Chapter 11

Matrix Algebra and Random Matrices

11.1. BACKGROUND TO MULTIVARIATE STATISTICS

11.1.1. Contrasts Between Multivariate and Univariate Statistics

Much of the material in the first 10 chapters of this book has pertained to analysis of univariate or one-dimensional data. That is, the analysis methods presented were oriented primarily toward scalar data values and their distributions. However, in many practical situations data sets are composed of vector observations. In such cases each data record consists of simultaneous values for multiple quantities. Such data sets are known as *multivariate*. Examples of multivariate atmospheric data include simultaneous observations of multiple variables at one location, or an atmospheric field as represented by a set of grid-point values at a particular time.

Univariate methods can be, and are, applied to individual scalar elements of multivariate data observations. The distinguishing attribute of multivariate methods is that both the joint behavior of the multiple simultaneous values, as well as the variations of the individual data elements, are considered. The remaining chapters of this book present introductions to some of the multivariate methods that are used most commonly with atmospheric data. These include approaches to data reduction and structural simplification, characterization and summarization of multiple dependencies, predictions of subsets of the variables from the remaining ones, and grouping and classification of the multivariate observations.

Multivariate methods are more difficult to understand and implement than univariate methods. Notationally, they require use of matrix algebra to make the presentation and mathematical manipulations tractable. The elements of matrix algebra that are necessary to understand the subsequent material are presented briefly in Section 11.3.

The complexities of multivariate data and the methods that have been devised to deal with them dictate that all but the very simplest multivariate analyses will be implemented using a computer. Enough detail is included here for readers comfortable with numerical methods to be able to implement the analyses themselves. However, many readers will use statistical software for this purpose, and the material in this portion of this book should help to understand what these computer programs are doing, and why.

11.1.2. Organization of Data and Basic Notation

In conventional univariate statistics, each datum or observation is a single number, or scalar. In multivariate statistics each datum is a collection of simultaneous observations of $K \geq 2$ scalar values. For both notational and computational convenience, these multivariate observations are arranged in an ordered

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00011-0

list known as a *vector*, with a boldface single symbol being used to represent the entire collection, for example,

$$\boldsymbol{x}^{\mathrm{T}} = \left[x_1, x_2, x_3, \cdots, x_K\right]. \tag{11.1}$$

The superscript "T" on the left-hand side has a specific meaning that will be explained in Section 11.3, but for now we can safely ignore it. Because the K individual values are arranged horizontally, Equation 11.1 is called a *row vector*, and each of the positions within it corresponds to one of the K scalars whose simultaneous relationships will be considered. It can be convenient to visualize (for $K=2$ or 3) or imagine (for higher dimensions) a data vector geometrically, as a point in a K-dimensional space, or as an arrow whose tip position is defined by the listed scalars and whose base is at the origin. Depending on the nature of the data, this abstract geometric space may correspond to a phase- or state-space (see Section 8.1.2), or some subset of the dimensions (a *subspace*) of such a space.

A univariate data set consists of a collection of n scalar observations x_i, $i=1, \ldots, n$. Similarly, a multivariate data set consists of a collection of n data vectors $\boldsymbol{x}_i$, $i=1, \ldots, n$. Again for both notational and computational convenience this collection of data vectors can be arranged into a rectangular array of numbers having n rows, each corresponding to one multivariate observation, and with each of the K columns containing all n observations of one of the variables. This arrangement of the $n \times K$ numbers in the multivariate data set is called a *data matrix*,

$$[X] = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \boldsymbol{x}_3^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ x_{3,1} & x_{3,2} & \cdots & x_{3,K} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix}. \tag{11.2}$$

Here n row-vector observations of the form shown in Equation 11.1 have been stacked vertically, or subjected to *row binding*, to yield a rectangular array, called a *matrix*, with n rows and K columns. An equally valid view is that the K univariate data sets in each column have been subjected to *column binding* to produce $[X]$. Conventionally, the first of the two subscripts of the scalar elements of a matrix denotes the row number, and the second indicates the column number so, for example, $x_{3,2}$ is the third of the n observations of the second of the K variables. In this book matrices, such as $[X]$, will be denoted using square brackets, as a pictorial reminder that the symbol within represents a rectangular array.

The data matrix $[X]$ in Equation 11.2 corresponds exactly to a conventional data table or spreadsheet display, in which each column pertains to one of the variables considered, and each row represents one of the n observations or cases. Its contents can also be visualized or imagined geometrically within an abstract K-dimensional space, with each of the n rows defining a single point. The simplest example is a data matrix for bivariate data, which has n rows and $K=2$ columns. The pair of numbers in each of the rows locates a point on the Cartesian plane. The collection of these n points on the plane defines a scatterplot of the bivariate data.

11.1.3. Multivariate Extensions of Common Univariate Statistics

Just as the data vector in Equation 11.1 is the multivariate extension of a scalar datum, multivariate sample statistics can be expressed using the notation of vectors and matrices. The most common of these

is the multivariate sample mean, which is just a vector of the K individual scalar sample means (Equation 3.2), arranged in the same order as the elements of the underlying data vectors,

$$\overline{\boldsymbol{x}}^{\mathrm{T}} = \left[\frac{1}{n}\sum_{i=1}^{n} x_{i,1},\ \frac{1}{n}\sum_{i=1}^{n} x_{i,2},\ \cdots,\ \frac{1}{n}\sum_{i=1}^{n} x_{i,K} \right] = [\overline{x}_1,\ \overline{x}_2,\ \cdots,\ \overline{x}_K]. \tag{11.3}$$

As before, the boldface symbol on the left-hand side of Equation 11.3 indicates a vector quantity, and the double-subscripted variables in the first equality are indexed according to the same convention as in Equation 11.2.

The multivariate extensions of the sample standard deviation (Equation 3.6), or (much more commonly, its square) the sample variance, are a little more complicated because all pairwise relationships among the K variables need to be considered. In particular, the multivariate extension of the sample variance is the collection of covariances between all possible pairs of the K variables,

$$\mathrm{Cov}(x_k, x_\ell) = s_{k,\ell} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{i,k} - \overline{x}_k\right)\left(x_{i,\ell} - \overline{x}_\ell\right), \tag{11.4}$$

which is equivalent to the numerator of Equation 3.22. If the two variables are the same, that is, if $k = \ell$, then Equation 11.4 defines the sample variance, ${s_k}^2 = s_{k,k}$, or the square of Equation 3.6.

Although the notation $s_{k,k}$ for the sample variance of the kth variable may seem a little strange at first, it is conventional in multivariate statistics, and is also convenient from the standpoint of arranging the covariances calculated according to Equation 11.4 into a square array called the *sample covariance matrix*,

$$[S] = \begin{bmatrix} s_{1,1} & s_{1,2} & s_{1,3} & \cdots & s_{1,K} \\ s_{2,1} & s_{2,2} & s_{2,3} & \cdots & s_{2,K} \\ s_{3,1} & s_{3,2} & s_{3,3} & \cdots & s_{3,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{K,1} & s_{K,2} & s_{K,3} & \cdots & s_{K,K} \end{bmatrix}. \tag{11.5}$$

That is, the covariance $s_{k,\ell}$ is displayed in the kth row and ℓth column of the covariance matrix. The sample covariance matrix, or sample *variance-covariance matrix*, is directly analogous to the sample (Pearson) correlation matrix (see Figure 3.29), with the relationship between corresponding elements of the two matrices being given by Equation 3.28, that is, $r_{k,l} = s_{k,l}/(s_{k,k}s_{l,l})^{1/2}$. The K covariances $s_{k,k}$ in the diagonal positions between the upper-left and lower-right corners of the sample covariance matrix are simply the K sample variances. The remaining, off-diagonal, elements are covariances among unlike variables, and the values below and to the left of the diagonal positions duplicate the values above and to the right.

The variance-covariance matrix is also known as the *dispersion matrix*, because it describes how the underlying data are dispersed around their (vector) mean in the K-dimensional space defined by the K variables. The diagonal elements are the individual variances, which index the degree to which the data are spread out in directions parallel to the K coordinate axes for this space, and the covariances in the off-diagonal positions describe the extent to which the cloud of data points is oriented at angles to these axes. The matrix $[S]$ is the sample estimate of the population dispersion matrix $[\Sigma]$, which appears in the probability density function for the multivariate normal distribution (Equation 12.1).

11.2. MULTIVARIATE DISTANCE

It was pointed out in the previous section that a data vector can be regarded as a point in the K-dimensional geometric space whose coordinate axes correspond to the K variables being simultaneously represented. Many multivariate statistical approaches are based on, and/or can be interpreted in terms of, distances within this K-dimensional space. Any number of distance measures can be defined (see Section 16.1.2), but two of these are of particular importance.

11.2.1. Euclidean Distance

The conventional *Euclidean distance* is perhaps the easiest and most intuitive distance measure, because it corresponds to our ordinary experience in the three-dimensional world. Euclidean distance is easiest to visualize in two dimensions, where it can easily be seen as a consequence of the Pythagorean theorem, as illustrated in Figure 11.1. Here two points, $\boldsymbol{x}$ and $\boldsymbol{y}$, located by the dots, define the hypotenuse of a right triangle whose other two legs are parallel to the two data axes. The Euclidean distance $||\boldsymbol{y}-\boldsymbol{x}||=||\boldsymbol{x}-\boldsymbol{y}||$ is obtained by taking the square root of the sum of the squared lengths of the other two sides.

Euclidean distance generalizes directly to $K \geq 3$ dimensions even though the corresponding geometric space may be difficult or impossible to imagine. In particular,

$$\|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{\sum_{k=1}^{K} (x_k - y_k)^2}. \tag{11.6}$$

Distance between a point $\boldsymbol{x}$ and the origin can also be calculated using Equation 11.6 by substituting a vector of K zeros (which locates the origin in the corresponding K-dimensional space) for the vector $\boldsymbol{y}$.

It is often mathematically convenient to work in terms of squared distances. No information is lost in so doing, because distance ordinarily is regarded as necessarily nonnegative, so that squared distance is a monotonic and invertible transformation of ordinary dimensional distance (e.g., Equation 11.6). In

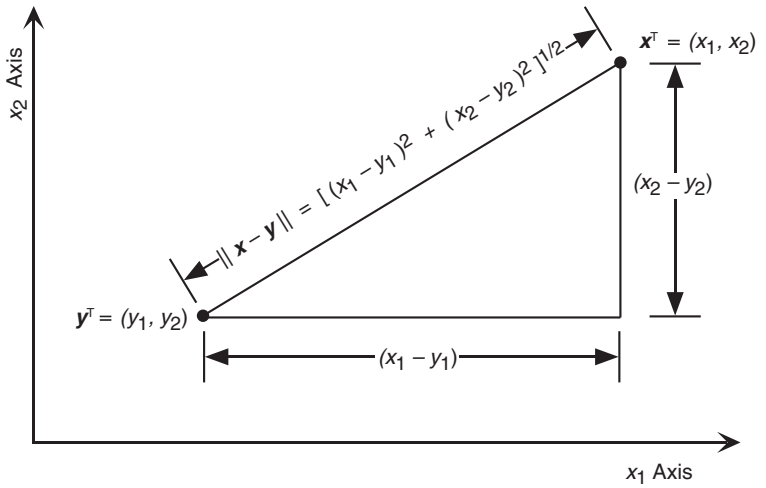

FIGURE 11.1 Illustration of the Euclidean distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $K=2$ dimensions using the Pythagorean theorem.

addition, the square-root operation is avoided. Points $\boldsymbol{x}$ at a constant squared distance $C^2 = ||\boldsymbol{x} - \boldsymbol{y}||^2$ from a fixed point $\boldsymbol{y}$ define a circle on the plane centered at $\boldsymbol{y}$ with radius C for $K = 2$ dimensions, a sphere in a volume with radius C for $K = 3$ dimensions, and a hypersphere with radius C within a K-dimensional hypervolume for $K > 3$ dimensions.

11.2.2. Mahalanobis (Statistical) Distance

Euclidean distance treats the separation of pairs of points in a K-dimensional space equally, regardless of their relative orientation. However, it will be very useful to interpret distances between points in terms of statistical dissimilarity or unusualness, and in this sense point separations in some directions are more unusual than others. The context for unusualness is established by a (K-dimensional, joint) probability distribution for the data points, which may be characterized using the scatter of a finite sample, or using a parametric probability density function.

Figure 11.2 illustrates the issues in $K = 2$ dimensions. Figure 11.2a shows a statistical context established by the scatter of points $\boldsymbol{x}^{\mathrm{T}} = (x_1, x_2)$. The distribution is centered at the origin, and the standard deviation of x_1 is approximately three times that of x_2, that is, $s_1 \approx 3s_2$. The orientation of the point cloud along one of the axes reflects the fact that the two variables x_1 and x_2 are essentially uncorrelated (the points in fact have been drawn from a bivariate Gaussian distribution with $\rho = 0$, see Section 4.4.2). Because of this difference in dispersion, a given Euclidean distance between a pair of points in the horizontal is less unusual than is the same distance in the vertical, relative to this data scatter. Although point A is closer to the center of the distribution according to Euclidean distance, it is more unusual than point B in the context established by the point cloud, and so is statistically further from the origin.

