Digital Image Processing

Second Edition

Review Material

Rafael C. Gonzalez
Richard E. Woods

Prentice Hall

Upper Saddle River, NJ 07458

www.prenhall.com/gonzalezwoods

or

www.imageprocessingbook.com

Preface

The purpose of this brief review is to provide the reader with the necessary background to follow material in the book dealing with matrices and vectors, probability, and linear systems. The review is divided into three main sections, each dealing one of the three preceding topics. The following material is highly focused to the needs of someone needing a refresher in one of these topics. Whenever possible, we have used the same notation employed in the text.

Contents

1 A Brief Review of Matrices and Vectors

The purpose of this short document is to provide the reader with background sufficient to follow the discussions in *Digital Image Processing*, 2nd ed., by Gonzalez and Woods. The notation is the same that we use in the book.

1.1 Matrices

Introductory Definitions

We begin with the definition of a matrix. An $m \times n$ (read "m by n") *matrix*, denoted by $\mathbf{A}$, is a rectangular array of *entries* or *elements* (numbers, or symbols representing numbers) enclosed typically by square brackets. In this notation, m is the number of horizontal rows and n the number of vertical columns in the array. Sometimes m and n are referred to as the *dimensions* or *order* of the matrix, and we say that matrix $\mathbf{A}$ has dimensions m by n or is of order m by n. We use the following notation to represent an $m \times n$ matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where a_{ij} represents the (i, j)-th entry.

If $m = n$, then $\mathbf{A}$ is a *square* matrix. If $\mathbf{A}$ is square and $a_{ij} = 0$ for all $i \neq j$, and not all a_{ii} are zero, the matrix is said to be *diagonal*. In other words, a diagonal matrix is a square matrix in which all elements not on the main diagonal are zero. A diagonal matrix in which all diagonal elements are equal to 1 is called the *identity* matrix, typically denoted by $\mathbf{I}$. A matrix in which all elements are 0 is called the *zero* or *null* matrix, typically denoted by $\mathbf{0}$. The *trace* of a matrix $\mathbf{A}$ (not necessarily diagonal),

denoted tr($\mathbf{A}$), is the sum of the elements in the main diagonal of $\mathbf{A}$. Two matrices $\mathbf{A}$ and $\mathbf{B}$ are equal if and only if they have the same number of rows and columns, and $a_{ij} = b_{ij}$ for all i and j.

The *transpose* of an $m \times n$ matrix $\mathbf{A}$, denote $\mathbf{A}^T$, is an $n \times m$ matrix obtained by interchanging the rows and columns of $\mathbf{A}$. That is, the first row of $\mathbf{A}$ becomes the first column of $\mathbf{A}^T$, the second row of $\mathbf{A}$ becomes the second column of $\mathbf{A}^T$, and so on. A square matrix for which $\mathbf{A}^T = \mathbf{A}$ is said to be *symmetric*.

Any matrix $\mathbf{X}$ for which $\mathbf{XA} = \mathbf{I}$ and $\mathbf{AX} = \mathbf{I}$ is called the *inverse* of $\mathbf{A}$. Usually, the inverse of $\mathbf{A}$ is denoted $\mathbf{A}^{-1}$. Although numerous procedures exist for computing the inverse of a matrix, the procedure usually is to use a computer program for this purpose, so we will not dwell on this topic here. The interested reader can consult any book an matrix theory for extensive theoretical and practical discussions dealing with matrix inverses. A matrix that possesses an inverse in the sense just defined is called a *nonsingular* matrix.

Associated with matrix inverses is the computation of the determinant of a matrix. Although the determinant is a scalar, its definition is a little more complicated than those discussed in the previous paragraphs. Let $\mathbf{A}$ be an $m \times m$ (square) matrix. The (i, j)-*minor* of $\mathbf{A}$, denoted M_{ij}, is the determinant of the $(m-1) \times (m-1)$ matrix formed by deleting the ith row and the jth column of $\mathbf{A}$. The (i, j)-*cofactor* of $\mathbf{A}$, denoted C_{ij}, is $(-1)^{i+j} M_{ij}$. The determinant of a 1×1 matrix $[\alpha]$, denoted $\det([\alpha])$, is $\det([\alpha]) = \alpha$. Finally, we define the determinant of an $m \times m$ matrix $\mathbf{A}$ as

$$\det(\mathbf{A}) = \sum_{j=1}^{m} a_{1j} C_{1j}.$$

In other words, the determinant of a (square) matrix is the sum of the products of the elements in the first row of the matrix and the cofactors of the first row. As is true of inverses, determinants usually are obtained using a computer.

Basic Matrix Operations

Let c be a real or complex number (often called a *scalar*). The *scalar multiple* of scalar c and matrix $\mathbf{A}$, denoted $c\mathbf{A}$, is obtained by multiplying every elements of $\mathbf{A}$ by c. If $c = -1$, the scalar multiple is called the *negative* of $\mathbf{A}$.

Assuming that they have the same number of rows and columns, the *sum* of two matrices $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} + \mathbf{B}$, is the matrix obtained by adding the corresponding elements

of the two matrices. In other words, the sum is an $m \times n$ matrix whose (i, j)-th element is $a_{ij} + b_{ij}$.Similarly, the *difference* of two matrices, denoted $\mathbf{A} - \mathbf{B}$, has elements $a_{ij} - b_{ij}$.

The product, $\mathbf{AB}$, of $m \times n$ matrix $\mathbf{A}$ and $p \times q$ matrix $\mathbf{B}$, is an $m \times q$ matrix $\mathbf{C}$ whose (i, j)-th element is formed by multiplying the entries across the *i*th row of $\mathbf{A}$ times the entries down the *j*th column of $\mathbf{B}$. In other words,

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{pj}$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, q$. We see from the preceding equation that matrix multiplication is defined only if n and p are equal. Also, as will be shown shortly, the sum of products just described is equal to the so-called inner product of rows of $\mathbf{A}$ with columns of $\mathbf{B}$. Finally, we note that division of one matrix by another is not defined.

Example 1.1 Suppose that

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 3 & 2 \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 1 \end{bmatrix}.$$

Then,

$$\mathbf{C} = \begin{bmatrix} -1 & -4 & 2 \\ 5 & 12 & 14 \end{bmatrix}.$$

Later in this discussion, we will make use of matrix products in which matrix $\mathbf{B}$ has only one column. A simple example is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{bmatrix}.$$

Also of special interest are products in which matrices consist of only one row or one column, appropriately called *row* and *column matrices*, respectively. In subsequent discussions we refer to these as *row vectors* or *column vectors*, respectively, and denote them by lowercase bold letters, with the understanding that they are row or column matrices. For example, consider two column vectors $\mathbf{a}$ and $\mathbf{b}$, of dimension $m \times 1$, as follows:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

and

$$\mathbf{a}^T = [a_1, a_2, \cdots, a_m]$$

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Keeping in mind the matrix dimensions required for matrix products defined above, the product of $\mathbf{a}$ and $\mathbf{b}$ is a 1×1 matrix, given by

$$\begin{aligned} \mathbf{a}^T\mathbf{b} &= \mathbf{b}^T\mathbf{a} = a_1b_1 + a_2b_2 + \cdots + a_mb_m \\ &= \sum_{i=1}^{m} a_ib_i. \end{aligned}$$

This particular product is often called the *dot-* or *inner product* of two vectors. We have much more to say about this in the following section. □

1.2 Vectors and Vector Spaces

Vectors

As introduced in the previous section, we refer to an $m \times 1$ *column matrix* as a *column vector*. Such a vector assumes geometric meaning when we associate geometrical properties with its elements. For example, consider the familiar two-dimensional (Euclidean) space in which a point is represented by its (x, y) coordinates. These coordinates can be expressed in terms of a column vector as follows:

$$\mathbf{u} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Then, for example, point $(1, 2)$ becomes the specific vector

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Geometrically, we represent this vector as a directed line segment from the origin to point $(1, 2)$. In three-dimensional space the vector would have components (x, y, z). In m-dimensional space we run out of letters and use the same symbol with subscripts to represent the elements of a vector. That is, an m-dimensional vector is represented as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}.$$

When expressed in the form of these column matrices, arithmetic operations between vectors follow the same rules as they do for matrices. The product of a vector by scalar is obtained simply by multiplying every element of the vector by the scalar. The sum of two vectors $\mathbf{x}$ and $\mathbf{y}$ is formed by the addition of corresponding elements ($x_1 + y_1$, $x_2 + y_2$, and so on), and similarly for subtraction. Multiplication of two vectors is as defined in Example 1. Division of one vector by another is not defined.

