_1)\} \\ &\quad + E\{2\sin(1.5n_1 + \phi_2)2\sin(1.5n_2 + \phi_2)\} \\ &= \tfrac{1}{2}\cos[0.1\pi(n_1 - n_2)] + 2\cos[1.5(n_1 - n_2)] \end{aligned}$$

or

$$r_x(l) = \tfrac{1}{2}\cos 0.1\pi l + 2\cos 1.5l \qquad l = n_1 - n_2$$

Thus the line spectrum $R_{\omega_k}^{(x)}$ is given by

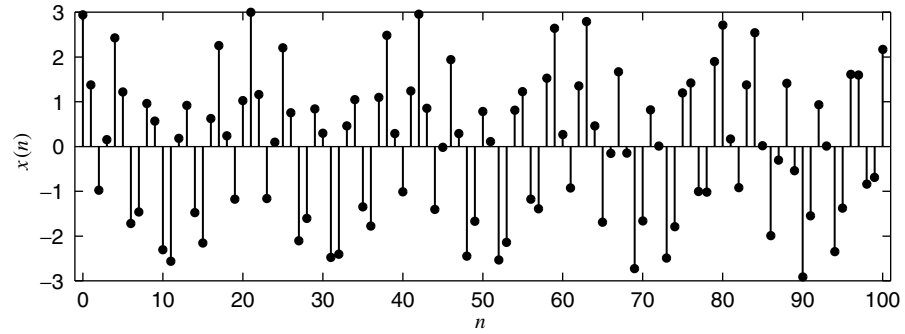
$$R_{\omega_k}^{(x)} = \begin{cases} 1 & \omega_1 = -1.5 \\ \frac{1}{4} & \omega_2 = -0.1\pi \\ \frac{1}{4} & \omega_3 = 0.1\pi \\ 1 & \omega_4 = 1.5 \end{cases}$$

and the power spectrum $R_x(e^{j\omega})$ is given by

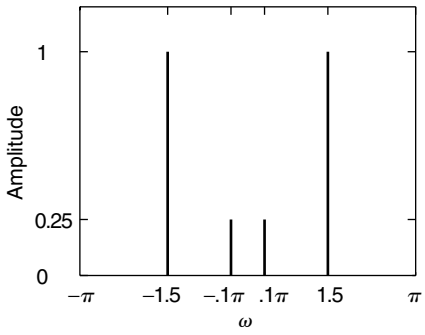
$$R_x(e^{j\omega}) = 2\pi\delta(\omega + 1.5) + \frac{\pi}{2}\delta(\omega + 0.1\pi) + \frac{\pi}{2}\delta(\omega - 0.1\pi) + 2\pi\delta(\omega - 1.5)$$

The line spectrum of $x(n)$ is shown in Figure 3.9(b) and the corresponding power spectrum in Figure 3.9(c).

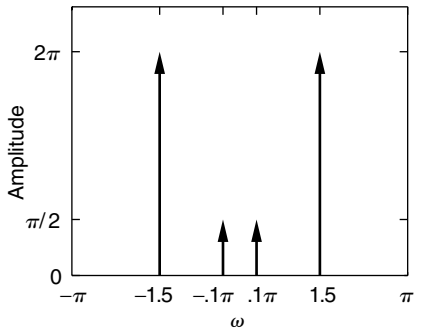
The harmonic process is predictable because any given realization is a sinusoidal sequence with fixed amplitude, frequency, and phase. We stress that the independence of the phases is required to guarantee the stationarity of $x(n)$ in (3.3.50). The uniform distribution of the phases is necessary to make $x(n)$ a stationary process (see Problem 3.9). The harmonic process (3.3.50), in general, is non-Gaussian; however, it becomes Gaussian if the amplitudes A_k are random variables with a Rayleigh distribution (Porat 1994).



(*a*) Sample function



(*b*) Line spectrum



(*c*) Power spectrum

FIGURE 3.9
The time and frequency-domain description of the harmonic process in Example 3.3.5.

EXAMPLE 3.3.6. Consider a complex-valued process given by

$$x(n) = Ae^{j\omega_0 n} = |A|e^{j(\omega_0 n+\phi)}$$

where A is a complex-valued random variable and ω_0 is constant. The mean of $x(n)$

$$E\{x(n)\} = E\{A\}e^{j\omega_0 n}$$

can be constant only if $E\{A\} = 0$. If $|A|$ is constant and ϕ is uniformly distributed on $[0, 2\pi]$, then we have $E\{A\} = |A|E\{e^{j\phi}\} = 0$. In this case the autocorrelation is

$$r_x(n_1, n_2) = E\{Ae^{j(\omega_0 n_1+\phi)}A^*e^{-j(\omega_0 n_2+\phi)}\} = |A|^2 e^{j(n_1-n_2)\omega_0}$$

Since the mean is constant and the autocorrelation depends on the difference $l \triangleq n_1 - n_2$, the process is wide-sense stationary.

The above example can be generalized to harmonic processes of the form

$$x(n) = \sum_{k=1}^{M} A_k e^{j(\omega_k n+\phi_k)} \tag{3.3.54}$$

where M, $\{A_k\}_1^M$, and $\{\omega_k\}_1^M$ are constants and $\{\phi_k\}_1^M$ are pairwise independent random variables uniformly distributed in the interval $[0, 2\pi]$. The autocorrelation sequence is

$$r_x(l) = \sum_{k=1}^{M} |A_k|^2 e^{j\omega_k l} \tag{3.3.55}$$

and the power spectrum consists of M impulses with amplitudes $2\pi|A_k|^2$ at frequencies ω_k. If the amplitudes $\{A_k\}_{k=1}^M$ are random variables, mutually independent of the random phases, the quantity $|A_k|^2$ is replaced by $E\{|A_k|^2\}$.

Cross-power spectral density

The cross-power spectral density of two zero-mean and jointly stationary stochastic processes provides a description of their statistical relations in the frequency domain and is defined as the DTFT of their cross-correlation, that is,

$$R_{xy}(e^{j\omega}) = \sum_{l=-\infty}^{\infty} r_{xy}(l)e^{-j\omega l} \tag{3.3.56}$$

The cross-correlation $r_{xy}(l)$ can be recovered by the inverse DTFT

$$r_{xy}(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_{xy}(e^{j\omega})e^{j\omega l}\,\mathrm{d}\omega \tag{3.3.57}$$

The cross-spectrum $R_{xy}(e^{j\omega})$ is, in general, a complex function of ω. From $r_{xy}(l) = r_{yx}^*(-l)$ it follows that

$$R_{xy}(e^{j\omega}) = R_{yx}^*(e^{j\omega}) \tag{3.3.58}$$

This implies that $R_{xy}(e^{j\omega})$ and $R_{yx}(e^{j\omega})$ have the same magnitude but opposite phase.

The normalized cross-spectrum

$$\mathcal{G}_{xy}(e^{j\omega}) \triangleq \frac{R_{xy}(e^{j\omega})}{\sqrt{R_x(e^{j\omega})}\sqrt{R_y(e^{j\omega})}} \tag{3.3.59}$$

is called the *coherence function*. Its squared magnitude

$$|\mathcal{G}_{xy}(e^{j\omega})|^2 = \frac{|R_{xy}(e^{j\omega})|^2}{R_x(e^{j\omega})R_y(e^{j\omega})} \tag{3.3.60}$$

is known as the *magnitude square coherence* (*MSC*) and can be thought of as a sort of correlation coefficient in the frequency domain. If $x(n) = y(n)$, then $\mathcal{G}_{xy}(e^{j\omega}) = 1$ (maximum correlation) whereas if $x(n)$ and $y(n)$ are uncorrelated, then $R_{xy}(l) = 0$ and hence $\mathcal{G}_{xy}(e^{j\omega}) = 0$. In other words, $0 \le |\mathcal{G}_{xy}(e^{j\omega})| \le 1$.

Complex spectral density functions

If the sequences $r_x(l)$ and $r_{xy}(l)$ are absolutely summable within a certain ring of the complex z plane, we can obtain their z-transforms

$$R_x(z) = \sum_{l=-\infty}^{\infty} r_x(l)z^{-l} \tag{3.3.61}$$

$$R_{xy}(z) = \sum_{l=-\infty}^{\infty} r_{xy}(l)z^{-l} \tag{3.3.62}$$

which are known as the *complex spectral density* and *complex cross-spectral density* functions, respectively. If the unit circle, defined by $z = e^{j\omega}$, is within the region of convergence of the above summations, then

$$R_x(e^{j\omega}) = R_x(z)|_{z=e^{j\omega}} \tag{3.3.63}$$

$$R_{xy}(e^{j\omega}) = R_{xy}(z)|_{z=e^{j\omega}} \tag{3.3.64}$$

The correlation and power spectral density properties of random sequences are summarized in Table 3.1.

EXAMPLE 3.3.7. Consider the random sequence given in Example 3.3.4 with autoPSD in (3.3.42)

$$R_x(e^{j\omega}) = \frac{1-a^2}{1+a^2-2a\cos\omega} \qquad |a| < 1$$

Determine the complex autoPSD $R_x(z)$.

Solution. The complex autoPSD is given by $R_x(z) = R_x(e^{j\omega})|_{e^{j\omega}=z}$. Since

$$\cos\omega = \frac{e^{j\omega} + e^{-j\omega}}{2} = \left.\frac{z + z^{-1}}{2}\right|_{z=e^{j\omega}}$$

we obtain

$$R_x(z) = \frac{1 - a^2}{1 + a^2 - 2a\left(\dfrac{z + z^{-1}}{2}\right)} = \frac{(a - a^{-1})z^{-1}}{1 - (a + a^{-1})z^{-1} + z^{-2}} \qquad |a| < |z| < \frac{1}{|a|}$$

Now the inverse z-transform of $R_x(z)$ determines the autocorrelation sequence $r_x(l)$, that is,

$$R_x(z) = \frac{(a - a^{-1})z^{-1}}{1 - (a + a^{-1})z^{-1} + z^{-2}} = \frac{(a - a^{-1})z^{-1}}{(1 - az^{-1})(1 - a^{-1}z^{-1})}$$

$$= \frac{1}{(1 - az^{-1})} - \frac{1}{(1 - a^{-1}z^{-1})} \qquad |a| < |z| < |a|^{-1}$$

or

$$r_x(l) = a^l u(l) + (a^{-1})^l u(-l - 1) = a^{|l|} \tag{3.3.65}$$

This approach can be used to determine autocorrelation sequences from autoPSD functions.

Table 3.1 provides a summary of correlation and spectral properties of stationary random sequences.

TABLE 3.1
Summary of correlation and spectral properties of stationary random sequences.

Definitions	
Mean value	$\mu_x = E\{x(n)\}$
Autocorrelation	$r_x(l) = E\{[x(n)x^*(n-l)\}$
Autocovariance	$\gamma_x(l) = E\{[x(n) - \mu_x][x(n-l) - \mu_x]^*\}$
Cross-correlation	$r_{xy}(l) = E\{x(n)y^*(n-l)\}$
Cross-covariance	$\gamma_{xy}(l) = E\{[x(n) - \mu_x][y(n-l) - \mu_y]^*\}$
Power spectral density	$R_x(e^{j\omega}) = \sum_{l=-\infty}^{\infty} r_x(l)e^{-j\omega l}$
Cross-power spectral density	$R_{xy}(e^{j\omega}) = \sum_{l=-\infty}^{\infty} r_{xy}(l)e^{-j\omega l}$
Magnitude square coherence	$\lvert\mathcal{G}_{xy}(e^{j\omega})\rvert^2 = \lvert R_{xy}(e^{j\omega})\rvert^2 / [R_x(e^{j\omega})R_y(e^{j\omega})]$

Interrelations
$\gamma_x(l) = r_x(l) - \lvert\mu_x\rvert^2$
$\gamma_{xy}(l) = r_{xy}(l) - \mu_x\mu_y^*$

Properties	
Autocorrelation	**Auto-PSD**
$r_x(l)$ is nonnegative definite	$R_x(e^{j\omega}) \geq 0$ and real
$r_x(l) = r_x^*(-l)$	$R_x(e^{j\omega}) = R_x(e^{-j\omega})$ [real $x(n)$]
$\lvert r_x(l)\rvert \leq r_x(0)$	$R_x(z) = R_x^*(1/z^*)$
$\lvert\rho_x(l)\rvert \leq 1$	$R_x(z) = R_x(z^{-1})$ [real $x(n)$]
Cross-correlation	**Cross-PSD**
$r_{xy}(l) = r_{yx}^*(-l)$	
$\lvert r_{xy}(l)\rvert \leq [r_x(0)r_y(0)]^{1/2} \leq \frac{1}{2}[r_x(0) + r_y(0)]$	$R_{xy}(z) = R_{yx}^*(1/z^*)$ $0 \leq \lvert\mathcal{G}_{xy}(e^{j\omega})\rvert \leq 1$
$\lvert\rho_{xy}(l)\rvert \leq 1$	

3.4 LINEAR SYSTEMS WITH STATIONARY RANDOM INPUTS

This section deals with the processing of stationary random sequences using linear, time-invariant (LTI) systems. We focus on expressing the second-order statistical properties of the output in terms of the corresponding properties of the input and the characteristics of the system.

3.4.1 Time-Domain Analysis

The first question to ask when we apply a random signal to a system is, Just what is the meaning of such an operation? We ask this because a random process is not just a single sequence but an ensemble of sequences (see Section 3.3). However, since each realization of the stochastic process is a deterministic signal, it is an acceptable input producing an output that is clearly a single realization of the output stochastic process. For an LTI system, each pair of input-output realizations is described by the convolution summation

$$y(n,\zeta) = \sum_{k=-\infty}^{\infty} h(k)x(n-k,\zeta) \tag{3.4.1}$$

If the sum in the right side of (3.4.1) exists for all ζ such that $\Pr\{\zeta\} = 1$, then we say that we have almost-everywhere convergence or convergence with probability 1 (Papoulis 1991). The existence of such convergence is ruled by the following theorem (Brockwell and Davis 1991).

THEOREM 3.2. If the process $x(n,\zeta)$ is stationary with $E\{|x(n,\zeta)|\} < \infty$ and if the system is BIBO-stable, that is, $\sum_{-\infty}^{\infty} |h(k)| < \infty$, then the output $y(n,\zeta)$ of the system in (3.4.1) converges absolutely with probability 1, or

$$y(n,\zeta) = \sum_{k=-\infty}^{\infty} h(k)x(n-k,\zeta) \qquad \text{for all } \zeta \in \mathcal{A}, \Pr\{\mathcal{A}\} = 1 \tag{3.4.2}$$

and is stationary. Furthermore, if $E\{|x(n,\zeta)|^2\} < \infty$, then $E\{|y(n,\zeta)|^2\} < \infty$ and $y(n,\zeta)$ converges in the mean square to the same limit and is stationary.

A less restrictive condition of finite energy on the system impulse response $h(n)$ also guarantees the mean square existence of the output process, as stated in the following theorem.

THEOREM 3.3. If the process $x(n,\zeta)$ is zero-mean and stationary with $\sum_{l=-\infty}^{\infty} |r_x(l)| < \infty$, and if the system (3.4.1) satisfies the condition

$$\sum_{n=-\infty}^{\infty} |h(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, \mathrm{d}\omega < \infty \tag{3.4.3}$$

then the output $y(n,\zeta)$ converges in the mean square sense and is stationary.

The above two theorems are applicable when input processes have finite variances. However, IID sequences with α-stable distributions have infinite variances. If the impulse response of the system in (3.4.1) decays fast enough, then the following theorem (Brockwell and Davis 1991) guarantees the absolute convergence of $y(n,\zeta)$ with probability 1. These issues are of particular importance for inputs with high variability and are discussed in Section 3.3.5.

THEOREM 3.4. Let $x(n,\zeta)$ be an IID sequence of random variables with α-stable distribution, $0 < \alpha < 2$. If the impulse response $h(n)$ satisfies

$$\sum_{n=-\infty}^{\infty} |h(n)|^{\delta} < \infty \qquad \text{for some } \delta \in (0,\alpha)$$

then the output $y(n,\zeta)$ in (3.4.1) converges absolutely with probability 1.

Clearly, a complete description of the output stochastic process $y(n)$ requires the computation of an infinite number of convolutions. Thus, a better alternative would be to determine the statistical properties of $y(n)$ in terms of the statistical properties of the input and the characteristics of the system. For Gaussian signals, which are used very often in practice, first- and second-order statistics are sufficient.

Output mean value. If $x(n)$ is stationary, its first-order statistic is determined by its mean value μ_x. To determine the mean value of the output, we take the expected value of both sides of (3.4.1):

$$\mu_y = \sum_{k=-\infty}^{\infty} h(k)E\{x(n-k)\} = \mu_x \sum_{k=-\infty}^{\infty} h(k) = \mu_x H(e^{j0}) \tag{3.4.4}$$

Since μ_x and $H(e^{j0})$ are constant, μ_y is also constant. Note that $H(e^{j0})$ is the dc gain of the spectrum.