Because the points in Figure 11.2a are uncorrelated, a distance measure that reflects unusualness in the context of the data scatter can be defined simply as

$$D^2 = \frac{(x_1 - \bar{x}_1)^2}{s_{1,1}} + \frac{(x_2 - \bar{x}_2)^2}{s_{2,2}}, \tag{11.7}$$

which is a special case of the *Mahalanobis distance* between the point $\boldsymbol{x}^{\mathrm{T}} = (x_1, x_2)$ and the origin (because the two sample means are zero) when variations in the $K = 2$ dimensions are uncorrelated. For convenience Equation 11.7 is expressed as a squared distance, and it is equivalent to the ordinary squared Euclidean distance after the transformation that divides each element of the data vector by its respective standard deviation (recall that, e.g., $s_{1,1}$ is the sample variance of x_1). Another interpretation of

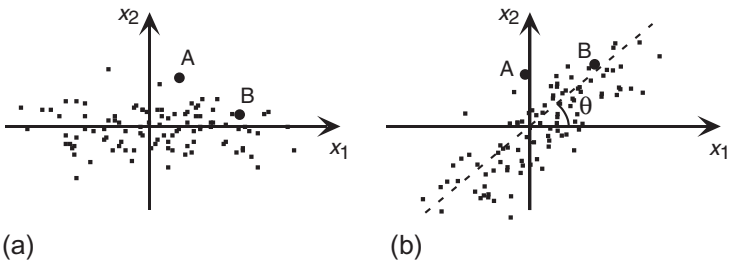

FIGURE 11.2 Distance in the context of data scatters centered at the origin. (a) The standard deviation of x_1 is approximately three times larger than the standard deviation of x_2. Point A is closer to the origin in terms of Euclidean distance, but point B is less unusual relative to the data scatter, and so is closer in statistical distance. (b) The same points rotated through an angle $\theta = 40$ degrees.

Equation 11.7 is as the sum of the two squared standardized anomalies, or z-scores (Equation 3.27). In either case, the importance ascribed to a distance along one of the axes is inversely proportional to the data scatter, or uncertainty, in that direction. Consequently point A is further from the origin than point B in Figure 11.2a when measured according to the Mahalanobis distance.

For a fixed Mahalanobis distance D^2, Equation 11.7 defines an ellipse of constant statistical distance on the plane, and that ellipse is also a circle if $s_{1,1}=s_{2,2}$. Generalizing Equation 11.7 to three dimensions by adding a third term for x_3, the set of points at a fixed distance D^2 constitute an ellipsoid that will be spherical if all three variances are equal, blimp-like if two variances are nearly equal but smaller than the third, and disk-like if two variances are nearly equal and larger than the third.

In general the variables within a multivariate data vector $\boldsymbol{x}$ will not be uncorrelated, and these correlations must also be accounted for when defining distances in terms of a data scatter or probability density. Figure 11.2b illustrates the situation in two dimensions, in which the points from Figure 11.2a have been rotated around the origin through an angle $\theta=40$ degrees, which results in the two variables being relatively strongly positively correlated. Again point B is closer to the origin in a statistical sense, although in order to calculate the actual Mahalanobis distances in terms of the variables x_1 and x_2 it would be necessary to use an equation of the form

$$D^2 = a_{1,1}(x_1-\bar{x}_1)^2 + 2a_{1,2}(x_1-\bar{x}_1)(x_2-\bar{x}_2) + a_{2,2}(x_2-\bar{x}_2)^2. \tag{11.8}$$

Analogous expressions of this kind for the Mahalanobis distance in K dimensions would involve $K(K+1)/2$ terms. Even in only two dimensions the coefficients $a_{1,1}$, $a_{1,2}$, and $a_{2,2}$ are fairly complicated functions of the rotation angle θ and the three covariances $s_{1,1}$, $s_{1,2}$, and $s_{2,2}$. For example,

$$\begin{aligned} a_{1,1} = &\frac{\cos^2(\theta)}{s_{1,1}\cos^2(\theta) - 2s_{1,2}\sin(\theta)\cos(\theta) + s_{2,2}\sin^2(\theta)} \\ &+ \frac{\sin^2(\theta)}{s_{2,2}\cos^2(\theta) - 2s_{1,2}\sin(\theta)\cos(\theta) + s_{1,1}\sin^2(\theta)}. \end{aligned} \tag{11.9}$$

Do not study this equation at all closely. It is here to help convince you, if that is even required, that conventional scalar notation is hopelessly impractical for expressing the mathematical ideas necessary to multivariate statistics. Matrix notation and matrix algebra, which will be reviewed in the next section, are practical necessities for taking the development further. Section 11.4 will resume the statistical development using matrix algebra notation, including revisiting the Mahalanobis distance in Section 11.4.4.

11.3. MATRIX ALGEBRA REVIEW

The mathematical mechanics of dealing simultaneously with multiple variables and their mutual correlations are greatly simplified by use of matrix notation, and a set of computational rules called *matrix algebra* or *linear algebra*. The notation for vectors and matrices was briefly introduced in Section 11.1.2. Matrix algebra is the toolkit used to mathematically manipulate these notational objects. A brief review of this subject, sufficient for the multivariate techniques described in the following chapters, is presented in this section. More complete introductions are readily available elsewhere (e.g., Golub and van Loan, 1996; Strang, 1988).

11.3.1. Vectors

The vector is a fundamental component of matrix algebra notation. It is essentially nothing more than an ordered list of scalar variables, or ordinary numbers, that are called the elements of the vector. The number of elements, also called the vector's dimension, will depend on the situation at hand. A familiar meteorological example is the two-dimensional horizontal wind vector, whose two elements are the eastward wind speed u, and the northward wind speed v.

Vectors already have been introduced in Equation 11.1, and as previously noted will be indicated using boldface type. A vector with only $K = 1$ element is just an ordinary number, or scalar. Unless otherwise indicated, vectors will be regarded as being *column vectors*, which means that their elements are arranged vertically. For example, the column vector $\boldsymbol{x}$ would consist of the elements $x_1, x_2, x_3, \ldots, x_K$, arranged as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix}. \tag{11.10}$$

These same elements can be arranged horizontally, as in Equation 11.1, which is a row vector. Column vectors are transformed to row vectors, and vice versa, through an operation called *transposing* the vector. The transpose operation is denoted by the superscript "T," so that we can write the vector $\boldsymbol{x}$ in Equation 11.10 as the row vector $\boldsymbol{x}^{\mathrm{T}}$ in Equation 11.1, which is pronounced "$\boldsymbol{x}$-transpose." The transpose of a column vector is useful for notational consistency within certain matrix operations. It is also useful for typographical purposes, as it allows a vector to be written on a horizontal line of text.

Addition of two or more vectors with the same dimension is straightforward. *Vector addition* is accomplished by adding the corresponding elements of the two vectors, for example

$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_K \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \\ \vdots \\ x_K + y_K \end{bmatrix}. \tag{11.11}$$

Subtraction is accomplished analogously. This operation reduces to ordinary scalar addition or subtraction when the two vectors have dimension $K = 1$. Addition and subtraction of vectors with different dimensions are not defined.

Multiplying a vector by a scalar results in a new vector whose elements are simply the corresponding elements of the original vector multiplied by that scalar. For example, multiplying the vector $\boldsymbol{x}$ in Equation 11.10 by a scalar constant c yields

$$c\boldsymbol{x} = \begin{bmatrix} c\,x_1 \\ c\,x_2 \\ c\,x_3 \\ \vdots \\ c\,x_K \end{bmatrix}. \tag{11.12}$$

Two vectors of the same dimension can be multiplied using an operation called the *dot product* or *inner product*. This operation consists of multiplying together each of the K like pairs of vector elements, and then summing these K products. That is,

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = [x_1,\ x_2,\ x_3,\ \cdots,\ x_K]\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_K \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 + \cdots + x_Ky_K \tag{11.13}$$
$$= \sum_{k=1}^{K} x_k y_k.$$

This vector multiplication has been written as the product of a row vector on the left and a column vector on the right in order to be consistent with the operation of matrix multiplication, which will be presented in Section 11.3.2. As will be seen, the dot product is in fact a special case of matrix multiplication, and (unless $K = 1$) the order of vector and matrix multiplication is important: in general the multiplications $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ and $\boldsymbol{y}\,\boldsymbol{x}^{\mathrm{T}}$ and their matrix generalizations yield entirely different results. Equation 11.13 also shows that vector multiplication can be expressed in component form using summation notation. Expanding vector and matrix operations in component form can be useful if the calculation is to be programmed for a computer, depending on the programming language.

As noted previously, a vector can be visualized as a point in K-dimensional space. The Euclidean length of a vector in that space is the ordinary distance between the point and the origin. Length is a scalar quantity that can be computed using the dot product, as

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \left(\sum_{k=1}^{K} x_k^2\right)^{1/2}. \tag{11.14}$$

Equation 11.14 is sometimes known as the *Euclidean norm* of the vector $\boldsymbol{x}$. Figure 11.1, with $\boldsymbol{y} = \boldsymbol{0}$ as the origin, illustrates that this length is simply an application of the Pythagorean theorem. A common use of Euclidean length is in the computation of the total horizontal wind speed from the horizontal velocity vector $\boldsymbol{v}^{\mathrm{T}} = [u, v]$, according to $v_{\mathrm{H}} = (u^2 + v^2)^{1/2}$. However, Equation 11.14 generalizes to arbitrarily high K as well.

The angle θ between two vectors is also computed using the dot product,

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|}\right). \tag{11.15}$$

This relationship implies that two vectors are perpendicular if their dot product is zero, since $\cos^{-1}(0) = 90$ degrees. Mutually perpendicular vectors are also called *orthogonal*.

The magnitude of the *projection* (or "length of the shadow") of a vector $\boldsymbol{x}$ onto a vector $\boldsymbol{y}$ is also a function of the dot product, given by

$$L_{x,y} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{y}\|}. \tag{11.16}$$

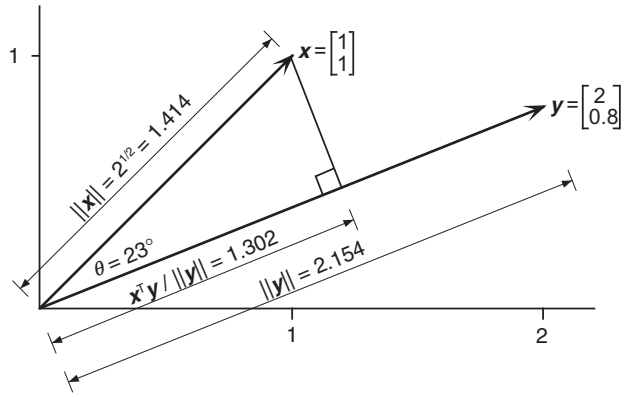

FIGURE 11.3 Illustration of the concepts of vector length (Equation 11.14), the angle between two vectors (Equation 11.15), and the projection of one vector onto another (Equation 11.16); for the two vectors $\boldsymbol{x}^{\mathrm{T}} = [1, 1]$ and $\boldsymbol{y}^{\mathrm{T}} = [2, 0.8]$.

The geometric interpretations of these three computations of length, angle, and projection are illustrated in Figure 11.3, for the vectors $\boldsymbol{x}^{\mathrm{T}} = [1, 1]$ and $\boldsymbol{y}^{\mathrm{T}} = [2, 0.8]$. The length of $\boldsymbol{x}$ is simply $||\boldsymbol{x}|| = (1^2+1^2)^{1/2} = \sqrt{2}$, and the length of $\boldsymbol{y}$ is $||\boldsymbol{y}|| = (2^2+0.8^2)^{1/2} = 2.154$. Since the dot product of the two vectors is $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = (1)(2)+(1)(0.8) = 2.8$, the angle between them is $\theta = \cos^{-1}(2.8/(2.154\sqrt{2})) = 23$ degrees, and the length of the projection of $\boldsymbol{x}$ onto $\boldsymbol{y}$ is $2.8/2.154 = 1.302$.

11.3.2. Matrices

A matrix is a two-dimensional rectangular array of numbers having I rows and J columns. The *dimension* of a matrix is specified by these numbers of rows and columns. A matrix dimension is written $(I \times J)$, and pronounced "I by J." Matrices are denoted here by uppercase letters surrounded by square brackets. Sometimes, for notational clarity, a parenthetical expression for the dimension of a matrix will be written directly below it. The *elements* of a matrix are the individual variables or numerical values occupying the particular row-and-column positions. The matrix elements are identified notationally by two subscripts; the first of these identifies the row number and the second identifies the column number. Equation 11.2 shows a $(n \times K)$ data matrix, and Equation 11.5 shows a $(K \times K)$ covariance matrix, with the subscripting convention illustrated.

A vector is a special case of a matrix, having a single row or a single column, and matrix operations are applicable also to vectors. A K-dimensional row vector is a $(1 \times K)$ matrix, and a column vector is a $(K \times 1)$ matrix. Just as a $K = 1$dimensional vector is also a scalar, so too is a (1×1) matrix.