Vector Spaces

Definition of a vector space is both intuitive and straightforward. A *vector space* is defined as a nonempty set V of entities called *vectors* and associated *scalars* that satisfy the conditions outlined in A through C below. A vector space is *real* if the scalars are real numbers; it is *complex* if the scalars are complex numbers.

Condition A: There is in V an operation called *vector addition*, denoted $\mathbf{x} + \mathbf{y}$, that satisfies:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for all vectors $\mathbf{x}$ and $\mathbf{y}$ in the space.
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ for all $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.
3. There exists in V a unique vector, called the *zero vector*, and denoted $\mathbf{0}$, such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ and $\mathbf{0} + \mathbf{x} = \mathbf{x}$ for all vectors $\mathbf{x}$.
4. For each vector $\mathbf{x}$ in V, there is a unique vector in V, called the negation of $\mathbf{x}$, and denoted $-\mathbf{x}$, such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ and $(-\mathbf{x}) + \mathbf{x} = \mathbf{0}$.

Condition B: There is in V an operation called *multiplication by a scalar* that associates with each scalar c and each vector $\mathbf{x}$ in V a unique vector called the *product of* c *and* $\mathbf{x}$, denoted by $c\mathbf{x}$ and $\mathbf{x}c$, and which satisfies:

1. $c(d\mathbf{x}) = (cd)\mathbf{x}$ for all scalars c and d, and all vectors $\mathbf{x}$.
2. $(c + d)\mathbf{x} = c\mathbf{x} + d\mathbf{x}$ for all scalars c and d, and all vectors $\mathbf{x}$.
3. $c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$ for all scalars c and all vectors $\mathbf{x}$ and $\mathbf{y}$.

Condition C: $1\mathbf{x} = \mathbf{x}$ for all vectors $\mathbf{x}$.

We are interested particularly in real vector spaces of real $m \times 1$ column matrices, with vector addition and multiplication by scalars being as defined earlier for matrices. We shall denote such spaces by $\Re^m$. Using the notation introduced previously, vectors (col-

umn matrices) in $\Re^m$ are written as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}.$$

Example 1.2 The vector space with which we are most familiar is the two-dimensional real vector space $\Re^2$, in which we make frequent use of graphical representations for operations such as vector addition, subtraction, and multiplication by a scalar. For instance, consider the two vectors

$$\mathbf{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Using the rules of matrix addition and subtraction we have

$$\mathbf{a} + \mathbf{b} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

and

$$\mathbf{a} - \mathbf{b} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Figure 1.1 shows the familiar graphical representation of these operations, as well as multiplication of vector **a** by scalar $c = -0.5$. □

Consider two real vector spaces V_0 and V such that: (1) Each element of V_0 is also an element of V (i.e., V_0 is a subset of V). (2) Operations on elements of V_0 are the same as on elements of V. Under these conditions, V_0 is said to be a *subspace* of V.

A *linear combination* of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is an expression of the form

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$$

where the α's are scalars.

A vector $\mathbf{v}$ is said to be *linearly dependent* on a set, S, of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ if and only if $\mathbf{v}$ can be written as a linear combination of these vectors. Otherwise, $\mathbf{v}$ is *linearly independent* of the set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

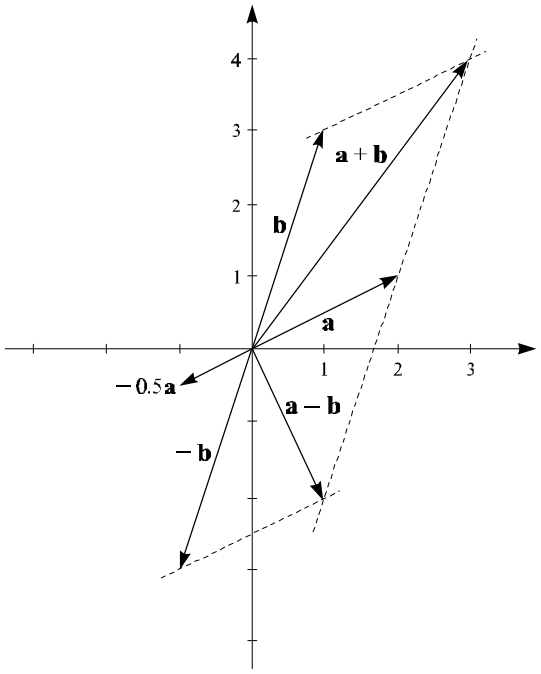

Figure 1.1

A set S of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in V is said to *span* some subspace V_0 of V if and only if S is a subset of V_0 and every vector $\mathbf{v}_0$ in V_0 is linearly dependent on the vectors in S. The set S is said to be a *spanning set* for V_0. A *basis* for a vector space V is a linearly independent spanning set for V. The number of vectors in the basis for a vector space is called the *dimension* of the vector space. If, for example, the number of vectors in the basis is n, we say that the vector space is n-dimensional.

An important aspect of the concepts just discussed lies in the representation of any vector in $\Re^m$ as a linear combination of the basis vectors. For example, any vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

in $\Re^3$ can be represented as a linear combination of the basis vectors

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Vector Norms

A *vector norm* on a vector space V is a function that assigns to each vector $\mathbf{v}$ in V a nonnegative real number, called the *norm* of $\mathbf{v}$, denoted by $\|\mathbf{v}\|$. By definition, the norm satisfies the following conditions:

1. $\|\mathbf{v}\| > 0$ for $\mathbf{v} \neq \mathbf{0}$; $\|\mathbf{0}\| = 0$,
2. $\|c\mathbf{v}\| = |c| \, \|\mathbf{v}\|$ for all scalars c and vectors $\mathbf{v}$, and
3. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

There are numerous norms that are used in practice. In our work, the norm most often used is the so-called 2-*norm*, which, for a vector $\mathbf{x}$ in real $\Re^m$, space is defined as

$$\|\mathbf{x}\| = \left[x_1^2 + x_2^2 + \cdots + x_m^2\right]^{1/2}.$$

The reader will recognize this expression as the Euclidean distance from the origin to point $\mathbf{x}$, which gives this expression the familiar name of the *Euclidean norm*. The expression also is recognized as the length of a vector $\mathbf{x}$, with origin at point $\mathbf{0}$. Based on the multiplication of two column vectors discussed earlier, we see that the norm also can be written as

$$\|\mathbf{x}\| = \left[\mathbf{x}^T\mathbf{x}\right]^{1/2}.$$

The well known Cauchy-Schwartz inequality states that

$$\left|\mathbf{x}^T\mathbf{y}\right| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|.$$

In words, this result states that the absolute value of the inner product of two vectors never exceeds the product of the norms of the vectors. This result is used in several places in the book. Another well-known result used in the book is the expression

$$\cos\,\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

where θ is the angle between vectors $\mathbf{x}$ and $\mathbf{y}$, from which we have that the inner product of two vectors can be written as

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\| \, \|\mathbf{y}\| \cos\,\theta.$$

Thus, the inner product of two vectors can be expressed as a function of the norms of the vectors and the angle between the vectors.