Input-output cross-correlation. If we take complex conjugate of (3.4.1), premultiply it by $x(n+l)$, and take the expectation of both sides, we have

$$E\{x(n+l)y^*(l)\} = \sum_{k=-\infty}^{\infty} h^*(k)E\{x(n+l)x^*(n-k)\}$$

or
$$r_{xy}(l) = \sum_{k=-\infty}^{\infty} h^*(k)r_{xx}(l+k) = \sum_{m=-\infty}^{\infty} h^*(-m)r_{xx}(l-m)$$

Hence,
$$r_{xy}(l) = h^*(-l) * r_{xx}(l) \tag{3.4.5}$$

Similarly,
$$r_{yx}(l) = h(l) * r_{xx}(l) \tag{3.4.6}$$

Output autocorrelation. Postmultiplying both sides of (3.4.1) by $y^*(n-l)$ and taking the expectation, we obtain

$$E\{y(n)y^*(n-l)\} = \sum_{k=-\infty}^{\infty} h(k)E\{x(n-k)y^*(n-l)\} \tag{3.4.7}$$

or
$$r_{yy}(l) = \sum_{k=-\infty}^{\infty} h(k)r_{xy}(l-k) = h(l) * r_{xy}(l) \tag{3.4.8}$$

From (3.4.5) and (3.4.8) we get

$$r_y(l) = h(l) * h^*(-l) * r_x(l) \tag{3.4.9}$$

or
$$r_y(l) = r_h(l) * r_x(l) \tag{3.4.10}$$

where
$$r_h(l) \triangleq h(l) * h^*(-l) = \sum_{n=-\infty}^{\infty} h(n)h^*(n-l) \tag{3.4.11}$$

is the autocorrelation of the impulse response and is called the system correlation sequence.

Since μ_y is constant and $r_y(l)$ depends only on the lag l, the response of a stable system to a stationary input is also a stationary process. A careful examination of (3.4.10) shows that *when a signal $x(n)$ is filtered by an LTI system with impulse response $h(n)$ its autocorrelation is "filtered" by a system with impulse response equal to the autocorrelation of its impulse response*, as shown in Figure 3.10.

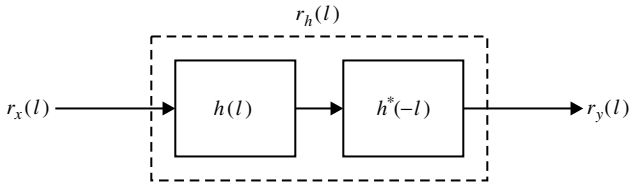


FIGURE 3.10
An equivalent LTI system for autocorrelation filtration.

Output power. The power $E\{|y(n)|^2\}$ of the output process $y(n)$ is equal to $r_y(0)$, which from (3.4.9) and (3.4.10) and the symmetry property of $r_x(l)$ is

$$\begin{aligned} P_y = r_y(0) &= r_h(l) * r_x(l)|_{l=0} \\ &= \sum_{k=-\infty}^{\infty} r_h(k) r_x(-k) = \sum_{k=-\infty}^{\infty} [h(k) * h^*(-k)] r_x(k) \\ &= \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h(m) h^*(m-k) r_x(k) \qquad (3.4.12) \\ &= \sum_{k=-\infty}^{\infty} r_h(k) r_x(k) \qquad (3.4.13) \end{aligned}$$

or for FIR filters with $\mathbf{h} = [h(0)\ h(1)\ \cdots\ h(M-1)]^T$, (3.4.12) can be written as

$$P_y = \mathbf{h}^H \mathbf{R}_x \mathbf{h} \qquad (3.4.14)$$

Finally, we note that when $\mu_x = 0$, we have $\mu_y = 0$ and $\sigma_y^2 = P_y$.

Output probability density function. Finding the probability density of the output of an LTI system is very difficult, except in some special cases. Thus, if $x(n)$ is a Gaussian process, then the output is also a Gaussian process with mean and autocorrelation given by (3.4.4) and (3.4.10). Also if $x(n)$ is IID, the probability density of the output is obtained by noting that $y(n)$ is a weighted sum of independent random variables. Indeed, the probability density of the sum of independent random variables is the convolution of their probability densities or the products of their characteristic functions. Thus if the input process is an IID stable process then the output process is also stable whose probability density can be computed by using characteristic functions.

3.4.2 Frequency-Domain Analysis

To obtain the output autoPSD and complex autoPSD, we recall that if $H(z) = \mathcal{Z}\{h(n)\}$, then, for real $h(n)$,

$$\mathcal{Z}\{h^*(-n)\} = H^*\left(\frac{1}{z^*}\right) \qquad (3.4.15)$$

From (3.4.5), (3.4.6), and (3.4.9) we obtain

$$R_{xy}(z) = H^*\left(\frac{1}{z^*}\right) R_x(z) \qquad (3.4.16)$$

$$R_{yx}(z) = H(z) R_x(z) \qquad (3.4.17)$$

and

$$R_y(z) = H(z) H^*\left(\frac{1}{z^*}\right) R_x(z) \qquad (3.4.18)$$

For a stable system, the unit circle $z = e^{j\omega}$ lies within the ROCs of $H(z)$ and $H(z^{-1})$. Thus,

$$R_{xy}(e^{j\omega}) = H^*(e^{j\omega})R_x(e^{j\omega}) \tag{3.4.19}$$

$$R_{yx}(e^{j\omega}) = H(e^{j\omega})R_x(e^{j\omega}) \tag{3.4.20}$$

and

$$R_y(e^{j\omega}) = H(e^{j\omega})H^*(e^{j\omega})R_x(e^{j\omega}) \tag{3.4.21}$$

or

$$R_y(e^{j\omega}) = |H(e^{j\omega})|^2 R_x(e^{j\omega}) \tag{3.4.22}$$

Thus, if we know the input and output autocorrelations or autospectral densities, we can determine the magnitude response of a system, but not its phase response. Only cross-correlation or cross-spectral densities can provide phase information [see (3.4.19) and (3.4.20)].

It can easily be shown that the power of the output is

$$E\{|y(n)|^2\} = r_{yy}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 R_x(e^{j\omega})\,\mathrm{d}\omega \tag{3.4.23}$$

$$= \sum_{l=-\infty}^{\infty} r_x(l)r_h(l) \tag{3.4.24}$$

which is equivalent to (3.4.13).

Consider now a narrowband filter with frequency response

$$H(e^{j\omega}) = \begin{cases} 1 & \omega_c - \dfrac{\triangle\omega}{2} \le \omega \le \omega_c + \dfrac{\triangle\omega}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{3.4.25}$$

The power of the filter output is

$$E\{|y(n)|^2\} = \frac{1}{2\pi}\int_{\omega_c-\triangle\omega/2}^{\omega_c+\triangle\omega/2} R_x(e^{j\omega})\,\mathrm{d}\omega \simeq \frac{\Delta\omega}{\pi} R_x(e^{j\omega_c}) \tag{3.4.26}$$

assuming that $\triangle\omega$ is sufficiently small and that $R_x(e^{j\omega})$ is continuous at $\omega = \omega_c$. Since $E\{|y(n)|^2\} \ge 0$, $R_x(e^{j\omega_c})$ is also nonnegative for all ω_c and $\triangle\omega$, hence

$$R_x(e^{j\omega}) \ge 0 \qquad -\pi \le \omega \le \pi \tag{3.4.27}$$

Hence, the PSD $R_x(e^{j\omega})$ is nonnegative definite for any random sequence $x(n)$ real (or complex). Furthermore, $R_x(e^{j\omega})\,\mathrm{d}\omega/(2\pi)$, has the interpretation of power, or $R_x(e^{j\omega})$ is a power density as a function of frequency (in radians per sample). Table 3.2 shows various input-output relationships in both the time and frequency domains.

TABLE 3.2
Second-order moments of stationary random sequences processed by linear, time-invariant systems.

Time domain	Frequency domain	z Domain
$y(n) = h(n) * x(n)$	Not available	Not available
$r_{yx}(l) = h(l) * r_x(l)$	$R_{yx}(e^{j\omega}) = H(e^{j\omega})R_x(e^{j\omega})$	$R_{yx}(z) = H(z)R_x(z)$
$r_{xy}(l) = h^*(-l) * r_x(l)$	$R_{xy}(e^{j\omega}) = H^*(e^{j\omega})R_x(e^{j\omega})$	$R_{xy}(z) = H^*(1/z^*)R_x(z)$
$r_y(l) = h(l) * r_{xy}(l)$	$R_y(e^{j\omega}) = H(e^{j\omega})R_{xy}(e^{j\omega})$	$R_y(z) = H(z)R_{xy}(z)$
$r_y(l) = h(l) * h^*(-l) * r_x(l)$	$R_y(e^{j\omega}) = \lvert H(e^{j\omega})\rvert^2 R_x(e^{j\omega})$	$R_y(z) = H(z)H^*(1/z^*)R_x(z)$

3.4.3 Random Signal Memory

Given the "zero-memory" process $w(n) \sim \mathrm{IID}(0, \sigma_w^2)$, we can introduce dependence by passing it though an LTI system. The extent and degree of the imposed dependence are dictated by the shape of the system's impulse response. The probability density of $w(n)$ is

not explicitly involved. Suppose now that we are given the resulting linear process $x(n)$, and we want to quantify its memory. For processes with finite variance we can use the *correlation length*

$$L_c = \frac{1}{r_x(0)} \sum_{l=0}^{\infty} r_x(l) = \sum_{l=0}^{\infty} \rho_x(l)$$

which equals the area under the normalized autocorrelation sequence curve and shows the maximum distance at which two samples are significantly correlated.

An IID process has no memory and is completely described by its first-order density. A linear process has memory introduced by the impulse response of the generating system. If $w(n)$ has finite variance, the memory of the process is determined by the autocorrelation of the impulse response because $r_x(l) = \sigma_w^2 r_h(l)$. Also, the higher-order densities of the process are nonzero. Thus, the variability of the output—that is, what amplitudes the signal takes, how often, and how fast the amplitude changes from sample to sample—is the combined effect of the input probability density and the system memory.

DEFINITION 3.9. A stationary process $x(n)$ with finite variance is said to have *long memory* if there exist constants α, $0 < \alpha < 1$, and $C_r > 0$ such that

$$\lim_{l \to \infty} \frac{1}{C_r \sigma_x^2} r_x(l) l^{\alpha} = 1$$

This implies that the autocorrelation has fat or heavy tails, that is, asymptotically decays as a power law

$$\rho_x(l) \simeq C_r |l|^{-\alpha} \qquad \text{as } l \to \infty$$

and slowly enough that

$$\sum_{l=-\infty}^{\infty} \rho_x(l) = \infty$$

that is, a long-memory process has infinite correlation length. If

$$\sum_{l=-\infty}^{\infty} \rho_x(l) < \infty$$

we say that that the process has *short memory*. This is the case for autocorrelations that decay exponentially, for example, $\rho_x(l) = a^{|l|}$, $-1 < a < 1$.

An equivalent definition of long memory can be formulated in terms of the power spectrum (Beran 1994; Samorodnitsky and Taqqu 1994).

DEFINITION 3.10. A stationary process $x(n)$ with finite variance is said to have *long memory* if there exist constants β, $0 < \beta < 1$, and $C_R > 0$ such that

$$\lim_{\omega \to 0} \frac{1}{C_R \sigma_x^2} R_x(e^{j\omega}) |\omega|^{\beta} = 1$$

This asymptotic definition implies that

$$R_x(e^{j\omega}) \simeq \frac{C_R \sigma_x^2}{|\omega|^{\beta}} \qquad \text{as } \omega \to 0$$

and

$$R_x(0) = \sum_{l=-\infty}^{\infty} r_x(l) = \infty$$

The first-order density determines the mean value and the variance of a process, whereas the second-order density determines the autocorrelation and power spectrum. There is a coupling between the probability density and the autocorrelation or power spectrum of a

process. However, this coupling is not extremely strong because there are processes that have different densities and the same autocorrelation. Thus, we can have random signal models with short or long memory and low or high variability. Random signal models are discussed in Chapters 4 and 12.

3.4.4 General Correlation Matrices

We first begin with the properties of general correlation matrices. Similar properties apply to covariance matrices.

PROPERTY 3.4.1. The correlation matrix of a random vector $\mathbf{x}$ is conjugate symmetric or Hermitian, that is,

$$\mathbf{R}_{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}^H \tag{3.4.28}$$

Proof. This follows easily from (3.2.19).

PROPERTY 3.4.2. The correlation matrix of a random vector $\mathbf{x}$ is nonnegative definite (n.n.d.); or for every nonzero complex vector $\mathbf{w} = [w_1 \; w_2 \; \cdots \; w_M]^T$, the quadratic form $\mathbf{w}^H\mathbf{R}_{\mathbf{x}}\mathbf{w}$ is nonnegative, that is,

$$\mathbf{w}^H\mathbf{R}_{\mathbf{x}}\mathbf{w} \geq 0 \tag{3.4.29}$$

Proof. To prove (3.4.29), we define the dot product

$$\alpha = \mathbf{w}^H\mathbf{x} = \mathbf{x}^T\mathbf{w}^* = \sum_{k=1}^{M} w_k^* x_k \tag{3.4.30}$$

The mean square value of the random variable α is

$$E\{|\alpha|^2\} = E\{\mathbf{w}^H\mathbf{x}\mathbf{x}^H\mathbf{w}\} = \mathbf{w}^H E\{\mathbf{x}\mathbf{x}^H\}\mathbf{w} = \mathbf{w}^H\mathbf{R}_{\mathbf{x}}\mathbf{w} \tag{3.4.31}$$

Since $E\{|\alpha|^2\} \geq 0$, if follows that $\mathbf{w}^H\mathbf{R}_{\mathbf{x}}\mathbf{w} \geq 0$. We also note that a matrix is called *positive definite* (p.d.) if $\mathbf{w}^H\mathbf{R}_{\mathbf{x}}\mathbf{w} > 0$.

Eigenvalues and eigenvectors of R

For a Hermitian matrix $\mathbf{R}$ we wish to find an $M \times 1$ vector $\mathbf{q}$ that satisfies the condition

$$\mathbf{R}\mathbf{q} = \lambda\mathbf{q} \tag{3.4.32}$$

where λ is a constant. This condition implies that the linear transformation performed by matrix $\mathbf{R}$ does not change the direction of vector $\mathbf{q}$. Thus $\mathbf{R}\mathbf{q}$ is a *direction-invariant* mapping. To determine the vector $\mathbf{q}$, we write (3.4.32) as

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{q} = \mathbf{0} \tag{3.4.33}$$

where $\mathbf{I}$ is the $M \times M$ identity matrix and $\mathbf{0}$ is an $M \times 1$ vector of zeros. Since $\mathbf{q}$ is arbitrary, the only way (3.4.33) is satisfied is if the determinant of $\mathbf{R} - \lambda\mathbf{I}$ equals zero, that is,

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0 \tag{3.4.34}$$

This equation is an Mth-order polynomial in λ and is called the *characteristic equation* of $\mathbf{R}$. It has M roots $\{\lambda_i\}_{i=1}^M$, called *eigenvalues*, which, in general, are distinct. If (3.4.34) has repeated roots, then $\mathbf{R}$ is said to have *degenerate* eigenvalues. For each eigenvalue λ_i we can satisfy (3.4.32)

$$\mathbf{R}\mathbf{q}_i = \lambda_i\mathbf{q}_i \qquad i = 1, \ldots, M \tag{3.4.35}$$

where the $\mathbf{q}_i$ are called *eigenvectors* of $\mathbf{R}$. Therefore, the $M \times M$ matrix $\mathbf{R}$ has M eigenvectors. To uniquely determine $\mathbf{q}_i$, we use (3.4.35) along with the normality condition that $\|\mathbf{q}_i\| = 1$. A MATLAB function `[Lambda,Q] = eig(R)` is available to compute eigenvalues and eigenvectors of $\mathbf{R}$.

There are further properties of the autocorrelation matrix $\mathbf{R}$ based on its eigenanalysis, which we describe below. Consider a matrix $\mathbf{R}$ that is Hermitian and nonnegative definite ($\mathbf{w}^H\mathbf{R}\mathbf{w} \geq 0$) with eigenvalues $\{\lambda_i\}_{i=1}^M$ and eigenvectors $\{\mathbf{q}_i\}_{i=1}^M$.

PROPERTY 3.4.3. The matrix $\mathbf{R}^k (k = 1, 2, \ldots)$ has eigenvalues $\lambda_1^k, \lambda_2^k, \ldots, \lambda_M^k$.

Proof. See Problem 3.16.

PROPERTY 3.4.4. If the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_M$ are distinct, the corresponding eigenvectors $\{\mathbf{q}_i\}_{i=1}^M$ are *linearly independent*.