A matrix with the same number of rows and columns, such as $[S]$ in Equation 11.5, is called a *square* matrix. The elements of a square matrix for which the subscript values $i = j$ are located on the diagonal between the upper left to the lower-right corners, and are called *diagonal* elements. Correlation matrices $[R]$ (see Figure 3.29) are square matrices having all 1's on the diagonal. A square matrix for which $a_{i,j} = a_{j,i}$ for all values of i and j is called *symmetric*. Correlation and covariance matrices are symmetric because the correlation between variable i and variable j is identical to the correlation between variable j and variable i. The *identity matrix* $[I]$, consisting of 1's on the diagonal and zeros everywhere else, is another important square, symmetric matrix,

$$[I] = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{11.17}$$

An identity matrix can be constructed for any (square) dimension. When the identity matrix appears in an equation it can be assumed to be of appropriate dimension for the relevant matrix operations to be defined. The identity matrix is a special case of a *diagonal matrix*, whose off-diagonal elements are all zeros.

The *transpose* operation is defined for any matrix, including the special case of vectors. The transpose of a matrix is obtained in general by exchanging row and column indices, not by a 90° rotation as might have been anticipated from a comparison of Equations 11.1 and 11.10. Geometrically, the transpose operation is like a reflection across the matrix diagonal, which extends downward and to the right from the upper, left-hand element. For example, the relationship between the (3×4) matrix $[B]$ and its transpose, the (4×3) matrix $[B]^{\mathrm{T}}$, is illustrated by comparing

$$\underset{(3\times4)}{[B]} = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \varepsilon & \varsigma & \eta & \theta \\ \iota & \kappa & \lambda & \mu \end{bmatrix} \tag{11.18a}$$

and

$$\underset{(4\times3)}{[B]^{\mathrm{T}}} = \begin{bmatrix} \alpha & \varepsilon & \iota \\ \beta & \varsigma & \kappa \\ \gamma & \eta & \lambda \\ \delta & \theta & \mu \end{bmatrix}. \tag{11.18b}$$

Equation 11.18 also illustrates the convention of indicating the matrix dimension parenthetically, beneath the matrix symbol. If a square matrix $[A]$ is symmetric, then $[A]^{\mathrm{T}} = [A]$.

Multiplication of a matrix by a scalar is the same as for vectors and is accomplished by multiplying each element of the matrix by the scalar,

$$c\,[D] = c \begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} = \begin{bmatrix} c\,d_{1,1} & c\,d_{1,2} \\ c\,d_{2,1} & c\,d_{2,2} \end{bmatrix}. \tag{11.19}$$

Similarly, matrix addition and subtraction are accomplished by performing these operations on the elements in corresponding row and column positions, and are defined only for matrices of identical dimension. For example, the sum of two (2×2) matrices would be computed as

$$[D] + [E] = \begin{bmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \end{bmatrix} + \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix} = \begin{bmatrix} d_{1,1} + e_{1,1} & d_{1,2} + e_{1,2} \\ d_{2,1} + e_{2,1} & d_{2,2} + e_{2,2} \end{bmatrix}. \tag{11.20}$$

Matrix multiplication is defined between two matrices if the number of columns in the left matrix is equal to the number of rows in the right matrix. Thus not only is matrix multiplication not commutative (i.e., $[A]\,[B] \neq [B]\,[A]$, in general), but multiplication of two matrices in reverse order is not even defined

unless the two have complementary row and column dimensions. The product of a matrix multiplication is another matrix, the row dimension of which is the same as the row dimension of the left matrix, and the column dimension of which is the same as the column dimension of the right matrix. That is, multiplying a $(I \times J)$ matrix $[A]$ (on the left) and a $(J \times K)$ matrix $[B]$ (on the right) yields a $(I \times K)$ matrix $[C]$. In effect, the middle dimension J is "multiplied out."

Consider the case where $I=2$, $J=3$, and $K=2$. In terms of the individual matrix elements, the matrix multiplication $[A]\,[B]=[C]$ expands to

$$\underset{(2\times 3)}{\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}} \underset{(3\times 2)}{\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{3,1} & b_{3,2} \end{bmatrix}} = \underset{(2\times 2)}{\begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix}}, \tag{11.21a}$$

where

$$[C] = \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix} = \begin{bmatrix} a_{1,1}b_{1,1}+a_{1,2}b_{2,1}+a_{1,3}b_{3,1} & a_{1,1}b_{1,2}+a_{1,2}b_{2,2}+a_{1,3}b_{1,3} \\ a_{2,1}b_{1,1}+a_{2,2}b_{2,1}+a_{2,3}b_{3,1} & a_{2,1}b_{1,2}+a_{2,2}b_{2,2}+a_{2,3}b_{3,2} \end{bmatrix}. \tag{11.21b}$$

The individual components of $[C]$ as written out in Equation 11.21b may look confusing at first exposure. In understanding matrix multiplication, it is helpful to realize that each element of the product matrix $[C]$ is simply the dot product, as defined in Equation 11.13, of one of the rows in the left matrix $[A]$ and one of the columns in the right matrix $[B]$. In particular, the number occupying the ith row and kth column of the matrix $[C]$ is exactly the dot product between the row vector comprising the ith row of $[A]$ and the column vector comprising the kth column of $[B]$. Equivalently, matrix multiplication can be written in terms of the individual matrix elements using summation notation,

$$c_{i,k} = \sum_{j=1}^{J} a_{i,j} b_{j,k}; i = 1,\ldots,I; k = 1,\ldots,K. \tag{11.22}$$

Figure 11.4 illustrates the procedure graphically, for one element of the matrix $[C]$ resulting from the multiplication $[A]\,[B]=[C]$.

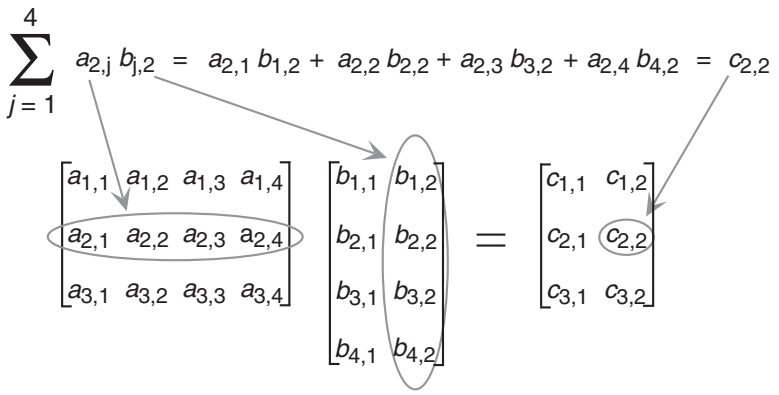

FIGURE 11.4 Graphical illustration of matrix multiplication as the dot product of the ith row of the left-hand matrix with the jth column of the right-hand matrix, yielding the element in the ith row and jth column of the matrix product.

The identity matrix (Equation 11.17) is so named because it functions as the multiplicative identity—that is, $[A]\,[I]=[A]$, and $[I]\,[A]=[A]$ regardless of the dimension of $[A]$—although in the former case $[I]$ is a square matrix with the same number of columns as $[A]$, and in the latter its dimension is the same as the number of rows in $[A]$.

The dot product, or inner product (Equation 11.13), is one application of matrix multiplication to vectors. But the rules of matrix multiplication also allow multiplication of two vectors in the opposite order, which is called the *outer product*. In contrast to the inner product, which is a $(1\times K)\times(K\times 1)$ matrix multiplication yielding a (1×1) scalar, the outer product of two vectors of the same dimension K is a $(K\times 1)\times(1\times K)$ matrix multiplication, yielding a $(K\times K)$ square matrix. For example, for $K=3$,

$$\boldsymbol{xy}^{\mathrm{T}}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}\begin{bmatrix}y_1, & y_2, & y_3\end{bmatrix}=\begin{bmatrix}x_1y_1 & x_1y_2 & x_1y_3\\x_2y_1 & x_2y_2 & x_2y_3\\x_3y_1 & x_3y_2 & x_3y_3\end{bmatrix}. \tag{11.23}$$

It is not necessary for two vectors forming an outer product to have the same dimension, because as vectors they have common ("inner") dimension 1. The outer product is sometimes known as the *dyadic product*, or *tensor product*, and the operation is sometimes indicated using a circled "x," that is, $\boldsymbol{xy}^{\mathrm{T}}=\boldsymbol{x}\otimes\boldsymbol{y}$.

The *trace* of a square matrix is simply the sum of its diagonal elements, that is,

$$\mathrm{tr}[A]=\sum_{k=1}^{K}a_{k,k}, \tag{11.24}$$

for the $(K\times K)$ matrix $[A]$. For the $(K\times K)$ identity matrix, $\mathrm{tr}[I]=K$.

The *determinant* of a square matrix is a scalar quantity defined as

$$\det[A]=|A|=\sum_{k=1}^{k}a_{1,k}\left|A_{1,k}\right|(-1)^{1+k}, \tag{11.25}$$

where $[A_{1,k}]$ is the $(K-1\times K-1)$ matrix formed by deleting the first row and kth column of $[A]$, and $a_{1,k}$ is the element in the original matrix at the intersection of the deleted row and deleted column. The absolute value notation for the matrix determinant suggests that this operation produces a scalar that is in some sense a measure of the magnitude of the matrix. The definition in Equation 11.25 is recursive, so for example computing the determinant of a $(K\times K)$ matrix requires that K determinants of reduced $(K-1\times K-1)$ matrices be calculated first, and so on, until reaching and $|\mathrm{A}|=a_{1,1}$ for $K=1$. Accordingly the process is quite tedious and is usually best left to a computer. However, in the (2×2) case,

$$\det_{(2\times 2)}[A]=\det\begin{bmatrix}a_{1,1} & a_{1,2}\\a_{2,1} & a_{2,2}\end{bmatrix}=a_{11}a_{2,2}-a_{1,2}a_{2,1}. \tag{11.26}$$

The matrix generalization of arithmetic division exists for square matrices that have a property known as *full rank*, or *nonsingularity*. This condition can be interpreted to mean that the matrix does not contain redundant information, in the sense that none of the rows can be constructed from linear combinations of the other rows. Considering each row of a nonsingular matrix as a vector, it is impossible to construct vector sums of rows multiplied by scalar constants that equal any one of the other rows. These same conditions applied to the columns also imply that the matrix is nonsingular. Nonsingular matrices have nonzero determinant.

Nonsingular square matrices are *invertible*. That a matrix $[A]$ is invertible means that another matrix $[B]$ exists such that

$$[A][B] = [B][A] = [I]. \tag{11.27}$$

It is then said that $[B]$ is the inverse of $[A]$, or $[B] = [A]^{-1}$; and that $[A]$ is the inverse of $[B]$, or $[A] = [B]^{-1}$. Loosely speaking, $[A]\,[A]^{-1}$ indicates division of the matrix $[A]$ by itself, and so yields the (matrix) identity $[I]$. Inverses of (2×2) matrices are easy to compute by hand, using

$$[A]^{-1} = \frac{1}{\det[A]} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix} = \frac{1}{a_{1,1}a_{2,2} - a_{2,1}a_{1,2}} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix}. \tag{11.28}$$

The name of this matrix is pronounced "A inverse." Explicit formulas for inverting matrices of higher dimension also exist, but quickly become very cumbersome as the dimensions get larger. Computer algorithms for inverting matrices are widely available, and as a consequence matrices with dimension higher than two or three are rarely inverted by hand. An important exception is the inverse of a diagonal matrix, which is simply another diagonal matrix whose nonzero elements are the reciprocals of the diagonal elements of the original matrix. If $[A]$ is symmetric (frequently in statistics, symmetric matrices are inverted), then $[A]^{-1}$ is also symmetric.

Table 11.1 lists some additional properties of arithmetic operations with matrices that have not been specifically mentioned in the foregoing.

TABLE 11.1 Some Elementary Properties of Arithmetic Operations With Matrices

Distributive multiplication by a scalar	$c\,([A][B]) = (c\,[A])\,[B] = [A]\,(c\,[B])$
Distributive matrix multiplication	$[A]\,([B]+[C]) = [A][B]+[A][C]$
	$([A]+[B])\,[C] = [A][C]+[B][C]$
Associative matrix multiplication	$[A]\,([B][C]) = ([A][B])\,[C]$
Inverse of a matrix product	$([A][B])^{-1} = [B]^{-1}[A]^{-1}$, $([A][B][C])^{-1} = [C]^{-1}[B]^{-1}[A]^{-1}$, etc.
Transpose of a matrix product	$([A][B])^{\mathrm{T}} = [B]^{\mathrm{T}}\,[A]^{\mathrm{T}}$, $([A][B][C])^{\mathrm{T}} = [C]^{\mathrm{T}}[B]^{\mathrm{T}}[A]^{\mathrm{T}}$, etc.
Combining matrix transpose and inverse	$([A]^{-1})^{\mathrm{T}} = ([A]^{\mathrm{T}})^{-1}$

Example 11.1. Computation of the Covariance and Correlation Matrices

The covariance matrix $[S]$ was introduced in Equation 11.5, and the correlation matrix $[R]$ was introduced in Figure 3.29 as a device for compactly representing the mutual correlations among K variables. The correlation matrix for the January 1987 data in Table A.1 (with the unit diagonal elements and the symmetry implicit) is presented in Table 3.5. The computation of the covariances in Equation 11.4 and of the correlations in Equation 3.29 can also be expressed in the notation of matrix algebra.