From the preceding results we have the definition that two vectors in $\Re^m$ are *orthogonal* if and only if their inner product is zero. Two vectors are *orthonormal* if, in addition to being orthogonal, the length of each vector is 1. From the concepts just discussed,

we see that an arbitrary vector $\mathbf{a}$ is turned into a vector $\mathbf{a}_n$ of unit length by performing the operation $\mathbf{a}_n = \mathbf{a}/\|\mathbf{a}\|$. Clearly, then, $\|\mathbf{a}_n\| = 1$. A *set* of vectors is said to be an *orthogonal set* if every two vectors in the set are orthogonal. A set of vectors is *orthonormal* if every two vectors in the set are orthonormal.

Some Important Aspects of Orthogonality

Let $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be an orthogonal or orthonormal basis in the sense defined in the previous section. Then, an important result in vector analysis is that any vector $\mathbf{v}$ can be represented with respect to the orthogonal basis B as

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$$

where the coefficients are given by

$$\begin{aligned} \alpha_i &= \frac{\mathbf{v}^T\mathbf{v}_i}{\mathbf{v}_i^T\mathbf{v}_i} \\ &= \frac{\mathbf{v}^T\mathbf{v}_i}{\|\mathbf{v}_i\|^2}. \end{aligned}$$

The key importance of this result is that, if we represent a vector as a linear combination of orthogonal or orthonormal basis vectors, we can determine the coefficients directly from simple inner product computations. It is possible to convert a linearly independent spanning set of vectors into an orthogonal spanning set by using the well-known Gram-Schmidt process. There are numerous programs available that implement the Gram-Schmidt and similar processes, so we will not dwell on the details here.

1.3 Eigenvalues and Eigenvectors

Properties of eigenvalues and eigenvectors are used extensively in *Digital Image Processing*, 2nd ed.. The following discussion is a brief overview of material fundamental to a clear understanding of the relevant material discussed in the book. We will limit discussion to real numbers, but the following results also are applicable to complex numbers.

Definition: The *eigenvalues* of a real matrix $\mathbf{M}$ are the real numbers λ for which there is a nonzero vector $\mathbf{e}$ such that $\mathbf{M}\mathbf{e} = \lambda\mathbf{e}$. The *eigenvectors* of $\mathbf{M}$ are the nonzero vectors $\mathbf{e}$ for which there is a real number λ such that $\mathbf{M}\mathbf{e} = \lambda\mathbf{e}$. If $\mathbf{M}\mathbf{e} = \lambda\mathbf{e}$ for $\mathbf{e} \neq \mathbf{0}$, then $\mathbf{e}$ is an eigenvector of $\mathbf{M}$ *associated with* eigenvalue λ, and vice versa. The eigenvectors and corresponding eigenvalues of $\mathbf{M}$ constitute the *eigensystem* of $\mathbf{M}$. Numerous theoretical and truly practical results in the application of matrices and vectors stem from this beautifully simple definition.

Example 1.3 Consider the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

It is easy to verify that $\mathbf{M}\mathbf{e}_1 = \lambda_1\mathbf{e}_1$ and $\mathbf{M}\mathbf{e}_2 = \lambda_2\mathbf{e}_2$ for $\lambda_1 = 1$, $\lambda_2 = 2$ and

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In other words, $\mathbf{e}_1$ is an eigenvector of $\mathbf{M}$ with associated eigenvalue λ_1, and similarly for $\mathbf{e}_2$ and λ_2. □

The following properties, which we give without proof, are essential background in the use of vectors and matrices in digital image processing. In each case, we assume a real matrix of order $m \times m$ although, as stated earlier, these results are equally applicable to complex numbers. We focus on real quantities simply because they play the dominant role in our work.

1. If $\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$, $q \leq m$, is set of *distinct* eigenvalues of $\mathbf{M}$, and $\mathbf{e}_i$ is an eigenvector of $\mathbf{M}$ with corresponding eigenvalue λ_i, $i = 1, 2, \ldots, q$, then $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_q\}$ is a *linearly independent* set of vectors. Note an important implication of this property: If an $m \times m$ matrix $\mathbf{M}$ has m distinct eigenvalues, its eigenvectors will constitute an orthogonal (orthonormal) set, which means that any m-dimensional vector can be expressed as a linear combination of the eigenvectors of $\mathbf{M}$.

2. The numbers along the main diagonal of a diagonal matrix are equal to its eigenvalues. It is not difficult to show using the definition $\mathbf{M}\mathbf{e} = \lambda\mathbf{e}$ that the eigenvectors can be written by inspection when $\mathbf{M}$ is diagonal.

3. A real, symmetric $m \times m$ matrix $\mathbf{M}$ has a set of m linearly independent eigenvectors that may be chosen to form an orthonormal set. This property is of particular importance when dealing with *covariance matrices* (e.g., see Section 11.4 and our review of probability) which are real and symmetric.

4. A corollary of Property 3 is that the eigenvalues of an $m \times m$ real symmetric matrix are real, and the associated eigenvectors may be chosen to form an orthonormal set of m vectors.

5. Suppose that $\mathbf{M}$ is a real, symmetric $m \times m$ matrix, and that we form a matrix $\mathbf{A}$ whose *rows* are the m orthonormal eigenvectors of $\mathbf{M}$. Then, the product $\mathbf{A}\mathbf{A}^T = \mathbf{I}$ because the rows of $\mathbf{A}$ are orthonormal vectors. (Recall from the discussion on matrix multiplication that the product of two matrices is formed by the inner product of the rows of one matrix with the column of the other. Since the rows of $\mathbf{A}$ and columns of $\mathbf{A}^T$ are orthonormal, their inner products are either 0 or 1). Thus, we see

that $\mathbf{A}^{-1} = \mathbf{A}^T$ when matrix $\mathbf{A}$ is formed as was just described.

6. Consider matrices $\mathbf{M}$ and $\mathbf{A}$ as defined in 5. Then, the product $\mathbf{D} = \mathbf{AMA}^{-1} = \mathbf{AMA}^T$ is a diagonal matrix whose elements along the main diagonal are the eigenvalues of $\mathbf{M}$. The eigenvectors of $\mathbf{D}$ are the same as the eigenvectors of $\mathbf{M}$.

Example 1.4 Suppose that we have a random population of vectors, denoted by $\{\mathbf{x}\}$, with covariance matrix (see the following chapter on a review of probability):

$$\mathbf{C_x} = E\{(\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^T\}$$

where E is the expected value operator and $\mathbf{m_x}$ is the mean of the population. Covariance matrices are real, square, symmetric matrices which, from Property 3, are known to have a set of orthonormal eigenvectors.

Suppose that we perform a transformation of the form $\mathbf{y} = \mathbf{Ax}$ on each vector $\mathbf{x}$, where the rows of $\mathbf{A}$ are the orthonormal eigenvectors of $\mathbf{C_x}$. The covariance matrix of the population $\{\mathbf{y}\}$ is

$$\begin{aligned}
\mathbf{C_y} &= E\{(\mathbf{y} - \mathbf{m_y})(\mathbf{y} - \mathbf{m_y})^T\} \\
&= E\{(\mathbf{Ax} - \mathbf{Am_x})(\mathbf{Ax} - \mathbf{Am_x})^T\} \\
&= E\{\mathbf{A}(\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^T\mathbf{A}^T\} \\
&= \mathbf{A}E\{(\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^T\}\mathbf{A}^T \\
&= \mathbf{AC_xA}^T
\end{aligned}$$

where $\mathbf{A}$ was factored out of the expectation operator because it is a constant matrix.