Proof. This property can be proved by using Property 3.4.3. If there exists M not-all-zero scalars $\{\alpha_i\}_{i=1}^M$, such that

$$\sum_{i=1}^{M} \alpha_i \mathbf{q}_i = \mathbf{0} \tag{3.4.36}$$

then the eigenvectors $\{\mathbf{q}_i\}_{i=1}^M$ are said to be *linearly dependent*. Assume that (3.4.36) is true for some not-all-zero scalars $\{\alpha_i\}_{i=1}^M$ and that the eigenvalues $\{\lambda_i\}_{i=1}^M$ are distinct. Now multiply (3.4.36) repeatedly by $\mathbf{R}^k$, $k = 0, \ldots, M-1$ and use Property 3.4.3 to obtain

$$\sum_{i=1}^{M} \alpha_i \mathbf{R}^k \mathbf{q}_i = \sum_{i=1}^{M} \alpha_i \lambda_i^k \mathbf{q}_i = \mathbf{0} \qquad k = 0, \ldots, M-1 \tag{3.4.37}$$

which can be arranged in a matrix format for $i = 1, \ldots, M$ as

$$\begin{bmatrix} \alpha_1\mathbf{q}_1 & \alpha_2\mathbf{q}_2 & \alpha_3\mathbf{q}_3 & \ldots & \alpha_M\mathbf{q}_M \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \ldots & \lambda_1^{M-1} \\ 1 & \lambda_2 & \lambda_2^2 & \ldots & \lambda_2^{M-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_M & \lambda_M^2 & \ldots & \lambda_M^{M-1} \end{bmatrix} = \mathbf{0} \tag{3.4.38}$$

Since all the λ_i are distinct, the matrix containing the λ_i in (3.4.38) above is nonsingular. This matrix is called a *Vandermonde* matrix. Therefore, premultiplying both sides of (3.4.38) by the inverse of the Vandermonde matrix, we obtain

$$[\alpha_1\mathbf{q}_1 \; \alpha_2\mathbf{q}_2 \; \alpha_3\mathbf{q}_3 \; \ldots \; \alpha_M\mathbf{q}_M] = \mathbf{0} \tag{3.4.39}$$

Since eigenvectors $\{\mathbf{q}_i\}_{i=1}^M$ are not zero vectors, the only way (3.4.39) can be satisfied is if all $\{\alpha_i\}_{i=1}^M$ are zero. This implies that (3.4.36) cannot be satisfied for any set of not-all-zero scalars $\{\alpha_i\}_{i=1}^M$, which further implies that $\{\mathbf{q}_i\}_{i=1}^M$ are linearly independent.

PROPERTY 3.4.5. The eigenvalues $\{\lambda_i\}_{i=1}^M$ are real and *nonnegative*.

Proof. From (3.4.35), we have

$$\mathbf{q}_i^H \mathbf{R} \mathbf{q}_i = \lambda_i \mathbf{q}_i^H \mathbf{q}_i \qquad i = 1, 2, \ldots, M \tag{3.4.40}$$

Since R is positive semidefinite, the quadratic form $\mathbf{q}_i^H \mathbf{R} \mathbf{q}_i \geq 0$. Also since $\mathbf{q}_i^H \mathbf{q}_i$ is an inner product, $\mathbf{q}_i^H \mathbf{q}_i > 0$. Hence

$$\lambda_i = \frac{\mathbf{q}_i^H \mathbf{R} \mathbf{q}_i}{\mathbf{q}_i^H \mathbf{q}_i} \geq 0 \qquad i = 1, 2, \ldots, M \tag{3.4.41}$$

Furthermore, if $\mathbf{R}$ is positive definite, then $\lambda_i > 0$ for all $1 \leq i \leq M$. The quotient in (3.4.41) is a useful quantity and is known as the *Raleigh quotient* of vector $\mathbf{q}_i$.

PROPERTY 3.4.6. If the eigenvalues $\{\lambda_i\}_{i=1}^M$ are distinct, then the corresponding eigenvectors are orthogonal to one another, that is,

$$\lambda_i \neq \lambda_j \Rightarrow \mathbf{q}_i^H \mathbf{q}_j = 0 \qquad \text{for } i \neq j \tag{3.4.42}$$

Proof. Consider (3.4.35). We have

$$\mathbf{R}\mathbf{q}_i = \lambda_i \mathbf{q}_i \tag{3.4.43}$$

and

$$\mathbf{R}\mathbf{q}_j = \lambda_j \mathbf{q}_j \tag{3.4.44}$$

for some $i \neq j$. Premultiplying both sides of (3.4.43) by $\mathbf{q}_j^H$, we obtain

$$\mathbf{q}_j^H \mathbf{R}\mathbf{q}_i = \mathbf{q}_j^H \lambda_i \mathbf{q}_i = \lambda_i \mathbf{q}_j^H \mathbf{q}_i \tag{3.4.45}$$

Taking the conjugate transpose of (3.4.44), using the Hermitian property (3.4.28) of $\mathbf{R}$, and using the realness Property 3.4.5 of eigenvalues, we get

$$\mathbf{q}_j^H \mathbf{R} = \lambda_j \mathbf{q}_j^H \tag{3.4.46}$$

Now postmultiplying (3.4.46) by q_i and comparing with (3.4.45), we conclude that

$$\lambda_i \mathbf{q}_j^H \mathbf{q}_i = \lambda_j \mathbf{q}_j^H \mathbf{q}_i \qquad \text{or} \qquad (\lambda_i - \lambda_j)\mathbf{q}_j^H \mathbf{q}_i = 0 \tag{3.4.47}$$

Since the eigenvalues are assumed to be distinct, the only way (3.4.47) can be satisfied is if $\mathbf{q}_j^H \mathbf{q}_i = 0$ for $i \neq j$, which further proves that the corresponding eigenvectors are orthogonal to one another.

PROPERTY 3.4.7. Let $\{\mathbf{q}_i\}_{i=1}^M$ be an orthonormal set of eigenvectors corresponding to the distinct eigenvalues $\{\lambda_i\}_{i=1}^M$ of an $M \times M$ correlation matrix $\mathbf{R}$. Then $\mathbf{R}$ can be diagonalized as follows:

$$\mathbf{\Lambda} = \mathbf{Q}^H \mathbf{R} \mathbf{Q} \tag{3.4.48}$$

where the orthonormal matrix $\mathbf{Q} \triangleq [\mathbf{q}_1 \ \cdots \ \mathbf{q}_M]$ is known as an *eigenmatrix* and $\mathbf{\Lambda}$ is an $M \times M$ diagonal eigenvalue matrix, that is,

$$\mathbf{\Lambda} \triangleq \text{diag}(\lambda_1, \ldots, \lambda_M) \tag{3.4.49}$$

Proof. Arranging the vectors in (3.4.35) in a matrix format, we obtain

$$[\mathbf{R}\mathbf{q}_1 \ \mathbf{R}\mathbf{q}_2 \ \cdots \ \mathbf{R}\mathbf{q}_M] = [\lambda_1 \mathbf{q}_1 \ \lambda_2 \mathbf{q}_2 \ \cdots \ \lambda_M \mathbf{q}_M]$$

which, by using the definitions of $\mathbf{Q}$ and $\mathbf{\Lambda}$, can be further expressed as

$$\mathbf{R}\mathbf{Q} = \mathbf{Q}\mathbf{\Lambda} \tag{3.4.50}$$

Since $\mathbf{q}_i$, $i = 1, \ldots, M$, is an orthonormal set of vectors, the eigenmatrix $\mathbf{Q}$ is unitary, that is, $\mathbf{Q}^{-1} = \mathbf{Q}^H$. Now premultiplying both sides of (3.4.50) by $\mathbf{Q}^H$, we obtain the desired result.

This diagonalization of the autocorrelation matrix plays an important role in filtering and estimation theory, as we shall see later. From (3.4.48) the correlation matrix $\mathbf{R}$ can also be written as

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^H + \cdots + \lambda_M \mathbf{q}_M \mathbf{q}_M^H = \sum_{m=1}^{M} \lambda_m \mathbf{q}_m \mathbf{q}_m^H \tag{3.4.51}$$

which is known as the *spectral theorem*, or *Mercer's theorem*. If $\mathbf{R}$ is positive definite (and hence invertible), its inverse is given by

$$\mathbf{R}^{-1} = (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H)^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^H = \sum_{m=1}^{M} \frac{1}{\lambda_m} \mathbf{q}_m \mathbf{q}_m^H \tag{3.4.52}$$

because $\mathbf{\Lambda}$ is a diagonal matrix.

PROPERTY 3.4.8. The trace of $\mathbf{R}$ is the summation of all eigenvalues, that is,

$$\text{tr}(\mathbf{R}) = \sum_{i=1}^{M} \lambda_i \tag{3.4.53}$$

Proof. See Problem 3.17.

PROPERTY 3.4.9. The determinant of $\mathbf{R}$ is equal to the product of all eigenvalues, that is,

$$\det \mathbf{R} = |\mathbf{R}| = \prod_{i=1}^{M} \lambda_i = |\mathbf{\Lambda}| \tag{3.4.54}$$

Proof. See Problem 3.18.

PROPERTY 3.4.10. Determinants of $\mathbf{R}$ and $\mathbf{\Gamma}$ are related by

$$|\mathbf{R}| = |\mathbf{\Gamma}|(1 + \boldsymbol{\mu}_x^H \mathbf{\Gamma} \boldsymbol{\mu}_x) \tag{3.4.55}$$

Proof. See Problem 3.19.

3.4.5 Correlation Matrices from Random Processes

A stochastic process can also be represented as a random vector, and its second-order statistics given by the mean vector and the correlation matrix. Obviously, these quantities are functions of the index n. Let an $M \times 1$ random vector $\mathbf{x}(n)$ be derived from the random process $x(n)$ as follows:

$$\mathbf{x}(n) \triangleq [x(n)\ x(n-1)\ \cdots\ x(n-M+1)]^T \tag{3.4.56}$$

Then its mean is given by an $M \times 1$ vector

$$\boldsymbol{\mu}_x(n) = [\mu_x(n)\ \mu_x(n-1)\ \cdots\ \mu_x(n-M+1)]^T \tag{3.4.57}$$

and the correlation by an $M \times M$ matrix

$$\mathbf{R}_x(n) = \begin{bmatrix} r_x(n,n) & \cdots & r_x(n,n-M+1) \\ \vdots & \ddots & \vdots \\ r_x(n-M+1,n) & \cdots & r_x(n-M+1,n-M+1) \end{bmatrix} \tag{3.4.58}$$

Clearly, $\mathbf{R}_x(n)$ is Hermitian since $r_x(n-i,n-j) = r_x^*(n-j,n-i)$, $0 \le i, j \le M-1$. This vector representation will be useful when we discuss optimum filters.

Correlation matrices of stationary processes

The correlation matrix $\mathbf{R}_x(n)$ of a general stochastic process $x(n)$ is a Hermitian $M \times M$ matrix defined in (3.4.58) with elements $r_x(n-i,n-j) = E\{x(n-i)x^*(n-j)\}$. For stationary processes this matrix has an interesting additional structure. First, $\mathbf{R}_x(n)$ is a constant matrix $\mathbf{R}_x$; then using (3.3.24), we have

$$r_x(n-i,n-j) = r_x(j-i) = r_x(l \triangleq j-i) \tag{3.4.59}$$

Finally, by using conjugate symmetry $r_x(l) = r_x^*(-l)$, the matrix $\mathbf{R}_x$ is given by

$$\mathbf{R}_x = \begin{bmatrix} r_x(0) & r_x(1) & r_x(2) & \cdots & r_x(M-1) \\ r_x^*(1) & r_x(0) & r_x(1) & \cdots & r_x(M-2) \\ r_x^*(2) & r_x^*(1) & r_x(0) & \cdots & r_x(M-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_x^*(M-1) & r_x^*(M-2) & r_x^*(M-3) & \cdots & r_x(0) \end{bmatrix} \tag{3.4.60}$$

It can be easily seen that $\mathbf{R}_x$ is Hermitian and Toeplitz.[†] Thus, the autocorrelation matrix of a stationary process is Hermitian, nonnegative definite, and Toeplitz. Note that $\mathbf{R}_x$ is not persymmetric because elements along the main antidiagonal are not equal, in general.

[†]A matrix is called *Toeplitz* if the elements along each diagonal, parallel to the main diagonal, are equal.

Eigenvalue spread and spectral dynamic range

The ill conditioning of a matrix $\mathbf{R}_x$ increases with its condition number $\mathcal{X}(\mathbf{R}_x) = \lambda_{\max}/\lambda_{\min}$. When $\mathbf{R}_x$ is a correlation matrix of a stationary process, then $\mathcal{X}(R_x)$ is bounded from above by the dynamic range of the PSD $R_x(e^{j\omega})$ of the process $x(n)$. The larger the spread in eigenvalues, the wider (or less flat) the variation of the PSD function. This is also related to the dynamic range or to the data spread in $x(n)$ and is a useful measure in practice. This result is given by the following theorem, in which we have dropped the subscript of $R_x(e^{j\omega})$ for clarity.

THEOREM 3.5. Consider a zero-mean stationary random process with autoPSD

$$R(e^{j\omega}) = \sum_{l=-\infty}^{\infty} r(l)e^{-j\omega l}$$

then $$\min_{\omega} R(e^{j\omega}) \le \lambda_i \le \max_{\omega} R(e^{j\omega}) \qquad \text{for all } i = 1, 2, \ldots, M \tag{3.4.61}$$

Proof. From (3.4.41) we have

$$\lambda_i = \frac{\mathbf{q}_i^H \mathbf{R}\mathbf{q}_i}{\mathbf{q}_i^H \mathbf{q}_i} \tag{3.4.62}$$

Consider the quadratic form

$$\mathbf{q}_i^H \mathbf{R}\mathbf{q}_i = \sum_{k=1}^{M}\sum_{l=1}^{M} q_i(k) r(l-k) q_i(l)$$

where $\mathbf{q}_i = [q_i(1)\ q_i(2)\ \cdots\ q_i(M)]^T$. Using (3.3.41) and the stationarity of the process, we obtain

$$\begin{aligned} \mathbf{q}_i^H \mathbf{R}\mathbf{q}_i &= \frac{1}{2\pi}\sum_k \sum_l q_i^*(k) q_i(l) \int_{-\pi}^{\pi} R(e^{j\omega}) e^{j\omega(l-k)}\, \mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} R(e^{j\omega}) \left[\sum_{k=1}^{M} q_i^*(k) e^{-j\omega k}\right]\left[\sum_{l=1}^{M} q_i(l) e^{j\omega l}\right] \mathrm{d}\omega \end{aligned} \tag{3.4.63}$$

or $$\mathbf{q}_i^H \mathbf{R}\mathbf{q}_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} R(e^{j\omega}) |Q(e^{j\omega})|^2 \mathrm{d}\omega \tag{3.4.64}$$

Similarly, we have

$$\mathbf{q}_i^H \mathbf{q}_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} |Q(e^{j\omega})|^2\, \mathrm{d}\omega \tag{3.4.65}$$

Substituting (3.4.64) and (3.4.65) in (3.4.62), we obtain

$$\lambda_i = \frac{\displaystyle\int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 R(e^{j\omega})\, \mathrm{d}\omega}{\displaystyle\int_{-\pi}^{\pi} |Q(e^{j\omega})|^2\, \mathrm{d}\omega} \tag{3.4.66}$$

However, since $R(e^{j\omega}) \ge 0$, we have the following inequality:

$$\begin{aligned} \min_{\omega} R(e^{j\omega}) \int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 \mathrm{d}\omega &\le \int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 R(e^{j\omega})\, \mathrm{d}\omega \\ &\le \max_{\omega} R(e^{j\omega}) \int_{-\pi}^{\pi} |Q(e^{j\omega})|^2 \mathrm{d}\omega \end{aligned}$$

from which we easily obtain the desired result. The above result also implies that

$$\mathcal{X}(\mathbf{R}) \triangleq \frac{\lambda_{\max}}{\lambda_{\min}} \le \frac{\max_{\omega} R(e^{j\omega})}{\min_{\omega} R(e^{j\omega})} \tag{3.4.67}$$

which becomes equality as $M \to \infty$.

3.5 WHITENING AND INNOVATIONS REPRESENTATION

In many practical and theoretical applications, it is desirable to represent a random vector (or sequence) with a linearly equivalent vector (or sequence) consisting of uncorrelated components. If $\mathbf{x}$ is a correlated random vector and if $\mathbf{A}$ is a nonsingular matrix, then the linear transformation

$$\mathbf{w} = \mathbf{A}\mathbf{x} \tag{3.5.1}$$

results in a random vector $\mathbf{w}$ that contains the same "information" as $\mathbf{x}$, and hence random vectors $\mathbf{x}$ and $\mathbf{w}$ are said to be linearly equivalent. Furthermore, if $\mathbf{w}$ has uncorrelated components and A is lower-triangular, then each component w_i of $\mathbf{w}$ can be thought of as *adding* "new" information (or *innovation*) to $\mathbf{w}$ that is not present in the remaining components. Such a representation is called an *innovations representation* and provides additional insight into the understanding of random vectors and sequences. Additionally, it can simplify many theoretical derivations and can result in computationally efficient implementations.

Since $\mathbf{\Gamma_w}$ must be a diagonal matrix, we need to diagonalize the Hermitian, positive definite matrix $\mathbf{\Gamma_x}$ through the transformation matrix $\mathbf{A}$. There are two approaches to this diagonalization. One approach is to use the eigenanalysis presented in Section 3.4.4, which results in the well-known Karhunen-Loève (KL) transform. The other approach is to use triangularization methods from linear algebra, which leads to the LDU (UDL) and LU (UL) decompositions. These vector techniques can be further extended to random sequences that give us the KL expansion and the spectral factorizations, respectively.