One way to begin the computation is with the $(n \times K)$ data matrix $[X]$ (Equation 11.2). Each row of this matrix is a vector, consisting of one observation for each of K variables. The number of these rows is the same as the sample size, n, so $[X]$ is just an ordinary data table such as Table A.1. In Table A.1 there are $K = 6$ variables (excluding the column containing the dates), each simultaneously observed on $n = 31$ occasions. An individual data element $x_{i,k}$ is the ith observation of the kth variable. For example, in Table A.1, $x_{4,6}$ would be the Canandaigua minimum temperature (19°F) observed on 4 January.

Define the $(n \times n)$ matrix [1], whose elements are all equal to 1. The $(n \times K)$ matrix of anomalies (in the meteorological sense of variables with their means subtracted), or centered data $[X']$ is then

$$[X'] = [X] - \frac{1}{n}[1][X]. \tag{11.29}$$

(Note that some authors use the prime notation to indicate matrix transpose, but the superscript "T" has been used for this purpose throughout this book, to avoid confusion.) The second term in Equation 11.29 is a $(n \times K)$ matrix containing the sample means. Each of its n rows is the same and consists of the K sample means in the same order as the corresponding variables appear in each row of $[X]$.

Multiplying $[X']$ by the transpose of itself, and dividing by $n-1$, yields the sample covariance matrix,

$$[S] = \frac{1}{n-1}[X']^{\mathrm{T}}[X']. \tag{11.30}$$

This is the same symmetric $(K \times K)$ matrix as in Equation 11.5, whose diagonal elements are the sample variances of the K variables, and whose other elements are the covariances among all possible pairs of the K variables. The operation in Equation 11.30 corresponds to the summation in the numerator of Equation 3.28.

Now define the $(K \times K)$ diagonal matrix $[D]$, whose diagonal elements are the sample standard deviations of the K variables. That is, $[D]$ consists of all zeros except for the diagonal elements, whose values are the square roots of the corresponding elements of $[S]$: $d_{k,k} = \sqrt{s_{k,k}}$, $k = 1, \ldots, K$. The correlation matrix can then be computed from the covariance matrix using

$$[R] = [D]^{-1}[S][D]^{-1}. \tag{11.31}$$

Since $[D]$ is diagonal, its inverse is the diagonal matrix whose elements are the reciprocals of the sample standard deviations on the diagonal of $[D]$. The matrix multiplication in Equation 11.31 corresponds to division by the standard deviations in Equation 3.29.

Note that the correlation matrix $[R]$ is equivalently the covariance matrix of the standardized variables (or standardized anomalies) z_k (Equation 3.27). That is, dividing the anomalies x_k' by their standard deviations $\sqrt{s_{k,k}}$ nondimensionalizes the variables and results in their having unit variance (1's on the diagonal of $[R]$) and covariances equal to their correlations. In matrix notation this can be seen by substituting Equation 11.30 into Equation 11.31 to yield

$$\begin{aligned}[R] &= \frac{1}{n-1}[D]^{-1}[X']^{\mathrm{T}}[X'][D]^{-1} \\ &= \frac{1}{n-1}[Z]^{\mathrm{T}}[Z]\end{aligned}, \tag{11.32}$$

where $[Z]$ is the $(n \times K)$ matrix whose rows are the vectors of standardized variables $\boldsymbol{z}$, analogously to the matrix $[X']$ of the anomalies. The first line of Equation 11.32 converts the matrix $[X']$ to the matrix $[Z]$ by dividing each element by its standard deviation, $d_{k,k}$. Comparing Equation 11.32 and 11.30 shows that $[R]$ is indeed the covariance matrix for the standardized variables $\boldsymbol{z}$.

It is also possible to formulate the computation of the covariance and correlation matrices in terms of outer products of vectors. Define the ith of n (column) vectors of anomalies

$$\boldsymbol{x}_i' = \boldsymbol{x}_i - \overline{\boldsymbol{x}}_i, \tag{11.33}$$

where the vector (sample) mean is the transpose of any of the rows of the matrix that is subtracted on the right-hand side of Equation 11.29 or, equivalently the transpose of Equation 11.3. Also let the corresponding standardized anomalies (the vector counterpart of Equation 3.27) be

$$\boldsymbol{z}_i = [D]^{-1}\boldsymbol{x}_i', \tag{11.34}$$

where $[D]$ is again the diagonal matrix of standard deviations. Equation 11.34 is called the *scaling transformation*, and simply indicates division of all the values in a data vector by their respective standard deviations. The covariance matrix can then be computed in a way that is notationally analogous to the usual computation of the scalar variance (Equation 3.6, squared),

$$[S] = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{x}_i'^{\mathrm{T}}, \tag{11.35}$$

and, similarly, the correlation matrix is

$$[R] = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}. \tag{11.36}$$

◇

Example 11.2. Multiple Linear Regression Expressed in Matrix Notation

The discussion of multiple linear regression in Section 7.3.1 indicated that the relevant mathematics are most easily expressed and solved using matrix algebra. In this notation, the expression for the predictand y as a function of the predictor variables x_i (Equation 7.25) becomes

$$\boldsymbol{y} = [X]\,\boldsymbol{b}, \tag{11.37a}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,K} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}. \tag{11.37b}$$

Here $\boldsymbol{y}$ is a $(n \times 1)$ matrix (i.e., a vector) of the n observations of the predictand, $[X]$ is a $(n \times K+1)$ data matrix containing the values of the predictors, and $\boldsymbol{b}^{\mathrm{T}} = [b_0, b_1, b_2, \ldots, b_K]$ is a $(K+1 \times 1)$ vector of the regression parameters. The data matrix in the regression context is similar to that in Equation 11.2, except that it has $K+1$ rather than K columns. This extra column is the leftmost column of $[X]$ in Equation 11.37 and consists entirely of 1's. Thus Equation 11.37 is a vector equation, with dimension $(n \times 1)$ on each side. It is actually n repetitions of Equation 7.25, once each for the n data records.

The normal equations (presented in Equation 7.6 for the simple case of $K=1$) are obtained by left-multiplying each side of Equation 11.37 by $[X]^{\mathrm{T}}$,

$$[X]^T \boldsymbol{y} = \left[X^T\right][X]\,\boldsymbol{b}, \tag{11.38a}$$

or

$$\begin{bmatrix} \Sigma y \\ \Sigma x_1 y \\ \Sigma x_2 y \\ \vdots \\ \Sigma x_K y \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & \Sigma x_2 & \cdots & \Sigma x_K \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 & \cdots & \Sigma x_1 x_K \\ \Sigma x_2 & \Sigma x_2 x_1 & \Sigma x_2^2 & \cdots & \Sigma x_2 x_K \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Sigma x_K & \Sigma x_K x_1 & \Sigma x_K x_2 & \cdots & \Sigma x_K^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}, \tag{11.38b}$$

where all the summations are over the n training data points. The symmetric $[X]^{\mathrm{T}}[X]$ matrix has dimension $(K+1 \times K+1)$. Each side of Equation 11.38 has dimension $(K+1 \times 1)$, and this equation actually represents $K+1$ simultaneous equations involving the $K+1$ unknown regression coefficients. Matrix algebra very commonly is used to solve sets of simultaneous linear equations such as these. One (relatively computationally inefficient) way to obtain the solution is to left-multiply both sides of Equation 11.38 by the inverse of the $[X]^{\mathrm{T}}[X]$ matrix. This operation is analogous to dividing both sides by this quantity, and yields

$$\begin{aligned} \left([X]^{\mathrm{T}}[X]\right)^{-1}[X]^{\mathrm{T}}\boldsymbol{y} &= \left([X]^{\mathrm{T}}[X]\right)^{-1}[X]^{\mathrm{T}}[X]\,\boldsymbol{b} \\ &= [X]^{-1}\left([X]^{\mathrm{T}}\right)^{-1}[X]^{\mathrm{T}}[X]\boldsymbol{b} \\ &= [X]^{-1}[I][X]\boldsymbol{b} \\ &= [I]\,\boldsymbol{b} \\ &= \boldsymbol{b} \end{aligned}, \tag{11.39}$$

which is the solution for the vector of regression parameters. The result for the inverse of a matrix product from Table 11.1 has been used in the second line. If there are no linear dependencies among the predictor variables, then the matrix $[X]^{\mathrm{T}}[X]$ is nonsingular, and its inverse will exist. Otherwise, regression software will be unable to compute Equation 11.39, and a suitable error message should be reported.

Variances and covariances for the joint sampling distribution of the $K+1$ regression parameters $\boldsymbol{b}^{\mathrm{T}}$, corresponding to Equations 7.18b and 7.19b, can also be calculated using matrix algebra. The $(K+1 \times K+1)$ covariance matrix, jointly for the intercept and the K regression coefficients, is

$$[S_{\boldsymbol{b}}] = \begin{bmatrix} s_{b_0}^2 & s_{b_0,b_1} & \cdots & s_{b_0,b_K} \\ s_{b_1,b_0} & s_{b_1}^2 & \cdots & s_{b_1,b_K} \\ s_{b_2,b_0} & s_{b_2,b_1} & \cdots & s_{b_2,b_K} \\ \vdots & \vdots & \ddots & \vdots \\ s_{b_K,b_0} & s_{b_K,b_1} & \cdots & s_{b_K}^2 \end{bmatrix} = s_e^2\left([X]^{\mathrm{T}}[X]\right)^{-1}. \tag{11.40}$$

As before, s_e^2 is the estimated residual variance,

$$s_e^2 = \frac{1}{n-K-1}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{b}^{\mathrm{T}}[X]^{\mathrm{T}}\boldsymbol{y}\right), \tag{11.41}$$

or MSE (as in Table 7.3). The diagonal elements of Equation 11.40 are the estimated variances of the sampling distributions of each of the elements of the parameter vector $\boldsymbol{b}$. The off-diagonal elements are the covariances among them, corresponding to (for covariances involving the intercept, b_0) the correlation in Equation 7.20. For sufficiently large sample sizes, the joint sampling distribution is multivariate normal (see Chapter 12) so Equation 11.40 fully defines its dispersion.

Similarly, the conditional variance of the sampling distribution of the multiple linear regression function, which is the multivariate extension of Equation 7.24, can be expressed in matrix form as

$$s_{\bar{y}|\boldsymbol{x}_0}^2 = s_e^2 \boldsymbol{x}_0^{\mathrm{T}}\left([X]^{\mathrm{T}}[X]\right)^{-1}\boldsymbol{x}_0, \tag{11.42}$$

and the prediction variance, corresponding to Equation 7.23 is

$$s_{\hat{y}|\boldsymbol{x}_0}^2 = s_e^2\left\{1 + \boldsymbol{x}_0^{\mathrm{T}}\left([X]^{\mathrm{T}}[X]\right)^{-1}\boldsymbol{x}_0\right\}. \tag{11.43}$$

Equations 11.42 and 11.43 both depend on the values of the predictors for which the regression function is evaluated, $\boldsymbol{x}_0^{\mathrm{T}} = [1, x_1, x_2, \ldots, x_K]$. ◇

A square matrix is called *orthogonal* if the vectors defined by its columns have unit lengths, and are mutually perpendicular (i.e., $\theta = 90$ degree according to Equation 11.15), and the same conditions hold for the vectors defined by its rows. In that case,

$$[A]^{\mathrm{T}} = [A]^{-1}, \tag{11.44a}$$

which implies that

$$[A][A]^T = [A]^T[A] = [I]. \tag{11.44b}$$

Orthogonal matrices are *unitary*, with this latter term encompassing also matrices that may have complex elements.

An *orthogonal transformation* is achieved by multiplying a vector by an orthogonal matrix. Considering a vector to define a point in K-dimensional space, an orthogonal transformation corresponds to a rigid rotation of the coordinate axes (and also a reflection, if the determinant is negative), resulting in a new basis (new set of coordinate axes) for the space. For example, consider $K = 2$ dimensions, and the orthogonal matrix

$$[T] = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \tag{11.45}$$

The lengths of both rows and both columns of this matrix are $\sin^2(\theta)+\cos^2(\theta)=1$ (Equation 11.14), and the angles between the two pairs of vectors are both 90 degrees (Equation 11.15), so $[T]$ is an orthogonal matrix.