From Property 6, we know that $\mathbf{C_y} = \mathbf{AC_xA}^T$ is a diagonal matrix with the eigenvalues of $\mathbf{C_x}$ along its main diagonal. Recall that the elements along the main diagonal of a covariance matrix are the variances of the components of the vectors in the population. Similarly, the off diagonal elements are the covariances of the components of these vectors The fact that the covariance $\mathbf{C_y}$ is diagonal means that the elements of the vectors in the population $\{\mathbf{y}\}$ are uncorrelated (their covariances are 0). Thus, we see that application of the linear transformation $\mathbf{y} = \mathbf{Ax}$ involving the eigenvectors of $\mathbf{C_x}$ decorrelates the data, and the elements of $\mathbf{C_y}$ along its main diagonal give the variances of the components of the $\mathbf{y}$'s *along the eigenvectors*. Basically, what has been accomplished here is a coordinate transformation that aligns the data along the eigenvectors of the covariance matrix of the population.

The preceding concepts are illustrated in Fig. 1.2. Figure 1.2(a) shows a data population $\{\mathbf{x}\}$ in two dimensions, along with the eigenvectors of $\mathbf{C_x}$ (the black dot is the mean). The result of performing the transformation $\mathbf{y} = \mathbf{A}(\mathbf{x} - \mathbf{m_x})$ on the $\mathbf{x}$'s is shown in

Fig. 1.2(b). The fact that we subtracted the mean from the $\mathbf{x}$'s caused the $\mathbf{y}$'s to have zero mean, so the population is centered on the coordinate system of the transformed data. It is important to note that all we have done here is make the eigenvectors the new coordinate system (y_1, y_2). Because the covariance matrix of the $\mathbf{y}$'s is diagonal, this in fact also decorrelated the data. The fact that the main data spread is along $\mathbf{e}_1$ is due to the fact that the rows of the transformation matrix $\mathbf{A}$ were chosen according the order of the eigenvalues, with the first row being the eigenvector corresponding to the largest eigenvalue. □

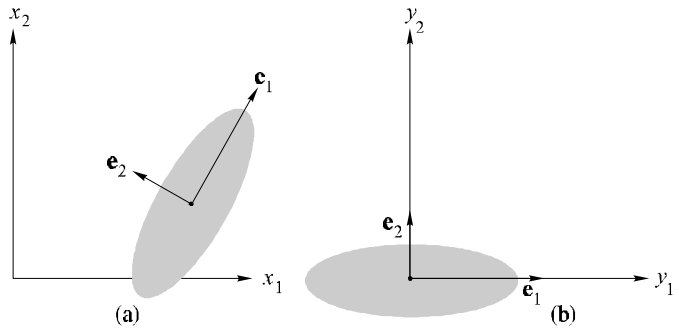

Figure 1.2

2 A Brief Review of Probability and Random Variables

The principal objective of the following material is to start with the basic principles of probability and to bring the reader to the level required to be able to follow all probability-based developments in the book.

2.1 Sets and Set Operations

Probability events are modeled as sets, so it is customary to begin a study of probability by defining sets and some simple operations among sets.

Sets

Informally, a *set* is a collection of *objects*, with each object in a set often referred to as an *element* or *member* of the set. Familiar examples include the set of all image processing books in the world, the set of prime numbers, and the set of planets circling the sun. Typically, sets are represented by uppercase letters, such as A, B, and C, and members of sets by lowercase letters, such as a, b, and c. We denote the fact that an element a belongs to set A by

$$a \in A$$

If a is not an element of A then we write

$$a \notin A.$$

A set can be specified by listing all of its elements, or by listing properties common to all elements. For example, suppose that I is the set of all integers. A set B consisting the first five nonzero integers is specified using the notation

$$B = \{1, 2, 3, 4, 5\}.$$

The set of all integers less than 10 is specified using the notation

$$C = \{c \in I \mid c < 10\}$$

which we read as "C is the set of integers such that each members of the set is less than 10." The "such that" condition is denoted by the symbol "$|$" and, as is shown in the previous two equations, the elements of the set are enclosed by curly brackets. The set with no elements is called the *empty* or *null* set, which we denote by $\emptyset$.

Two sets A and B are said to be equal if and only if they contain the same elements. Set equality is denoted by

$$A = B.$$

If the elements of two sets are not the same, we say that the sets are not equal, and denote this by

$$A \neq B.$$

If every element of B is also an element of A, we say that B is a *subset* of A:

$$B \subseteq A$$

where the equality is included to account for the case in which A and B have the same elements. If A contains more elements than B, then B is said to be a *proper subset* of A, and we use the notation

$$B \subset A.$$

Finally, we consider the concept of a *universal set*, which we denote by U and define to be the set containing all elements of interest in a given situation. For example, in an experiment of tossing a coin, there are two possible (realistic) outcomes: heads or tails. If we denote heads by H and tails by T, the universal set in this case is $\{H, T\}$. Similarly, the universal set for the experiment of throwing a single die has six possible outcomes, which normally are denoted by the face value of the die, so in this case $U = \{1, 2, 3, 4, 5, 6\}$. For obvious reasons, the universal set is frequently called the *sample space*, which we denote by S. It then follows that, for any set A, we assume that $\emptyset \subseteq A \subseteq S$, and for any element a, $a \in S$ and $a \notin \emptyset$.

Some Basic Set Operations

The operations on sets associated with basic probability theory are straightforward. The *union* of two sets A and B, denoted by

$$A \cup B$$

is the set of elements that are either in A or in B, or in both. In other words,

$$A \cup B = \{z \mid z \in A \text{ or } z \in B\}.$$

Similarly, the *intersection* of sets A and B, denoted by

$$A \cap B$$

is the set of elements common to both A and B; that is,

$$A \cap B = \{z \mid z \in A \text{ and } z \in B\}.$$

Two sets having no elements in common are said to be *disjoint* or *mutually exclusive*, in which case

$$A \cap B = \emptyset.$$

The *complement* of set A is defined as

$$A^c = \{z \mid z \notin A\}.$$

Clearly, $(A^c)^c = A$. Sometimes the complement of A is denoted as $\overline{A}$.

The difference of two sets A and B, denoted $A - B$, is the set of elements that belong to A, but not to B. In other words,

$$A - B = \{z \mid z \in A,\ z \notin B\}.$$

It is easily verified that $(A - B) = A \cap B^c$.

The union operation is applicable to multiple sets. For example the union of sets $A_1, A_2, \ldots A_n$ is the set of points that belong to at least one of these sets. Similar comments apply to the intersection of multiple sets.

Table 2.1 summarizes without proof several important relationships between sets. Proofs for these relationships are found in most books dealing with elementary set theory.

Table 2.1

Some Important Set Relationships
$S^c = \emptyset;\ \emptyset^c = S;$
$A \cup A^c = S;\ A \cap A^c = \emptyset$
$A \cup \emptyset = A;\ A \cap \emptyset = \emptyset;\ S \cup \emptyset = S;\ S \cap \emptyset = \emptyset$
$A \cup A = A;\ A \cap A = A;\ A \cup S = S;\ A \cap S = A$
$A \cup B = B \cup A;\ A \cap B = B \cap A$
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
$(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$
$(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$

It often is quite useful to represent sets and sets operations in a so-called *Venn diagram*, in which S is represented as a rectangle, sets are represented as areas (typically circles), and points are associated with elements. The following example shows various uses of Venn diagrams.