3.5.1 Transformations Using Eigendecomposition

Let $\mathbf{x}$ be a random vector with mean vector $\boldsymbol{\mu}_\mathbf{x}$ and covariance matrix $\mathbf{\Gamma_x}$. The linear transformation

$$\mathbf{x}_0 = \mathbf{x} - \boldsymbol{\mu}_\mathbf{x} \tag{3.5.2}$$

results in a zero-mean vector $\mathbf{x}_0$ with correlation (and covariance) matrix equal to $\mathbf{\Gamma_x}$. This transformation shifts the origin of the M-dimensional coordinate system to the mean vector. We will now consider the zero-mean random vector $\mathbf{x}_0$ for further transformations.

Orthonormal transformation

Let $\mathbf{Q_x}$ be the eigenmatrix of $\mathbf{\Gamma_x}$, and let us choose $\mathbf{Q}_\mathbf{x}^H$ as our linear transformation matrix $\mathbf{A}$ in (3.2.32). Consider

$$\mathbf{w} = \mathbf{Q}_\mathbf{x}^H \mathbf{x}_0 = \mathbf{Q}_\mathbf{x}^H (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) \tag{3.5.3}$$

Then

$$\boldsymbol{\mu}_\mathbf{w} = \mathbf{Q}_\mathbf{x}^H (E\{\mathbf{x}_0\}) = \mathbf{0} \tag{3.5.4}$$

and from (3.2.39) and (3.4.48)

$$\mathbf{\Gamma_w} = \mathbf{R_w} = E\{\mathbf{Q}_\mathbf{x}^H \mathbf{x}_0 \mathbf{x}_0^H \mathbf{Q_x}\} = \mathbf{Q}_\mathbf{x}^H \mathbf{\Gamma_x} \mathbf{Q_x} = \mathbf{\Lambda_x} \tag{3.5.5}$$

Since $\mathbf{\Lambda_x}$ is diagonal, $\mathbf{\Gamma_w}$ is also diagonal, and hence this transformation has some interesting properties:

1. The random vector $\mathbf{w}$ has zero mean, and its components are mutually uncorrelated (and hence orthogonal). Furthermore, if $\mathbf{x}$ is $\mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Gamma_x})$, then $\mathbf{w}$ is $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda_x})$ with independent components.
2. The variances of random variables $w_i, i = 1, \ldots, M$, are equal to the eigenvalues of $\mathbf{\Gamma_x}$.
3. Since the transformation matrix $\mathbf{A} = \mathbf{Q}_\mathbf{x}^H$ is orthonormal, the transformation is called an *orthonormal transformation* and the distance measure

$$d^2(\mathbf{x}_0) \triangleq \mathbf{x}_0^H \mathbf{\Gamma}_\mathbf{x}^{-1} \mathbf{x}_0 \tag{3.5.6}$$

is preserved under the transformation. This distance measure is also known as the *Mahalanobis distance*; and in the case of normal random vectors, it is related to the log-likelihood function.

4. Since $\mathbf{w} = \mathbf{Q}_{\mathbf{x}}^H(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$, we have

$$w_i = \mathbf{q}_i^H(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) = \|\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\| \cos[\measuredangle(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}, \mathbf{q}_i)] \qquad i = 1, \ldots, M \tag{3.5.7}$$

which is the projection of $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}$ onto the unit vector $\mathbf{q}_i$. Thus $\mathbf{w}$ represents $\mathbf{x}$ in a new coordinate system that is shifted to $\boldsymbol{\mu}_{\mathbf{x}}$ and spanned by $\mathbf{q}_i, i = 1, \ldots, M$. A geometric interpretation of this transformation for a two-dimensional case is shown in Figure 3.11, which shows a contour of $d^2(\mathbf{x}_0) = \mathbf{x}^H\boldsymbol{\Gamma}_{\mathbf{x}}^{-1}\mathbf{x} = \mathbf{w}^H\boldsymbol{\Lambda}_{\mathbf{x}}^{-1}\mathbf{w}$ in the $\mathbf{x}$ and $\mathbf{w}$ coordinate systems ($\mathbf{w} = Q_{\mathbf{x}}^H\mathbf{x}$).

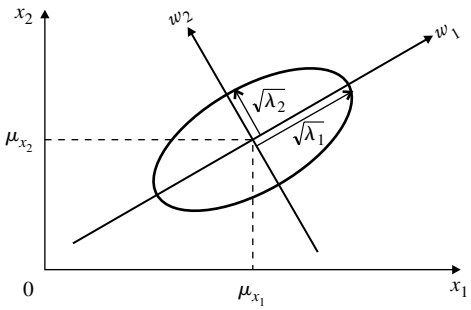


FIGURE 3.11
Orthogonal transformation in two dimensions.

Isotropic transformation

In the above orthonormal transformation, the autocorrelation matrix $\mathbf{R}_{\mathbf{w}}$ is diagonal but not an identity matrix $\mathbf{I}$. This can be achieved by an additional linear mapping of $\boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}$. Let

$$\mathbf{y} = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{w} = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{Q}_{\mathbf{x}}^H\mathbf{x}_0 = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{Q}_{\mathbf{x}}^H(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \tag{3.5.8}$$

Then

$$\mathbf{R}_{\mathbf{y}} = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{Q}_{\mathbf{x}}^H\boldsymbol{\Gamma}_{\mathbf{x}}\mathbf{Q}_{\mathbf{x}}\boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2} = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2} = \mathbf{I} \tag{3.5.9}$$

This is called an *isotropic transformation* because *all* components of $\mathbf{y}$ are zero-mean, uncorrelated random variables with unit variance.[†] The geometric interpretation of this transformation for a two-dimensional case is shown in Figure 3.12. It clearly shows that there is not only a shift and rotation but also a scaling of the coordinate axis so that the distribution is equal in all directions, that is, it is direction-invariant. Because the transformation $\mathbf{A} = \boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{Q}_{\mathbf{x}}^H$ is orthogonal but not orthonormal, the distance measure $d^2(\mathbf{x}_0)$ is not preserved under this mapping. Since the correlation matrix after this transformation is an identity matrix $\mathbf{I}$, it is invariant under any orthonormal mapping, that is,

$$\mathbf{Q}^H\mathbf{I}\mathbf{Q} = \mathbf{Q}^H\mathbf{Q} = \mathbf{I} \tag{3.5.10}$$

This fact can be used for simultaneous diagonalization of two Hermitian matrices.

EXAMPLE 3.5.1. Consider a stationary sequence with correlation matrix

$$\mathbf{R}_x = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$$

where $-1 < a < 1$. The eigenvalues

$$\lambda_1 = 1 + a \qquad \lambda_2 = 1 - a$$

[†] In the literature, an isotropic transformation is also known as a *whitening* transformation. We believe that this terminology is not accurate because both vectors $\mathbf{Q}_{\mathbf{x}}^H\mathbf{x}_0$ and $\boldsymbol{\Lambda}_{\mathbf{x}}^{-1/2}\mathbf{Q}_{\mathbf{x}}^H\mathbf{x}_0$ have uncorrelated coefficients.

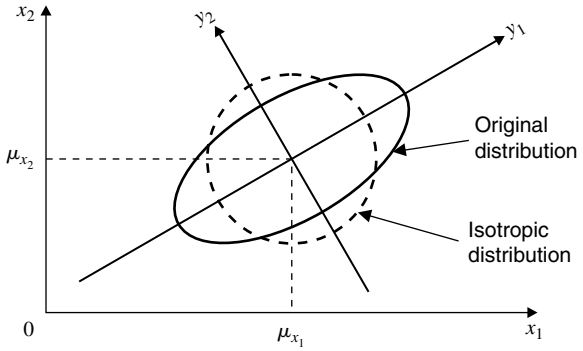


FIGURE 3.12
Isotropic transformation in two dimensions.

are obtained from the characteristic equation

$$\det(\mathbf{R}_x - \lambda\mathbf{I}) = \det\begin{bmatrix} 1-\lambda & a \\ a & 1-\lambda \end{bmatrix} = (1-\lambda)^2 - a^2 = 0$$

To find the eigenvector $\mathbf{q}_1$, we solve the linear system

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}\begin{bmatrix} q_1^{(1)} \\ q_2^{(1)} \end{bmatrix} = (1+a)\begin{bmatrix} q_1^{(1)} \\ q_2^{(1)} \end{bmatrix}$$

which gives $q_1^{(1)} = q_2^{(1)}$. Similarly, we find that $q_1^{(2)} = -q_2^{(2)}$. If we normalize both vectors to unit length, we obtain the eigenvectors

$$\mathbf{q}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathbf{q}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

From the above results we see that $\det \mathbf{R}_x = 1 - a^2 = \lambda_1\lambda_2$ and $\mathbf{Q}^H\mathbf{Q} = \mathbf{I}$, where $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2]$.

3.5.2 Transformations Using Triangular Decomposition

The linear transformations discussed above were based on diagonalization of hermitian matrices through eigenvalue-eigenvector decomposition. These are useful in many detection and estimation problems. Triangular matrix decomposition leads to transformations that result in *causal* or *anticausal* linear filtering of associated sequences. Hence these mappings play an important role in linear filtering. There are two such decompositions: the *lower-diagonal-upper* (*LDU*) one leads to causal filtering while the *upper-diagonal-lower* (*UDL*) one results in anticausal filtering.

Lower-diagonal-upper decomposition

Any Hermitian, positive definite matrix $\mathbf{R}$ can be factored as (Goulob and Van Loan 1989)

$$\mathbf{R} = \mathbf{L}\mathbf{D}_L\mathbf{L}^H \tag{3.5.11}$$

or equivalently

$$\mathbf{L}^{-1}\mathbf{R}\mathbf{L}^{-H} = \mathbf{D}_L \tag{3.5.12}$$

where $\mathbf{L}$ is a *unit lower triangular* matrix, $\mathbf{D}_L$ is a diagonal matrix with positive elements, and $\mathbf{L}^H$ is a *unit upper triangular* matrix. The MATLAB function `[L,D]=ldlt(R)`, given in Section 5.2, computes the LDU decomposition.

Since $\mathbf{L}$ is unit lower triangular, we have $\det \mathbf{R} = \prod_{i=1}^{M} \xi_i^l$, where $\xi_1^l, \ldots, \xi_M^l$ are the diagonal elements of $\mathbf{D}_L$. If we define the linear transformation

$$\mathbf{w} = \mathbf{L}^{-1}\mathbf{x} \triangleq \mathbf{B}\mathbf{x} \tag{3.5.13}$$

we find that

$$\mathbf{R}_{\mathbf{w}} = E\{\mathbf{w}\mathbf{w}^H\} = \mathbf{L}^{-1}E\{\mathbf{x}\mathbf{x}^H\}\mathbf{L}^{-H} = \mathbf{L}^{-1}\mathbf{R}\mathbf{L}^{-H} = \mathbf{D}_L \tag{3.5.14}$$

Clearly, the components of $\mathbf{w}$ are orthogonal, and the elements $\xi_1^l, \ldots, \xi_M^l$ are their second moments. Therefore, this transformation appears to be similar to the orthogonal one. However, the vector $\mathbf{w}$ is not obtained as a simple rotation of $\mathbf{x}$. To understand this mapping, we first note that $\mathbf{B} = \mathbf{L}^{-1}$ is also a unit lower triangular matrix (Goulob and Van Loan 1989). Then we can write (3.5.13) as

$$\begin{bmatrix} w_1 \\ \vdots \\ w_i \\ \vdots \\ w_M \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ b_{i1} & \cdots & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ b_{M1} & \cdots & b_{Mi} & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_M \end{bmatrix} \tag{3.5.15}$$

where b_{ik} are elements of $\mathbf{B}$. From (3.5.15) we conclude that w_i is a linear combination of $x_k, k \leq i$, that is,

$$w_i = \sum_{k=1}^{i} b_{ik}x_k \qquad 1 \leq i \leq M \tag{3.5.16}$$

If the signal vector $\mathbf{x}$ consists of consecutive samples of a discrete-time stochastic process $x(n)$, that is,

$$\mathbf{x} = [x(n)\ x(n-1)\ \cdots\ x(n-M+1)]^T \tag{3.5.17}$$

then (3.5.16) can be interpreted as a causal linear filtering of the random sequence (see Chapter 2). This transformation will be used extensively in optimum linear filtering and prediction problems.

A similar LDU decomposition of autocovariance matrices can be performed by following the identical steps above. In this case, the components of the transformed vector $\mathbf{w}$ are uncorrelated, and the elements ξ_i^l, $1 \leq i \leq M$, of $\mathbf{D}_L$ are variances.

Upper-diagonal-lower decomposition

This diagonalization is almost identical to the previous one and involves factorization of a Hermitian, positive definite matrix into an upper-diagonal-lower form. It is given by

$$\mathbf{R} = \mathbf{U}\mathbf{D}_U\mathbf{U}^H \tag{3.5.18}$$

or equivalently

$$\mathbf{U}^{-1}\mathbf{R}\mathbf{U}^{-H} = \mathbf{D}_U = \operatorname{diag}(\xi_1^u, \ldots, \xi_M^u) \tag{3.5.19}$$

in which the matrix $\mathbf{U}$ is unit upper triangular, the matrix $\mathbf{U}^H$ is unit lower triangular, and the matrix $\mathbf{D}_U$ is diagonal with positive elements. Note that $\mathbf{U}^H \neq \mathbf{L}$ and $\mathbf{D}_U \neq \mathbf{D}_L$. Following the same analysis as above, we have $\det \mathbf{R} = \det \mathbf{D}_U = \prod_{i=1}^{M} \xi_i^u$. Since $\mathbf{A} = \mathbf{U}^{-1}$ is unit upper triangular in the transformation $\mathbf{w} = \mathbf{U}^{-1}\mathbf{x}$, the components of $\mathbf{w}$ are orthogonal and are obtained by linear combinations of $x_k, k \geq i$, that is,

$$w_i = \sum_{k=i}^{M} l_{ik}x_k \qquad 1 \leq i \leq M \tag{3.5.20}$$

This represents an anticausal filtering of a random sequence if $\mathbf{x}$ is a signal vector. Table 3.3 compares and contrasts orthogonal and triangular decompositions. We note that the LDU decomposition does not have the nice geometric interpretation (rotation of the coordinate system) of the eigendecomposition transformation.

Generation of real-valued random vectors with given second-order moments. Suppose that we want to generate M samples, say, $x_1, x_2, \ldots, x_M$, of a real-valued random vector $\mathbf{x}$ with mean $\mathbf{0}$ and a given symmetric and positive definite autocorrelation matrix $\mathbf{R}_{\mathbf{x}}$.

TABLE 3.3
Comparison of orthogonal and triangular decompositions for zero-mean random vectors.

Orthogonal decomposition	Triangular decomposition
$\mathbf{R} = E\{\mathbf{x}\mathbf{x}^H\}$	$\mathbf{R} = E\{\mathbf{x}\mathbf{x}^H\}$
$\mathbf{R}\mathbf{q}_i = \lambda_i \mathbf{q}_i$	
$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M]$	$\mathbf{L}$ = unit lower triangular
$\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$	$\mathbf{D} = \text{diag}\{\xi_1, \xi_2, \ldots, \xi_M\}$
$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H = \sum_{i=1}^{M} \lambda_i \mathbf{q}_i \mathbf{q}_i^H$	$\mathbf{R} = \mathbf{L}\mathbf{D}\mathbf{L}^H$
$\mathbf{\Lambda} = \mathbf{Q}^H \mathbf{R} \mathbf{Q}$	$\mathbf{D} = \mathbf{L}^{-1}\mathbf{R}\mathbf{L}^{-H}$
$\mathbf{R}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^H = \sum_{i=1}^{M} \frac{1}{\lambda_i} \mathbf{q}_i \mathbf{q}_i^H$	$\mathbf{R}^{-1} = \mathbf{L}^{-H}\mathbf{D}^{-1}\mathbf{L}^{-1}$
$\mathbf{\Lambda}^{-1} = \mathbf{Q}^H \mathbf{R}^{-1} \mathbf{Q}$	$\mathbf{D}^{-1} = \mathbf{L}^{-H}\mathbf{R}^{-1}\mathbf{L}^{-1}$
$\det \mathbf{R} = \det \mathbf{\Lambda} = \prod_{i=1}^{M} \lambda_i$	$\det \mathbf{R} = \det \mathbf{D} = \prod_{i=1}^{M} \xi_i$
$\text{tr}\, \mathbf{R} = \text{tr}\, \mathbf{\Lambda} = \sum_{i=1}^{M} \lambda_i$	

Whitening (noncausal)	Whitening (causal)
$\mathbf{w} = \mathbf{Q}^H \mathbf{x}$	$\mathbf{w} = \mathbf{L}^{-1}\mathbf{x}$
$E\{\mathbf{w}\mathbf{w}^H\} = \mathbf{\Lambda}$	$E\{\mathbf{w}\mathbf{w}^H\} = \mathbf{D}$

The innovations representation given in this section suggests three approaches to generate samples of such a random vector. The general approach is to factor $\mathbf{R}_\mathbf{x}$, using either the orthonormal or the triangularization transformation, to obtain the diagonal matrix ($\mathbf{\Lambda}_\mathbf{x}$ or $\mathbf{D}_L^{(\mathbf{x})}$ or $\mathbf{D}_U^{(\mathbf{x})}$), generate M samples of an IID sequence with the obtained diagonal variances, and then transform these samples by using the inverse transformation matrix ($\mathbf{Q}_\mathbf{x}$ or $\mathbf{L}_\mathbf{x}$ or $\mathbf{U}_\mathbf{x}$). We hasten to add that, in general, the original distribution of the IID samples will not be preserved unless the samples are jointly normal. Therefore, in the following discussion, we assume that a normal pseudorandom number generator is used to generate M independent samples of $\mathbf{w}$. The three methods are as follows.