Multiplication of a vector $\boldsymbol{x}$ by the transpose of this matrix corresponds to a rigid counterclockwise rotation of the coordinate axes through an angle θ. Consider the point $\boldsymbol{x}^{\mathrm{T}} = (1, 1)$ in Figure 11.5. Left-multiplying it by $[T]^{\mathrm{T}}$, with $\theta = 72$ degrees, yields the point in a new coordinate system (dashed axes)

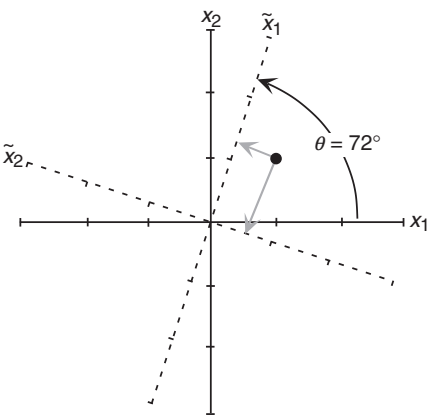

FIGURE 11.5 The point $\boldsymbol{x}^{\mathrm{T}}=(1,\ 1)$, when subjected to an orthogonal rotation of the coordinate axes through an angle of $\theta=72$ degrees, is transformed to the point $\tilde{\boldsymbol{x}}^{\mathrm{T}}=(1.26,\ -0.64)$ in the new basis (dashed coordinate axes).

$$\begin{aligned}\tilde{\boldsymbol{x}} &= \begin{bmatrix} \cos\left(72^\circ\right) & \sin\left(72^\circ\right) \\ -\sin\left(72^\circ\right) & \cos\left(72^\circ\right) \end{bmatrix} \boldsymbol{x} \\ &= \begin{bmatrix} .309 & .951 \\ -.951 & .309 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} .309 + .951 \\ -.951 + .309 \end{bmatrix} = \begin{bmatrix} 1.26 \\ -0.64 \end{bmatrix}.\end{aligned} \tag{11.46}$$

Because the rows and columns of an orthogonal matrix all have unit length, orthogonal transformations preserve length. That is, they do not compress or expand the (rotated) coordinate axes. In terms of (squared) Euclidean length (Equation 11.14),

$$\begin{aligned}\tilde{\boldsymbol{x}}^{\mathrm{T}}\tilde{\boldsymbol{x}} &= \left([T]^{\mathrm{T}}\boldsymbol{x}\right)^{\mathrm{T}}\left([T]^{\mathrm{T}}\boldsymbol{x}\right) \\ &= \boldsymbol{x}^{\mathrm{T}}[T][T]^{\mathrm{T}}\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathrm{T}}[\boldsymbol{I}]\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\end{aligned}. \tag{11.47}$$

The result for the transpose of a matrix product from Table 11.1 has been used in the second line, and Equation 11.44 has been used in the third.

11.3.3. Eigenvalues and Eigenvectors of a Square Matrix

An *eigenvalue* λ, and an *eigenvector*, $\boldsymbol{e}$ of a square matrix $[A]$ are a scalar and nonzero vector, respectively, satisfying the equation

$$[A]\,\boldsymbol{e} = \lambda\,\boldsymbol{e}, \tag{11.48a}$$

or equivalently

$$([A] - \lambda[I])\,\boldsymbol{e} = \boldsymbol{0}, \tag{11.48b}$$

where $\boldsymbol{0}$ is a vector consisting entirely of zeros. For every eigenvalue and eigenvector pair that can be found to satisfy Equation 11.48, any scalar multiple of the eigenvector, $c\boldsymbol{e}$, will also satisfy the equation together with that eigenvalue. Consequently, for definiteness it is usual to require that eigenvectors have unit length,

$$\|\boldsymbol{e}\| = 1. \tag{11.49}$$

This restriction removes the ambiguity only up to a change in sign, since if a vector $\boldsymbol{e}$ satisfies Equation 11.48 then its negative, $-\boldsymbol{e}$ will also.

If $[A]$ is nonsingular there will be K eigenvalue-eigenvector pairs λ_k and $\boldsymbol{e}_k$ with nonzero eigenvalues, where K is the number of rows and columns in $[A]$. Each eigenvector will be dimensioned $(K \times 1)$. If $[A]$ is singular at least one of its eigenvalues will be zero, with the corresponding eigenvector(s) being arbitrary. Synonymous terminology that is sometimes also used for eigenvalues and eigenvectors includes *characteristic values* and *characteristic vectors*, *latent values* and *latent vectors*, and *proper values* and *proper vectors*.

Because each eigenvector is defined to have unit length, the dot product of any eigenvector with itself is one. If, in addition, the matrix $[A]$ is symmetric, then its eigenvectors are mutually orthogonal, so that

$$\boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{e}_j = \begin{cases} 1\ , & i = j \\ 0\ , & i \neq j \end{cases}. \tag{11.50}$$

Orthogonal vectors of unit length are said to be *orthonormal*. (This terminology has nothing to do with the Gaussian or "normal" distribution.) The orthonormality property is analogous to Equation 10.66, expressing the orthogonality of the sine and cosine functions.

For many statistical applications, eigenvalues and eigenvectors are calculated for real (not containing complex or imaginary numbers) symmetric matrices, such as covariance or correlation matrices. Eigenvalues and eigenvectors of such matrices have a number of important and remarkable properties. The first of these properties is that their eigenvalues and eigenvectors are real valued. Also, as just noted, the eigenvectors of symmetric matrices are orthogonal. That is, their dot products with each other are zero, so that they are mutually perpendicular in K-dimensional space.

Often the $(K \times K)$ matrix $[E]$ is formed, the K columns of which are the eigenvectors $\boldsymbol{e}_k$. That is,

$$[E] = [\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \boldsymbol{e}_3,\ \cdots,\ \boldsymbol{e}_K]. \tag{11.51}$$

Because of the orthogonality and unit length of the eigenvectors of symmetric matrices, the matrix $[E]$ is orthogonal, having the properties expressed in Equation 11.44. The orthogonal transformation $[E]^{\mathrm{T}}\boldsymbol{x}$ defines a rigid rotation of the K-dimensional coordinate axes of $\boldsymbol{x}$, called an *eigenspace*. This space covers the same "territory" as the original coordinates, but using the different set of axes defined by the solutions to Equation 11.48. Metaphorically, and in two dimensions, the underlying landscape has not changed, but a compass would indicate that some different direction is north.

The K eigenvalue-eigenvector pairs contain the same information as the matrix $[A]$ from which they were computed, and so can be regarded as a transformation of $[A]$. This equivalence can be expressed, again for $[A]$ symmetric, as the *spectral decomposition*, or *Jordan decomposition*,

$$[A] = [E][\Lambda][E]^{\mathrm{T}} \tag{11.52a}$$

$$= [E] \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_K \end{bmatrix} [E]^{\mathrm{T}}, \tag{11.52b}$$

so that $[\Lambda]$ denotes a diagonal matrix whose nonzero elements are the K eigenvalues of $[A]$. It is illuminating to consider also the equivalent of Equation 11.52 in summation notation,

$$[A] = \sum_{k=1}^{K} \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k^{\mathrm{T}} \tag{11.53a}$$

$$= \sum_{k=1}^{K} \lambda_k [E_k]. \tag{11.53b}$$

The outer product of each eigenvector with itself in Equation 11.53a defines a matrix $[E_k]$. Equation 11.53b shows that the original matrix $[A]$ can be recovered as a weighted sum of these $[E_k]$ matrices, where the weights are the corresponding eigenvalues. Hence the spectral decomposition of a matrix is analogous to the Fourier decomposition of a function or data series (Equation 10.62a), with the eigenvalues playing the role of the Fourier amplitudes and the $[E_k]$ matrices corresponding to the cosine functions.

Other consequences of the equivalence of the information on the two sides of Equation 11.52 pertain to the eigenvalues. The first of these is

$$\mathrm{tr}[A] = \sum_{k=1}^{K} a_{k,k} = \sum_{k=1}^{K} \lambda_k = \mathrm{tr}[\Lambda]. \tag{11.54}$$

This relationship is especially important when $[A]$ is a covariance matrix, in which case its diagonal elements $a_{k,k}$ are the K variances. Equation 11.54 says the sum of these variances is equal to the sum of the eigenvalues of the covariance matrix.

The second consequence of Equation 11.52 for the eigenvalues is

$$\det[A] = \prod_{k=1}^{K} \lambda_k = \det[\Lambda], \tag{11.55}$$

which is consistent with the property that at least one of the eigenvalues of a singular matrix (having zero determinant) will be zero. A real symmetric matrix with all eigenvalues positive is called *positive definite*.

The matrix of eigenvectors $[E]$ has the property that it *diagonalizes* the original symmetric matrix $[A]$ from which the eigenvectors and eigenvalues were calculated. Left-multiplying Equation 11.52a by $[E]^{\mathrm{T}}$, right-multiplying by $[E]$, and using the orthogonality of $[E]$ yields

$$[E]^{\mathrm{T}}[A][E] = [\Lambda]. \tag{11.56}$$

That is, multiplication of $[A]$ on the left by $[E]^{\mathrm{T}}$ and on the right by $[E]$ produces the diagonal matrix of eigenvalues $[\Lambda]$.

There is also a strong connection between the eigenvalues λ_k and eigenvectors $\boldsymbol{e}_k$ of a nonsingular symmetric matrix, and the corresponding quantities λ_k^* and $\boldsymbol{e}_k^*$ of its inverse. The eigenvectors of matrix-inverse pairs are the same—that is, $\boldsymbol{e}_k^* = \boldsymbol{e}_k$ for each k—and the corresponding eigenvalues are reciprocals, $\lambda_k^* = \lambda_k^{-1}$. Therefore the eigenvector of $[A]$ associated with its largest eigenvalue is the same as the eigenvector of $[A]^{-1}$ associated with its smallest eigenvalue, and vice versa. One way to see this analytically is to subject both sides of Equation 11.52a to matrix inversion, and then use the property for the inverse of a matrix product in Table 11.1 and the orthonormality property (Equation 11.44) of an eigenvector matrix, to derive

$$\begin{aligned} ([S])^{-1} &= \left([E]^{\mathrm{T}}[\Lambda][E]\right)^{-1}, \\ &= [E]^{\mathrm{T}}[\Lambda]^{-1}[E] \end{aligned} \tag{11.57}$$

recalling that the inverse of a diagonal matrix is also diagonal with elements that are reciprocals of the original matrix.

The extraction of eigenvalue-eigenvector pairs from matrices is a computationally demanding task, particularly as the dimensionality of the problem increases. It is possible but very tedious to do the computations by hand if $K=2$, 3, or 4, using the equation

$$\det([A]-\lambda[I])=0. \tag{11.58}$$

This calculation requires first solving a Kth-order polynomial for the K eigenvalues, and then solving K sets of K simultaneous equations to obtain the eigenvectors. In general, however, widely available computer algorithms for calculating very close numerical approximations to eigenvalues and eigenvectors are used. These computations can also be done within the framework of the singular value decomposition (see Section 11.3.5).

Example 11.3. Eigenvalues and Eigenvectors of a (2×2) Symmetric Matrix and Its Inverse
The symmetric matrix

$$[A]=\begin{bmatrix}185.47 & 110.84\\ 110.84 & 77.58\end{bmatrix} \tag{11.59}$$

has as its eigenvalues $\lambda_1=254.76$ and $\lambda_2=8.29$, with corresponding eigenvectors $\boldsymbol{e}_1{}^{\mathrm{T}}=[0.848, 0.530]$ and $\boldsymbol{e}_2{}^{\mathrm{T}}=[-0.530, 0.848]$. It is easily verified that both eigenvectors are of unit length. Their dot product is zero, which indicates that the two vectors are perpendicular, or orthogonal.

The matrix of eigenvectors is therefore

$$[E]=\begin{bmatrix}0.848 & -0.530\\ 0.530 & 0.848\end{bmatrix}, \tag{11.60}$$

and the original matrix can be recovered using the eigenvalues and eigenvectors (Equations 11.52 and 11.53) as

$$[A]=\begin{bmatrix}185.47 & 110.84\\ 110.84 & 77.58\end{bmatrix}=\begin{bmatrix}.848 & -.530\\ .530 & .848\end{bmatrix}\begin{bmatrix}254.76 & 0\\ 0 & 8.29\end{bmatrix}\begin{bmatrix}.848 & .530\\ -.530 & .848\end{bmatrix} \tag{11.61a}$$

$$=254.76\begin{bmatrix}.848\\ .530\end{bmatrix}[.848 \quad .530]+8.29\begin{bmatrix}-.530\\ .848\end{bmatrix}[-.530 \quad .848] \tag{11.61b}$$

$$=254.76\begin{bmatrix}.719 & .449\\ .449 & .281\end{bmatrix}+8.29\begin{bmatrix}.281 & -.449\\ -.449 & .719\end{bmatrix} \tag{11.61c}$$

Equation 11.61a expresses the spectral decomposition of $[A]$ in the form of Equation 11.52, and Equations 11.61b and 11.61c show the same decomposition in the form of Equation 11.53.

The matrix of eigenvectors diagonalizes the original matrix $[A]$ according to

$$\begin{aligned}[E]^{\mathrm{T}}[A][E]&=\begin{bmatrix}.848 & .530\\ -.530 & .848\end{bmatrix}\begin{bmatrix}185.47 & 110.84\\ 110.84 & 77.58\end{bmatrix}\begin{bmatrix}.848 & -.530\\ .530 & .848\end{bmatrix}\\ &=\begin{bmatrix}254.0 & 0\\ 0 & 8.29\end{bmatrix}=[\Lambda].\end{aligned} \tag{11.62}$$

The sum of the eigenvalues, $254.76+8.29=263.05$, equals the sum of the diagonal elements of the original $[A]$ matrix, $185.47+77.58=263.05$.

Applying Equation 11.29 to the matrix $[A]$ in Equation 11.59 yields its inverse

$$[A]^{-1} = \begin{bmatrix} .03688 & -.05270 \\ -.05270 & .08818 \end{bmatrix}. \tag{11.63}$$

The leading (i.e., largest) eigenvalue of $[A]^{-1}$ is then $\lambda_2^* = 1/8.29 = .1206$, and its last (i.e., smallest) eigenvalue is $\lambda_1^* = 1/254.76 = .003925$. The eigenvectors are the same as those shown in the columns of Equation 11.59, although conventionally the ordering of these columns would be reversed for $[A]^{-1}$ because the first column is paired with the largest eigenvalue. ◇

11.3.4. Square Roots of a Symmetric Matrix

Consider two square matrices of the same order, $[A]$ and $[B]$. If the condition

$$[A] = [B][B]^{\mathrm{T}} \tag{11.64}$$

holds, then $[B]$ multiplied by itself yields $[A]$, so $[B]$ is said to be a "square root" of $[A]$, or $[B] = [A]^{1/2}$. Unlike the square roots of scalars, the square root of a symmetric matrix is not uniquely defined. That is, there are any number of matrices $[B]$ that can satisfy Equation 11.64, although two algorithms are used most frequently to find solutions for it.