Example 2.1 Figure 2.1 shows various examples of Venn diagrams. The shaded areas are the result (sets of points) of the operations indicated in the figure. The diagrams in

the top row are self explanatory. The diagrams in the bottom row are used to prove the validity of the expression

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) - A \cap B \cap C$$

which is used in the proof of some probability relationships.□

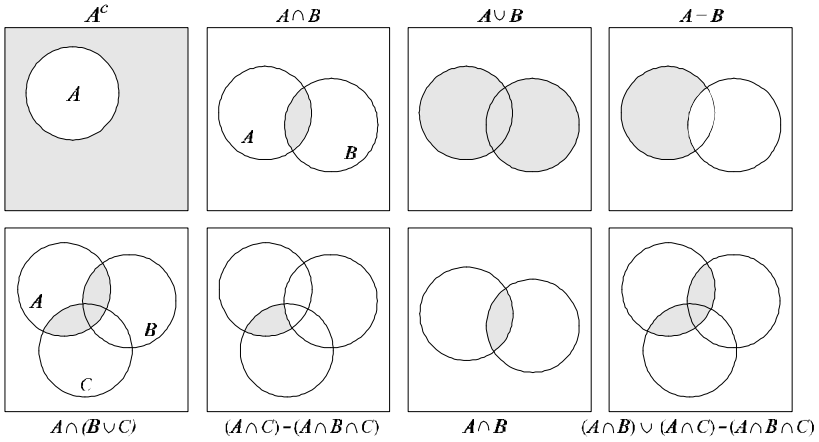

Figure 2.1

2.2 Relative Frequency and Probability

A *random experiment* is an experiment in which it is not possible to predict the outcome. Perhaps the best known random experiment is the tossing of a coin. Assuming that the coin is not biased, we are used to the concept that, on average, half the tosses will produce heads (H) and the others will produce tails (T). This is intuitive and we do not question it. In fact, few of us have taken the time to verify that this is true. If we did, we would make use of the concept of *relative frequency*. Let n denote the total number of tosses, n_H the number of heads that turn up, and n_T the number of tails. Clearly,

$$n_H + n_T = n.$$

Dividing both sides by n gives us

$$\frac{n_H}{n} + \frac{n_T}{n} = 1.$$

The term n_H/n is called the *relative frequency* of the event we have denoted by H, and similarly for n_T/n. If we performed the tossing experiment a large number of times, we would find that each of these relative frequencies tends toward a stable, limiting value. We call this value the *probability* of the event, and denoted it by $P(\text{event})$. In the current discussion the probabilities of interest are $P(H)$ and $P(T)$. We know in this case that $P(H) = P(T) = 1/2$. Note that the event of an experiment need not signify

a single outcome. For example, in the tossing experiment we could let D denote the event "heads or tails," (note that the event is now a *set*) and the event E, "neither heads nor tails." Then, $P(D) = 1$ and $P(E) = 0$.

The first important property of P is that, for an event A,

$$0 \leq P(A) \leq 1.$$

That is, the probability of an event is a positive number bounded by 0 and 1. For the certain event, S,

$$P(S) = 1.$$

Here the certain event means that the outcome is from the universal or sample set, S. Similarly, we have that for the impossible event, S^c

$$P(S^c) = 0.$$

This is the probability of an event being outside the sample set. In the example given at the end of the previous paragraph, $S = D$ and $S^c = E$.

Consider a case with the possibilities that events A or B or both or *neither* can occur. The *event* that either events A or B or both have occurred is simply the union of A and B (recall from two paragraphs back that events can be *sets*). Earlier, we denoted the union of two sets by $A \cup B$. One often finds the equivalent notation $A + B$ used interchangeably in discussions on probability. Similarly, the event that both A and B occurred is given by the intersection of A and B, which we denoted earlier by $A \cap B$. The equivalent notation AB is used much more frequently to denote the occurrence of both events in an experiment.

Suppose that we conduct our experiment n times. Let n_1 be the number of times that only event A occurs; n_2 the number of times that B occurs; n_3 the number of times that AB occurs; and n_4 the number of times that neither A nor B occur. Clearly, $n_1 + n_2 + n_3 + n_4 = n$. Using these numbers we obtain the following relative frequencies:

$$\begin{aligned} \frac{n_A}{n} &= \frac{n_1 + n_3}{n} \\ \frac{n_B}{n} &= \frac{n_2 + n_3}{n} \\ \frac{n_{AB}}{n} &= \frac{n_3}{n} \end{aligned}$$

and

$$\begin{aligned} \frac{n_{A \cup B}}{n} &= \frac{n_1 + n_2 + n3}{n} \\ &= \frac{(n_1 + n_3) + (n_2 + n_3) - n_3}{n} \\ &= \frac{n_A}{n} + \frac{n_B}{n} - \frac{n_{AB}}{n}. \end{aligned}$$

Using the previous definition of probability based on relative frequencies we have the important result

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

If A and B are mutually exclusive it follows that the set AB is empty and, consequently, $P(AB) = 0$.

The relative frequency of event A occurring, given that event B has occurred, is given by

$$\begin{aligned} \frac{n_{A/B}}{n} &= \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n}} \\ &= \frac{n_3}{n_2 + n_3}. \end{aligned}$$

This *conditional probability* is denoted by $P(A/B)$, where we note the use of the symbol "$/$"to denote conditional occurrence. It is common terminology to refer to $P(A/B)$ as "the probability of A given B." Similarly, the relative frequency of B occurring, given that A has occurred is

$$\begin{aligned} \frac{n_{B/A}}{n} &= \frac{\frac{n_{AB}}{n}}{\frac{n_A}{n}} \\ &= \frac{n_3}{n_1 + n_3}. \end{aligned}$$

We call this relative frequency "the probability of B given A," and denote it by $P(B/A)$. A little manipulation of the preceding results yields the following important relationships

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

and

$$P(AB) = P(A)P(B/A) = P(B)P(A/B).$$

The second expression may be written as

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

which is known as *Bayes' theorem*, so named after the 18th century mathematician Thomas Bayes.

Example 2.2 Suppose that we want to extend the expression

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

to three variables, A, B, and C. Recalling that AB is the same as $A \cap B$, we replace B by $B \cup C$ in the preceding equation to obtain

$$P(A \cup B \cup C) = P(A) + P(B \cup C) - P(A \cap [B \cup C]).$$

The second term in the right can be written as

$$P(B \cup C) = P(B) + P(C) - P(BC).$$

From Table 2.1, we know that $A \cap [B \cup C] = (A \cap B) \cup (A \cap C)$, so

$$\begin{aligned} P(A \cap [B \cup C]) &= P([A \cap B] \cup [A \cap C]) \\ &= P(AB \cup AC) \\ &= P(AB) + P(AC) - P(ABC). \end{aligned}$$

Collecting terms gives us the final result

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

Proceeding in a similar fashion gives

$$P(ABC) = P(A)P(B/A)P(C/AB).$$

The preceding approach can be used to generalize these expressions to N events. □

If A and B are *statistically independent*, then $P(B/A) = P(B)$ and it follows that

$$\begin{aligned} P(A/B) &= P(A) \\ P(B/A) &= P(B) \end{aligned}$$

and

$$P(AB) = P(A)P(B).$$

It was stated earlier that if sets (events) A and B are mutually exclusive, then $A \cap B = \emptyset$ from which it follows that $P(AB) = P(A \cap B) = 0$. As was just shown, the two sets are statistically independent if $P(AB) = P(A)P(B)$, which we assume to be nonzero in general. Thus, we conclude that for two events to be statistically independent, they *cannot be* mutually exclusive.