Eigendecomposition approach. First factor $\mathbf{R}_\mathbf{x}$ as $\mathbf{R}_\mathbf{x} = \mathbf{Q}_\mathbf{x}\mathbf{\Lambda}_\mathbf{x}\mathbf{Q}_\mathbf{x}^H$. Then generate $\mathbf{w}$, using the distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_\mathbf{x})$. Finally, compute the desired vector $\mathbf{x}$, using $\mathbf{x} = \mathbf{Q}_\mathbf{x}\mathbf{w}$.

LDU triangularization approach. First factor $\mathbf{R}_\mathbf{x}$ as $\mathbf{R}_\mathbf{x} = \mathbf{L}_\mathbf{x}\mathbf{D}_\mathbf{L}^{(\mathbf{x})}\mathbf{L}_\mathbf{x}^H$. Then generate $\mathbf{w}$, using the distribution $\mathcal{N}(\mathbf{0}, \mathbf{D}_\mathbf{L}^{(\mathbf{x})})$. Finally, compute the desired vector $\mathbf{x}$, using $\mathbf{x} = \mathbf{L}_\mathbf{x}\mathbf{w}$.[†]

UDL triangularization approach. First factor $\mathbf{R}_\mathbf{x}$ as $\mathbf{R}_\mathbf{x} = \mathbf{U}_\mathbf{x}\mathbf{D}_\mathbf{U}^{(\mathbf{x})}\mathbf{U}_\mathbf{x}^H$. Then generate $\mathbf{w}$, using the distribution $\mathcal{N}(\mathbf{0}, \mathbf{D}_\mathbf{U}^{(\mathbf{x})})$. Finally, compute the desired vector $\mathbf{x}$, using $\mathbf{x} = \mathbf{U}_\mathbf{x}\mathbf{w}$.

Additional discussion and more complete treatment on the generation of random vectors are given in Johnson (1994).

3.5.3 The Discrete Karhunen-Loève Transform

In many signal processing applications, it is convenient to represent the samples of a random signal in another set of numbers (or coefficients) so that this new representation possesses some useful properties. For example, for coding purposes we want to transform a signal

[†] If we use the Cholesky decomposition $\mathbf{R}_\mathbf{x} = \tilde{\mathbf{L}}_\mathbf{x}\tilde{\mathbf{L}}_\mathbf{x}^H$, where $\tilde{\mathbf{L}}_\mathbf{x} = \{\mathbf{D}_\mathbf{L}^{(\mathbf{x})}\}^{1/2}\mathbf{L}_\mathbf{x}$, then $\mathbf{w} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ will generate $\mathbf{x}$ with the given correlation $\mathbf{R}_x$, using $\mathbf{x} = \tilde{\mathbf{L}}_\mathbf{x}\mathbf{w}$.

so that its energy is concentrated in only a few coefficients (which are then transmitted); or for optimal filtering purposes we may want uncorrelated samples so that the filtering complexity is reduced or the signal-to-noise ratio is enhanced. A general approach is to expand a signal as a linear combination of orthogonal basis functions so that components of the signal with respect to basis functions do not interfere with one another. There are several such basis functions; the most widely known is the set of complex exponentials used in DTFT (or DFT) that are used in linear filtering, as we discussed in Section 3.4. Other examples are functions used in discrete cosine transform, discrete sine transform, Haar transform, etc., which are useful in coding applications (Jain 1989).

As discussed in this section, a set of orthogonal basis functions for which the signal components are statistically uncorrelated to one another is based on the second-order properties of the random process and, in particular, on the diagonalization of its covariance matrix. It is also an optimal representation of the signal in the sense that it provides a representation with the *smallest mean square error* among all other orthogonal transforms. This has applications in the analysis of random signals as well as in coding. This transform was first suggested by Karhunen and Loève for continuous random processes. It was extended to discrete random signals by Hotelling and is also known as the Hotelling transform. In keeping with the current nomenclature, we will call it the *discrete Karhunen-Loève transform* (*DKLT*) (Fukunaga 1990).

Development of the DKLT

Let $\mathbf{x} = [x_1\ x_2\ \cdots\ x_M]^T$ be a zero-mean[†] random vector with autocorrelation matrix $\mathbf{R_x}$. We want to represent $\mathbf{x}$ using the linear transformation

$$\mathbf{w} = \mathbf{A}^H\mathbf{x} \qquad \mathbf{A}^{-1} = \mathbf{A}^H \tag{3.5.21}$$

where $\mathbf{A}$ is a unitary matrix. Then

$$\mathbf{x} = \mathbf{A}\mathbf{w} = \sum_{i=1}^{M} w_i\mathbf{a}_i \qquad \mathbf{a}_i^H\mathbf{a}_j = 0 \qquad i \neq j \tag{3.5.22}$$

Let us represent $\mathbf{x}$ using the first m, $1 \leq m \leq M$, components of $\mathbf{w}$, that is,

$$\hat{\mathbf{x}} \triangleq \sum_{i=1}^{m} w_i\mathbf{a}_i \qquad 1 \leq m \leq M \tag{3.5.23}$$

Then from (3.5.22) and (3.5.23), the error between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is given by

$$\mathbf{e}_m \triangleq \mathbf{x} - \hat{\mathbf{x}} = \sum_{i=1}^{M} w_i\mathbf{a}_i - \sum_{i=1}^{m} w_i\mathbf{a}_i = \sum_{i=m+1}^{M} w_i\mathbf{a}_i \tag{3.5.24}$$

and hence the mean-squared error (MSE) is

$$E_m \triangleq E\{\mathbf{e}_m^H\mathbf{e}_m\} = \sum_{i=m+1}^{M} \mathbf{a}_i^H E\{|w_i|^2\}\mathbf{a}_i = \sum_{i=m+1}^{M} E\{|w_i|^2\}\mathbf{a}_i^H\mathbf{a}_i \tag{3.5.25}$$

Since from (3.5.21) $w_i = \mathbf{a}_i^H\mathbf{x}$, we have $E\{|w_i|^2\} = \mathbf{a}_i^H\mathbf{R_x}\mathbf{a}_i$. Now we want to determine the matrix $\mathbf{A}$ that will minimize the MSE E_m subject to $\mathbf{a}_i^H\mathbf{a}_i = 1, i = m+1, \ldots, M$ so that from (3.5.25)

$$E_m = \sum_{i=m+1}^{M} E\{|w_i|^2\} = \sum_{i=m+1}^{M} \mathbf{a}_i^H\mathbf{R_x}\mathbf{a}_i \qquad \mathbf{a}_i^H\mathbf{a}_i = 1 \qquad i = m+1, \ldots, M \tag{3.5.26}$$

[†]If the mean is not zero, then we perform the transformation on the mean-subtracted vector, using the covariance matrix.

This optimization can be done by using the Lagrange multiplier approach (Appendix B); that is, we minimize

$$\sum_{i=m+1}^{M} \mathbf{a}_i^H \mathbf{R}_\mathbf{x} \mathbf{a}_i + \sum_{i=m+1}^{M} \lambda_i (1 - \mathbf{a}_i^H \mathbf{a}_i) \qquad i = m+1, \ldots, M$$

Hence after setting the gradient equal to zero,

$$\nabla_{\mathbf{a}_i} \left[\sum_{i=m+1}^{M} \mathbf{a}_i^H \mathbf{R}_\mathbf{x} \mathbf{a}_i + \sum_{i=m+1}^{M} \lambda_i (1 - \mathbf{a}_i^H \mathbf{a}_i) \right] = (\mathbf{R}_\mathbf{x} \mathbf{a}_i)^* - (\lambda_i \mathbf{a}_i)^* = 0 \qquad (3.5.27)$$

we obtain

$$\mathbf{R}_\mathbf{x} \mathbf{a}_i = \lambda_i \mathbf{a}_i \qquad i = m+1, \ldots, M$$

which is equivalent to (3.4.35) in the eigenanalysis of Section 3.4.4. Hence λ_i is the eigenvalue, and the corresponding $\mathbf{a}_i$ is the eigenvector of $\mathbf{R}_\mathbf{x}$. Clearly, since $1 \le m \le M$, the transformation matrix $\mathbf{A}$ should be chosen as the eigenmatrix $\mathbf{Q}$. Hence

$$\begin{bmatrix} \uparrow \\ \mathbf{w} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \longleftarrow & \mathbf{q}_1^H & \longrightarrow \\ \longleftarrow & \mathbf{q}_2^H & \longrightarrow \\ \vdots & \vdots & \vdots \\ \longleftarrow & \mathbf{q}_M^H & \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ \mathbf{x} \\ \downarrow \end{bmatrix}$$

or more concisely

$$\mathbf{w} = \mathbf{Q}^H \mathbf{x} \qquad (3.5.28)$$

provides an orthonormal transformation so that the transformed vector $\mathbf{w}$ is a zero-mean, uncorrelated random vector with autocorrelation $\mathbf{\Lambda}$. This transformation is called the DKLT, and its inverse relationship (or synthesis) is given by

$$\begin{bmatrix} \uparrow \\ \mathbf{x} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_M \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ \mathbf{w} \\ \downarrow \end{bmatrix} \qquad (3.5.29)$$

or

$$\mathbf{x} = \mathbf{Q}\mathbf{w} = \mathbf{q}_1 w_1 + \mathbf{q}_2 w_2 + \cdots + \mathbf{q}_M w_M \qquad (3.5.30)$$

From Section 3.5.1, the geometric interpretation of this transformation is that $\{w_k\}_1^M$ are projections of the vector $\mathbf{x}$ with respect to the rotated coordinate system of $\{\mathbf{q}_k\}_1^M$. The eigenvalues λ_i also have an interesting interpretation, as we shall see in the following representation.

Optimal reduced-basis representation

Generally we would expect any transformation to provide only few meaningful components so that we can use only those basis vectors resulting in a smaller representation error. To determine this *reduced-basis representation* property of the DKLT, let us use first $K < M$ eigenvectors (instead of all $\mathbf{q}_i$). Then from (3.5.26), we have

$$E_K = \sum_{i=K+1}^{M} \lambda_i \qquad (3.5.31)$$

In other words, the MSE in the reduced-basis representation, when the first K basis vectors are used, is the sum of the remaining eigenvalues (which are never negative). Therefore, to obtain a minimum MSE (that is, an optimum) representation, the procedure is to choose K eigenvectors corresponding to the K *largest* eigenvalues.

Application in data compression. The DKLT is a transformation on a random vector that produces a zero-mean, uncorrelated vector and that can minimize the mean square representation error. One of its popular applications is data compression in communications

and, in particular, in speech and image coding. Suppose we want to send a sample function of a speech process $x_c(t)$. If we sample this waveform and obtain M samples $\{x(n)\}_0^{M-1}$, then we need to send M data values. Instead, if we analyze the correlation of $\{x(n)\}_0^{M-1}$ and determine that M values can be approximated by a smaller K numbers of w_i and the corresponding $\mathbf{q}_i$, then we can compute these K data values $\{w_i\}_1^K$ at the transmitter and send them to the receiver through the communication channel. At the receiver, we can reconstruct $\{x(n)\}_0^{M-1}$ by using (3.5.23), as shown in Figure 3.13. Obviously, both the transmitter and receiver must have the information about the eigenvectors $\{\mathbf{q}_i\}_1^M$. A considerable amount of compression is achieved if K is much smaller than M.

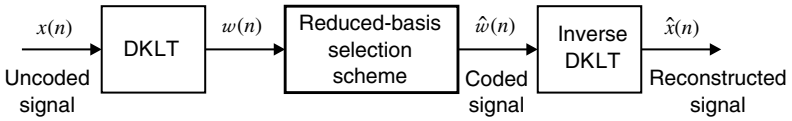


FIGURE 3.13
Signal coding scheme using the DKLT.

Periodic random sequences

As we noted in the previous section, the correlation matrix of a stationary process is Toeplitz. If the autocorrelation sequence of a random process is periodic with fundamental period M, its correlation matrix becomes *circulant*. All rows (columns) of a circulant matrix are obtained by circular rotation of its first row (column). Using (3.4.60) and the periodicity relation $r_x(l) = r_x(l - M)$, we obtain

$$\mathbf{R}_x = \begin{bmatrix} r_x(0) & r_x(1) & r_x(2) & \cdots & r_x(M-1) \\ r_x(M-1) & r_x(0) & r_x(1) & \cdots & r_x(M-2) \\ r_x(M-2) & r_x(M-1) & r_x(0) & \cdots & r_x(M-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_x(1) & r_x(2) & r_x(3) & \cdots & r_x(0) \end{bmatrix} \tag{3.5.32}$$

which is a circulant matrix. We note that a circulant matrix is Toeplitz but not vice versa.

If we define the M-point DFT of the periodic sequence $r_x(l)$

$$\tilde{R}_x(k) = \sum_{l=0}^{M-1} r_x(l) W_M^{kl} \tag{3.5.33}$$

where $W_M \triangleq e^{-j2\pi/M}$, and the vector

$$\mathbf{w}_k \triangleq \frac{1}{\sqrt{M}}[1 \; W_M^k \; W_M^{2k} \; \cdots \; W_M^{(M-1)k}]^T \qquad 0 \le k \le M-1 \tag{3.5.34}$$

we can easily see that multiplying the first row of $\mathbf{R}_x$ by the vector $\mathbf{w}_k$ results in $\tilde{R}_x(k)/\sqrt{M}$. Using $W_M^{-k} = W_M^{(M-1)k}$, we find that the product of the second row by $\mathbf{w}_k$ is equal to $\tilde{R}_x(k)W_M^k/\sqrt{M}$. In general, the ith row by $\mathbf{w}_k$ gives $\tilde{R}_x(k)W_M^{(i-1)k}/\sqrt{M}$. Therefore, we have

$$\mathbf{R}_x \mathbf{w}_k = \tilde{R}_x(k)\mathbf{w}_k \qquad 0 \le k \le M-1 \tag{3.5.35}$$

which shows that the normalized DFT vectors $\mathbf{w}_k$ are the eigenvectors of the circulant matrix $\mathbf{R}_x$ with as corresponding eigenvalues the DFT coefficients $\tilde{R}_x(k)$. Therefore, *the DFT provides the DKLT of periodic random sequences*. We recall that $\tilde{R}_x(k)$ are samples of the DTFT $R_x(e^{j2\pi k/M})$ of the finite-length sequence $r_x(l)$, $0 \le l \le M-1$.

If we define the $M \times M$ matrix

$$\mathbf{W} \triangleq [\mathbf{w}_0 \ \mathbf{w}_1 \ \cdots \ \mathbf{w}_{M-1}] \tag{3.5.36}$$

we can show that

$$\mathbf{W}^H\mathbf{W} = \mathbf{W}\mathbf{W}^H = \mathbf{I} \tag{3.5.37}$$

that is, the matrix $\mathbf{W}$ is unitary. The set of equations (3.5.35) can be written as

$$\mathbf{W}^H\mathbf{R}_x\mathbf{W} = \text{diag}\{\tilde{R}_x(0), \tilde{R}_x(1), \ldots, \tilde{R}_x(M-1)\} \tag{3.5.38}$$

which shows that the DFT performs the diagonalization of circulant matrices. Although there is no fast algorithm for the diagonalization of general Toeplitz matrices, in many cases we can use the DFT to approximate the DKLT of stationary random sequences. The approximation is adequate if the correlation becomes negligible for $|l| > M$, which is the case for many stationary processes. This explains the fact that the eigenvectors of a Toeplitz matrix resemble complex exponentials for large values of M. The DKLT also can be extended to handle the representation of random sequences. These issues are further explored in Therrien (1992), Gray (1972), and Fukunaga (1990).

3.6 PRINCIPLES OF ESTIMATION THEORY

The key assumption underlying our discussion up to this point was that the probability distributions associated with the problem under consideration were known. As a result, all required probabilities, autocorrelation sequences, and PSD functions either could be derived from a set of assumptions about the involved random processes or were given a priori. However, in most practical applications, this is the exception rather than the rule. Therefore, the properties and parameters of random variables and random processes should be obtained by collecting and analyzing finite sets of measurements. In this section, we introduce some basic concepts of estimation theory that will be used repeatedly in the rest of the book. Complete treatments of estimation theory can be found in Kay (1993), Helstrom (1995), Van Trees (1968), and Papoulis (1991).