If $[A]$ is of full rank, a lower-triangular matrix $[B]$ satisfying Equation 11.64 can be found using the *Cholesky decomposition* of $[A]$. (A *lower-triangular* matrix has zeros above and to the right of the main diagonal, i.e., $b_{i,j}=0$ for $i<j$.) Beginning with

$$b_{1,1} = \sqrt{a_{1,1}} \tag{11.65}$$

as the only nonzero element in the first row of $[B]$, the Cholesky decomposition proceeds iteratively, by calculating the nonzero elements of each of the subsequent rows, i, of $[B]$ in turn according to

$$b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{i,k} b_{j,k}}{b_{j,j}}, j = 1, \ldots, i-1 \tag{11.66a}$$

and

$$b_{i,i} = \left[a_{i,i} - \sum_{k=1}^{i-1} b_{i,k}^2 \right]^{1/2}. \tag{11.66b}$$

It is a good idea to do these calculations in double precision in order to minimize the accumulation roundoff errors that can lead to a division by zero in Equation 10.64a for large matrix dimension K, even if $[A]$ is of full rank.

Some authors (e.g., Golub and van Loan, 1996) define the matrix square root more restrictively, requiring $[A] = [B][B] = [B]^2$ for $[B]$ to be a square root of $[A]$. A symmetric square-root matrix $[B]$ will satisfy both this condition and that in Equation 11.64. Such matrices can be found using the eigenvalues and eigenvectors of $[A]$ when $[A]$ is symmetric and is computable even if $[A]$ is not of full rank. Using the spectral decomposition (Equation 11.52) for $[B]$,

$$[B] = [A]^{1/2} = [E][\Lambda]^{1/2}[E]^{\mathrm{T}}, \tag{11.67}$$

where $[E]$ is the matrix of eigenvectors for both $[A]$ and $[B]$ (i.e., they are the same vectors). The matrix $[\Lambda]$ contains the eigenvalues of $[A]$, which are the squares of the eigenvalues of $[B]$ on the diagonal of

$[\Lambda]^{1/2}$. That is, $[\Lambda]^{1/2}$ is the diagonal matrix with elements $\lambda_k^{1/2}$, where the λ_k are the eigenvalues of $[A]$. Equation 11.67 is still defined even if some of these eigenvalues are zero, so this method can be used to find a square root for a matrix that is not of full rank. Note that $[\Lambda]^{1/2}$ also conforms to both definitions of a square-root matrix, since $[\Lambda]^{1/2}\,([\Lambda]^{1/2})^{\mathrm{T}} = [\Lambda]^{1/2}\,[\Lambda]^{1/2} = ([\Lambda]^{1/2})^2 = [\Lambda]$. The square-root decomposition in Equation 11.67 is more tolerant of roundoff error than the Cholesky decomposition when the matrix dimension is large, because (computationally, as well as truly) zero eigenvalues do not produce undefined arithmetic operations.

Equation 11.67 can be extended to find the square root of a matrix inverse, $[A]^{-1/2}$, if $[A]$ is symmetric and of full rank. Because a matrix has the same eigenvectors as its inverse, so also will it have the same eigenvectors as the square root of its inverse. Accordingly,

$$[A]^{-1/2} = [E][\Lambda]^{-1/2}[E]^{\mathrm{T}}, \tag{11.68}$$

where $[\Lambda]^{-1/2}$ is the diagonal matrix with elements $\lambda_k^{-1/2}$, the reciprocals of the square roots of the eigenvalues of $[A]$. The implications of Equation 11.68 are those that would be expected, that is, $[A]^{-1/2}\,([A]^{-1/2})^{\mathrm{T}} = [A]^{-1}$, and $[A]^{-1/2}\,([A]^{1/2})^{\mathrm{T}} = [I]$.

Example 11.4. Square Roots of a Matrix and Its Inverse

The symmetric matrix $[A]$ in Equation 11.59 is of full rank, since both of its eigenvalues are positive. Therefore a lower-triangular square-root matrix $[B] = [A]^{1/2}$ can be computed using the Cholesky decomposition. Equation 11.66 yields $b_{1,1} = (a_{1,1})^{1/2} = 185.47^{1/2} = 13.619$ as the only nonzero element of the first row ($i = 1$) of $[B]$. Because $[B]$ has only one additional row, Equations 11.66 need to be applied only once each. Equation 11.66a yields $b_{2,1} = (a_{1,1} - 0)/b_{1,1} = 110.84/13.619 = 8.139$. Zero is subtracted in the numerator of Equation 11.66a for $b_{2,1}$ because there are no terms in the summation. (If $[A]$ had been a (3×3) matrix, Equation 11.66a would be applied twice for the third ($i = 3$) row: the first of these applications, for $b_{3,1}$, would again have no terms in the summation; but when calculating $b_{3,2}$ there would be one term corresponding to $k = 1$.) Finally, the calculation indicated by Equation 11.66b is $b_{2,2} = \left(a_{2,2} - b^2_{2,1}\right)^{1/2} = \left(77.58 - 8.139^2\right)^{1/2} = 3.367$. The Cholesky lower-triangular square-root matrix for $[A]$ is thus

$$[B] = [A]^{1/2} = \begin{bmatrix} 13.619 & 0 \\ 8.139 & 3.367 \end{bmatrix}, \tag{11.69}$$

which can be verified as a valid square root of $[A]$ through the matrix multiplication $[B][B]^{\mathrm{T}}$.

A symmetric square-root matrix for $[A]$ can be computed using its eigenvalues and eigenvectors from Example 11.3, and Equation 11.67:

$$\begin{aligned} [B] = [A]^{1/2} &= [E][\Lambda]^{1/2}[E]^{\mathrm{T}} \\ &= \begin{bmatrix} .848 & -.530 \\ .530 & .848 \end{bmatrix} \begin{bmatrix} \sqrt{254.76} & 0 \\ 0 & \sqrt{8.29} \end{bmatrix} \begin{bmatrix} .848 & .530 \\ -.530 & .848 \end{bmatrix}. \\ &= \begin{bmatrix} 12.286 & 5.879 \\ 5.879 & 6.554 \end{bmatrix} \end{aligned} \tag{11.70}$$

This matrix also can be verified as a valid square root of $[A]$ by calculating $[B][B]^{\mathrm{T}} = [B]^2$.

Equation 11.68 allows calculation of a square-root matrix for the inverse of $[A]$,

$$
\begin{aligned}
[A]^{-1/2} &= [E][\Lambda]^{-1/2}[E]^{\mathrm{T}} \\
&= \begin{bmatrix} .848 & -.530 \\ .530 & .848 \end{bmatrix} \begin{bmatrix} 1/\sqrt{254.76} & 0 \\ 0 & 1/\sqrt{8.29} \end{bmatrix} \begin{bmatrix} .848 & .530 \\ -.530 & .848 \end{bmatrix}. \\
&= \begin{bmatrix} .1426 & -.1279 \\ -.1279 & .2674 \end{bmatrix}
\end{aligned} \tag{11.71}
$$

This is also a symmetric matrix. The matrix product $[A]^{-1/2}\,([A]^{-1/2})^{\mathrm{T}} = [A]^{-1/2}\,[A]^{-1/2} = [A]^{-1}$. The validity of Equation 11.71 can be checked by comparing the product $[A]^{-1/2}\,[A]^{-1/2}$ with $[A]^{-1}$ as calculated using Equation 11.28 or by verifying $[A][A]^{-1/2}\,[A]^{-1/2} = [A][A]^{-1} = [I]$. ◇

11.3.5. Singular Value Decomposition (SVD)

Equation 11.52 expresses the spectral decomposition of a symmetric square matrix. This decomposition can be extended to any $(n \times m)$ rectangular matrix $[A]$ with at least as many rows as columns $(n \geq m)$ using the *singular value decomposition* (SVD),

$$
\underset{(n\times m)}{[A]} = \underset{(n\times m)}{[L]}\ \underset{(m\times m)}{[\Omega]}\ \underset{(m\times m)}{[R]^{T}}, n \geq m. \tag{11.72}
$$

The m columns of $[L]$ are called the left *singular vectors*, and the m columns of $[R]$ (not its transpose) are called the right singular vectors. (Note that, in the context of SVD, $[R]$ does not denote a correlation matrix.) Both sets of vectors are mutually orthonormal, so $[L]^{\mathrm{T}}[L] = [R]^{\mathrm{T}}[R] = [R][R]^{\mathrm{T}} = [I]$, with dimension $(m \times m)$. The matrix $[\Omega]$ is diagonal, with nonnegative diagonal elements that are called the *singular values* of $[A]$. Equation 11.72 is sometimes called the "thin" SVD, in contrast to an equivalent expression in which the dimension of $[L]$ is $(n \times n)$, and the dimension of $[\Omega]$ is $(n \times m)$, but with the last $n - m$ rows of $[\Omega]$ containing all zeros so that the last $n - m$ columns of $[L]$ are arbitrary.

If $[A]$ is square and symmetric, then Equation 11.72 reduces to Equation 11.52, with $[L] = [R] = [E]$, and $[\Omega] = [\Lambda]$. It is therefore possible to compute eigenvalues and eigenvectors for symmetric matrices using an SVD algorithm from a package of matrix-algebra computer routines, which are widely available (e.g., Press et al., 1986). Analogously to Equation 11.53 for the spectral decomposition of a symmetric square matrix, Equation 11.72 can be expressed as a summation of weighted outer products of the left and right singular vectors,

$$
[A] = \sum_{i=1}^{m} \omega_i \boldsymbol{\ell}_i \boldsymbol{r}_i^{\mathrm{T}}. \tag{11.73}
$$

Even if $[A]$ is not symmetric, there is a connection between the SVD and the eigenvalues and eigenvectors of both $[A]^{\mathrm{T}}[A]$ and $[A][A]^{\mathrm{T}}$, both of which matrix products are square (with dimensions $(m \times m)$ and $(n \times n)$, respectively) and symmetric. Specifically, the columns of $[R]$ are the $(m \times 1)$ eigenvectors of $[A]^{\mathrm{T}}[A]$, the columns of $[L]$ are the $(n \times 1)$ eigenvectors of $[A][A]^{\mathrm{T}}$. The respective singular values are the square roots of the corresponding eigenvalues, i.e., $\omega_i^2 = \lambda_i$.

Example 11.5. Eigenvalues and Eigenvectors of a Covariance Matrix Using SVD
Consider the (31×2) matrix $(30)^{-1/2}[X']$, where $[X']$ is the matrix of anomalies (Equation 11.29) for the minimum temperature data in Table A.1. The SVD of this matrix can be used to obtain the eigenvalues and eigenvectors of the sample covariance matrix for these data, without first explicitly computing $[S]$ (if $[S]$ is already known, SVD can also be used to compute the eigenvalues and eigenvectors, through the equivalence of Equations 11.72 and 11.52).

The SVD of $(30)^{-1/2}[X']$, in the form of Equation 11.72, is

$$\frac{1}{\sqrt{30}}[X'] = \underset{(31\times 2)}{\begin{bmatrix} 1.09 & 1.42 \\ 2.19 & 1.42 \\ 1.64 & 1.05 \\ \vdots & \vdots \\ 1.83 & 0.51 \end{bmatrix}} = \underset{(31\times 2)}{\begin{bmatrix} .105 & .216 \\ .164 & .014 \\ .122 & .008 \\ \vdots & \vdots \\ .114 & -.187 \end{bmatrix}} \underset{(2\times 2)}{\begin{bmatrix} 15.961 & 0 \\ 0 & 2.879 \end{bmatrix}} \underset{(2\times 2)}{\begin{bmatrix} .848 & .530 \\ -.530 & .848 \end{bmatrix}}. \tag{11.74}$$

The reason for multiplying the anomaly matrix $[X']$ by $30^{-1/2}$ should be evident from Equation 11.30: the product $(30^{-1/2}\,[X']^{\mathrm{T}})\,(30^{-1/2}\,[X'])=(n-1)^{-1}\,[X']^{\mathrm{T}}[X']$ yields the covariance matrix $[S]$ for these data, which is the same as the matrix $[A]$ in Equation 11.59. Because the matrix of right singular vectors $[R]$ contains the eigenvectors for the product of the matrix on the left-hand side of Equation 11.74, left-multiplied by its transpose, the matrix $[R]^{\mathrm{T}}$ on the far right of Equation 11.74 is the same as the (transpose of) the matrix $[E]$ in Equation 11.60. Similarly the squares of the singular values in the diagonal matrix $[\Omega]$ in Equation 11.74 are the corresponding eigenvalues, for example, $\omega^2{}_1 = 15.961^2 = \lambda_1 = 254.7$.