For three events A, B, and C to be independent, it must be true that

$$\begin{aligned} P(AB) &= P(A)P(B) \\ P(AC) &= P(A)P(C) \\ P(BC) &= P(B)P(C) \end{aligned}$$

and

$$P(ABC) = P(A)P(B)P(C).$$

In general, for N events to be statistically independent, it must be true that, for all combinations $1 \le i \le j \le k \le \cdots \le N$

$$\begin{aligned} P(A_iA_j) &= P(A_i)P(A_j) \\ P(A_iA_jA_k) &= P(A_i)P(A_j)P(A_k) \\ &\vdots \\ P(A_1A_2 \cdots A_N) &= P(A_1)P(A_2) \cdots P(A_N). \end{aligned}$$

Example 2.3 (a) An experiment consists of throwing a single die twice. The probability

of any of the six faces, 1 through 6, coming up in either experiment is 1/6. Suppose that we want to find the probability that a 2 comes up, followed by a 4. These two events are statistically independent (the second event does not depend on the outcome of the first). Thus, letting A represent a 2 and B a 4,

$$P(AB) = P(A)P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

We would have arrived at the same result by defining "2 followed by 4" to be a single event, say C. The sample set of all possible outcomes of two throws of a die is 36. Then, $P(C) = 1/36$.

(b) Consider now an experiment in which we draw *one* card from a standard card deck of 52 cards. Let A denote the event that a king is drawn, B denote the event that a queen or jack is drawn, and C the event that a diamond-face card is drawn. A brief review of the previous discussion on relative frequencies would show that

$$\begin{aligned} P(A) &= \frac{4}{52}, \\ P(B) &= \frac{8}{52}, \end{aligned}$$

and

$$P(C) = \frac{13}{52}.$$

Furthermore,

$$P(AC) = P(A \cap C) = P(A)P(C) = \frac{1}{52}$$

and

$$P(BC) = P(B \cap C) = P(B)P(C) = \frac{2}{52}.$$

Events A and B are mutually exclusive (we are drawing only one card, so it would be impossible to draw a king and a queen or jack simultaneously). Thus, it follows from the preceding discussion that $P(AB) = P(A \cap B) = 0$ [and also that $P(AB) \neq P(A)P(B)$].

(c) As a final experiment, consider the deck of 52 cards again, and let A_1, A_2, A_3, and A_4 represent the events of drawing an ace in each of four successive draws. If we replace the card drawn before drawing the next card, then the events are statistically independent and it follows that

$$\begin{aligned} P(A_1A_2A_3A_4) &= P(A_1)P(A_2)P(A_3)P(A_4) \\ &= \left[\frac{4}{52}\right]^4 \approx 3.5 \times 10^{-5}. \end{aligned}$$

Suppose now that we do not replace the cards that are drawn. The events then are no

longer statistically independent. With reference to the results in Example 2.2, we write

$$\begin{aligned} P(A_1A_2A_3A_4) &= P(A_1)P(A_2A_3A_4/A_1) \\ &= P(A_1)P(A_2/A_1)P(A_3A_4/A_1A_2) \\ &= P(A_1)P(A_2/A_1)P(A_3/A_1A_2)P(A_4/A_1A_2A_3) \\ &= \frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}\cdot\frac{1}{49} \approx 3.7\times 10^{-6}. \end{aligned}$$

Thus we see that not replacing the drawn card reduced our chances of drawing fours successive aces by a factor of close to 10. This significant difference is perhaps larger than might be expected from intuition. □

2.3 Random Variables

Random variables often are a source of confusion when first encountered. This need not be so, as the concept of a random variable is in principle quite simple. A *random variable*, x, is a real-valued function *defined* on the events of the sample space, S. In words, for each event in S, there is a real number that is the corresponding value of the random variable. Viewed yet another way, a random variable maps each event in S onto the real line. That is it. A simple, straightforward definition.

Part of the confusion often found in connection with random variables is the fact that they are *functions*. The notation also is partly responsible for the problem. In other words, although typically the notation used to denote a random variable is as we have shown it here, x, or some other appropriate variable, to be strictly formal, a random variable should be written as a function $x(\cdot)$ where the argument is a specific event being considered. However, this is seldom done, and, in our experience, trying to be formal by using function notation complicates the issue more than the clarity it introduces. Thus, we will opt for the less formal notation, with the warning that it must be keep clearly in mind that random variables are functions.

Example 2.4 Consider again the experiment of drawing a single card from a standard deck of 52 cards. Suppose that we define the following events. A: a heart; B: a spade; C: a club; and D: a diamond, so that $S = \{A, B, C, D\}$. A random variable is easily defined by letting $x = 1$ represent event A, $x = 2$ represent event B, and so on.

As a second illustration, consider the experiment of throwing a single die and observing the value of the up-face. We can define a random variable as the numerical outcome of the experiment (i.e., 1 through 6), but there are many other possibilities. For example, a binary random variable could be defined simply by letting $x = 0$ represent the event

that the outcome of throw is an even number and $x = 1$ otherwise.

Note the important fact in the examples just given that the probability of the events have not changed; all a random variable does is map events onto the real line. □

Thus far we have been concerned with random variables whose values are discrete. To handle *continuous random variables* we need some additional tools. In the discrete case, the probabilities of events are numbers between 0 and 1. When dealing with continuous quantities (which are not denumerable) we can no longer talk about the "probability of an event" because that probability is zero. This is not as unfamiliar as it may seem. For example, given a continuous function we know that the area of the function between two limits a and b is the integral from a to b of the function. However, the area *at a point* is zero because the integral from,say, a to a is zero. We are dealing with the same concept in the case of continuous random variables.

Thus, instead of talking about the probability of a specific value, we talk about the probability that the value of the random variable lies in a specified range. In particular, we are interested in the probability that the random variable is less than or equal to (or, similarly, greater than or equal to) a specified constant a. We write this as

$$F(a) = P(x \leq a).$$

If this function is given for all values of a (i.e., $-\infty < a < \infty$), then the values of random variable x have been defined. Function F is called the *cumulative probability distribution function* or simply the *cumulative distribution function* (cdf). The shortened term *distribution function* also is used.

It is important to point out that the notation we have used makes no distinction between a random variable and the values it assumes. If confusion is likely to arise, we can use more formal notation in which we let capital letters denote the random variable and lowercase letters denote its values. For example, the cdf using this notation would be written as $F_X(x) = P(X \leq x)$. When confusion is not likely, the cdf often is written simply as $F(x)$. This notation will be used in the following discussion when speaking generally about the cdf of a random variable.

Due to the fact that it is a probability, the cdf has the following properties:

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. $0 \leq F(x) \leq 1$
4. $F(x_1) \leq F(x_2) \quad \text{if} \quad x_1 < x_2$

5. $P(x_1 < x \leq x_2) = F(x_2) - F(x_1)$
6. $F(x^+) = F(x)$,

where $x^+ = x + \varepsilon$, with ε being a positive, infinitesimally small number.

The *probability density function* (pdf) of random variable x is defined as the derivative of the cdf

$$p(x) = \frac{dF(x)}{dx}.$$

The term *density function* is commonly used also. The pdf satisfies the following properties:

1. $p(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} p(x)dx = 1$
3. $F(x) = \int_{-\infty}^{x} p(\alpha)d\alpha$, where α is a dummy variable
4. $P(x_1 < x \leq x_2) = \int_{x_1}^{x_2} p(x)dx$.

We point out that the preceding concepts are applicable to discrete random variables. In the discrete case, we have a finite number of events and we talk about probabilities, rather than probability density functions. Also, we replace the integrals by summations and sometimes subscript the random variables. For example, in the case of a discrete variable with N possible values we would denote the probabilities by $P(x_i), i = 1, 2, \ldots, N$. In Section 3.3 of the book we used the notation $p(r_k), k = 0, 1, \ldots, L-1$, to denote the histogram of an image with L possible gray levels, $r_k, k = 0, 1, \ldots, L-1$, where $p(r_k)$ is the probability of the kth gray level (random event) occurring. Clearly, the discrete random variables in this case are gray levels. It generally is clear from the context whether one is working with continuous or discrete random variables, and whether the use of subscripting is necessary for clarity. Also, uppercase letters (e.g., P) are frequently used to distinguish between probabilities and probability density functions (e.g., p) when they are used together in the same discussion.