3.6.1 Properties of Estimators

Suppose that we collect N observations $\{x(n)\}_0^{N-1}$ from a stationary stochastic process and use them to estimate a parameter θ (which we assume to be real-valued) of the process using some function $\hat{\theta}[\{x(n)\}_0^{N-1}]$. The same results can be used for a set of measurements $\{x_k(n)\}_1^N$ obtained from N sensors sampling stochastic processes with the same distributions. The function $\hat{\theta}[\{x(n)\}_0^{N-1}]$ is known as an *estimator* whereas the value taken by the estimator, using a particular set of observations, is called a *point estimate* or simply an *estimate*. The intention of the estimator design is that the estimate should be as close to the true value of the parameter as possible. However, if we use another set of observations or a different number of observations from the same set, it is highly unlikely that we will obtain the same estimate. As an example of an estimator, consider estimating the mean μ_x of a stationary process $x(n)$ from its N observations $\{x(n)\}_0^{N-1}$. Then the natural estimator is a simple arithmetic average of these observations, given by

$$\hat{\mu}_x = \hat{\theta}[\{x(n)\}_0^{N-1}] = \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{3.6.1}$$

Similarly, a natural estimator of the variance σ_x^2 of the process $x(n)$ would be

$$\hat{\sigma}_x^2 = \hat{\theta}[\{x(n)\}_0^{N-1}] = \frac{1}{N}\sum_{n=0}^{N-1}[x(n) - \hat{\mu}_x]^2 \tag{3.6.2}$$

If we repeat this procedure a large number of times, we will obtain a large number of estimates, which can be used to generate a histogram showing the distribution of the estimates. Before the collection of observations, we would like to describe all sets of data that can be obtained by using the random variables $\{x(n,\zeta)\}_0^{N-1}$. The obtained set of N observations $\{x(n)\}_0^{N-1}$ can thus be regarded as one realization of the random variables $\{x(n,\zeta)\}_0^{N-1}$ defined on an N-dimensional sample space. In this sense, the estimator $\hat{\theta}[\{x(n,\zeta)\}_0^{N-1}]$ becomes a random variable whose distribution can be obtained from the joint distribution of the random variables $\{x(n,\zeta)\}_0^{N-1}$. This distribution is called the *sampling distribution* of the estimator and is a fundamental concept in estimation theory because it provides all the information we need to evaluate the quality of an estimator.

The sampling distribution of a "good" estimator should be concentrated as closely as possible about the parameter that it estimates. To determine how "good" an estimator is and how different estimators of the same parameter compare with one another, we need to determine their sampling distributions. Since it is *not* always possible to derive the exact sampling distributions, we have to resort to properties that use the lower-order moments (mean, variance, mean square error) of the estimator.

Bias of estimator. The *bias* of an estimator $\hat{\theta}$ of a parameter θ is defined as

$$B(\hat{\theta}) \triangleq E[\hat{\theta}] - \theta \tag{3.6.3}$$

while the *normalized* bias is defined as

$$\varepsilon_{\mathrm{b}} \triangleq \frac{B(\hat{\theta})}{\theta} \qquad \theta \neq 0 \tag{3.6.4}$$

When $B(\hat{\theta}) = 0$, the estimator is said to be *unbiased* and the pdf of the estimator is centered exactly at the true value θ. Generally, one should select estimators that are unbiased such as the mean estimator in (3.6.1) or very nearly unbiased such as the variance estimator in (3.6.2). However, it is not always wise to select an unbiased estimator, as we will see below and in Section 5.2 on the estimation of autocorrelation sequences.

Variance of estimator. The *variance* of the estimator $\hat{\theta}$ is defined by

$$\mathrm{var}(\hat{\theta}) = \sigma_{\hat{\theta}}^2 \triangleq E\{|\hat{\theta} - E\{\hat{\theta}\}|^2\} \tag{3.6.5}$$

which measures the spread of the pdf of $\hat{\theta}$ around its average value. Therefore, one would select an estimator with the smallest variance. However, this selection is not always compatible with the small bias requirement. As we will see below, reducing variance may result in an increase in bias. Therefore, a balance between these two conflicting requirements is required, which is provided by the mean square error property. The *normalized standard deviation* (also called the coefficient of variation) is defined by

$$\varepsilon_{\mathrm{r}} \triangleq \frac{\sigma_{\hat{\theta}}}{\theta} \qquad \theta \neq 0 \tag{3.6.6}$$

Mean square error. The *mean square error* (*MSE*) of the estimator is given by

$$\mathrm{MSE}(\theta) = E\{|\hat{\theta} - \theta|^2\} = \sigma_{\hat{\theta}}^2 + |B_{\hat{\theta}}|^2 \tag{3.6.7}$$

Indeed, we have

$$\begin{aligned}\mathrm{MSE}(\theta) &= E\{|\theta - E\{\hat{\theta}\} - (\hat{\theta} - E\{\hat{\theta}\})|^2\} \\ &= E\{|\theta - E\{\hat{\theta}\}|^2\} + E\{|\hat{\theta} - E\{\hat{\theta}\}|^2\} \qquad (3.6.8)\\ &\quad -(\theta - E\{\hat{\theta}\})E\{(\hat{\theta} - E\{\hat{\theta}\})^*\} - (\theta - E\{\hat{\theta}\})^* E\{\hat{\theta} - E\{\hat{\theta}\}\} \\ &= |\theta - E\{\hat{\theta}\}|^2 + E\{|\hat{\theta} - E\{\hat{\theta}\}|^2\} \qquad (3.6.9)\end{aligned}$$

which leads to (3.6.7) by using (3.6.3) and (3.6.5). Ideally, we would like to minimize the MSE, but this minimum is not always zero. Hence minimizing variance can increase the bias. The *normalized MSE* is defined as

$$\varepsilon \triangleq \frac{\text{MSE}(\theta)}{\theta} \qquad \theta \neq 0 \tag{3.6.10}$$

Cramér-Rao lower bound. If it is possible to minimize the MSE when the bias is zero, then clearly the variance is also minimized. Such estimators are called *minimum variance unbiased* estimators, and they attain an important minimum bound on the variance of the estimator, called the *Cramér-Rao lower bound* (CRLB), or *minimum variance bound*. If $\hat{\theta}$ is unbiased, then it follows that $E\{\hat{\theta} - \theta\} = 0$, which may be expressed as

$$\int_{-\infty}^{\infty} \cdots \int (\hat{\theta} - \theta) f_{\mathbf{x};\theta}(\mathbf{x}; \theta)\, d\mathbf{x} = 0 \tag{3.6.11}$$

where $\mathbf{x}(\zeta) = [x_1(\zeta), x_2(\zeta), \ldots, x_N(\zeta)]^T$ and $f_{\mathbf{x};\theta}(\mathbf{x}; \theta)$ is the joint density of $\mathbf{x}(\zeta)$, which depends on a fixed but unknown parameter θ. If we differentiate (3.6.11) with respect to θ, assuming real-valued $\hat{\theta}$, we obtain

$$0 = \int_{-\infty}^{\infty} \cdots \int \frac{\partial}{\partial\theta}[(\hat{\theta} - \theta) f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]\, d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int (\hat{\theta} - \theta) \frac{\partial f_{\mathbf{x};\theta}(\mathbf{x}; \theta)}{\partial\theta}\, d\mathbf{x} - 1 \tag{3.6.12}$$

Using the fact

$$\frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} = \frac{1}{f_{\mathbf{x};\theta}(\mathbf{x}; \theta)} \frac{\partial f_{\mathbf{x};\theta}(\mathbf{x}; \theta)}{\partial\theta}$$

or

$$\frac{\partial f_{\mathbf{x};\theta}(\mathbf{x}; \theta)}{\partial\theta} = \frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} f_{\mathbf{x};\theta}(\mathbf{x}; \theta) \tag{3.6.13}$$

and substituting (3.6.13) in (3.6.12), we get

$$\int_{-\infty}^{\infty} \cdots \int \left\{ (\hat{\theta} - \theta) \frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} \right\} f_{\mathbf{x};\theta}(\mathbf{x}; \theta)\, d\mathbf{x} = 1 \tag{3.6.14}$$

Clearly, the left side of (3.6.14) is simply the expectation of the expression inside the brackets, that is,

$$E\left\{ (\hat{\theta} - \theta) \frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} \right\} = 1 \tag{3.6.15}$$

Using the *Cauchy-Schwarz inequality* (Papoulis 1991; Stark and Woods 1994) $|E\{x(\zeta)y(\zeta)\}|^2 \leq E\{|x(\zeta)|^2\}E\{|y(\zeta)|^2\}$, we obtain

$$E\{(\hat{\theta} - \theta)^2\} E\left\{ \left(\frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} \right)^2 \right\} \geq E^2 \left\{ (\hat{\theta} - \theta) \frac{\partial \ln[f_{\mathbf{x};\theta}(\mathbf{x}; \theta)]}{\partial\theta} \right\} = 1 \tag{3.6.16}$$

The first term on the left-hand side is the variance of the estimator $\hat{\theta}$ since it is unbiased. Hence

$$\text{var}(\hat{\theta}) \geq \frac{1}{E\{[\partial \ln f_{\mathbf{x};\theta}(\mathbf{x}; \theta)/\partial\theta]^2\}} \tag{3.6.17}$$

which is one form of the CRLB and can also be expressed as

$$\text{var}(\hat{\theta}) \geq -\frac{1}{E\{\partial^2 \ln f_{\mathbf{x};\theta}(\mathbf{x}; \theta)/\partial\theta^2\}} \tag{3.6.18}$$

The function $\ln f_{\mathbf{x};\theta}(\mathbf{x}; \theta)$ is called the *log likelihood function* of θ. The CRLB expresses the minimum error variance of any estimator $\hat{\theta}$ of θ in terms of the joint density $f_{\mathbf{x};\theta}(\mathbf{x}; \theta)$

of observations. Hence every unbiased estimator must have a variance greater than a certain number. An unbiased estimate that satisfies the CRLB (3.6.18) with equality is called an *efficient* estimate. If such an estimate exists, then it can be obtained as a unique solution to the likelihood equation

$$\frac{\partial \ln f_{\mathbf{x};\theta}(\mathbf{x};\theta)}{\partial \theta} = 0 \tag{3.6.19}$$

The solution of (3.6.19) is called the *maximum likelihood* (*ML*) estimate. Note that if the efficient estimate does not exist, then the ML estimate will not achieve the lower bound and hence it is difficult to ascertain how closely the variance of any estimate will approach the bound. The CRLB can be generalized to handle the estimation of vector parameters (Therrien 1992).

Consistency of estimator. If the MSE of the estimator can be made to approach zero as the sample size N becomes large, then from (3.6.7) both the bias and the variance will tend to zero. Then the sampling distribution will tend to concentrate about θ, and eventually as $N \to \infty$, the sampling distribution will become an impulse at θ. This is an important and desirable property, and the estimator that possesses it is called a *consistent* estimator.

Confidence interval. If we know the sampling distribution of an estimator, we can use the observations to compute an interval that has a specified probability of covering the unknown true parameter value. This interval is called a *confidence interval*, and the coverage probability is called the *confidence level*. When we interpret the meaning of confidence intervals, it is important to remember that it is the interval that is the random variable, and not the parameter. This concept will be explained in the sequel by means of specific examples.

3.6.2 Estimation of Mean

The natural estimator of the mean μ_x of a stationary sequence $x(n)$ from the observations $\{x(n)\}_0^{N-1}$ is the *sample mean*, given by

$$\hat{\mu}_x = \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{3.6.20}$$

The estimate $\hat{\mu}_x$ is a random variable that depends on the number and values of the observations. Changing N or the set of observations will lead to another value for $\hat{\mu}_x$. Since the mean of the estimator is given by

$$E\{\hat{\mu}_x\} = \mu_x \tag{3.6.21}$$

the estimator $\hat{\mu}_x$ is unbiased. If $x(n) \sim \mathrm{WN}(\mu_x, \sigma_x^2)$, we have

$$\mathrm{var}(\hat{\mu}_x) = \frac{\sigma_x^2}{N} \tag{3.6.22}$$

because the samples of the process are uncorrelated random variables. This variance, which is a measure of the estimator's quality, increases if $x(n)$ is nonwhite.

Indeed, for a correlated random sequence, the variance of $\hat{\mu}_x$ is given by (see Problem 3.30)

$$\mathrm{var}(\hat{\mu}_x) = N^{-1}\sum_{l=-N}^{N}\left(1 - \frac{|l|}{N}\right)\gamma_x(l) \le N^{-1}\sum_{l=-N}^{N}|\gamma_x(l)| \tag{3.6.23}$$

where $\gamma_x(l)$ is the covariance sequence of $x(n)$. If $\gamma_x(l) \to 0$ as $l \to \infty$, then $\mathrm{var}(\hat{\mu}_x) \to 0$ as $N \to \infty$ and hence $\hat{\mu}_x$ is a consistent estimator of μ_x. If $\sum_{l=-\infty}^{\infty}|\gamma_x(l)| < \infty$, then

from (3.6.23)

$$\lim_{N\to\infty} N \operatorname{var}(\hat{\mu}_x) = \lim_{N\to\infty} \sum_{l=-N}^{N} \left(1 - \frac{|l|}{N}\right) \gamma_x(l) = \sum_{l=-\infty}^{\infty} \gamma_x(l) \tag{3.6.24}$$

The expression for $\operatorname{var}(\hat{\mu}_x)$ in (3.6.23) can also be put in the form (see Problem 3.30)

$$\operatorname{var}(\hat{\mu}_x) = \frac{\sigma_x^2}{N}[1 + \Delta_N(\rho_x)] \tag{3.6.25}$$

where
$$\Delta_N(\rho_x) = 2\sum_{l=1}^{N} \left(1 - \frac{l}{N}\right) \rho_x(l) \qquad \rho_x(l) = \frac{\gamma_x(l)}{\sigma_x^2} \tag{3.6.26}$$

When $\Delta_N(\rho_x) \geq 0$, the variance of the estimator increases as the amount of correlation among the samples of $x(n)$ increases. This implies that as the correlation increases, we need more samples to retain the quality of the estimate because each additional sample carries "less information." For this reason the estimation of long-memory processes and processes with infinite variance is extremely difficult.

Sampling distribution. If we know the joint pdf of the random variables $\{x(n)\}_0^{N-1}$, we can determine, at least in principle, the pdf of $\hat{\mu}_x$. For example, if it is assumed that the observations are IID as $\mathcal{N}(\mu_x, \sigma_x^2)$ then from (3.6.21) and (3.6.22), it can be seen that $\hat{\mu}_x$ is normal with mean μ_x and variance σ_x^2/N, that is,

$$f_{\hat{\mu}_x}(\hat{\mu}_x) = \frac{1}{\sqrt{2\pi}(\sigma_x/\sqrt{N})} \exp\left[-\frac{1}{2}\left(\frac{\hat{\mu}_x - \mu_x}{\sigma_x/\sqrt{N}}\right)^2\right] \tag{3.6.27}$$

which is the sampling distribution of the mean. If N is large, then from the central limit theorem, the sampling distribution of the sample mean (3.6.27) is usually very close to the normal distribution, even if the individual distributions are not normal.

If we know the standard deviation σ_x, we can compute the probability

$$\Pr\left\{\mu_x - k\frac{\sigma_x}{\sqrt{N}} < \hat{\mu}_x < \mu_x + k\frac{\sigma_x}{\sqrt{N}}\right\} \tag{3.6.28}$$

that the random variable $\hat{\mu}_x$ is within a certain interval specified by two fixed quantities. A simple rearrangement of the above inequality leads to

$$\Pr\left\{\hat{\mu}_x - k\frac{\sigma_x}{\sqrt{N}} < \mu_x < \hat{\mu}_x + k\frac{\sigma_x}{\sqrt{N}}\right\} \tag{3.6.29}$$

which gives the probability that the fixed quantity μ_x lies between the two random variables $\hat{\mu}_x - k\sigma_x/\sqrt{N}$ and $\hat{\mu}_x + k\sigma_x/\sqrt{N}$. Hence (3.6.29) provides the probability that an interval with fixed length $2k\sigma_x/\sqrt{N}$ and randomly centered at the estimated mean includes the true mean. If we choose k so that the probability defined by (3.6.29) is equal to 0.95, the interval is known as the 95 percent confidence interval. To understand the meaning of this reasoning, we stress that for each set of measurements we compute a confidence interval that either contains or does not contain the true mean. However, if we repeat this process for a large number of observation sets, about 95 percent of the obtained confidence intervals will include the true mean. We stress that by no means does this imply that a confidence interval includes the true mean with probability 0.95.

If the variance σ_x^2 is unknown, then it has to be determined from the observations. This results in two modifications of (3.6.29). First, σ_x is replaced by

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{n=0}^{N-1}[x(n) - \hat{\mu}_x]^2 \tag{3.6.30}$$

which implies that the center and the length of the confidence interval are different for each set of observations. Second, the random variable $(\hat{\mu}_x - \mu_x)/(\hat{\sigma}_x/\sqrt{N})$ is distributed according to *Student's t distribution with* $v = N - 1$ *degrees of freedom* (Parzen 1960), which tends to a Gaussian for large values of N. In these cases, the factor k in (3.6.29) is replaced by the appropriate value t of Student's distribution, using $N - 1$ degrees of freedom, for the desired level of confidence.