The right-singular vectors of $(n-1)^{-1/2}\,[X']$ are the eigenvectors of the (2×2) covariance matrix $[S]=(n-1)^{-1}\,[X']^{\mathrm{T}}[X']$. The left singular vectors in the matrix $[L]$ are eigenvectors of the (31×31) matrix $(n-1)^{-1}\,[X][X]^{\mathrm{T}}$. This matrix actually has 31 eigenvectors, but only two of them (the two shown in Equation 11.74) are associated with nonzero eigenvalues. It is in this sense, of truncating the zero eigenvalues and their associated irrelevant eigenvectors, that Equation 11.74 is an example of a thin SVD. ◇

The SVD is a versatile tool with a variety of applications. One of these is maximum covariance analysis (MCA), to be described in Section 14.3. Sometimes MCA is confusingly called "SVD analysis," even though SVD is merely the computational tool used to calculate a MCA.

11.4. RANDOM VECTORS AND MATRICES

11.4.1. Expectations and Other Extensions of Univariate Concepts

Just as ordinary random variables are scalar quantities, a random vector (or random matrix) is a vector (or matrix) whose entries are random variables. The purpose of this section is to extend the rudiments of matrix algebra presented in Section 11.3 to include statistical ideas.

A vector $\boldsymbol{x}$ whose K elements are the random variables x_k is a random vector. The expected value of this random vector is also a vector, called the vector mean, whose K elements are the individual expected values (i.e., probability-weighted averages) of the corresponding random variables. If all the x_k are continuous variables,

$$\boldsymbol{\mu} = \begin{bmatrix} \int_{-\infty}^{\infty} x_1 f_1(x_1)\,dx_1 \\ \int_{-\infty}^{\infty} x_2 f_2(x_2)\,dx_2 \\ \vdots \\ \int_{-\infty}^{\infty} x_K f_K(x_K)\,dx_K \end{bmatrix}. \tag{11.75}$$

If some or all of the K variables in $\boldsymbol{x}$ are discrete, the corresponding elements of $\boldsymbol{\mu}$ will be sums in the form of Equation 4.13.

The properties of expectations listed in Equation 4.15 extend also to vectors and matrices in ways that are consistent with the rules of matrix algebra. If c is a scalar constant, $[X]$ and $[Y]$ are random matrices with the same dimensions (and which may be random vectors if one of their dimensions is 1), and $[A]$ and $[B]$ are constant (nonrandom) matrices,

$$E(c[X]) = c\,E([X]), \tag{11.76a}$$

$$E([X] + [Y]) = E([X]) + E([Y]), \tag{11.76b}$$

$$E([A][X][B]) = [A]\,E([X])\,[B], \tag{11.76c}$$

$$E([A][X] + [B]) = [A]\,E([X]) + [B]. \tag{11.76d}$$

The (population, or generating-process) covariance matrix, corresponding to the sample estimate $[S]$ in Equation 11.5, is the matrix expected value

$$\underset{(K\times K)}{[\Sigma]} = E\left(\underset{(K\times 1)}{[\boldsymbol{x}-\boldsymbol{\mu}]}\,\underset{(1\times K)}{[\boldsymbol{x}-\boldsymbol{\mu}]^{\mathrm{T}}} \right) \tag{11.77a}$$

$$= E\left(\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_K-\mu_K) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \cdots & (x_2-\mu_2)(x_K-\mu_K) \\ \vdots & \vdots & \ddots & \vdots \\ (x_K-\mu_K)(x_1-\mu_1) & (x_K-\mu_K)(x_2-\mu_2) & \cdots & (x_K-\mu_K)^2 \end{bmatrix} \right) \tag{11.77b}$$

$$= \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,K} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K,1} & \sigma_{K,2} & \cdots & \sigma_{K,K} \end{bmatrix}. \tag{11.77c}$$

The diagonal elements of Equation 11.74 are the scalar (population or generating-process) variances, which would be computed (for continuous variables) using Equation 4.21 with $g(x_k)=(x_k-\mu_k)^2$ or, equivalently, Equation 4.22. The off-diagonal elements are the covariances, which would be computed using the double integrals

$$\sigma_{k,\ell} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_k - \mu_k)(x_\ell - \mu_\ell) f_{k,\ell}(x_k, x_\ell)\, dx_\ell\, dx_k, \tag{11.78}$$

which is analogous to the summation in Equation 11.4 for the sample covariances. Here $f_{k,\ell}(x_k, x_\ell)$ is the joint (bivariate) PDF for x_k and x_ℓ. Analogously to Equation 4.22b for the scalar variance, an equivalent expression for the (population) covariance matrix is

$$[\Sigma] = E\left(\boldsymbol{x}\boldsymbol{x}^T\right) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}. \tag{11.79}$$

11.4.2. Partitioning Vectors and Matrices

In some settings it is natural to define collections of variables that segregate into two or more groups. Simple examples are one set of L predictands together with a different set of $K-L$ predictors, or sets of two or more variables, each observed simultaneously at some large number of locations or gridpoints. In such cases it is often convenient and useful to maintain these distinctions notationally, by partitioning the corresponding vectors and matrices.

Partitions are indicated by thin or dashed lines in the expanded representation of vectors and matrices. These indicators of partitions are imaginary lines, in the sense that they have no effect whatsoever on the matrix algebra as applied to the larger vectors or matrices. For example, consider a $(K \times 1)$ random vector $\boldsymbol{x}$ that consists of one group of L variables and another group of $K-L$ variables,

$$\boldsymbol{x}^T = \left[x_1 \quad x_2 \quad \cdots \quad x_L \mid x_{L+1} \quad x_{L+2} \quad \cdots \quad x_K\right], \tag{11.80a}$$

which would have expectation

$$E\left(\boldsymbol{x}^T\right) = \boldsymbol{\mu}^T = \left[\mu_1 \quad \mu_2 \quad \cdots \quad \mu_L \mid \mu_{L+1} \quad \mu_{L+2} \quad \cdots \quad \mu_K\right], \tag{11.80b}$$

exactly as Equation 11.75, except that both $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are partitioned as (i.e., composed of a concatenation of) a $(L \times 1)$ vector and a $(K-L \times 1)$ vector.

The covariance matrix of $\boldsymbol{x}$ in Equation 11.80 would be computed in exactly the same way as indicated in Equation 11.77, with the partitions being carried forward:

$$[\Sigma] = E\left([\boldsymbol{x} - \boldsymbol{\mu}][\boldsymbol{x} - \boldsymbol{\mu}]^{\mathrm{T}}\right) \tag{11.81a}$$

$$= \left[\begin{array}{cccc|cccc} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,L} & \sigma_{1,L+1} & \sigma_{1,L+2} & \cdots & \sigma_{1,K} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,L} & \sigma_{2,L+1} & \sigma_{2,L+2} & \cdots & \sigma_{2,K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \sigma_{L,1} & \sigma_{L,2} & \cdots & \sigma_{L,L} & \sigma_{L,L+1} & \sigma_{L,L+2} & \cdots & \sigma_{L,K} \\ \hline \sigma_{L+1,1} & \sigma_{L+1,2} & \cdots & \sigma_{L+1,L} & \sigma_{L+1,L+1} & \sigma_{L+1,L+2} & \cdots & \sigma_{L+1,K} \\ \sigma_{L+2,1} & \sigma_{L+2,2} & \cdots & \sigma_{L+2,L} & \sigma_{L+2,L+1} & \sigma_{L+2,L+2} & \cdots & \sigma_{L+2,K} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{K,1} & \sigma_{K,2} & \cdots & \sigma_{K,L} & \sigma_{K,L+1} & \sigma_{K,L+2} & \cdots & \sigma_{K,K} \end{array}\right], \tag{11.81b}$$

$$= \left[\begin{array}{c|c} [\Sigma_{1,1}] & [\Sigma_{1,2}] \\ \hline [\Sigma_{2,1}] & [\Sigma_{2,2}] \end{array}\right], \tag{11.81c}$$

so that the covariance matrix $[\Sigma]$ for a data vector $\boldsymbol{x}$ partitioned into two segments as in Equation 11.80 is itself partitioned into four submatrices. The $(L \times L)$ matrix $[\Sigma_{1,1}]$ is the covariance matrix for the first L variables, $[x_1, x_2, \ldots, x_L]^T$, and the $(K-L \times K-L)$ matrix $[\Sigma_{2,2}]$ is the covariance matrix for the last $K-L$ variables, $[x_{L+1}, x_{L+2}, \ldots, x_K]^T$. Both of these matrices have variances on the main diagonal and covariances among the variables within its respective group in the other positions.

The $(K-L \times L)$ matrix $[\Sigma_{2,1}]$ contains the covariances among all possible pairs of variables consisting of one member in the second group and the other member in the first group. Because it is not a full covariance matrix it does not contain variances along the main diagonal even if it is square, and in general it is not symmetric. The $(L \times K-L)$ matrix $[\Sigma_{1,2}]$ contains the same covariances among all possible pairs of variables having one member in the first group and the other member in the second group. Because the full covariance matrix $[\Sigma]$ is symmetric, $[\Sigma_{1,2}]^T = [\Sigma_{2,1}]$.

11.4.3. Linear Combinations

A *linear combination* is essentially a weighted sum of two or more of the variables $x_1, x_2, \ldots, x_K$ in a data vector $\boldsymbol{x}$. For example, the multiple linear regression in Equation 7.25 is a linear combination of the K regression predictors that yields a new variable, which in this case is the regression prediction. For simplicity, assume that the parameter $b_0 = 0$ in Equation 7.25 (this would be the case if the predictand and all predictors are expressed as anomalies). Then Equation 7.25 can be expressed in matrix notation as

$$y = \boldsymbol{b}^T \boldsymbol{x}, \tag{11.82}$$

where $\boldsymbol{b}^T = [b_1, b_2, \ldots, b_K]$ is the vector of parameters that are the weights in the weighted sum.

Usually in regression the predictors $\boldsymbol{x}$ are considered to be fixed constants rather than random variables. But consider now the case where $\boldsymbol{x}$ is a random vector with mean $\boldsymbol{\mu}_x$ and covariance $[\Sigma_x]$. The linear combination in Equation 11.82 will then also be a random variable. Extending Equation 4.14c for vector $\boldsymbol{x}$, with $g_j(x) = b_j x_j$, the mean of y will be

$$\mu_y = \sum_{k=1}^{K} b_k \mu_k, \tag{11.83}$$

where $\mu_k = \mathrm{E}(x_k)$. The variance of the linear combination is more complicated, both notationally and computationally, and involves the covariances among all pairs of the x's. For simplicity, suppose $K = 2$. Then,

$$\begin{aligned}
\sigma_y^2 &= \mathrm{Var}(b_1 x_1 + b_2 x_2) = E\left\{ [(b_1 x_1 + b_2 x_2) - (b_1 \mu_1 + b_2 \mu_2)]^2 \right\} \\
&= E\left\{ [b_1 (x_1 - \mu_1) + b_2 (x_2 - \mu_2)]^2 \right\} \\
&= E\left\{ b_1^2 (x_1 - \mu_1)^2 + b_2^2 (x_2 - \mu_2)^2 + 2 b_1 b_2 (x_1 - \mu_1)(x_2 - \mu_2) \right\} \\
&= b_1^2 E\left\{ (x_1 - \mu_1)^2 \right\} + b_2^2 E\left\{ (x_2 - \mu_2)^2 \right\} + 2 b_1 b_2 E\{ (x_1 - \mu_1)(x_2 - \mu_2) \} \\
&= b_1^2 \sigma_{1,1} + b_2^2 \sigma_{2,2} + 2 b_1 b_2 \sigma_{1,2}
\end{aligned} \tag{11.84}$$

This scalar result is fairly cumbersome, even though the linear combination is composed of only two random variables, and the extension to linear combinations of K random variables involves $K(K+1)/2$ terms. More generally, and much more compactly, in matrix notation Equations 11.83 and 11.84 become

$$\mu_y = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mu} \tag{11.85a}$$

and

$$\sigma_y^2 = \boldsymbol{b}^{\mathrm{T}}[\Sigma_x]\,\boldsymbol{b}. \tag{11.85b}$$

The quantities on the left-hand sides of Equations 11.85 are scalars, because the result of the single linear combination in Equation 11.82 is scalar. But consider simultaneously forming L linear combinations of the K random variables $\boldsymbol{x}$,

$$\begin{aligned} y_1 &= b_{1,1}x_1 + b_{1,2}x_2 + \cdots + b_{1,K}x_K \\ y_2 &= b_{2,1}x_1 + b_{2,2}x_2 + \cdots + b_{2,K}x_K \\ &\vdots \quad \vdots \quad \vdots \qquad\qquad \vdots \\ y_L &= b_{L,1}x_1 + b_{L,2}x_2 + \cdots + b_{L,K}x_K \end{aligned}, \tag{11.86a}$$

or

$$\underset{(L\times 1)}{\boldsymbol{y}} = \underset{(L\times K)}{[B]^T}\ \underset{(K\times 1)}{\boldsymbol{x}}. \tag{11.86b}$$

Here each row of $[B]^{\mathrm{T}}$ defines a single linear combination as in Equation 11.82, and collectively these L linear combinations define the random vector $\boldsymbol{y}$. Extending Equations 11.85 to the mean vector and covariance matrix of this collection of L linear combinations of $\boldsymbol{x}$,

$$\underset{(L\times 1)}{\boldsymbol{\mu}_y} = \underset{(L\times K)}{[B]^{\mathrm{T}}}\ \underset{(K\times 1)}{\boldsymbol{\mu}_x} \tag{11.87a}$$

and

$$\underset{(L\times L)}{\left[\Sigma_y\right]} = \underset{(L\times K)}{[B]^T}\ \underset{(K\times K)}{[\Sigma_x]}\ \underset{(K\times L)}{[B]}. \tag{11.87b}$$

Note that by using Equations 11.87 it is not actually necessary to explicitly compute the transformed variables in Equation 11.86 in order to find their mean and covariance, if the mean vector and covariance matrix of the x's are known.