We will have much more to say about probability density functions in the following sections. Before leaving this section, however, we point out that if a random variable x is transformed by a monotonic transformation function $T(x)$ to produce a new random variable y, the probability density function of y can be obtained from knowledge of $T(x)$ and the probability density function of x, as follows:

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

where the subscripts on the p's are used to denote the fact that they are different functions, and the vertical bars signify the absolute value. Recall that a function $T(x)$ is monotonically increasing if $T(x_1) < T(x_2)$ for $x_1 < x_2$, and monotonically decreasing if $T(x_1) > T(x_2)$ for $x_1 < x_2$. The preceding equation is valid if $T(x)$ is an increasing or decreasing monotonic function.

2.4 Expected Value and Moments

The *expected value* of a function $g(x)$ of a continuos random variable is defined as

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)dx.$$

If the random variable is discrete the definition becomes

$$E[g(x)] = \sum_{i=1}^{N} g(x_i)P(x_i).$$

The expected value is one of the operations used most frequently when working with random variables. For example, the expected value of random variable x is obtained by letting $g(x) = x$:

$$E[x] = \overline{x} = m = \int_{-\infty}^{\infty} xp(x)dx$$

when x is continuos and

$$E[x] = \overline{x} = m = \sum_{i=1}^{N} x_i P(x_i)$$

when x is discrete. The expected value of x is equal to its *average* (or *mean*) value, hence the use of the equivalent notation $\overline{x}$ and m.

The *variance* of a random variable, denoted by σ^2, is obtained by letting $g(x) = x^2$ which gives

$$\sigma^2 = E[x^2] = \int_{-\infty}^{\infty} x^2 p(x)dx$$

for continuous random variables and

$$\sigma^2 = E[x^2] = \sum_{i=1}^{N} x_i^2 P(x_i)$$

for discrete variables. Of particular importance is the variance of random variables that have been normalized by subtracting their mean. In this case, the variance is

$$\sigma^2 = E[(x-m)^2] = \int_{-\infty}^{\infty} (x-m)^2 p(x)dx$$

and

$$\sigma^2 = E[(x-m)^2] = \sum_{i=1}^{N} (x_i - m)^2 P(x_i)$$

for continuous and discrete random variables, respectively. The square root of the vari-

ance is called the *standard deviation*, and is denoted by σ.

We can continue along this line of thought and define the nth *central moment* of a continuous random variable by letting $g(x) = (x - m)^n$:

$$\mu_n = E[(x - m)^n] = \int_{-\infty}^{\infty} (x - m)^n p(x) dx$$

and

$$\mu_n = E[(x - m)^n] = \sum_{i=1}^{N} (x_i - m)^n P(x_i)$$

for discrete variables, where we assume that $n \geq 0$. Clearly, $\mu_0 = 1$, $\mu_1 = 0$, and $\mu_2 = \sigma^2$. The term *central* when referring to moments indicates that the mean of the random variables has been subtracted out. The moments defined above in which the mean is not subtracted out sometimes are called *moments about the origin*.

In image processing moments are used for a variety of purposes, including histogram processing, segmentation, and description. In general, moments are used to characterize the probability density function of a random variable. For example, the second, third, and fourth central moments are intimately related to the shape of the probability density function of a random variable. The second central moment (the centralized variance) is a measure of spread of values of a random variable about its mean value, the third central moment is a measure of skewness (bias to the left or right) of the values of x about the mean value, and the fourth moment is a relative measure of flatness (e.g., see Section 11.3.3). In general, knowing all the moments of a density specifies that density.

Example 2.5 Consider an experiment consisting of repeatedly firing a rifle at a target, and suppose that we wish to characterize the behavior of bullet impacts on the target in terms of whether we are shooting high or low.. We divide the target into an upper and lower region by passing a horizontal line through the bull's-eye. The events of interest are the vertical distances from the center of an impact hole to the horizontal line just described. Distances above the line are considered positive and distances below the line are considered negative. The distance is zero when a bullet hits the line.

In this case, we define a random variable directly as the value of the distances in our sample set. Computing the mean of the random variable will tell us whether, on average, we are shooting high or low. If the mean is zero, we know that the average of our shots are on the line. However, the mean does not tell us how far our shots deviated from the horizontal. The variance (or standard deviation) will give us an idea of the spread of the shots. A small variance indicates a tight grouping (with respect to the mean, and in the vertical position); a large variance indicates the opposite. Finally, a third moment

of zero would tell us that the spread of the shots is symmetric about the mean value, a positive third moment would indicate a high bias, and a negative third moment would tell us that we are shooting low more than we are shooting high with respect to the mean location. □

2.5 The Gaussian Probability Density Function

Because of its importance, we will focus in this tutorial on the Gaussian probability density function to illustrate many of the preceding concepts, and also as the basis for generalization to more than one random variable in Sections 2.6 and 2.7. The reader is referred to Section 5.2.2 of the book for examples of other density functions.

A random variable is called *Gaussian* if it has a probability density of the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/\sigma^2}$$

where m and σ are as defined in the previous section. The term *normal* also is used to refer to the Gaussian density. A plot and properties of this density function are given in Section 5.2.2 of the book.

From the discussion in Section 2.3 of this review, the cumulative distribution function corresponding to the Gaussian density is

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} p(x)dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(x-m)^2/\sigma^2} dx. \end{aligned}$$

which, as before, we interpret as the probability that the random variable lies between minus infinite and an arbitrary value x. This integral has no known closed-form solution, and it must be solved by numerical or other approximation methods. Extensive tables exist for the Gaussian cdf.

2.6 Several Random Variables

In Example 2.5 we used a single random variable to describe the behavior of rifle shots with respect to a horizontal line passing through the bull's-eye in the target. Although this is useful information, it certainly leaves a lot to be desired in terms of telling us how well we are shooting with respect to the center of the target. In order to do this we need *two* random variables that will map our events onto the xy-plane. It is not difficult to see how if we wanted to describe events in 3-D space we would need three random

variables. In general, we consider in this section the case of n random variables, which we denote by $x_1, x_2, \ldots, x_n$ (the use of n here is not related to our use of the same symbol to denote the nth moment of a random variable).

It is convenient to use vector notation when dealing with several random variables. Thus, we represent a vector random variable $\mathbf{x}$ as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Then, for example, the cumulative distribution function introduced earlier becomes

$$\begin{aligned} F(\mathbf{a}) &= F(a_1, a_2, \ldots, a_n) \\ &= P\{x_1 \leq a_1, x_2 \leq a_2, \ldots, x_n \leq a_n\} \end{aligned}$$

when using vectors. As before, when confusion is not likely, the cdf of a random variable vector often is written simply as $F(\mathbf{x})$. This notation will be used in the following discussion when speaking generally about the cdf of a random variable vector.

As in the single variable case, the probability density function of a random variable vector is defined in terms of derivatives of the cdf; that is,

$$\begin{aligned} p(\mathbf{x}) &= p(x_1, x_2, \ldots, x_n) \\ &= \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}. \end{aligned}$$

An example of a multivariable density will follow shortly. The expected value of a function of $\mathbf{x}$ is defined basically as before:

$$\begin{aligned} E[g(\mathbf{x})] &= E[g(x_1, x_2, \ldots, x_n)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n. \end{aligned}$$

Cases dealing with expectation operations involving *pairs* of elements of $\mathbf{x}$ are particularly important. For example, the *joint moment* (about the origin) of *order* kq between variables x_i and x_j is

$$\eta_{kq}(i, j) = E[x_i^k x_j^q] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i^k x_j^q p(x_i, x_j) dx_i dx_j.$$

When working with any two random variables (any two elements of $\mathbf{x}$) it is common practice to simplify the notation by using x and y to denote the random variables. In

this case the joint moment just defined becomes

$$\eta_{kq} = E[x^k y^q] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^q p(x,y)dxdy.$$

It is easy to see that η_{k0} is the kth moment of x and η_{0q} is the qth moment of y, as discussed in Section 2.4.