If the observations are normal but not IID, then from (3.6.25), the mean estimator $\hat{\mu}_x$ is normal with mean μ and variance $(\sigma_x^2/N)[1 + \Delta_N(\rho_x)]$. It is now easy to construct exact confidence intervals for $\hat{\mu}_x$ if $\rho_x(l)$ is known, and approximate confidence intervals if $\rho_x(l)$ is to be estimated from the observations. For large N, the variance $\text{var}(\hat{\mu}_x)$ can be approximated by

$$\begin{aligned} \text{var}(\hat{\mu}_x) &= \frac{\sigma_x^2}{N}[1 + \Delta_N(\rho_x)] \\ &\simeq \frac{\sigma_x^2}{N}\left[1 + 2\sum_1^N \rho_x(l)\right] \\ &\triangleq \frac{v}{N} \qquad v = \sigma_x^2\left\{1 + 2\sum_1^N \rho_x(l)\right\} \end{aligned} \tag{3.6.31}$$

and hence an approximate 95 percent confidence interval for $\hat{\mu}_x$ is given by

$$\left(\hat{\mu}_x - 1.96\sqrt{\frac{v}{N}}, \hat{\mu}_x + 1.96\sqrt{v/N}\right) \tag{3.6.32}$$

This means that, on average, the above interval will enclose the true value μ_x on 95 percent of occasions. For many practical random processes (especially those modeled as ARMA processes), the result in (3.6.32) is a good approximation.

EXAMPLE 3.6.1. Consider the AR(1) process

$$x(n) = ax(n-1) + w(n) \qquad -1 < a < 1$$

where $w(n) \sim \text{WN}(0, \sigma_w^2)$. We wish to compute the variance of the mean estimator $\hat{\mu}_x$ of the process $x(n)$. Using straightforward calculations, we obtain

$$\mu_x = 0 \qquad \sigma_x^2 = \frac{\sigma_w^2}{1 - a^2} \qquad \text{and} \qquad \rho_x(l) = a^{|l|}$$

From (3.6.26) we evaluate the term

$$\Delta_N(\rho) = \frac{2a}{1-a}\left[1 - \frac{1}{N(1-a)} + \frac{a^N}{N(1-a)}\right] \simeq \frac{2a}{1-a} \qquad \text{for } N \gg 1$$

When $a \to 1$, that is, when the dependence between the signal samples increases, then the factor $\Delta_N(\rho)$ takes large values and the quality of estimator decreases drastically. Similar conclusions can be drawn using the approximation (3.6.31)

$$v = \left(1 + 2\sum_1^\infty a^l\right)\frac{\sigma_w^2}{1-a^2} = \frac{\sigma_w^2}{(1-a)^2}$$

We will next verify these results using two Monte Carlo simulations: one for $a = 0.9$, which represents high correlations among samples, and the other for $a = 0.1$. Using a Gaussian pseudorandom number generator with mean 0 and variance $\sigma_w^2 = 1$, we generated $N = 100$ samples of the AR(1) process $x(n)$. Using v in (3.6.31) and (3.6.32), we next computed the confidence intervals. For $a = 0.9$, we obtain

$$v = 100 \qquad \text{and} \qquad \text{confidence interval: } (\hat{\mu}_x - 1.96, \hat{\mu}_x + 1.96)$$

and for $a = 0.1$, we obtain

$$v = 1.2345 \qquad \text{and} \qquad \text{confidence interval: } (\hat{\mu}_x - 0.2178, \hat{\mu}_x + 0.2178)$$

Clearly, when the dependence between signal samples increases, the quality of the estimator decreases drastically and hence the confidence interval is wider. To have the same confidence interval, we should increase the number of samples N.

We next estimate the mean, using (3.6.20), and we repeat the experiment 10,000 times. Figure 3.14 shows histograms of the computed means for $a = 0.9$ and $a = 0.1$. The confidence intervals are also shown as dotted lines around the true mean. The histograms are approximately Gaussian in shape. The histogram for the high-correlation case is wider than that for the low-correlation case, which is to be expected. The 95 percent confidence intervals also indicate that very few estimates are outside the interval.

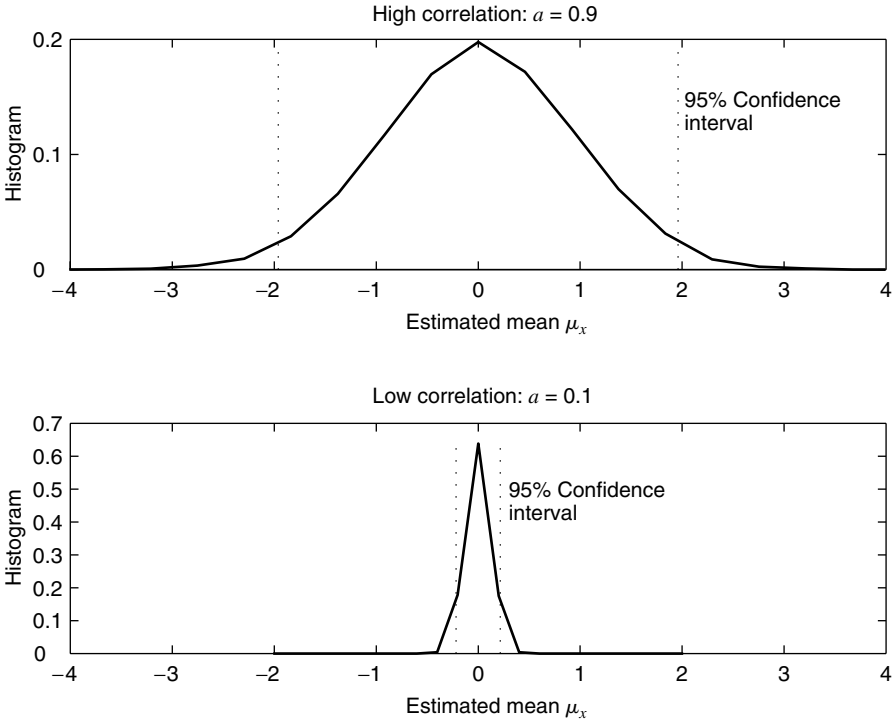


FIGURE 3.14
Histograms of mean estimates in Example 3.6.1.

3.6.3 Estimation of Variance

The natural estimator of the variance σ_x of a stationary sequence $x(n)$ from the observations $\{x(n)\}_0^{N-1}$ is the *sample variance*, given by

$$\hat{\sigma}_x^2 \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\{x(n) - \hat{\mu}_x\}^2 \tag{3.6.33}$$

By using the mean estimate $\hat{\mu}_x$ from (3.6.20), the mean of the variance estimator can be shown to equal (see Problem 3.31)

$$E\{\hat{\sigma}_x^2\} = \sigma_x^2 - \text{var}(\hat{\mu}_x) = \sigma_x^2 - \frac{1}{N}\sum_{l=-N}^{N}\left(1 - \frac{|l|}{N}\right)\gamma_x(l) \tag{3.6.34}$$

If the sequence $x(n)$ is uncorrelated, then

$$E\{\hat{\sigma}_x^2\} = \sigma_x^2 - \frac{\sigma_x^2}{N} = \left(\frac{N-1}{N}\right)\sigma_x^2 \tag{3.6.35}$$

From (3.6.34) or (3.6.35), it is obvious that the estimator in (3.6.33) is biased. If $\gamma_x(l) \to 0$ as $l \to \infty$, then $\text{var}(\hat{\mu}_x) \to 0$ as $N \to \infty$ and hence $\hat{\sigma}_x^2$ is an asymptotically unbiased estimator of σ_x^2. In practical applications, the variance estimate is nearly unbiased for large N. Note that if we use the actual mean μ_x in (3.6.33), then the resulting estimator is unbiased.

The general expression for the variance of the variance estimator is fairly complicated and requires higher-order moments. It can be shown that for either estimators

$$\text{var}(\hat{\sigma}_x^2) \approx \frac{\gamma_x^{(4)}}{N} \qquad \text{for large } N \tag{3.6.36}$$

where $\gamma_x^{(4)}$ is the fourth central moment of $x(n)$ (Brockwell and Davis 1991). Thus the estimator in (3.6.33) is also consistent.

Sampling distribution. In the case of the mean estimator, the sampling distribution involved the distribution of sums of random variables. The variance estimator involves the sum of the squares of random variables, for which the sampling distribution computation is complicated. For example, if there are N independent measurements from an $\mathcal{N}(0, 1)$ distribution, then the sampling distribution of the random variable

$$\chi_N^2 = x_1^2 + x_2^2 + \cdots + x_N^2 \tag{3.6.37}$$

is given by the *chi-squared distribution with N degrees of freedom*. The general form of χ_N^2 with ν degrees of freedom is

$$f_{\chi_\nu^2}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp\left(-\frac{x}{2}\right) \qquad 0 \le x \le \infty \tag{3.6.38}$$

where $\Gamma(\nu/2) = \int_0^\infty e^{-t} t^{\nu/2-1}\, dt$ is the gamma function with argument $\nu/2$.

For the variance estimator in (3.6.33), it can be shown (Parzen 1960) that $N\hat{\sigma}_x^2$ is distributed as chi squared with $\nu = N - 1$ *degrees of freedom*. This means that, for any set of N observations, there will only be $N - 1$ independent deviations $\{x(n) - \hat{\mu}_x\}$, since their sum is zero from the definition of the mean. Assuming that the observations are $\mathcal{N}(\mu, \sigma^2)$, the random variables $x(n)/\sigma$ will be $\mathcal{N}(\mu/\sigma, 1)$ and hence the random variable

$$\frac{N\hat{\sigma}_x^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} [x(n) - \hat{\mu}_x]^2 \tag{3.6.39}$$

will be chi squared distributed with $\nu = N - 1$. Therefore, using values of the chi-squared distribution, confidence intervals for the variance estimator can be computed. In particular, since $N\hat{\sigma}_x^2/\sigma^2$ is distributed as χ_ν^2, the 95 percent limits of the form

$$\Pr\left\{\chi_\nu\left(\frac{0.05}{2}\right) < N\hat{\sigma}_x^2/\sigma^2 \le \chi_\nu\left(1 - \frac{0.05}{2}\right)\right\} = 0.95 \tag{3.6.40}$$

can be obtained from chi-squared tables (Fisher and Yates 1938). By rearranging (3.6.40), the random variable $\sigma^2/\hat{\sigma}_x^2$ satisfies

$$\Pr\left\{\frac{N}{\chi_\nu(0.975)} < \frac{\sigma^2}{\hat{\sigma}_x^2} \le \frac{N}{\chi_\nu(0.025)}\right\} = 0.95 \tag{3.6.41}$$

Using $l_1 = N/\chi_\nu(0.975)$ and $l_2 = N/\chi_\nu(0.025)$, we see that (3.6.41) implies that

$$\Pr\{l_2\hat{\sigma}_x^2 \ge \sigma^2 \quad \text{and} \quad l_1\hat{\sigma}_x^2 < \sigma^2\} = 0.95 \tag{3.6.42}$$

Thus the 95 percent confidence interval based on the estimate $\hat{\sigma}_x^2$ is $(l_1\hat{\sigma}_x^2, l_2\hat{\sigma}_x^2)$. Note that this interval is sensitive to the validity of the normal assumption of random variables leading to (3.6.39). This is not the case for the confidence intervals for the mean estimates

because, thanks to the central limit theorem, the computation of the interval can be based on the normal assumption.

EXAMPLE 3.6.2. Consider again the AR(1) process given in Example 3.6.1:

$$x(n) = ax(n-1) + w(n) \qquad -1 < a < 1 \qquad w(n) \sim \text{WN}(0, 1)$$

with $$\mu_x = 0 \qquad \sigma_x^2 = \frac{\sigma_w^2}{1-a^2} \qquad \text{and} \qquad \rho_x(l) = a^{|l|} \tag{3.6.43}$$

We wish to compute the mean of the variance estimator $\hat{\sigma}_x^2$ of the process $x(n)$. From (3.6.34), we obtain

$$E[\hat{\sigma}_x^2] = \sigma_x^2 \left[1 - \frac{1}{N} \sum_{l=-N}^{N} \left(1 - \frac{|l|}{N} \right) a^{|l|} \right] \tag{3.6.44}$$

When $a \to 1$, that is, when the dependence between the signal samples increases, the mean of the estimate deviates significantly from the true value σ_x^2 and the quality of the estimator decreases drastically. For small dependence, the mean is very close to σ_x^2. These conclusions can be verified using two Monte Carlo simulations as before: one for $a = 0.9$, which represents high correlations among samples, and the other for $a = 0.1$. Using a Gaussian pseudorandom number generator with mean 0 and unit variance, we generated $N = 100$ samples of the AR(1) process $x(n)$. The computed parameters according to (3.6.43) and (3.6.44) are

$$a = 0.9: \qquad \sigma_x^2 = 5.2632 \qquad E\{\hat{\sigma}_x^2\} = 4.3579$$

$$a = 0.1: \qquad \sigma_x^2 = 1.0101 \qquad E\{\hat{\sigma}_x^2\} = 0.9978$$

We next estimate the variance by using (3.6.33) and repeat the experiment 10,000 times. Figure 3.15 shows histograms of computed variances for $a = 0.9$ and for $a = 0.1$. The computed

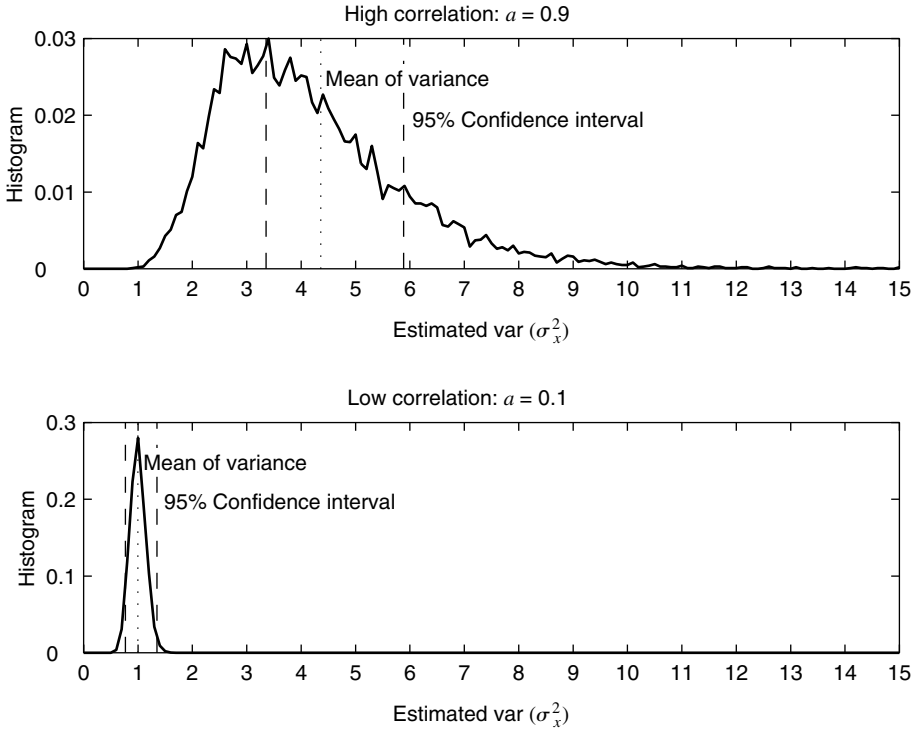


FIGURE 3.15
Histograms of variance estimates in Example 3.6.2.

means of the variance estimates are also shown as dotted lines. Clearly, the histogram is much wider for the high-correlation case and much narrower (almost symmetric and Gaussian) for the low-correlation case.

The 95 percent confidence intervals are given by $(l_1\hat{\sigma}_x^2, l_2\hat{\sigma}_x^2)$, where $l_1 = N/\chi_\nu(0.975)$ and $l_2 = N/\chi_\nu(0.025)$. The values of l_1 and l_2 are obtained from the chi-squared distribution curves (Jenkins and Watts 1968). For $N = 100$, $l_1 = 0.77$ and $l_2 = 1.35$; hence the 95 percent confidence intervals for σ_x^2 are

$$(0.77\hat{\sigma}_x^2, 1.35\hat{\sigma}_x^2)$$

also shown as dashed lines around the mean value $E\{\hat{\sigma}_x^2\}$. The confidence interval for the high-correlation case, $a = 0.9$, does not appear to be a good interval, which implies that the approximation leading to (3.6.42) is not a good one for this case. Such is not the case for $a = 0.1$.

3.7 SUMMARY

In this chapter we provided an overview of the basic theory of discrete-time stochastic processes. We began with the notion of a random variable as a mapping from the abstract probability space to the real space, extended it to random vectors as a collection of random variables, and introduced discrete-time stochastic processes as an indexed family (or time series) of random variables. A complete probabilistic description of these random objects requires the knowledge of joint distribution or density functions, which is difficult to acquire except in simple cases. Therefore, the emphasis was placed on description using joint moments of distributions, and, in particular, the emphasis was placed on the second-order moments, which are relatively easy to estimate or compute in practice.

We defined the mean and the variance to describe random variables, and we provided three useful models of random variables. For random vector description, we defined the mean vector and the autocorrelation matrix. Linear transformations of random vectors were discussed, using densities and correlation matrices. The normal random vector was then introduced as a useful model of a random vector. A particularly simple linear transformation, namely, the sum of independent random variables, was used to introduce random variables with stable and infinitely divisible distributions. To describe stochastic processes, we proceeded to define mean and autocorrelation sequences. In many applications, the concept of stationary of random processes is a useful one that reduces the computational complexity. Assuming time invariance on the first two moments, we defined a wide-sense stationary (WSS) process in which the mean is a constant and correlation between random variables at two distinct times is a function of time difference or lag. The rest of the chapter was devoted to the analysis of WSS processes.