Example 11.6. Mean Vector and Covariance Matrix for a Pair of Linear Combinations
Example 11.5 showed that the matrix in Equation 11.59 is the covariance matrix for the Ithaca and Canandaigua minimum temperature data in Table A.1. The mean vector for these data is $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_{\mathrm{Ith}}, \mu_{\mathrm{Can}}) = (13.0, 20.2)$. Consider now two linear combinations of these minimum temperature data in the form of Equation 11.45, with $\theta = 32$ degrees. That is, each of the two rows of $[T]^{\mathrm{T}}$ defines a linear combination (Equation 11.82), which can be expressed jointly as in Equation 11.86b. Together, these two linear combinations are equivalent to a transformation that corresponds to a counterclockwise rotation of the coordinate axes through the angle θ. That is, each vector $\boldsymbol{y} = [T]^{\mathrm{T}}\boldsymbol{x}$ would locate the same point, but in the framework of the rotated coordinate system.

One way to find the mean and covariance for the transformed points, $\boldsymbol{\mu}_y$ and $[\Sigma_y]$, would be to carry out the transformation for all $n = 31$ point pairs, and then to compute the mean vector and covariance matrix for the transformed data set. However, knowing the mean and covariance of the underlying $\boldsymbol{x}$'s it is straightforward and much easier to use Equation 11.87 to obtain

$$\boldsymbol{\mu}_y = \begin{bmatrix} \cos 32^\circ & \sin 32^\circ \\ -\sin 32^\circ & \cos 32^\circ \end{bmatrix} \boldsymbol{\mu}_x = \begin{bmatrix} .848 & .530 \\ -.530 & .848 \end{bmatrix} \begin{bmatrix} 13.0 \\ 20.2 \end{bmatrix} = \begin{bmatrix} 21.7 \\ 10.2 \end{bmatrix} \tag{11.88a}$$

and

$$\begin{aligned} \left[\Sigma_y\right] = [T]^{\mathrm{T}}[\Sigma_x][T] &= \begin{bmatrix} .848 & .530 \\ -.530 & .848 \end{bmatrix} \begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.58 \end{bmatrix} \begin{bmatrix} .848 & -.530 \\ .530 & .848 \end{bmatrix} \\ &= \begin{bmatrix} 254.76 & 0 \\ 0 & 8.29 \end{bmatrix} \end{aligned} . \tag{11.88b}$$

The rotation angle $\theta = 32$ degrees is evidently a special one for these data, as it produces a pair of transformed variables $\boldsymbol{y}$ that are uncorrelated. In fact this transformation is exactly the same as in Equation 11.62, which was expressed in terms of the eigenvectors of $[\Sigma_x]$. ◇

Just as the mean and variance of a linear combination can be expressed and computed without actually calculating the linear combinations, the covariance of two linear combinations can similarly be computed, using

$$\mathrm{Cov}\left([A]^{\mathrm{T}}\boldsymbol{x}_1, [B]^{\mathrm{T}}\boldsymbol{x}_2\right) = [A]^T \left[\Sigma_{1,2}\right] [B]. \tag{11.89}$$

Here $[\Sigma_{1,2}]$ is the matrix of covariances between the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which is the upper right-hand quadrant of Equation 11.81, when the vector $\boldsymbol{x}$ has been partitioned into the subvectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. If $[A]^{\mathrm{T}}$ and $[B]^{\mathrm{T}}$ are vectors (and so dimensioned $(1 \times L)$ and $(1 \times K - L)$, respectively), Equation 11.89 will yield the scalar covariance between the single pair of linear combinations. If $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \boldsymbol{x}$, then Equation 11.89 becomes

$$\mathrm{Cov}\left([A]^{\mathrm{T}}\boldsymbol{x}, [B]^{\mathrm{T}}\boldsymbol{x}\right) = [A]^{\mathrm{T}} [\Sigma] [B], \tag{11.90}$$

where, as before $[\Sigma]$ is the covariance matrix for $\boldsymbol{x}$.

11.4.4. Mahalanobis Distance, Revisited

Section 11.2.2 introduced the Mahalanobis, or statistical, distance as a way to measure differences or unusualness within the context established by an empirical data scatter or an underlying multivariate probability density. If the K variables in the data vector $\boldsymbol{x}$ are mutually uncorrelated, the (squared) Mahalanobis distance takes the simple form of the sum of the squared standardized anomalies z_k, as indicated in Equation 11.7 for $K = 2$ variables. When some or all of the variables are correlated the Mahalanobis distance accounts for the correlations as well, although as noted in Section 11.2.2 the notation is prohibitively complicated in scalar form. In matrix notation, the Mahalanobis distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in their K-dimensional space is

$$D^2 = [\boldsymbol{x} - \boldsymbol{y}]^{\mathrm{T}} [S]^{-1} [\boldsymbol{x} - \boldsymbol{y}], \tag{11.91}$$

where $[S]$ is the covariance matrix in the context of which the distance is being calculated.

If the dispersion defined by $[S]$ involves zero correlation among the K variables, it is not difficult to see that Equation 11.91 reduces to Equation 11.7 (in two dimensions, with obvious extension to higher

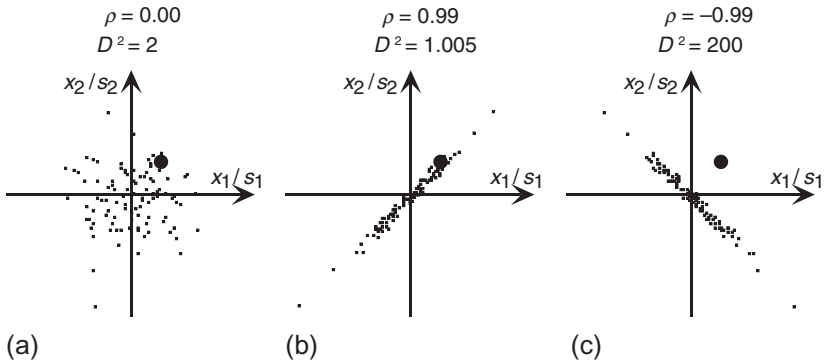

FIGURE 11.6 The point $z^T=(1, 1)$ (large dot) in the contexts of data scatters with (a) zero correlation, (b) correlation 0.99, and (c) correlation -0.99. Mahalanobis distances, D^2, to the origin are drastically different in these three cases.

dimensions). In that case, $[S]$ is diagonal, and its inverse is also diagonal with elements $(s_{k,k})^{-1}$, so Equation 11.91 would reduce to $D^2=\Sigma_k\ (x_k-y_k)^2/s_{k,k}$. This observation underscores one important property of the Mahalanobis distance, namely, that different intrinsic scales of variability for the K variables in the data vector do not confound D^2, because each is divided by its standard deviation before squaring. If $[S]$ is diagonal, the Mahalanobis distance is the same as the Euclidean distance after dividing each variable by its standard deviation.

The second salient property of the Mahalanobis distance is that it accounts for the redundancy in information content among correlated variables. Again, this concept is easiest to see in two dimensions. Two strongly correlated variables provide very nearly the same information, and ignoring strong correlations when calculating statistical distance (i.e., using Equation 11.7 when the correlation is not zero), effectively double-counts the contribution of the (nearly) redundant second variable. The situation is illustrated in Figure 11.6, which shows the standardized point $z^T=(1, 1)$ in the contexts of three very different point clouds. In Figure 11.6a the correlation reflected by the circular point cloud is zero, so it is equivalent to use Equation 11.7 to calculate the Mahalanobis distance to the origin (which is also the vector mean of the point cloud), after having accounted for possibly different scales of variation for the two variables by dividing by the respective standard deviations. That distance is $D^2=2$ (corresponding to an ordinary Euclidean distance of $\sqrt{2}=1.414$). The correlation between the two variables in Figure 11.6b is 0.99, so that one or the other of the two variables provides nearly the same information as both together: z_1 and z_2 are nearly the same variable. Using Equation 11.91 the Mahalanobis distance to the origin is $D^2=1.005$, which is only slightly more than if only one of the two nearly redundant variables had been considered alone, and substantially smaller than the distance appropriate to the context of the scatter in Figure 11.6a.

Finally, Figure 11.6c shows a very different situation, in which the correlation is -0.99. Here the point (1, 1) is extremely unusual in the context of the data scatter, and using Equation 11.91 we find $D^2=200$. That is, it is extremely far from the origin relative to the dispersion of the point cloud, and this unusualness is reflected by the very large Mahalanobis distance. The point (1, 1) in Figure 11.6c is a *multivariate outlier*. Visually it is well removed from the point scatter in two dimensions. But relative to either of the two univariate distributions it is a quite ordinary point that is relatively close to (one standard deviation from) each scalar mean, so that it would not stand out as unusual when applying standard scalar EDA methods to the two variables individually. It is an outlier in the sense that it does

not behave like the scatter of the negatively correlated point cloud, in which large values of x_1/s_1 are associated with small values of x_2/s_2, and vice versa. The large Mahalanobis distance to the center (vector mean) of the point cloud identifies it as a multivariate outlier.

Equation 11.91 is an example of what is called a *quadratic form*. It is quadratic in the vector $\boldsymbol{x}-\boldsymbol{y}$, in the sense that this vector is multiplied by itself, together with scaling constants in the symmetric matrix $[S]^{-1}$. In $K=2$ dimensions a quadratic form written in scalar notation takes the form of Equation 11.7 if the symmetric matrix of scaling constants is diagonal, and in the form of Equation 11.8 if it is not. Equation 11.91 emphasizes that quadratic forms can be interpreted as squared distances, and as such it is generally desirable for them to be nonnegative, and furthermore strictly positive if the vector being squared is not zero. This latter condition is met if the symmetric matrix of scaling constants is positive definite, so that all its eigenvalues are positive.

Finally, it was noted in Section 11.2.2 that Equation 11.7 describes ellipses of constant distance D^2. The ellipses described by Equation 11.7, corresponding to zero correlations in the matrix $[S]$ in Equation 11.91, have their axes aligned with the coordinate axes. Equation 11.91 also describes ellipses of constant Mahalanobis distance D^2, whose axes are rotated away from the directions of the coordinate axes to the extent that some or all of the correlations in $[S]$ are nonzero. In such cases the axes of the ellipses of constant D^2 are aligned in the directions of the eigenvectors of $[S]$, as will be seen in Section 12.1.

11.5. EXERCISES

11.1. Calculate the matrix product $[A][E]$, using the values in Equations 11.59 and 11.60.

11.2. Derive the regression equation produced in Example 7.1, using matrix notation.

11.3. Calculate the angle between the two eigenvectors of the matrix $[A]$ in Equation 11.59.

11.4. Verify through matrix multiplication that both $[T]$ in Equation 11.45, and its transpose, are orthogonal matrices.

11.5. Show that Equation 11.67 produces a valid square root.

11.6. Assuming all the relevant matrix and vector dimensions are compatible, simplify:

$$([C](\overline{\boldsymbol{x}}-\boldsymbol{\mu}))^{\mathrm{T}}\left(\frac{1}{n}[C][S][C]^{\mathrm{T}}\right)^{-1}([C](\overline{\boldsymbol{x}}-\boldsymbol{\mu}))$$

11.7. The $(K\times K)$ square matrix [E] contains eigenvectors of a covariance matrix as its columns. Solve for the $(\mathrm{K}\times 1)$ vector $\boldsymbol{x}$:

$$\boldsymbol{u}=[E]^{\mathrm{T}}\boldsymbol{x}$$

11.8. The eigenvalues and eigenvectors of the covariance matrix for the Ithaca and Canandaigua maximum temperatures in Table A.1 are $\lambda_1 = 118.8$ and $\lambda_2 = 2.60$, and $\boldsymbol{e}_1{}^{\mathrm{T}} = [.700, .714]$ and $\boldsymbol{e}_2{}^{\mathrm{T}} = [-.714, .700]$, where the first element of each vector corresponds to the Ithaca temperature.

a. Find the covariance matrix $[S]$, using its spectral decomposition.
b. Find $[S]^{-1}$ using its eigenvalues and eigenvectors.
c. Find $[S]^{-1}$ using the result of part (a), and Equation 11.28.
d. Find a symmetric $[S]^{1/2}$.
e. Find the Mahalanobis distance between the observations for 1 January and 2 January.

11.9. a. Use the Pearson correlations in Table 3.5 and the standard deviations from Table A.1 to compute the covariance matrix [S] for the four temperature variables in Table A.1.
b. Consider the average daily temperatures defined by the two linear combinations:
$y_1 = 0.5$ (Ithaca Max)+0.5 (Ithaca Min)
$y_2 = 0.5$ (Canandaigua Max)+0.5 (Canandaigua Min)
Find $\boldsymbol{\mu}_y$ and [S_y] without actually computing the individual $\boldsymbol{y}$ values.