The moment $\eta_{11} = E[xy]$ is called the *correlation* of x and y. As discussed in Chapters 4 and 12 of the book, correlation is an important concept in image processing. In fact, it is important in most areas of signal processing, where typically it is given a special symbol, such as R_{xy}:

$$R_{xy} = \eta_{11} = E[xy] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)dxdy.$$

If the condition

$$R_{xy} = E[x]E[y]$$

holds, then the two random variables are said to be *uncorrelated*. From the discussion in Section 2.2, we know that if x and y are statistically independent, then $p(x,y) = p(x)p(y)$, in which case we write

$$R_{xy} = \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy = E[x]E[y].$$

Thus, we see that if two random variables are statistically independent then they are also uncorrelated. The converse of this statement is not true in general.

The *joint central moment* of *order* kq involving random variables x and y is defined as

$$\begin{aligned}\mu_{kq} &= E[(x-m_x)^k(y-m_y)^q] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-m_x)^k(y-m_y)^q p(x,y)dxdy\end{aligned}$$

where $m_x = E[x]$ and $m_y = E[y]$ are the means of x and y, as defined earlier. We note that

$$\mu_{20} = E[(x-m_x)^2]$$

and

$$\mu_{02} = E[(y-m_y)^2]$$

are the variances of x and y, respectively. The moment μ_{11}

$$\begin{aligned}\mu_{11} &= E[(x-m_x)(y-m_y)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-m_x)(y-m_y)p(x,y)dxdy\end{aligned}$$

is called the *covariance* of x and y. As in the case of correlation, the covariance is an important concept, usually given a special symbol such as C_{xy}. By direct expansion of the terms inside the expected value brackets, and recalling the $m_x = E[x]$ and $m_y = E[y]$, it is straightforward to show that

$$\begin{aligned} C_{xy} &= E[xy] - m_y E[x] - m_x E[y] + m_x my \\ &= E[xy] - E[x]E[y] \\ &= R_{xy} - E[x]E[y]. \end{aligned}$$

From our discussion on correlation, we see that the covariance is zero if the random variables are *either* uncorrelated *or* statistically independent. This is an important result worth remembering.

If we divide the covariance by the square root of the product of the variances we obtain

$$\begin{aligned} \gamma &= \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} \\ &= \frac{C_{xy}}{\sigma_x \sigma_y} \\ &= E\left[\frac{(x-m_x)}{\sigma_x}\frac{(y-m_y)}{\sigma_y}\right]. \end{aligned}$$

The quantity γ is called the *correlation coefficient* of random variables x and y. It can be shown that the correlation coefficient is in the range $-1 \leq \gamma \leq 1$ (see Problem 12.5). As discussed in Section 12.2.1, the correlation coefficient is used in image processing for matching.

2.7 The Multivariate Gaussian Density

As an illustration of a probability density function of more than one random variable, we consider the multivariate Gaussian probability density function, defined as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}[(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})]}$$

where n is the dimensionality (number of components) of the random vector $\mathbf{x}$, $\mathbf{C}$ is the covariance matrix (to be defined below), $|\mathbf{C}|$ is the determinant of matrix $\mathbf{C}$, $\mathbf{m}$ is the mean vector (also to be defined below) and T indicates transposition (see the review of matrices and vectors).

The mean vector is defined as

$$\mathbf{m} = E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix}$$

and the covariance matrix is defined as

$$\mathbf{C} = E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T].$$

The element of $\mathbf{C}$ are the covariances of the elements of $\mathbf{x}$, such that

$$c_{ij} = C_{x_i x_j} = E[(x_i - m_i)(x_j - m_j)]$$

where, for example, x_i is the *i*th component of $\mathbf{x}$ and m_i is the *i*th component of $\mathbf{m}$.

Covariance matrices are real and symmetric (see the review of matrices and vectors). The elements along the main diagonal of $\mathbf{C}$ are the variances of the elements $\mathbf{x}$, such that $c_{ii} = \sigma_{x_i}^2$. When all the elements of $\mathbf{x}$ are uncorrelated or statistically independent, $c_{ij} = 0$, and the covariance matrix becomes a diagonal matrix. If all the variances are equal, then the covariance matrix becomes proportional to the identity matrix, with the constant of proportionality being the variance of the elements of $\mathbf{x}$.

Example 2.6 Consider the following bivariate $(n = 2)$ Gaussian probability density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}|^{1/2}} e^{-\frac{1}{2}[(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x}-\mathbf{m})]}$$

with

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

and

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

where, because $\mathbf{C}$ is known to be symmetric, $c_{12} = c_{21}$. A schematic diagram of this density is shown in Fig. 2.2(a). Figure 2.2(b) is a horizontal slice of Fig. 2.2(a). From our review of vectors and matrices, we know that the main directions of data spread are in the directions of the eigenvectors of $\mathbf{C}$. Furthermore, if the variables are uncorrelated or statistically independent, the covariance matrix will be diagonal and the eigenvectors will be in the same direction as the coordinate axes x_1 and x_2 (and the ellipse shown in Fig. 2.2(b) would be oriented along the x_1- and x_2-axis). If, in addition, the variances along the main diagonal are equal, the density would be symmetrical in all directions (in the form of a bell) and Fig. 2.2(b) would be a circle. Note in Figs. 2.2(a) and (b) that the density is centered at the mean values (m_1, m_2). □

2.8 Linear Transformations of Random Vectors

As discussed in Section 1.3, a linear transformation of a vector $\mathbf{x}$ to produce a vector $\mathbf{y}$ is of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}.$$

Of particular importance in our work is the case when the rows of $\mathbf{A}$ are the eigenvectors of the covariance matrix. Because $\mathbf{C}$ is real and symmetric, we know from the discussion in Section 1.3 that it is always possible to find n orthonormal eigenvectors from which to form $\mathbf{A}$. The implications of this are discussed in considerable detail in Section 1.3, which we recommend should be reviewed as a conclusion to the present discussion.

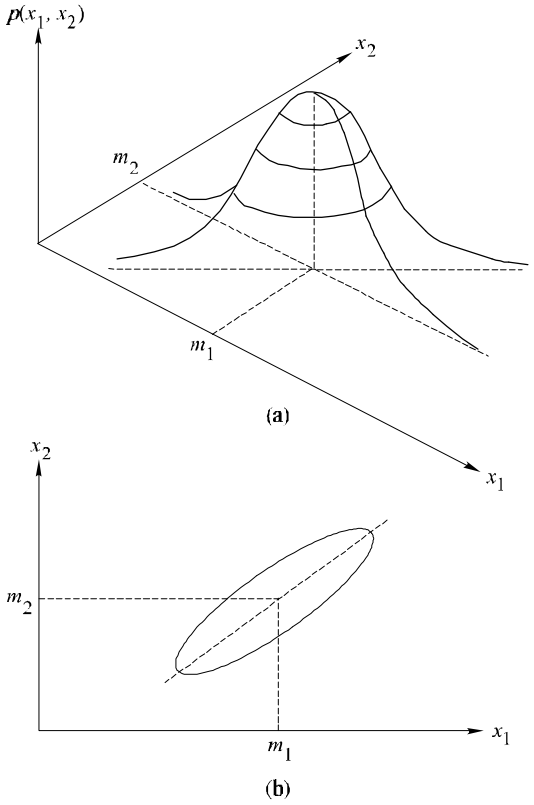

Figure 2.2