A stochastic process is generally observed in practice as a single sample function (a speech signal or a radar signal) from which it is necessary to estimate the first- and the second-order moments. This requires the notion of ergodicity, which provides a framework for the computation of statistical averages using time averages over a single realization. Although this framework requires theoretical results using mean square convergence, we provided a simple approach of using appropriate time averages. An important random signal characteristic called variability was introduced. The WSS processes were then described in the frequency domain using the power spectral density function, which is a physical quantity that can be measured in practice. Some random processes exhibiting flat spectral envelopes were analyzed including one of white noise. Since random processes are generally processed using linear systems, we described linear system operations with random inputs in both the time and frequency domains.

The properties of correlation matrices and sequences play an important role in filtering and estimation theory and were discussed in detail, including eigenanalysis. Another important random signal characteristic called memory was also introduced. Stationary random

signals were modeled using autocorrelation matrices, and the relationship between spectral flatness and eigenvalue spread was explored. These properties were used in an alternate representation of random vectors as well as processes using uncorrelated components which were based on diagonalization and triangularization of correlation matrices. This resulted in the discrete KL transform and KL expansion. These concepts will also be useful in later chapters on optimal filtering and adaptive filtering.

Finally, we concluded this chapter with the introduction of elementary estimation theory. After discussion of properties of estimators, two important estimators of mean and variance were treated in detail along with their sampling distributions. These topics will be useful in many subsequent chapters.

PROBLEMS

3.1 The exponential density function is given by

$$f_X(x) = \frac{1}{a} e^{-x/a} u(x) \tag{P.1}$$

where a is a parameter and $u(x)$ is a unit step function.

(a) Plot the density function for $a = 1$.
(b) Determine the mean, variance, skewness,and kurtosis of the Rayleigh random variable with $a = 1$. Comment on the significance of these moments in terms of the shape of the density function.
(c) Determine the characteristic function of the exponential pdf.

3.2 The Rayleigh density function is given by

$$f_X(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} u(x) \tag{P.2}$$

where σ is a parameter and $u(x)$ is a unit step function. Repeat Problem 3.1 for $\sigma = 1$.

3.3 Using the binomial expansion of $\{x(\zeta) - \mu_x\}^m$, show that the mth central moment is given by

$$M_m^{(x)} = \sum_{k=0}^{m} \binom{m}{k} (-1)^k \mu_x^k \xi_{m-k}^{(x)}$$

Similarly, show that

$$\xi_m^{(x)} = \sum_{k=0}^{m} \binom{m}{k} \mu_x^k M_{m-k}^{(x)}$$

3.4 Consider a zero-mean random variable $x(\zeta)$. Using (3.1.26), show that the first four cumulants of $x(\zeta)$ are given by (3.1.28) through (3.1.31).

3.5 A random vector $\mathbf{x}(\zeta) = [x_1(\zeta)\ x_2(\zeta)]^T$ has mean vector $\boldsymbol{\mu}_\mathbf{x} = [1\ 2]^T$ and covariance matrix

$$\boldsymbol{\Gamma}_\mathbf{x} = \begin{bmatrix} 4 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

This vector is transformed to another random vector $\mathbf{y}(\zeta)$ by the following linear transformation:

$$\begin{bmatrix} y_1(\zeta) \\ y_2(\zeta) \\ y_3(\zeta) \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ -1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1(\zeta) \\ x_2(\zeta) \end{bmatrix}$$

Determine (a) the mean vector $\boldsymbol{\mu}_\mathbf{y}$, ($b$) the autocovariance matrix $\boldsymbol{\Gamma}_\mathbf{y}$, and ($c$) the cross-correlation matrix $\mathbf{R}_{\mathbf{xy}}$.

3.6 Using the moment generating function, show that the linear transformation of a Gaussian random vector is also Gaussian.

3.7 Let $\{x_k(\zeta)\}_{k=1}^4$ be four IID random variables with exponential distribution (P.1) with $a = 1$. Let

$$y_k(\zeta) = \sum_{l=1}^{k} x_l(\zeta) \qquad 1 \le k \le 4$$

(*a*) Determine and plot the pdf of $y_2(\zeta)$.
(*b*) Determine and plot the pdf of $y_3(\zeta)$.
(*c*) Determine and plot the pdf of $y_4(\zeta)$.
(*d*) Compare the pdf of $y_4(\zeta)$ with that of the Gaussian density.

3.8 For each of the following, determine whether the random process is (1) WSS or (2) m.s. ergodic in the mean.

(*a*) $X(t) = A$, where A is a random variable uniformly distributed between 0 and 1.
(*b*) $X_n = A\cos\omega_0 n$, where A is a Gaussian random variable with mean 0 and variance 1.
(*c*) A Bernoulli process with $\Pr[X_n = 1] = p$ and $\Pr[X_n = -1] = 1 - p$.

3.9 Consider the harmonic process $x(n)$ defined in (3.3.50).

(*a*) Determine the mean of $x(n)$.
(*b*) Show that the autocorrelation sequence is given by

$$r_x(l) = \frac{1}{2}\sum_{k=1}^{M} |c_k|^2 \cos\omega_k l \qquad -\infty < l < \infty$$

3.10 Suppose that the random variables ϕ_k in the real-valued harmonic process model are distributed with a pdf $f_{\phi_k}(\phi_k) = (1 + \cos\phi_k)/(2\pi)$, $-\pi \le \phi_k \le \pi$. Is the resulting stochastic process stationary?

3.11 A stationary random sequence $x(n)$ with mean $\mu_x = 4$ and autocovariance

$$\gamma_x(n) = \begin{cases} 4 - |n| & |n| \le 3 \\ 0 & \text{otherwise} \end{cases}$$

is applied as an input to a linear shift-invariant (LSI) system whose impulse response $h(n)$ is

$$h(n) = u(n) - u(n-4)$$

where $u(n)$ is a unit step sequence. The output of this system is another random sequence $y(n)$. Determine (*a*) the mean sequence $\mu_y(n)$, (*b*) the cross-covariance $\gamma_{xy}(n_1, n_2)$ between $x(n_1)$ and $y(n_2)$, and (*c*) the autocovariance $\gamma_y(n_1, n_2)$ of the output process $y(n)$.

3.12 A causal LTI system, which is described by the difference equation

$$y(n) = \tfrac{1}{2}y(n-1) + x(n) + \tfrac{1}{3}x(n-1)$$

is driven by a zero-mean WSS process with autocorrelation $r_x(l) = 0.5^{|l|}$.

(*a*) Determine the PSD and the autocorrelation of the output sequence $y(n)$.
(*b*) Determine the cross-correlation $r_{xy}(l)$ and cross-PSD $R_{xy}(e^{j\omega})$ between the input and output signals.

3.13 A WSS process with PSD $R_x(e^{j\omega}) = 1/(1.64 + 1.6\cos\omega)$ is applied to a causal system described by the following difference equation

$$y(n) = 0.6y(n-1) + x(n) + 1.25x(n-1)$$

Compute (*a*) the PSD of the output and (*b*) the cross-PSD $R_{xy}(e^{j\omega})$ between input and output.

3.14 Determine whether the following matrices are valid correlation matrices:

$$(a)\ \mathbf{R}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad (b)\ \mathbf{R}_2 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix}$$

$$(c)\ \mathbf{R}_3 = \begin{bmatrix} 1 & 1-j \\ 1+j & 1 \end{bmatrix} \qquad (d)\ \mathbf{R}_4 = \begin{bmatrix} 1 & \frac{1}{2} & 1 \\ \frac{1}{2} & 2 & \frac{1}{2} \\ 1 & 1 & 1 \end{bmatrix}$$

3.15 Consider a normal random vector $\mathbf{x}(\zeta)$ with components that are mutually uncorrelated, that is, $\rho_{ij} = 0$. Show that (*a*) the covariance matrix $\Gamma_{\mathbf{x}}$ is diagonal and (*b*) the components of $\mathbf{x}(\zeta)$ are mutually independent.

3.16 Show that if a real, symmetric, and nonnegative definite matrix $\mathbf{R}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_M$, then the matrix $\mathbf{R}^k$ has eigenvalues $\lambda_1^k, \lambda_2^k, \ldots, \lambda_M^k$.

3.17 Prove that the trace of $\mathbf{R}$ is given by

$$\operatorname{tr} \mathbf{R} = \sum \lambda_i$$

3.18 Prove that the determinant of $\mathbf{R}$ is given by

$$\det \mathbf{R} = |\mathbf{R}| = \prod \lambda_i = |\mathbf{\Lambda}|$$

3.19 Show that the determinants of $\mathbf{R}$ and $\mathbf{\Gamma}$ are related by

$$\det \mathbf{R} = \det \mathbf{\Gamma}(1 + \boldsymbol{\mu}^H \mathbf{\Gamma} \boldsymbol{\mu})$$

3.20 Let $\mathbf{R}_{\mathbf{x}}$ be the correlation matrix of the vector $\mathbf{x} = [x(0)\ x(2)\ x(3)]^T$, where $x(n)$ is a zero-mean WSS process.

(*a*) Check whether the matrix $\mathbf{R}_{\mathbf{x}}$ is Hermitian, Toeplitz, and nonnegative definite.
(*b*) If we know the matrix $\mathbf{R}_{\mathbf{x}}$, can we determine the correlation matrix of the vector $\bar{\mathbf{x}} = [x(0)\ x(1)\ x(2)\ x(3)]^T$?

3.21 Using the nonnegativeness of $E\{[x(n+l) \pm x(n)]^2\}$, show that $r_x(0) \geq |r_x(l)|$ for all l.

3.22 Show that $r_x(l)$ is nonnegative definite, that is,

$$\sum_{l=1}^{M} \sum_{k=1}^{M} a_l r_x(l-k) a_k^* \geq 0 \qquad \forall M,\ \forall a_1, \ldots, a_M$$

3.23 Let $x(n)$ be a random process generated by the AP(1) system

$$x(n) = \alpha x(n-1) + w(n) \qquad n \geq 0 \qquad x(-1) = 0$$

where $w(n)$ is an IID$(0, \sigma_w^2)$ process.

(*a*) Determine the autocorrelation $r_x(n_1, n_2)$ function.
(*b*) Show that $r_x(n_1, n_2)$ asymptotically approaches $r_x(n_1 - n_2)$, that is, it becomes shift-invariant.

3.24 Let $\mathbf{x}$ be a random vector with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and autocorrelation $\mathbf{R}_{\mathbf{x}}$.

(*a*) Show that $\mathbf{y} = \mathbf{Q}^T \mathbf{x}$ transforms $\mathbf{x}$ to an uncorrelated component vector $\mathbf{y}$ if $\mathbf{Q}$ is the eigenmatrix of $\mathbf{R}_{\mathbf{x}}$.
(*b*) Comment on the geometric interpretation of this transformation.

3.25 The mean and the covariance of a Gaussian random vector $\mathbf{x}$ are given by, respectively,

$$\boldsymbol{\mu}_{\mathbf{x}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Gamma}_{\mathbf{x}} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

Plot the 1σ, 2σ, and 3σ concentration ellipses representing the contours of the density function in the (x_1, x_2) plane. *Hints:* The radius of an ellipse with major axis a (along x_1) and minor axis $b < a$ (along x_2) is given by

$$r^2 = \frac{a^2 b^2}{a^2 \sin^2\theta + b^2 \cos^2\theta}$$

where $0 \le \theta \le 2\pi$. Compute the 1σ ellipse specified by $a = \sqrt{\lambda_1}$ and $b = \sqrt{\lambda_2}$ and then rotate and translate each point $\mathbf{x}^{(i)} = [x_1^{(i)}\ x_2^{(i)}]^T$ using the transformation $\mathbf{w}^{(i)} = \mathbf{Q}_x \mathbf{x}^{(i)} + \boldsymbol{\mu}_x$.

3.26 Consider the process $x(n) = ax(n-1) + w(n)$, where $w(n) \sim \text{WN}(0, \sigma_w^2)$.

(*a*) Show that the $M \times M$ correlation matrix of the process is symmetric Toeplitz and is given by

$$\mathbf{R}_{\mathbf{x}} = \frac{\sigma_w^2}{1-a^2} \begin{bmatrix} 1 & a & \cdots & a^{m-1} \\ a & 1 & \cdots & a^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{m-1} & a^{m-2} & \cdots & 1 \end{bmatrix}$$

(*b*) Verify that

$$\mathbf{R}_{\mathbf{x}}^{-1} = \frac{1}{\sigma_w^2} \begin{bmatrix} 1 & -a & 0 & \cdots & 0 \\ -a & 1+a^2 & -a & \cdots & 0 \\ 0 & -a & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1+a^2 & -a \\ 0 & 0 & \cdots & -a & 1 \end{bmatrix}$$

(*c*) Show that if

$$\mathbf{L}_{\mathbf{x}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -a & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & -a & 1 \end{bmatrix}$$

then $\mathbf{L}_x^T \mathbf{R}_x \mathbf{L}_x = (1-a^2)\mathbf{I}$.

(*d*) For $\sigma_w^2 = 1$, $a = 0.95$, and $M = 8$ compute the DKLT and the DFT.

(*e*) Plot the eigenvalues of each transform in the same graph of the PSD of the process. Explain your findings.

(*f*) Plot the eigenvectors of each transform and compare the results.

(*g*) Repeat parts (*e*) and (*f*) for $M = 16$ and $M = 32$. Explain the obtained results.

(*h*) Repeat parts (*e*) to (*g*) for $a = 0.5$ and compare with the results obtained for $a = 0.95$.

3.27 Determine three different innovations representations of a zero-mean random vector $\mathbf{x}$ with correlation matrix

$$\mathbf{R}_x = \begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}$$

3.28 Verify that the eigenvalues and eigenvectors of the $M \times M$ correlation matrix of the process $x(n) = w(n) + bw(n-1)$, where $w(n) \sim \text{WN}(0, \sigma_w^2)$ are given by $\lambda_k = R_x(e^{j\omega_k})$, $q_n^{(k)} = \sin \omega_k n$, $\omega_k = \pi k/(M+1)$, where $k = 1, 2, \ldots, M$, (*a*) analytically and (*b*) numerically for $\sigma_w^2 = 1$ and $M = 8$. *Hint:* Plot the eigenvalues on the same graph with the PSD.

3.29 Consider the process $x(n) = w(n) + bw(n-1)$.

(a) Compute the DKLT for $M = 3$.

(b) Show that the variances of the DKLT coefficients are $\sigma_x^2(1+\sqrt{2}b)$, σ_x^2, and $\sigma_x^2(1-\sqrt{2}b)$.

3.30 Let $x(n)$ be a stationary random process with mean μ_x and covariance $\gamma_x(l)$. Let $\hat{\mu}_x = 1/N \sum_{n=0}^{N-1} x(n)$ be the sample mean from the observations $\{x(n)\}_{n=0}^{N-1}$.

(a) Show that the variance of $\hat{\mu}_x$ is given by

$$\text{var}(\hat{\mu}_x) = N^{-1} \sum_{l=-N}^{N} \left(1 - \frac{|l|}{N}\right) \gamma_x(l) \leq N^{-1} \sum_{l=-N}^{N} |\gamma_x(l)| \tag{P.3}$$

(b) Show that the above result (P.3) can be expressed as

$$\text{var}(\hat{\mu}_x) = \frac{\sigma_x^2}{N}[1 + \Delta_N(\rho_x)] \tag{P.4}$$

where $$\Delta_N(\rho_x) = 2\sum_{l=1}^{N} \left(1 - \frac{l}{N}\right) \rho_x(l) \qquad \rho_x(l) = \frac{\gamma_x(l)}{\sigma_x^2}$$

(c) Show that (P.3) reduces to $\text{var}(\hat{\mu}_x) = \sigma_x^2/N$ for a WN(μ_x, σ_x^2) process.

3.31 Let $x(n)$ be a stationary random process with mean μ_x, variance σ_x^2, and covariance $\gamma_x(l)$. Let

$$\hat{\sigma}_x^2 \triangleq \frac{1}{N} \sum_{n=0}^{N-1} [x(n) - \hat{\mu}_x]^2$$

be the sample variance from the observations $\{x(n)\}_{n=0}^{N-1}$.

(a) Show that the mean of $\hat{\sigma}_x^2$ is given by

$$E\{\hat{\sigma}_x^2\} = \sigma_x^2 - \text{var}(\hat{\mu}_x) = \sigma_x^2 - \frac{1}{N} \sum_{l=-N}^{N} \left(1 - \frac{|l|}{N}\right) \gamma_x(l)$$

(b) Show that the above result reduces to $E\{\hat{\sigma}^2\} = (N-1)\sigma_x^2/N$ for a WN(μ_x, σ_x^2) process.

3.32 The Cauchy distribution with mean μ is given by

$$f_x(x) = \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2} \qquad -\infty < x < \infty$$

Let $\{x_k(\zeta)\}_{i=k}^{N}$ be N IID random variables with the above distribution. Consider the mean estimator based on $\{x_k(\zeta)\}_{i=k}^{N}$

$$\hat{\mu}(\zeta) = \frac{1}{N} \sum_{k=1}^{N} x_k(\zeta)$$

Determine whether $\hat{\mu}(\zeta)$ is a consistent estimator of μ.